Beyond Pixel Simulation: Pathology Image Generation via Diagnostic Semantic Tokens and Prototype Control

Supplementary Material

Appendix Contents

A. Limitations

Text–Image Alignment Performance. As shown in Table 1, UNIPATH is SOTA on every downstream metric, on Real2Gen retrieval, and in the MLLM-as-Judge comparison, yet trails Show-o2 on the single CONCH CLIP-Score (0.348 vs. 0.357). We argue that this gap reflects a bias in the evaluator–model paradigm, rather than a true semantic alignment weakness.

- **Show-o2's semantic tokens are contrastive-distilled**. During training, Show-o2 loads SigLIP weights into its Semantic Layers and minimizes a distillation loss, forcing those tokens to mimic SigLIP patch features. Although generation itself uses flow matching, the resulting high-level tokens remain "SigLIP-style" at inference.
- **The evaluator is also contrastive.** CONCH [6] is a CLIP-family model whose similarity metric is the same cosine space SigLIP is trained in. A model whose internal tokens are pre-aligned to this space naturally receives a higher CLIP score.
- **UNIPATH uses MLLM-derived pathology semantics without SigLIP distillation.** Our HLS stream extracts diagnosis tokens from a frozen Patho-R1 MLLM. Patho-R1's vision encoder is CLIP-based, but no part of the generator is forced to match CLIP/SigLIP features; the DiT learns purely via flow-matching reconstruction. Hence, its latent space is optimised for morphological fidelity, not for cosine similarity with CLIP evaluators.
- **Cross-metric consistency.** UNIPATH leads on Real2Gen retrieval (closer in feature space to real WSIs), on MLLM-as-Judge human-preference scoring, and delivers the largest Tier 4 F1 gains. These orthogonal results confirm that the small CLIP-Score gap is an artefact of evaluator homology, not of inferior text–image alignment.

In summary, the lower CONCH CLIP-Score stems from Show-o2's SigLIP-distilled tokens matching the evaluator's contrastive space, whereas UNIPATH prioritises pathology-specific morphology and semantics, which better serve real diagnostic tasks.

Dependency on Prototype Bank. One of UNIPATH's strengths comes from the component-level control provided by the Prototype Stream (PS). This advantage, however, is highly dependent on the quality and coverage of our 8K instance prototype bank. If an extremely rare morphological component is not well-represented in our 8K bank, the PS stream cannot provide precise control for that concept.

B. Ethical Statement

This research strictly adheres to the relevant ethical guidelines for medical AI research.

Data Usage and Patient Privacy. All data used in this study (TCGA and HISTAI) are publicly available datasets intended for research. All data were fully anonymized and de-identified by the original providers prior to release and contain no Protected Health Information (PHI). Our usage strictly complies with the Data Use Agreements (DUAs) for both TCGA and HISTAI.

Potential for Misuse and Mitigation. We acknowledge that high-fidelity pathological image generation (*i.e.*, "medical deepfakes") carries a potential risk of misuse, such as attempting to interfere with clinical diagnostic workflows in extreme cases. We emphasize that UNIPATH is currently intended for research purposes only, with the design goals of (i) advancing controllable generation in pathology, (ii) providing controllable data augmentation for computational pathology, and (iii) serving as an educational tool. This model **must not** be used for any direct clinical diagnosis.

Algorithmic Bias. Our model's performance relies on the quality and distribution of our training data. Despite our efforts to add diversity (1.03M HISTAI) and balance our 68K subset (via K-means and elite sampling), our training data may still contain undiscovered biases (*e.g.*, in demographic representation across race, age, or sex). The model may learn and amplify these biases. Future work is required to specifically quantify and mitigate such biases.

C. Future Work

While UNIPATH marks significant progress in unifying pathology understanding and controllable synthesis, several key directions remain for future exploration.

Support for Higher Resolution and Broader Histological Context. The current UNIPATH model primarily operates on 384×384 pixel patches. While sufficient for capturing cell-level morphological features, this limits the model's ability to understand and generate larger-scale architectural patterns, such as complex glandular structures or tumor-stroma interactions. Future work should explore extending UNIPATH to higher resolutions or integrating a larger field of view, enabling the generation of images that are more histologically context-aware.

Controllable Pathological Image Editing. UNIPATH currently focuses on image generation from text prompts. A high-impact extension is to enable fine-grained editing of existing real pathology images. This can be framed as a "counterfactual synthesis" task — such as "adding moderate nuclear atypia" to a benign tissue image or "removing the specified inflammatory infiltrate." The MSC architecture of UNIPATH provides an ideal framework for this: the HLS stream could parse the editing instruction (*e.g.*, "increase mitotic figures"), while the PS stream could retrieve and inject the corresponding morphological prototypes to achieve the precise, localized modification.

Scaling the Prototype Bank. Our prototype-based control mechanism opens a promising avenue for future enhancement. While the curated 8K bank establishes the efficacy of this approach, we can further enhance the model's generative "vocabulary" by scaling this bank. Future work could explore using active learning or self-supervised methods to automatically mine and cluster novel, informative prototypes from large-scale, unlabeled datasets. This expansion would enable UNIPATH to synthesize an even greater diversity of morphological features with high precision, particularly for rare, long-tail pathological phenomena.

D. Detailed Implementation Details

D.1. Model Architecture Specifications

Generation Backbone (DiT) and Conditioning. Our generation backbone is a 0.6B-parameter DiT (Diffusion Transformer) designed in the PixArt [2] style. It comprises 28 Transformer layers, 16 attention heads, and a hidden dimension $d = 1152$. Our model employs a hybrid conditioning mechanism. The fused conditional vector C_{comp} is injected into every DiT layer via traditional cross-attention. In contrast, the timestep is handled separately, injected via the AdaLayerNorm-Single (AdaLN-S) mechanism to perform conditional normalization.

Understanding Backbone and VAE. Our backbone is based on the Patho-R1 (7B) [16] model, which is post-trained on pathology domain data from Qwen2.5-VL 7B [1]. The backbone remains fully frozen throughout all training stages. The VAE employed is the Stable Diffusion 3 [5] VAE, featuring an 8x downsampling factor.

MSC Module Implementation. In the Multi-Stream Control (MSC) module, the projection layers for the three streams (HLS, RTS, and PS), such as $\mathrm{MLP}_{\mathrm{DST}}$, are implemented as separate 2-layer Feed-Forward Networks (FFNs) with unshared weights. Each of these MLPs follows the same architecture: a linear projection from the input dimension to the hidden size, followed by a GELU activation, and a second linear projection from the hidden size back to the hidden size. The hidden size is 1152 (matching the DiT's hidden dimension). For the Prototype Stream (PS) retrieval, we provide the specific hyperparameters used for the hybrid strategy. The total number of prototypes is $K_m = 16$. For the Global Semantic Retrieval (U_g, Eq. 4), we set the retrieval tops $k_t = 4$ (Text) and $k_v = 4$ (Vision). For the Local Fine-grained Retrieval (U_l), we parse the four rarest keywords from the prompt and randomly sample 2 prototypes for each term, resulting in 8 local prototypes. The final set $\hat{U}$ is the union of U_g and U_l, clipped to 16 (Eq. 5).

