Universal-to-Specific: Dynamic Knowledge-Guided Multiple Instance Learning for Few-Shot Whole Slide Image Classification

Supplementary Material

1. Compared Methods

To evaluate the performance of DyKo, we benchmark it against nine state-of-the-art methods. These baselines comprise five canonical MIL architectures (ABMIL[2], TransMIL[6], RRTMIL[8], WiKG[3], and MiCo [4]) and four recent prompt-based approaches (TopMIL [5], ViLa-MIL [7], ConcepPath [9], and FOCUS [1]). The details of each method are as follows:

- **ABMIL:** ABMIL employs an attention mechanism to assign discriminative weights to individual instances, thereby identifying the most informative instances within a bag for classification.
- **TransMIL:** Applies the Transformer architecture to MIL, effectively capturing long-range dependencies between instances through a self-attention mechanism for better WSI classification.
- **RRTMIL:** RRTMIL constructs a hierarchical representation of instances to model both local and global contextual relationships, thereby improving classification accuracy at the bag level.
- **WiKG:** This method represents a WSI as a knowledge graph, treating cropped patches as nodes and leveraging head-to-tail patch embeddings to generate dynamic graph representations for WSI analysis.
- **MiCo:** MiCo introduces context-aware clustering to link dispersed tissue instances via semantic anchors, enhancing intra- and inter-tissue correlations for robust WSI analysis.
- **TopMIL:** A two-level prompt learning Multiple Instance Learning framework guided by pathology language prior knowledge, proposed to address few-shot weakly-supervised learning scenarios.
- **PatchGCN:** This method models a histopathology image as a graph of patches, utilizing a Graph Convolutional Network (GCN) to capture spatial adjacencies and morphological relationships.
- **ViLaMIL:** A dual-scale vision-language multiple instance learning framework that aligns dual-scale slide-level features with a visual descriptive text prompt to efficiently perform whole slide image classification.
- **ConcepPath:** ConcepPath leverages both expert concepts, induced from medical literature via GPT-4 , and complementary data-driven concepts to guide a two-stage hierarchical feature aggregation for WSI classification.
- **FOCUS:** A knowledge-enhanced adaptive visual compression framework that uniquely combines pathology foundation models with language prior knowledge to enable a focused analysis of diagnostically relevant regions by prioritizing discriminative WSI patches.

2. Details of Concept Feature Construction

As described in Section 3.3 of the main paper, we construct the concept feature set $\boldsymbol{C}$ specifically for each dataset using a filtering and encoding pipeline. The raw source of our concepts is the **Region-of-Interest (ROI) text** annotations from the Quilt-1M dataset. Unlike general captions, these ROI texts provide fine-grained lists of morphological entities (e.g., [”dense inflammation”, ”lymphocytes”, ”inflammatory cells”]) specific to local tissue regions. To ensure the relevance of the retrieved knowledge while preserving rich contextual details, we employ an ROI-based expansion filtering strategy followed by semantic clustering.

2.1. Keyword Filtering Rules

The initial candidate pool is generated by filtering the ROI text lists. We utilize a set of pre-defined, expert-validated keywords tailored to the anatomical site and cancer subtypes of the target dataset. These keywords cover the organ of interest (e.g., ”lung”, ”kidney”) and specific histological subtypes (e.g., ”adenocarcinoma”, ”clear cell”). The specific keywords used for each dataset are listed in Table 2.

2.2. Concept Selection Pipeline

The construction of the final concept set C involves the following steps:

1. **ROI-based Expansion Filtering:** We iterate through the ROI text lists in Quilt-1M. Each ROI entry consists of a list of concepts (e.g., *[”dense inflammation here”, ”lymphocytes”, ”inflammatory cells”]*). We apply a **”hit-one-keep-all”** strategy: if *any single concept* within an ROI list matches a keyword from Table 2 (case-insensitive matching), the *entire* list of concepts is retained and added to the candidate pool. This strategy ensures that we capture not only the target pathological entities but also their co-occurring morphological context (e.g., inflammation surrounding a tumor).
2. **Encoding:** All individual concept strings collected in the candidate pool are encoded into d-dimensional feature vectors using the text encoder of the TITAN foundation model.

Table 1. Key Mathematical Notations in DyKo.