Flow Matching Training and Inference. During the training stage, we adopt a Rectified Flow strategy. Specifically, we first sample $u \sim \mathcal{U}(0, 1)$, map it to timesteps and sigmas, and construct the noise-interpolated z_t accordingly. The model v_θ is trained to predict the target velocity $v_t = z_0 - z_1$. During the inference stage, we use the Euler solver with 30 function evaluations. We employ Classifier-Free Guidance with a guidance_scale of 3.0.

D.2. Training Hyperparameters

General Setup. Across both training stages, we used the AdamW optimizer with default betas ($\beta_1 = 0.9, \beta_2 = 0.999$), an epsilon of $1e^{-8}$, and a weight decay of 0.01. All training was conducted using mixed precision. All experiments were conducted on 16 NVIDIA H100 GPUs. All input images were processed to a resolution of 384×384.

Stage 1: Semantic Alignment (Pre-training). The model was pre-trained on 2.58M text-image pairs (excluding the 68K subset) for 10,000 steps using the Flow Matching (MSE) loss. We used a global batch size of 512. The learning rate was linearly warmed up for the first 2% of steps (200 steps) to a peak of 1e-4. It was then decayed using a cosine scheduler with a minimum learning rate of $1e^{-5}$.

Stage 2: High-Quality Fine-tuning. The model was subsequently fine-tuned on the 50K high-quality subset for 500 steps. We used a global batch size of 512 and a fixed learning rate of 2e-5 for this entire stage.

E. Dataset Detailed Analysis

E.1. Statistics of the 1.03M Corpus

Word Count. We analyzed the caption-length distributions of the 1.03M corpus before and after summarization using Qwen3-8B, as shown on the left side of Figure S1. The distributions exhibit unimodal and symmetric curves. In the original captions, the most frequent length is around 120 words, accounting for 9.6%. After refinement, the peak shifts to approximately 35 words, accounting for 14.9%, whereas captions longer than 60 words are virtually absent.

Word Frequency. We analyzed the word-frequency profiles of captions before and after cleaning the 1.03M corpus, which are presented as word clouds in the right panel of Figure S1. The word clouds indicate that the captions emphasize microscopic morphological features such as "cells," "nuclei," and "stroma," as well as diagnostic descriptors including "inflammatory" and "stained." A comparison of the two word clouds shows an increased prevalence of morphology-related terms and a marked reduction in non-informative tokens such as "which" and "image."

E.2. Analysis of the 68K Refined Subset

Word Count. We analyzed the caption lengths in the 68K Refined Subset, as shown on the left of the Figure S2. The distribution is symmetric and unimodal, with the most frequent length around 47 words, accounting for approximately 19%. Compared with the 1.03M Corpus, captions in the 68K subset are more extended, rarely shorter than 35 or longer than 55 words, due to the prompt-imposed 30–60-word constraint. This moderate length reduces redundancy while ensuring sufficient content to accurately describe the images, enabling the model to learn a broader range of knowledge.

Word Frequency. We also analyzed the word-frequency distribution of captions in the 68K Refined Subset, visualized as word clouds on the right side of Figure S2. The three subsets (8K, 10K, and 50K) exhibit a highly coherent vocabulary profile dominated by morphological, nuclear, cytoplasmic, stromal, and diagnostic descriptors. Unlike the 1M Corpus, where terms such as "nuclei," "cells," and "stroma" overwhelmingly dominate, the 68K Refined Subset exhibits a more balanced distribution of key pathological concepts. The captions in this refined subset are more detailed and make use of a richer and more uniformly distributed set of domain-specific terms, thereby providing higher-quality supervision that is advantageous for model evaluation and fine-tuning.

E.3. Spot-Check Validation of the 10K Test Set

To definitively validate the reliability of the "Gemini-2.5 Pro generation then GPT-5 review" automated pipeline, we additionally invited a domain-expert pathologist to conduct an independent spot-check quality control (QC) on 500 random samples from our 10K high-quality test set.

The reviewer's task was to assign each image-text pair to one of three categories. **3: Excellent** was defined as: The description is accurate, comprehensive, and professional, perfectly corresponding to all key pathological features in the image (Gold Standard). **2: Acceptable** was defined as: The description captures the main diagnostic features without factual errors, but may contain minor deficiencies, such as omitting a secondary feature, slight imprecision in non-critical terminology, or a minor deviation in descriptive focus (Still Usable for Evaluation). **1: Unusable** was defined as: The description contains severe factual errors, rendering it unsuitable as an evaluation benchmark, such as hallucinating key features not present in the image, misidentifying the primary cell type, or completely omitting the main diagnostic point of the image (Failure).

Upon reviewing the 500 random samples, the pathologist's evaluation was as follows: 43.4% of the image-text pairs were rated as 3: Excellent; 50.2% were rated as 2: Acceptable; and 6.4% were rated as 1: Unusable. This results in an Overall Usability Rate (*i.e.*, Excellent + Acceptable) of 93.6%. This extremely low "Unusable" rate (6.4%) strongly confirms the SOTA reliability of the automated data annotation pipeline we employed.

E.4. 1231-term Pathology Vocabulary

Our 1231-term pathology vocabulary, which was used to build the inverted index $\mathcal{I}$, is provided as a separate file ("vocabulary.txt") in the supplementary material bundle.

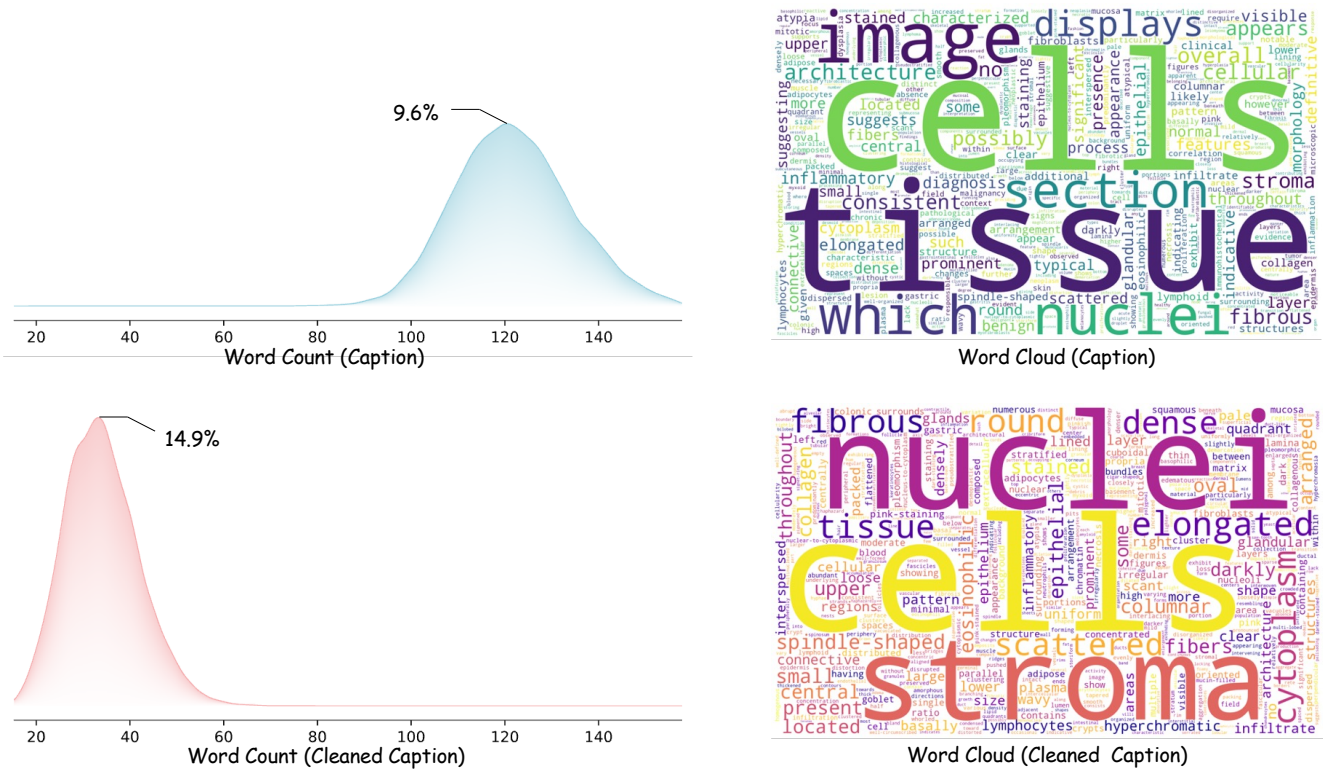


Figure S1. Visualization of the caption-length distribution generated by PathGen-LLaVa (top left) and its corresponding word-frequency cloud (top right), as well as the caption-length distribution after summarization by Qwen3-8B (bottom left) and the associated word-frequency cloud (bottom right).

Table S1. Overall results of models on the PathMMU **test set**. The best-performing MLLM in each subset for general and medical/pathology MLLM is **in-bold**, and the second-best performing MLLM is underlined. An asterisk (∗) indicates results copied from the original source for the non–open-sourced model.

	Unif. Model	Test Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
		Tiny (1131)	ALL (9483)	Tiny (281)	ALL (3068)	Tiny (210)	All (1661)	Tiny (255)	All (1938)	Tiny (208)	ALL (1007)	Tiny (177)	ALL (1809)
Expert Performance	-	71.8	-	72.9	-	71.5	-	69.0	-	68.3	-	78.9	-
General Multimodal Large Language Models													
LLaVA-1.5-13B	✘	39.1	37.7	45.2	41.4	39.6	40.1	35.0	39.9	46.5	43.4	25.9	23.7
LLaVA-OneVision-7B	✘	36.8	34.0	38.2	36.9	34.9	37.7	46.4	37.4	39.6	38.8	20.0	19.5
Qwen3-VL-8B-Instruct	✘	54.9	51.5	59.1	54.4	63.3	56.3	56.1	53.7	52.4	55.1	39.5	37.9
Qwen3-VL-30B-A3B-Instruct	✘	57.6	54.5	64.4	57.2	67.6	65.8	58.0	56.1	56.3	58.3	35.6	35.7
BLIP3o-8B	✔	40.0	37.6	40.5	40.6	42.6	39.2	47.5	42.3	40.7	38.7	24.7	25.3
Show-o2-7B	✔	40.8	40.4	43.7	48.8	43.8	40.4	42.4	41.9	42.8	40.4	28.2	24.4
BAGEL-14B	✔	57.3	52.3	63.0	56.2	61.4	59.6	60.8	54.1	64.4	61.3	29.9	31.9
GPT-4V-1106	-	53.8	49.9	59.7	53.9	58.5	53.4	60.6	53.9	47.6	52.6	36.6	33.9
Gemini-2.5 Pro	-	**69.0**	**68.0**	**74.5**	**71.8**	**72.3**	**68.9**	**72.4**	69.8	66.8	69.5	53.9	58.1
Medical / Pathology-specific Multimodal Large Language Models													
LLaVA-Med	✘	25.5	26.8	29.2	28.5	29.8	28.2	23.3	27.4	21.7	30.4	22.0	20.1
Quilt-LLaVA	✘	45.4	41.2	46.8	41.9	46.3	45.9	51.0	45.0	46.5	43.7	32.8	30.1
PathGen-LLaVA	✘	59.8	58.4	59.3	59.9	60.2	58.7	60.1	60.5	64.3	65.1	54.4	49.7
CPath-Omni*	✘	-	-	74.0	69.9	-	-	69.8	70.6	65.9	70.6	**75.7**	**79.0**
UniMedVL	✔	54.5	50.6	58.0	54.9	57.1	56.7	55.7	56.0	66.3	59.4	30.5	26.9
UNIPATH (Ours)	✔	68.3	65.7	72.9	66.4	67.9	68.4	70.1	**73.9**	**79.2**	**77.7**	46.1	46.6