Symbol	Description	Symbol	Description
1. Base Data & Features			
W_i	The i-th whole slide image.	d	Feature vector dimension.
N	Number of patches in a WSI.	X	Visual feature matrix ($N \times d$).
x_i	Feature vector for patch i ($1 \times d$).		
2. Knowledge & Prompts			
C	Concept feature set ($N_C \times d$).	n_{cls}	Number of target classes.
N_C	Number of concepts in knowledge base.	T_{static}	Fixed prompt embeddings ($t_1 \times d$).
t_1	Token length of static prompts.	T_{learn}	Learnable prompt embeddings ($t_2 \times d$).
t_2	Token length of learnable prompts.	T_{input}	Input prompt sequence ($(t_1 + t_2) \times d$).
T	Class-level prompts ($n_{cls} \times d$).		
3. WSI-Adaptive Knowledge Instantiation (WAKI) Module			
M	Number of visual clusters (prototypes).	$C_{u_j}^K$	Top-K concepts for cluster (prototype) j.
S_j	Set of patch features in cluster j.	$\alpha_{i,c}$	Attention score (patch i, concept c).
u_j	Centroid of cluster j (prototype) ($1 \times d$).	τ	Attention temperature.
K	Number of concepts for each prototypes to retrieve.	X'	Knowledge-instantiated features ($N \times d$).
x_i'	Knowledge-instantiated feature for patch i ($1 \times d$).		
4. Fusion & Classification			
F_{vis}	Visual stream output ($n_{cls} \times d$).	F_{fused}	Final WSI-level feature ($n_{cls} \times d$).
F_{con}	Conceptual stream output ($n_{cls} \times d$).		
5. Loss Functions & Optimization			
$\mathcal{L}_{CE}$	Cross-Entropy Loss (classification loss).	$\mathcal{L}_{SC}$	Structural Consistency Loss.
p_{vis}	Probability distribution from the visual cluster head.	$\mathcal{L}$	The final composite objective function.
p_{sem}	Probability distribution from the semantic cluster head.	λ	Hyperparameter to weight the $\mathcal{L}_{SC}$ loss.

3. **Clustering and Refinement:** The raw candidate pool may contain redundancies (e.g., "lymphocytes" and "lymphocytic infiltrate") or noise. To distill the most representative pathological concepts, we perform K-means clustering on the candidate embeddings with N_C clusters.
4. **Selection:** For each cluster, we select the single concept feature closest to the cluster centroid. These N_C centroid-aligned features constitute the final concept set $C \in \mathbb{R}^{N_C \times d}$ used in the WAKI module.

3. Pathology Prior Knowledge Prompt

This section details the specific pathological descriptions used as prior knowledge prompts for each diagnostic category across the evaluated datasets. These descriptions, generated by a Cluade-3.5-Sonnet, provide the model with domain-specific morphological characteristics.

3.1. CAMELYON16

Normal: Intact follicular architecture with distinct germinal centers containing tingible body macrophages. Small lymphocytes with scant cytoplasm and condensed chromatin arranged in organized zones. Well-defined capsule, subcapsular sinuses, and traversing blood vessels. No cellular atypia or architectural distortion.

Tumor: Architectural effacement by cohesive clusters of epithelial cells with pleomorphic nuclei, prominent nucleoli, and irregular nuclear membranes. Intracytoplasmic lumina, mitotic figures, and apoptotic bodies present. Desmoplastic stromal reaction with lymphatic invasion and extranodal extension.

3.2. NSCLC

Lung Squamous Cell Carcinoma: A whole slide image of lung squamous cell carcinoma at high resolution with visually descriptive characteristics of squamous cell differentiation, round structures with eosinophilic cytoplasm, distinct cell borders and abundant cytoplasm, enlarged nuclei, irregular nuclear shape, increased chromatin density.

Lung Adenocarcinoma: A whole slide image of lung adenocarcinoma at high resolution with visually descriptive characteristics of clear cytoplasm, round or oval nuclei, prominent nucleoli, rich vascularity, irregular blood vessels, intratumoral septa, and heterogeneity.

Table 2. The domain-specific keywords used to filter the Quilt-1M dataset for constructing the candidate concept pool.

Dataset	Filtering Keywords
CAMELYON16	'breast', 'lymph', 'lymph node', 'metastatic carcinoma', 'metastasis'
NSCLC	'lung', 'pulmonary', 'nsclc', 'adenocarcinoma', 'squamous cell carcinoma', 'bronchi', 'alveoli', 'lung squamous', 'lung adenocarcinoma'
UBC-OCEAN	'ovarian', 'ovary', 'adnexal', 'serous carcinoma', 'endometrioid', 'mucinous', 'clear cell carcinoma', 'high-grade serous', 'low-grade serous'
TCGA-RCC	'renal', 'kidney', 'rcc', 'renal cell carcinoma', 'chromophobe', 'chrcc', 'clear-cell', 'ccrcc', 'papillary', 'prcc'

3.3. UBC

Clear-cell Ovarian Carcinoma: Polygonal cells with clear cytoplasm, hobnail appearance, and prominent nucleoli. Tubulocystic or papillary architecture with hyalinized cores. Nuclear pleomorphism and irregular nuclear membranes present.