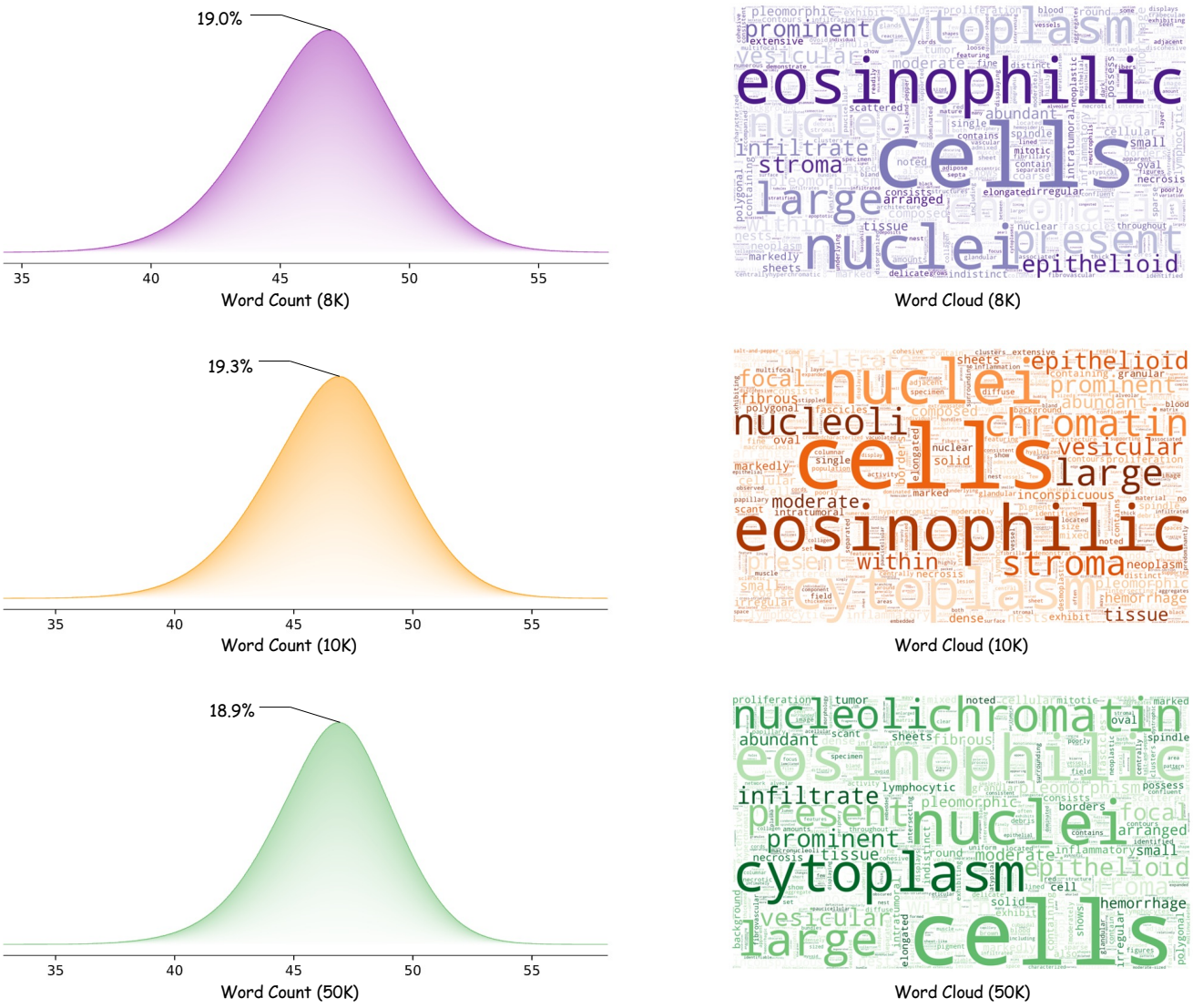


Figure S2. Visualization of the caption-length distributions and word-frequency clouds for the 8K, 10K, and 50K refined subsets (8K: top row; 10K: middle row; 50K: bottom row), with length distributions shown on the left and word-frequency clouds on the right.

E.5. Analysis of Data Leakage Risks

To ensure the integrity of our evaluation splits and proactively address potential concerns regarding overlap between the 8K Prototype Bank and the 10K Test Set, we conducted a rigorous validation. We computed the exhaustive pairwise visual feature similarity ($8,000 \times 10,000 = 80,000,000$ comparisons) between all images in the bank and all images in the test set. For this check, we used the UNI2-h extractor [4], which is the same high-performance backbone we employed for dataset-wide de-duplication in Section 3.1. The statistics confirmed that the sets are strictly disjoint: the average similarity was 0.1358 (standard deviation = 0.0686). Critically, the maximum similarity observed across all 80 million pairs was 0.9416. This maximum value (0.9416) is below our defined de-duplication threshold of 0.95 (as detailed in Section 3.1). This result strongly confirms that our prototype bank and test set are strictly disjoint. It therefore eliminates any risk of data leakage via the Prototype Stream (PS) retrieval, validating the integrity of our Tier 2 (Alignment) and Tier 3 (Control) evaluations.

F. Supplementary Quantitative Results

F.1. Full Understanding Capability

To validate the diagnostic understanding capability of UNIPATH, which is critical for our Multi-Stream Control (MSC) module, we evaluated its performance on the com-

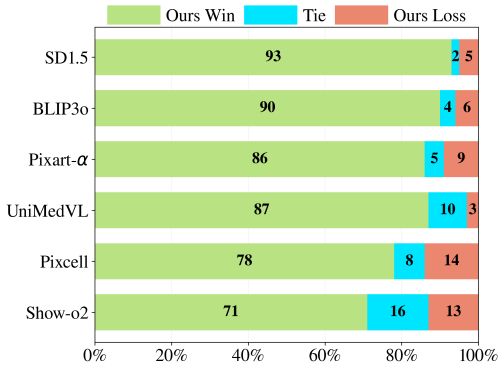


Figure S3. Gemini 2.5 Pro as Judge.

prehensive PathMMU benchmark [10]. The full results are presented in Table 3. As shown in the table, UNIPATH achieves an overall score of 65.7 on the full test set. This performance establishes UNIPATH as the SOTA among all evaluated open-source models, substantially outperforming other leading open-source pathology MLLMs, including PathGen-LLaVA (58.4) and the unified model UniMedVL (50.6). Furthermore, UNIPATH achieves the top score across all models (including closed-source systems) on the EduContent (73.9) and Atlas (77.7) sub-tasks. Its overall score also closely approaches that of top-tier closed-source models, such as Gemini-2.5 Pro (68.0). This strong understanding performance confirms that our frozen MLLM backbone provides the robust, phrasing-invariant semantics necessary to steer controllable generation.

F.2. Additional Fidelity & Alignment Results

In our main evaluation (Section 5), the text-image alignment metrics and Patho-FID/KID metrics were based on the CONCH and UNI2-h backbones, respectively. To further validate the robustness and generality of our findings, we conducted an additional evaluation using two entirely independent, external backbones not used anywhere in our model pipeline: Virchow2 [11] for Visual Fidelity and MUSK [12] for Text-Image Alignment. This analysis confirms that our model's superior performance is a genuine advantage and not an artifact of a specific evaluator. The results are shown in Table S2.

Visual Fidelity (Virchow2 Backbone). The results using the Virchow2 feature extractor strongly reinforce our main findings. UNIPATH achieves a FID of 484.38 and a KID of 0.192. This performance is not just SOTA, but represents a massive improvement over the next-best model, Pixcell (FID: 929.30, KID: 0.259). This confirms that the superior visual fidelity of UNIPATH is a genuine model advantage, not an artifact of the UNI2-h evaluator.

Text-Image Alignment (MUSK Backbone). The alignment metrics using the MUSK backbone provide a crucial, unbiased perspective.

- **CLIP-Score & T2I Retrieval:** Using MUSK, **Show-o2** achieves the highest CLIP-Score (0.545) and the best Report-Gen (T2I) retrieval metrics. This result is consistent with our main paper's findings (Table 1) and supports our hypothesis that Show-o2's architecture is well-optimized for general-purpose T2I alignment.
- **I2I Retrieval:** Critically, on the Real-Gen (I2I) retrieval task—which measures how close the generated images are to real images in this new feature space—UNIPATH achieves dominant SOTA performance across all four metrics (Recall@10/50, mAP@10/50). For instance, our mAP@10 (6.38) is nearly double that of the second-best Show-o2 (3.22).

Conclusion. These results, obtained from fully independent feature extractors, strongly validate our conclusions. Our model's superior visual fidelity and its ability to generate images that are most faithful to the real pathology manifold are thus demonstrated as robust and general findings, independent of the specific evaluator used.

F.3. Detailed Few-Shot Classification Results

We provide the detailed numerical results for the Tier 4: Downstream Task Utility evaluation (Kather-CRC-2016 few-shot classification) in Table S3. This table contains the precise Mean±Std (Weighted F1) scores and the absolute change (Δ) for all K-shot values (K=2, 4, 8, 16, 32). These are the raw data used to generate Figure 4 in the main text.

F.4. MLLM and Human Judge: Setup & Reliability

Human Expert Evaluation. Here, we detail the implementation and reliability analysis of our human expert evaluation. We employed a panel of three trained annotators to conduct a blind pairwise comparison (UNIPATH vs. Baseline) on 500 image-text pairs randomly sampled from the 10K test set. During the evaluation, annotators were shown a text prompt and two anonymized images and were tasked with choosing which image better matched the prompt, without knowing which image was generated by UNIPATH. To validate the reliability of this evaluation, we measured the inter-annotator agreement. As shown in the analysis output, the panel achieved an overall Fleiss' Kappa of 0.7509, indicating "substantial" agreement. The per-model agreement was also robust, ranging from "Moderate" to "Almost Perfect" (UniMedVL: $\kappa = 0.8833$; show-o2: $\kappa = 0.7998$; Pixcell: $\kappa = 0.7950$; PixArt: $\kappa = 0.6502$; SD15: $\kappa = 0.5708$; BLIP3o: $\kappa = 0.5672$). The aggregated win/loss/tie statistics from this reliable panel were used to generate the human expert results in the main paper.