Endometrioid Carcinoma: Glandular structures resembling endometrium, cribriform pattern, squamous differentiation. Oval nuclei with moderate pleomorphism and distinct nucleoli.

High-grade Serous Carcinoma: Marked nuclear atypia, complex papillae, slit-like spaces, and numerous mitoses. Cells show prominent nucleoli, irregular nuclear contours, and high nuclear-to-cytoplasmic ratio.

Low-grade serous carcinoma: Uniform nuclei, micropapillary architecture, psammoma bodies. Minimal nuclear atypia, regular nuclear membranes, and infrequent mitoses.

Mucinous carcinoma: Stratified columnar cells with intracellular mucin, goblet cells. Complex glandular architecture, gastrointestinal-type epithelium, and variable nuclear atypia.

3.4. TCGA-RCC

Clear-cell Renal Cell Carcinoma: Abundant clear cytoplasm due to glycogen/lipid content. Distinct cell membranes forming nests/acini with delicate vasculature. Nuclei centrally located, round to oval with occasional prominent nucleoli. Variable nuclear grade. Typical alveolar/acinar architecture with sheet-like growth.

Chromophobe Renal Cell Carcinoma: Large polygonal cells with distinct cell borders. Pale, reticulated cytoplasm. Prominent cell membranes. Characteristic perinuclear halos. Wrinkled "raisinoid" nuclei. Plant-like cell arrangement. Lacks prominent vasculature of clear-cell RCC.

Papillary Renal Cell Carcinoma: Papillary/tubulopapillary architecture with fibrovascular cores. Basophilic/eosinophilic cytoplasm. Foamy macrophages in papillary cores. Nuclear crowding/pseudostratification. Hemosiderin deposits.

4. Detailed Pathological Concept Sets

In this section, we provide the complete list of domain-specific pathological concepts retrieved from the Quilt-1M knowledge base for each dataset. These concepts were filtered and selected to construct the concept feature set C used in our DyKo framework.

Table 3. A random selection of 50 domain-specific pathological concepts from the constructed knowledge base for each dataset.(These concepts are derived from Quilt-1M and serve as the semantic basis for the WAKI module.)

Dataset	Concepts
CAMELYON16	leomorphic nuclei, prominent nucleoli, scant cytoplasm, mitotic figures, signet ring cells, apoptotic bodies, Atypical cells, blast-like cells, epithelial cells, epithelioid cells, foamy cells, Giant cells, hyperchromatic cells, multinucleated giant cells, neoplastic cells, plasmacytoid cell, Spindle-shaped cells, micropapillary architecture, solid areas of growth, glandular structures, cribriform pattern, cohesive clusters, Architectural distortion, Benign lobule, breast ducts, circumscribed lesion, ductal structures, Dysplasia, germinal centers, Glands infiltrating into the fat, Lymphoid follicle, Nests, lymphocytic infiltrate, Desmoplastic stromal reaction, adipose tissue, desmoplasia, Eosinophils, fibrosis, Granulomas visible under high power, plasma cells, Tumor-infiltrating lymphocytes, adenocarcinoma, Metastatic breast cancer, Carcinoma, Fibromatosis, invasion, lymph node, malignant, metastasis, necrosis.
NSCLC	pleomorphic squamous epithelial cells, Hyperchromatic nuclei, Signet ring adenocarcinoma cells, poorly differentiated tumor cells, Type II pneumocytes, Pleomorphic cells, small round blue cells, Multinucleated giant cell, Giant cells, Area of squamous differentiation, Cribriform pattern, Solid adenocarcinoma, NC ratio, mitotic figures, Loss of polarity, lepidic growth pattern, Glandular structures, Cribriform pattern, Micropapillary pattern, nests, Keratin pearls, solid areas, infiltrating glands, malignant glands, Tumor cells growing along the alveolar septa, pagetoid spread, crowding, Alveolar spaces, Residual normal pulmonary architecture, desmoplastic stroma reaction, lymphocytic infiltrate, Invasion, Vascular invasion, granulomatous inflammation, fibrosis, necrotizing granulomatous inflammation, Interstitial thickening, pigmented macrophages, infiltrative growth pattern, Adenocarcinoma, Squamous Cell Carcinoma, Small cell lung carcinoma, Adenocarcinoma in situ, Invasive adenocarcinoma, mucinous adenocarcinoma, Sarcoidosis, organizing pneumonia, interstitial lung disease, Granuloma, Metastasis, Keratin pearl formation,Area of squamous differentiation.
UBC-OCEAN	Intensely basophilic cells, adenoid cystic cells, hyperchromatic nuclei, Eosinophilic cytoplasm, columnar mucin-producing epithelial cells, apoptotic figures, intracellular pale mucin, mesothelial cells, Columnar epithelial cells, giant cells, Mitotic figures, Cytoplasmic clearing, Squamous cells, basal cells, Nuclear atypia, Loss of polarity, clear cells, cuboidal cells, Intranuclear inclusion, Marked nuclear pleomorphism, Spindle-shaped cells, hobnail cells, Rhabdoid change, micropapillary morphology, epithelial tumor islands, glandular like spaces, fibrovascular cores, cords, malignant glands, solid architecture, Nests of tumor cells, tubular pattern, Inflammatory exudate, hyalinizing stroma, fibrovascular stroma, fibrous stroma, Smooth muscle fibers, lymphocytic aggregates, plasma cells, adipose tissue, Blood vessels, lymphoplasmacytic infiltrates, Germinal centers, Endometrioid carcinoma, background of endometriosis, mucinous carcinoma, tumor necrosis, squamous cell carcinoma, Metastatic mucinous carcinoma, lymph node.
TCGA-RCC	clear cells, granular cytoplasm, chromophobe cells, spindle cell, nuclear atypia, nuclear grooves, foamy macrophages, distinct membranes, Perinuclear halo appearance, prominent nucleoli, Oncocytic cells, kidney-shaped nuclei, multinucleated giant cells, Nuclear hyperchromasia, intranuclear cytoplasmic inclusions, Rabdoid appearance, cuboidal epithelial cells, mitotic figures, pleomorphic nuclei, papillary architecture, Microcystic pattern, encapsulated lesion, tubular structures, trabecular pattern, Sarcomatoid areas, cribriform pattern, glomeruloid appearance, fibrovascular core, cysts, Sheet of clear cells, Nests, solid looking tumor, well-demarcated border, micropapillary carcinoma structure, fibrosis, hyaline arteriolosclerosis, increased vascularity, blood vessels, invasion, inflammatory infiltrate, desmoplastic reaction, adipose tissue, hemorrhage, edematous stroma, hemosiderin deposition, necrosis, psammoma bodies, glomerulus, perinephric fat, vascular invasion.

5. Detailed Pathological Concept Sets

In this section, we provide an overview of the domain-specific pathological concepts identified for

each dataset. These concepts serve as the semantic basis for the WAKI module, enabling the model to align visual features with established clinical terminology. To illustrate the granularity and domain relevance of our knowledge base, **a random selection of 50 representative concepts** for each dataset is presented in Table 3.

CAMELYON16 (Breast Cancer) The retrieved concept set for this dataset centers on breast cancer metastasis in lymph nodes. It contains morphological descriptions that align with the distinction between normal nodal architecture (e.g., *germinal centers*, *lymphoid follicles*) and metastatic tumor deposits (e.g., *macrometastasis*, *tumor-infiltrating lymphocytes*, *desmoplastic reaction*).

NSCLC (Lung Cancer) For Non-Small Cell Lung Cancer, the identified concepts encompass distinctive features relevant to both Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). The set highlights key architectural patterns, such as *glandular structures* and *lepidic growth* for adenocarcinoma, and *keratin pearls* and *intercellular bridges* for squamous cell carcinoma.

UBC-OCEAN (Ovarian Cancer) The concept set associated with ovarian carcinoma reflects the high morphological heterogeneity of its subtypes. It comprises specific cellular and architectural descriptors corresponding to High-grade Serous Carcinoma, Clear-cell Ovarian Carcinoma, and others, featuring terms like *papillary architecture*, *hobnail cells*, and *psammoma bodies*.

TCGA-RCC (Kidney Cancer) In the context of renal cell carcinoma, the concept feature set covers characteristics pertinent to Clear Cell (ccRCC), Papillary (pRCC), and Chromophobe (chRCC) subtypes. It prominently features cytoplasmic characteristics (e.g., *clear cytoplasm*, *perinuclear halos*) and tissue architectures (e.g., *nested patterns*, *prominent vasculature*) distinctive to these histotypes.

6. Discussion

6.1. Interpretability

As shown in Fig.1, we first visualize its analysis of a case from the CAMELYON16 dataset. The model's attention heatmap accurately localizes metastatic tumor regions, providing a coarse-grained validation of its diagnostic focus. Additionally, we conduct a patch-level t-SNE analysis on this WSI, where each patch is annotated as normal or tumor. We compare two versions of DyKo: one trained with our proposed SC loss and another trained without it. Without this constraint, the model exhibits severe semantic drift, where the knowledge-instantiated features become completely dissociated from their corresponding visual features. In contrast, the model trained with SC loss shows strong alignment between the visual and knowledge-instantiated feature spaces for both tumor and normal patches. This proves that our SC loss is crucial for preventing the model from learning spurious, ungrounded concepts.