Gemini-2.5 Pro as Judge. In the main paper, we presented the "as-Judge" results from GPT-5 and the human expert panel. For completeness, we provide a parallel evaluation using Gemini-2.5 Pro as the judge, following the exact same experimental setup. The results are presented

Table S2. Quantitative comparison of Visual Fidelity and Text–Image Alignment with merged T2I/I2I. FID/KID uses the Virchow2 extractor; Similarity and retrieval metrics use MUSK. ★ marks models fully fine-tuned on our large dataset. The best is **bold**, the second best is underlined.

	Unif. Model	Visual Fidelity ↓		Text–Image Alignment ↑				
		FID	KID	Sim.	Recall@10	Recall@50	mAP@10	mAP@50
Real Data	-	-	-	0.557	13.40/–	33.50/–	5.28/–	6.17/–
General Text to Image Generation Models								
SD1.5★ [9]	✘	1804.69	0.519	0.483	0.66/0.60	1.91/1.94	0.80/0.80	0.85/0.91
SDXL★ [8]	✘	2570.19	0.602	0.445	0.34/0.22	1.20/0.80	0.60/0.38	0.77/0.49
Pixart-α★ [2]	✘	2574.85	0.685	0.482	1.14/0.60	4.54/2.65	1.58/0.72	1.74/0.93
BLIP3o★ [3]	✔	2008.75	0.550	0.455	1.55/1.50	5.87/5.80	2.39/1.93	2.58/2.26
Show-o2★ [13]	✔	1398.52	0.415	**0.545**	**8.41**/2.71	**22.77**/9.74	**10.93**/3.22	**9.42**/3.11
Pathological / Medical Text to Image Generation Models								
Pixcell [15]	✘	929.30	0.259	0.524	-	-	-	-
PathLDM [14]	✘	1126.36	0.376	0.483	0.15/0.17	0.73/0.70	0.18/0.25	0.30/0.36
UniMedVL [7]	✔	1435.07	0.363	0.520	3.82/2.23	11.52/7.18	5.31/2.59	5.14/2.74
UNIPATH (Ours)	✔	**484.38**	**0.192**	0.538	7.55/**4.25**	21.61/**13.72**	10.22/**6.38**	8.93/**6.07**

Table S3. Few-shot downstream performance (Weighted F1) across different shots K. Values are Mean±Std (%); "Δ" *columns* report absolute change vs. original data in percentage points. Best is **bold**, second best is underlined.

Model	K = 2		K = 4		K = 8		K = 16		K = 32	
	Wgt. F1	Δ	Wgt. F1	Δ	Wgt. F1	Δ	Wgt. F1	Δ	Wgt. F1	Δ
Baselines (Only Real Data)										
Original Data	$67.34_{\pm 2.37}$	–	$76.42_{\pm 3.31}$	–	$81.43_{\pm 1.88}$	–	$83.85_{\pm 1.48}$	–	$86.88_{\pm 0.88}$	–
UNIPATH (Ours)	$50.31_{\pm 8.92}$	–	$69.33_{\pm 2.81}$	–	$76.89_{\pm 2.35}$	–	$78.30_{\pm 0.99}$	–	$81.05_{\pm 0.77}$	–
Data Augmented Comparisons (Real Data with Generated Data)										
SD1.5 [9]	$60.81_{\pm 5.31}$	−6.53	$71.14_{\pm 4.28}$	−5.28	$77.03_{\pm 2.17}$	−4.40	$81.94_{\pm 2.94}$	−1.91	$85.35_{\pm 1.13}$	−1.53
UniMedVL [7]	$63.86_{\pm 4.54}$	−3.48	$74.94_{\pm 3.04}$	−1.48	$80.69_{\pm 2.03}$	−0.74	$83.11_{\pm 1.11}$	−0.74	$86.75_{\pm 1.27}$	−0.13
Pixcell [15]	$67.06_{\pm 3.18}$	−0.28	$77.13_{\pm 2.16}$	+0.71	$81.00_{\pm 1.99}$	−0.43	$84.19_{\pm 1.18}$	+0.34	$86.93_{\pm 0.51}$	+0.05
Show-o2 [13]	$68.68_{\pm 5.31}$	+1.34	$77.70_{\pm 4.30}$	+1.28	$81.76_{\pm 2.35}$	+0.33	$84.51_{\pm 1.78}$	+0.66	$86.97_{\pm 1.04}$	+0.09
UNIPATH (Ours)	$\mathbf{69.65}_{\pm 4.35}$	+2.31	$\mathbf{79.14}_{\pm 1.82}$	+2.72	$\mathbf{82.22}_{\pm 1.10}$	+0.79	$\mathbf{85.39}_{\pm 1.37}$	+1.54	$\mathbf{87.15}_{\pm 0.61}$	+0.27

in Figure S3. As shown, Gemini-2.5 Pro's assessment is highly consistent with our other evaluations. It demonstrates a clear preference for UNIPATH against all baselines, preferring UNIPATH over the strongest baseline (Show-o2) in 71% of cases. This additional MLLM evaluation further corroborates our model's robust advantage in nuanced, human-aligned semantic understanding.

F.5. MSC Sensitivity and Ablation Studies

We evaluate the inference performance of the Prototype Stream (PS) using Patho-FID and CONCH CLIP-Score. Unlike the fixed parameters of the High-Level Semantics (HLS) stream, the PS architecture allows for dynamic adjustments to the prototype bank size (K_m) and retrieval strategies at inference without retraining.

Sensitivity to Prototype Quantity (K_m). We analyzed the trade-off between context sufficiency and information density by varying the prototype count $K_m \in \{0, 4, 8, 16, 32\}$ on the 10K Test Set. Throughout these experiments, we maintained a consistent allocation strategy: $K_m/4$ for global text retrieval, $K_m/4$ for global vision re-

Table S4. **Ablation on Retrieval Components.** Combining Global Hybrid retrieval with Local Sparse retrieval achieves the best trade-off, minimizing Patho-FID while maximizing CLIP-Score.