Building on this validated semantic alignment, DyKo effectively decomposes the WSI into distinct and clinically meaningful morphological clusters, each annotated with a list of top pathological concepts. Crucially, DyKo correctly identifies high-grade tumor areas (e.g., Cluster 4, 6, 9) and associates them with pathological concepts like *Metastatic breast cancer* and *adenocarcinoma*. Simultaneously, it distinguishes non-malignant but critical micro-environmental tissues (e.g., Cluster 0, 1, 2), such as *lymphocytic infiltrate* and *adipose tissue*. DyKo further demonstrates nuance by identifying complex features like *necrosis* within tumor clusters.

6.2. Effect of the Number of Visual Prototypes

We investigate the sensitivity of the model to the granularity of visual prototypes by evaluating configurations with $M \in \{5, 10, 15\}$. As detailed in Table 4, the configuration with $M = 10$ yields superior performance in the challenging 4-shot setting, achieving an AUC of 0.871. In the 8-shot and 16-shot scenarios, where supervision is more abundant, performance remains robust across all configurations with minimal variance. Consequently, we select $M = 10$ as the default choice, as it demonstrates the best trade-off across varying data regimes. This suggests that 10 prototypes effectively balance the need to capture sufficient morphological heterogeneity while avoiding the noise and fragmentation associated with excessive clustering.

6.3. Effects of Selected Concept Number

We evaluate the impact of the number of selected concepts K by testing values in the set $\{5, 10, 15, 20\}$. As detailed in Table 5, selecting the top 10 concepts consistently achieves optimal performance across all tested scenarios. This indicates that $K = 10$ strikes the best balance between capturing core semantic information and filtering out irrelevant noise. Specifically, performance degradation with fewer concepts (e.g., $K = 5$) suggests an incomplete capture of essential diagnostic evidence, while the decline observed with larger retrieval pools ($K = 15, 20$) indicates that including less relevant concepts introduces noise that hinders model discrimination.