Retrieval Configuration	Patho-FID ↓	CLIP-Score ↑
Global (Text-Only)	87.52	0.327
Global (Vision-Only)	83.10	0.336
Global (Hybrid)	82.04	0.342
Local	91.45	0.325
Global + Local (Ours)	**80.86**	**0.348**

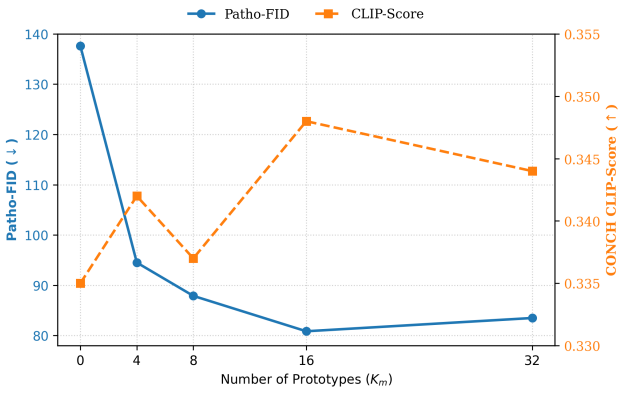


Figure S4. **Inference-time sensitivity of prototype quantity (K_m).** Our default $K_m = 16$ achieves the optimal trade-off between visual fidelity (Patho-FID, ↓) and semantic alignment (CLIP-Score, ↑).

trieval, and $K_m/2$ for local fine-grained retrieval (assigning 2 prototypes to each of the top $K_m/4$ parsed keywords). As illustrated in Figure S4, the baseline without prototype guidance ($K_m = 0$) exhibits significantly inferior fidelity and alignment, validating the necessity of the PS stream. Increasing K_m yields rapid gains up to $K_m = 16$, at which point our model achieves the optimal balance. Beyond this point, retrieving lower-ranked prototypes ($K_m = 32$) introduces irrelevant noise, diluting the conditioning signal and slightly degrading Patho-FID.

Ablation on Retrieval Components. We dissected the impact of our retrieval modules as reported in Table S4. Within the Global module, the Hybrid strategy (82.04 FID, 0.342 CLIP) consistently outperforms single-modality baselines, bridging the gap between Text-Only (87.52 FID) and Vision-Only (83.10 FID) retrieval. We also observe that relying solely on Local sparse retrieval yields the poorest fidelity (91.45 FID), indicating that sparse keywords alone lack sufficient generative context. However, the integration of Global and Local modules is transformative; the Full Strategy achieves the best overall performance (80.86 FID, 0.348 CLIP), confirming that fine-grained sparse guidance complements dense global context to maximize both visual fidelity and semantic alignment.

G. Supplementary Qualitative Results

G.1. PS Stream: Component-level Control

To visually validate the efficacy of our Prototype Stream (PS) in achieving component-level morphological control, we provide qualitative examples of its internal retrieval mechanism in Figure S5. As described in Section 4.2, our PS employs a hybrid retrieval strategy that combines Global Semantic Retrieval with Local Fine-grained Retrieval to capture both the holistic context and the specific morphological components of a prompt. Taking Figure S5a as an example, we illustrate the complete process for a complex "Generation Instruction."

- **Inputs and Generation:** The top-left panel shows the complex multi-part prompt, the original "Ground Truth" image, and our final "Generation Image." The generated image successfully synthesizes all specified pathological features, including "solid sheets," "marked pleomorphism," and "extensive hemorrhage," demonstrating high visual fidelity to the ground truth.
- **Global Semantic Retrieval:** The top-right panel shows the prototypes retrieved by the global strategy (both Text and Vision Feature Retrieval). These images capture the holistic gist or overall appearance of the prompt — such as the general pink/purple "H&E" color profile, high cellularity, and areas of hemorrhage.
- **Local Fine-grained Retrieval:** The bottom panel provides direct evidence of component-level control. Here, the prompt is parsed into specific keywords (*e.g.*, "arranged in solid sheets," "marked pleomorphism," "irregular contours," "extensive hemorrhage"). The inverted index ($\mathcal{I}$) then recalls prototypes that specifically and accurately match each individual component. For example, the prototypes for "extensive hemorrhage" are almost exclusively composed of red blood cells, while the prototypes for "marked pleomorphism" correctly show cells with high nuclear variation.

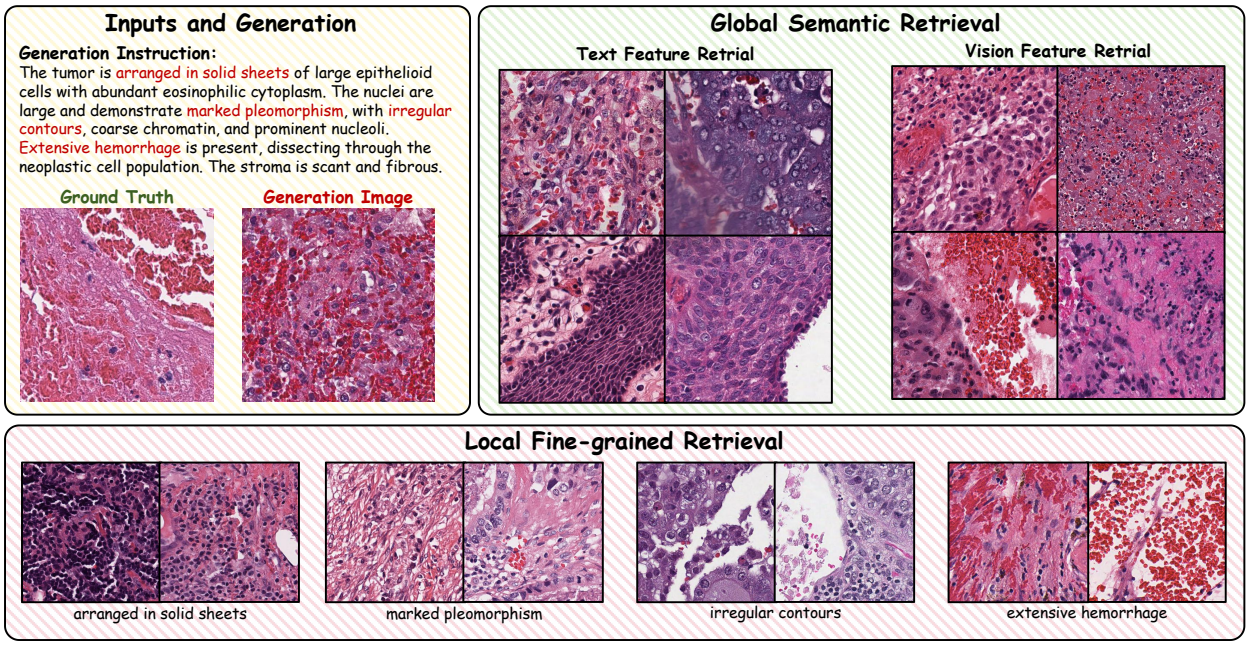
This visualization confirms that UNIPATH steers generation by combining these two complementary sets of prototypes, allowing it to render complex scenes with precise control over individual pathological components.

G.2. SOTA Model Comparisons

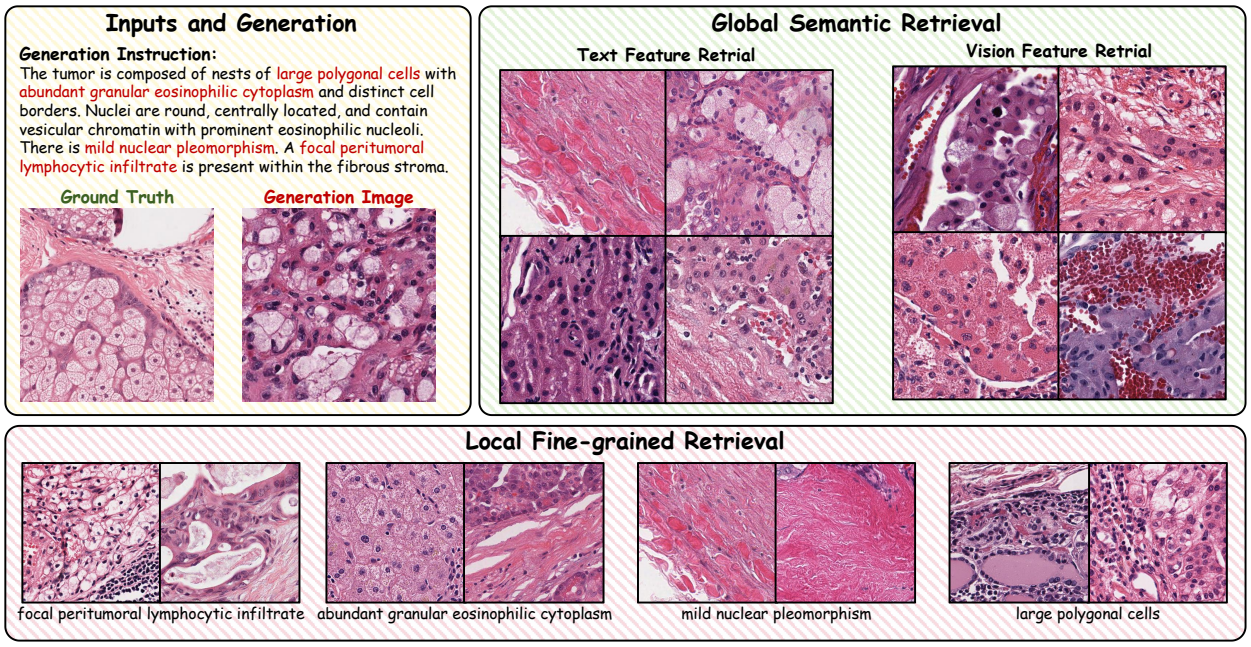
To complement the examples presented in the main text, Figure S6 presents additional qualitative comparison sets covering a broader range of prompts. Consistent with the observations in the main paper, baseline methods frequently exhibit partial concept omission, morphological inconsistencies, or visually implausible artifacts when handling prompts containing multiple fine-grained pathological attributes. In contrast, UNIPATH systematically preserves the entirety of the described features and renders them with higher morphological fidelity. These visual examples provide a more faithful demonstration of UNIPATH's performance, capturing semantic and morphological details that automated metrics such as CLIP-Score fail to reflect fully.

G.3. Gallery of Randomly Sampled Generations

To offer a complementary viewpoint on model behavior, Figure S7 centers solely on the visual quality of images generated by UNIPATH. We present a diverse set of sampled test-set cases, each paired with its corresponding Ground Truth image, spanning a broad spectrum of histopathologi-



(a) Case 1.



(b) Case 2.

Figure S5. Visualization of the UNIPATH Prototype Stream (PS) hybrid retrieval mechanism. Both (a) and (b) illustrate how component-level control is achieved by combining Global Semantic Retrieval (top right) and Local Fine-grained Retrieval (bottom).

cal appearances such as epithelial structures, adipose tissue, smooth or skeletal muscle, collagenous stroma, and inflammatory infiltrates. Across these diverse cases, the side-by-side comparison highlights that UNIPATH consistently produces images with high visual fidelity, well-preserved fine-grained morphological details, and realistic tissue textures, without introducing implausible artifacts. These qualitative examples provide a direct and intuitive assessment of generative realism that complements automated metrics such as FID, offering a more faithful reflection of the model's practical visual reliability.

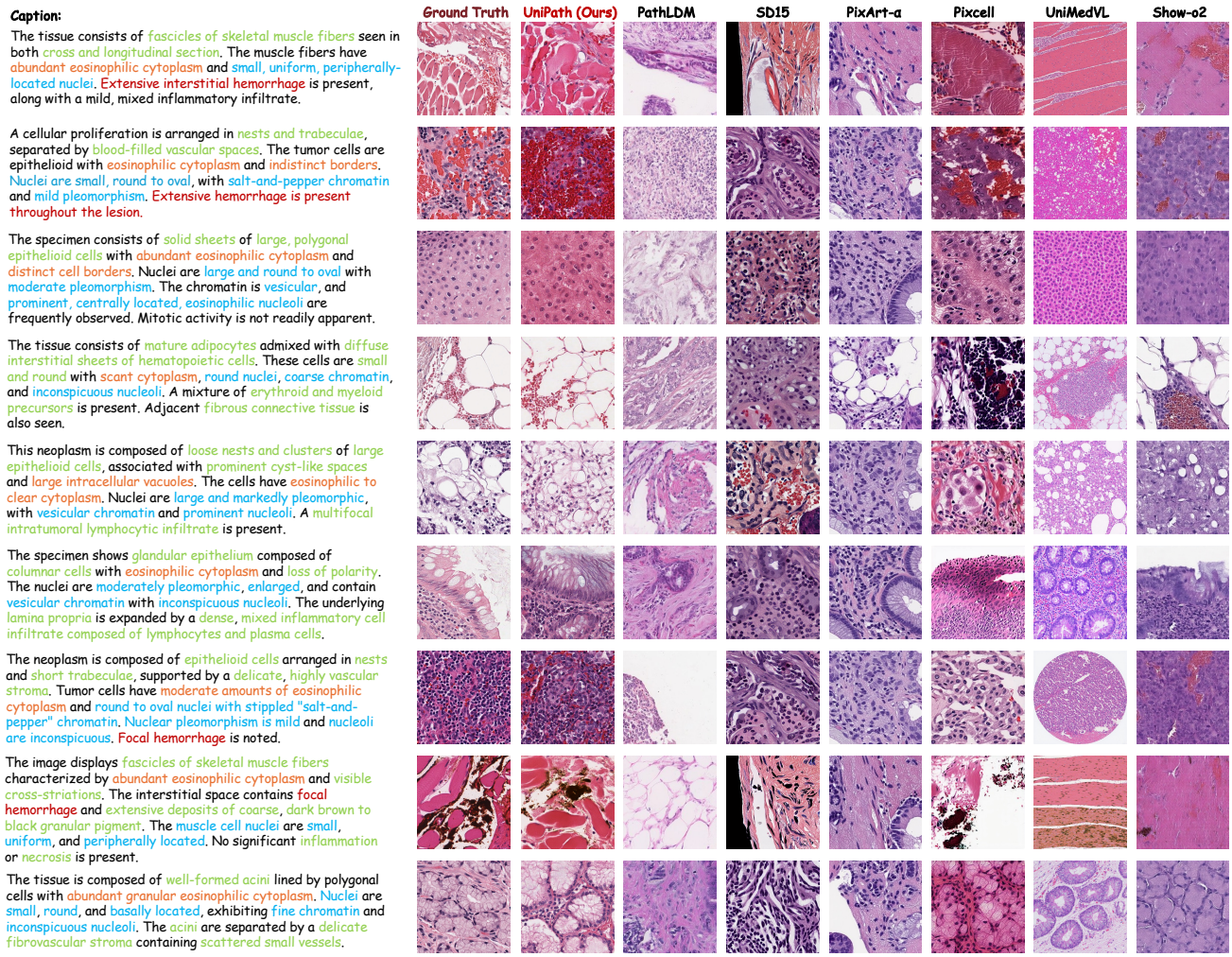


Figure S6. Comparison of pathology image generation results across UNIPATH, PathLDM, SD15, PixArt-α, Pixcell, UniMedVL, and Show-o2 under different input captions. Colors in the captions denote distinct pathological features: Tissue/Cell Type, Nuclear Features, Cytoplasm, Hemorrhage.