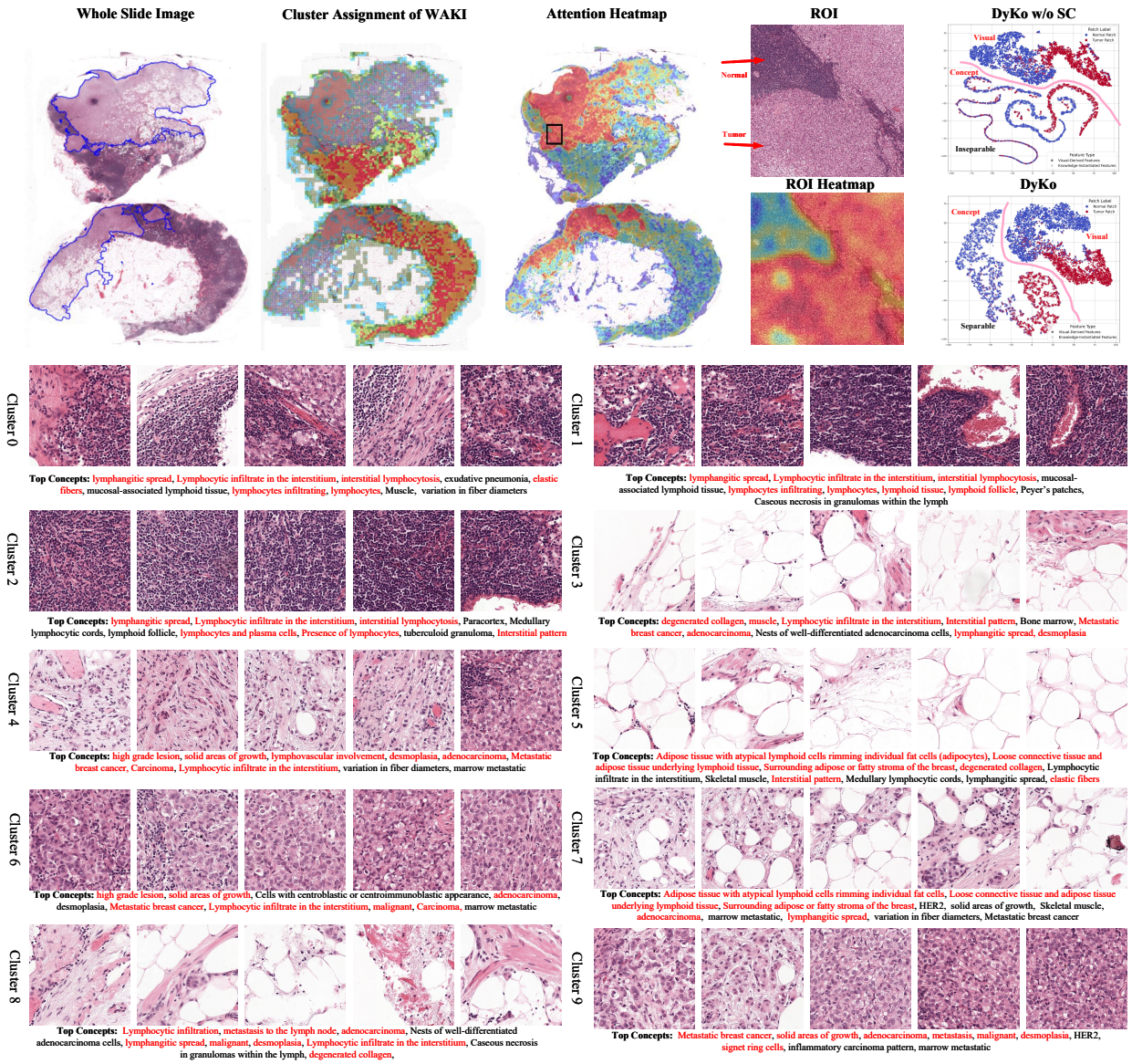

Figure 1. Qualitative interpretability analysis of DyKo. The top row illustrates the process from a WSI to an attention heatmap highlighting the tumor region of interest (ROI). The t-SNE visualizations compare DyKo's feature distributions with and without Structural Consistency (SC) loss. The bottom row presents 10 data-driven clusters, each comprising morphologically similar image patches. Concepts highlighted in red represent those that were reviewed and confirmed by a physician as being accurate and representative of the cluster's morphology. This process validates the model's ability to link its visual representations to established clinical terminology.

Table 4. Effect of the number of visual prototypes (M) on performance on the CAMELYON16 dataset.

M	4-shot			8-shot			16-shot		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
5	$0.839_{\pm 0.08}$	$0.798_{\pm 0.09}$	$0.822_{\pm 0.08}$	$0.960_{\pm 0.01}$	$0.938_{\pm 0.03}$	$0.944_{\pm 0.02}$	$0.962_{\pm 0.02}$	$0.937_{\pm 0.02}$	$0.942_{\pm 0.02}$
10	$0.871_{\pm 0.11}$	$0.827_{\pm 0.08}$	$0.841_{\pm 0.07}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.02}$	$0.930_{\pm 0.01}$	$0.961_{\pm 0.03}$	$0.943_{\pm 0.01}$	$0.947_{\pm 0.01}$
15	$0.865_{\pm 0.10}$	$0.827_{\pm 0.11}$	$0.839_{\pm 0.11}$	$0.951_{\pm 0.03}$	$0.932_{\pm 0.03}$	$0.938_{\pm 0.03}$	$0.962_{\pm 0.01}$	$0.943_{\pm 0.02}$	$0.947_{\pm 0.02}$

Table 5. Performance Comparison with Different Number of Selected Concepts on the CAMELYON16 Dataset.

K	4-shot			8-shot			16-shot		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
5	$0.831_{\pm 0.11}$	$0.792_{\pm 0.15}$	$0.833_{\pm 0.10}$	$0.949_{\pm 0.02}$	$0.916_{\pm 0.02}$	$0.924_{\pm 0.01}$	$0.960_{\pm 0.01}$	$0.935_{\pm 0.02}$	$0.939_{\pm 0.01}$
10	$0.871_{\pm 0.11}$	$0.827_{\pm 0.08}$	$0.841_{\pm 0.07}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.02}$	$0.930_{\pm 0.01}$	$0.961_{\pm 0.03}$	$0.943_{\pm 0.01}$	$0.947_{\pm 0.01}$
15	$0.851_{\pm 0.07}$	$0.807_{\pm 0.08}$	$0.826_{\pm 0.08}$	$0.947_{\pm 0.01}$	$0.911_{\pm 0.03}$	$0.920_{\pm 0.02}$	$0.956_{\pm 0.02}$	$0.924_{\pm 0.01}$	$0.930_{\pm 0.01}$
20	$0.798_{\pm 0.13}$	$0.752_{\pm 0.14}$	$0.779_{\pm 0.12}$	$0.938_{\pm 0.02}$	$0.907_{\pm 0.01}$	$0.917_{\pm 0.01}$	$0.954_{\pm 0.02}$	$0.906_{\pm 0.04}$	$0.915_{\pm 0.03}$

Table 6. Effects of Sample Number of Knowledge Base.

N_C	4-shot			8-shot			16-shot		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
100	$0.817_{\pm 0.11}$	$0.769_{\pm 0.12}$	$0.795_{\pm 0.09}$	$0.947_{\pm 0.03}$	$0.915_{\pm 0.03}$	$0.923_{\pm 0.02}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.01}$	$0.930_{\pm 0.01}$
500	$0.852_{\pm 0.08}$	$0.807_{\pm 0.08}$	$0.820_{\pm 0.08}$	$0.955_{\pm 0.01}$	$0.917_{\pm 0.02}$	$0.924_{\pm 0.02}$	$0.965_{\pm 0.02}$	$0.942_{\pm 0.01}$	$0.945_{\pm 0.01}$
1000	$0.871_{\pm 0.11}$	$0.827_{\pm 0.08}$	$0.841_{\pm 0.07}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.02}$	$0.930_{\pm 0.01}$	$0.961_{\pm 0.03}$	$0.943_{\pm 0.01}$	$0.947_{\pm 0.01}$
2000	$0.842_{\pm 0.08}$	$0.823_{\pm 0.10}$	$0.839_{\pm 0.10}$	$0.961_{\pm 0.02}$	$0.932_{\pm 0.01}$	$0.938_{\pm 0.01}$	$0.967_{\pm 0.02}$	$0.935_{\pm 0.02}$	$0.940_{\pm 0.01}$

Table 7. Performance Comparison of Different LLMs for Class-level Prompt Generation on the CAMELYON16 Dataset.

Models	4-shot			8-shot			16-shot		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
DeepSeek-R1	$0.830_{\pm 0.08}$	$0.819_{\pm 0.10}$	$0.841_{\pm 0.08}$	$0.951_{\pm 0.02}$	$0.912_{\pm 0.03}$	$0.919_{\pm 0.02}$	$0.957_{\pm 0.02}$	$0.934_{\pm 0.02}$	$0.940_{\pm 0.01}$
Gemini-2.5-Pro	$0.833_{\pm 0.08}$	$0.810_{\pm 0.12}$	$0.822_{\pm 0.12}$	$0.953_{\pm 0.02}$	$0.921_{\pm 0.03}$	$0.928_{\pm 0.03}$	$0.965_{\pm 0.02}$	$0.932_{\pm 0.03}$	$0.938_{\pm 0.03}$
GPT4.1	$0.861_{\pm 0.04}$	$0.817_{\pm 0.03}$	$0.831_{\pm 0.03}$	$0.954_{\pm 0.02}$	$0.915_{\pm 0.01}$	$0.924_{\pm 0.01}$	$0.958_{\pm 0.02}$	$0.919_{\pm 0.04}$	$0.926_{\pm 0.03}$
Cluade-3.5-Sonnet	$0.871_{\pm 0.11}$	$0.827_{\pm 0.08}$	$0.841_{\pm 0.07}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.02}$	$0.930_{\pm 0.01}$	$0.961_{\pm 0.03}$	$0.943_{\pm 0.01}$	$0.947_{\pm 0.01}$

Table 8. Comparison of Different Pathology Foundation Models on the CAMELYON16 Dataset.

Models	4-shot			8-shot			16-shot		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
QuiltNet	$0.789_{\pm 0.10}$	$0.771_{\pm 0.14}$	$0.791_{\pm 0.13}$	$0.902_{\pm 0.05}$	$0.881_{\pm 0.03}$	$0.892_{\pm 0.03}$	$0.931_{\pm 0.03}$	$0.921_{\pm 0.03}$	$0.928_{\pm 0.02}$
CONCH	$0.812_{\pm 0.09}$	$0.876_{\pm 0.10}$	$0.790_{\pm 0.10}$	$0.924_{\pm 0.02}$	$0.903_{\pm 0.06}$	$0.913_{\pm 0.05}$	$0.945_{\pm 0.02}$	$0.902_{\pm 0.02}$	$0.910_{\pm 0.02}$
TITAN	$0.871_{\pm 0.11}$	$0.827_{\pm 0.08}$	$0.841_{\pm 0.07}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.02}$	$0.930_{\pm 0.01}$	$0.961_{\pm 0.03}$	$0.943_{\pm 0.01}$	$0.947_{\pm 0.01}$

Table 9. Performance Comparison with Different Pathology Knowledge Bases on the CAMELYON16 Dataset.

Knowledge Bases	4-shot			8-shot			16-shot		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
PathologyOutlines	$0.859_{\pm 0.07}$	$0.821_{\pm 0.09}$	$0.837_{\pm 0.08}$	$0.952_{\pm 0.02}$	$0.919_{\pm 0.03}$	$0.926_{\pm 0.02}$	$0.960_{\pm 0.01}$	$0.933_{\pm 0.01}$	$0.938_{\pm 0.01}$
Quilt-1M	$0.871_{\pm 0.11}$	$0.827_{\pm 0.08}$	$0.841_{\pm 0.07}$	$0.956_{\pm 0.02}$	$0.923_{\pm 0.02}$	$0.930_{\pm 0.01}$	$0.961_{\pm 0.03}$	$0.943_{\pm 0.01}$	$0.947_{\pm 0.01}$

6.4. Effects of Knowledge Base Sample Size

We investigate the impact of the knowledge refinement sample size N_C by evaluating the set $\{100, 500, 1000, 2000\}$, with results detailed in Table 6. The analysis reveals that the optimal N_C is contingent upon the volume of available supervised data. In the data-scarce 4-shot regime, performance peaks at $N_C = 1000$, suggesting a saturation point where additional concepts may yield diminishing returns. Conversely, in the 8-shot and 16-shot settings, performance scales positively with N_C, achieving optimal results at

$N_C = 2000$. This indicates that as supervision increases, the model gains a stronger capacity to effectively filter and assimilate information from a larger body of external pathological knowledge1.

6.5. Effects of LLMs for Prompt Generation

We investigate how the choice of LLM used for the offline generation of class-level descriptions affects model performance. Specifically, we evaluate the quality of the static prompt embeddings ($\boldsymbol{T}_{static}$) derived from four advanced models: DeepSeek-R1, Gemini-Pro-2.5, GPT-4.1, and Claude-3.5-Sonnet. As shown in Table 7, the reasoning capability of the LLM significantly influences few-shot outcomes. In the data-scarce 4-shot setting, prompts generated by Claude-3.5-Sonnet yield the highest AUC (0.871), followed closely by GPT-4.1 (0.861). This suggests that superior LLMs provide richer and more accurate domain-specific morphological characteristics, creating a stronger semantic prior that is critical when visual evidence is limited. While Gemini-Pro-2.5 achieves the leading performance in the 16-shot setting (0.965 AUC), the consistent advantage of high-end models in low-shot regimes underscores the importance of high-fidelity linguistic initialization.

6.6. Impact of Pathology Foundation Models

We evaluate the feature extraction capabilities of three leading pathology foundation models (QuiltNet, CONCH, and TITAN) to determine their influence on few-shot learning performance. As presented in Table 8, TITAN demonstrates a definitive lead across all experimental configurations. In the challenging 4-shot setting, TITAN achieves an AUC of 0.871, significantly outperforming both CONCH (0.812) and QuiltNet (0.789). These findings confirm that a powerful, domain-aligned feature encoder is the cornerstone of achieving superior performance in data-scarce histopathology analysis.

6.7. Effects of Pathology Knowledge Base Source

We assess the influence of the external knowledge source by benchmarking concepts derived from the large-scale Quilt-1M dataset against those from the curated Pathology Outlines. As detailed in Table 9, while Quilt-1M consistently yields the highest performance (e.g., 0.871 AUC in the 4-shot setting), the model maintains highly competitive results with Pathology Outlines (0.859 AUC). This minimal performance gap, particularly in data-rich settings, validates the framework's generalization capability and robustness to different knowledge origins.

6.8. Computational Cost of WAKI

The computational overhead of the WAKI module was evaluated on the CAMELYON16 test set (129 WSIs, Table 10). On an RTX 3080Ti, average processing time per WSI is 0.180s (<1% of feature extraction time). Faiss retrieval takes only 0.003s. For large WSIs (˜19,500 patches), total WAKI time is ˜1.1s with peak GPU memory of ˜1.5 GB, allowing deployment on standard GPUs (e.g., 8GB VRAM).

Table 10. Computational cost of WAKI (per WSI). CAMELYON16 test set.

Component	Mean Time (s)	Max Time (s)	Peak GPU Mem (MB)†
Clustering (On-the-fly)	0.177	1.095	1496
Retrieval (Concept Matching)	0.003	0.007	1499
Total WAKI Overhead	**0.180**	**1.102**	**∼1500 (1.5GB)**

† WAKI runs **one K-means per WSI** at inference/training time, using **all patches** in the slide (no subsampling). Peak GPU Mem is reported as PyTorch "reserved" memory.

References

[1] Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15590–15600, 2025. 1

[2] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1

[3] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11323–11332, 2024. 1

[4] Junjian Li, Jin Liu, Hulin Kuang, Hailin Yue, Mengshen He, and Jianxin Wang. Mico: Multiple instance learning with context-aware clustering for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 376–385. Springer, 2025. 1

[5] Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. 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*, 36:67551–67564, 2023. 1

[6] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 1

[7] Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11248–11258, 2024. 1

[8] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu,

Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11343–11352, 2024. 1

[9] Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus CF Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. *npj Digital Medicine*, 7(1):383, 2024. 1