H. Prompt Engineering Details

This section presents several prompts used during dataset construction and evaluation. Specifically, **(i)** The instructions for generating initial captions for the 68K Refined Subset using Gemini, as well as the cross-validation prompts utilized by GPT-5 to independently review quality and factual accuracy, shown in Figure S8 (Gemini-2.5 Pro) and Figure S9 (GPT-5); **(ii)** The prompts used to comprehensively assess generation quality with MLLMs serving as judges, shown in Figure S10; **(iii)** The prompts used to filter pathology terminology with an LLM, shown in Figure S11.

H.1. Prompts for Re-annotation

For the construction of the 68K Refined Subset, we employed a two-stage prompt pipeline consisting of a Gemini-2.5 Pro-based caption generator and a GPT-5–based cross-modal validator. The entire process consumed 300M tokens, including both the input and output tokens.

Stage 1: Captioning with Gemini-2.5 Pro. The Gemini prompt instructs the model to inspect each H&E ROI and produce a structured JSON output containing Lite-schema labels, along with a 30–60-word morphological description, without any diagnosis. The prompt provides explicit enumeration rules (*e.g.*, nuclear size, pleomorphism, stromal reaction, types of inflammation) and a style specification that requires objective, declarative wording while forbidding diagnostic terms, negative-absence phrasing, or modality-related metadata. This design ensures that the initial captions remain focused on morphology, remain consistent, and can be used in later-generation tasks.

Stage 2: Verification with GPT-5. The second-stage

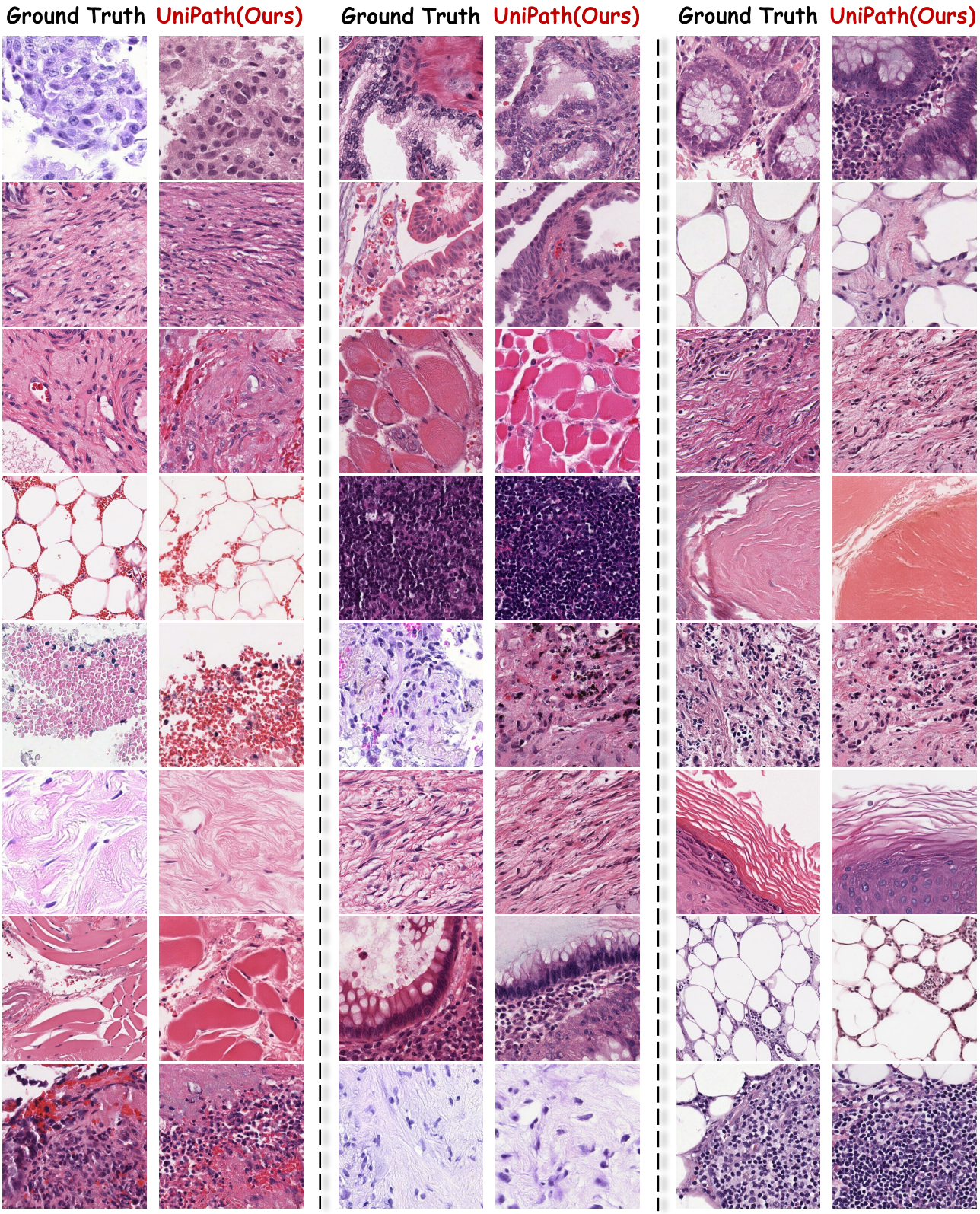


Figure S7. Comparative visualization of Ground-Truth and the corresponding pathology images synthesized by our model.

GPT-5 prompt performs strict factual verification of Gemini's output. It checks whether each label is visually supported at the given magnification and flags any ambiguous or contradicted features as errors. In addition to visual factuality, it enforces mandatory textual rules (*e.g.*, no diagnosis, correct style) and performs minimal schema checks to ensure alignment with the predefined JSON format. This cross-modal validator effectively removes inconsistent or hallucinated descriptions, making sure that only high-quality annotations enter the final dataset.

H.2. Prompts for MLLM-as-Judge

To systematically evaluate the alignment between generated histopathology images and textual descriptions, we designed a cross-modal evaluation prompt. The prompt enables comparison between the images generated by UNIPATH and those of other baseline models for a given caption. Its functionality includes identifying visual features in the caption, assessing whether each image feature is supported or contradicted, and producing a quantitative alignment judgment (WIN / TIE / LOSS) based on a predefined hierarchy of histological features (*e.g.*, Architecture, Cytology, Nuclear features).

H.3. Prompts for Pathology Vocabulary Filtering

We also designed a rule-driven prompt to curate pathology feature phrases for downstream modeling. The prompt enables a model to evaluate a list of short text phrases, acting as a board-certified anatomic pathologist and computational pathology NLP expert. Its functionality includes determining whether each phrase represents a complete and discriminative histopathologic feature according to a precision-first hierarchy (*e.g.*, nuclear, cytoplasmic, cellular lineage, architectural, cytologic atypia, qualified inflammatory or hemorrhagic features). The output preserves the original input order and formatting, is strictly plain text, and contains no additional explanation or modification, ensuring reproducibility and direct applicability for downstream modeling.

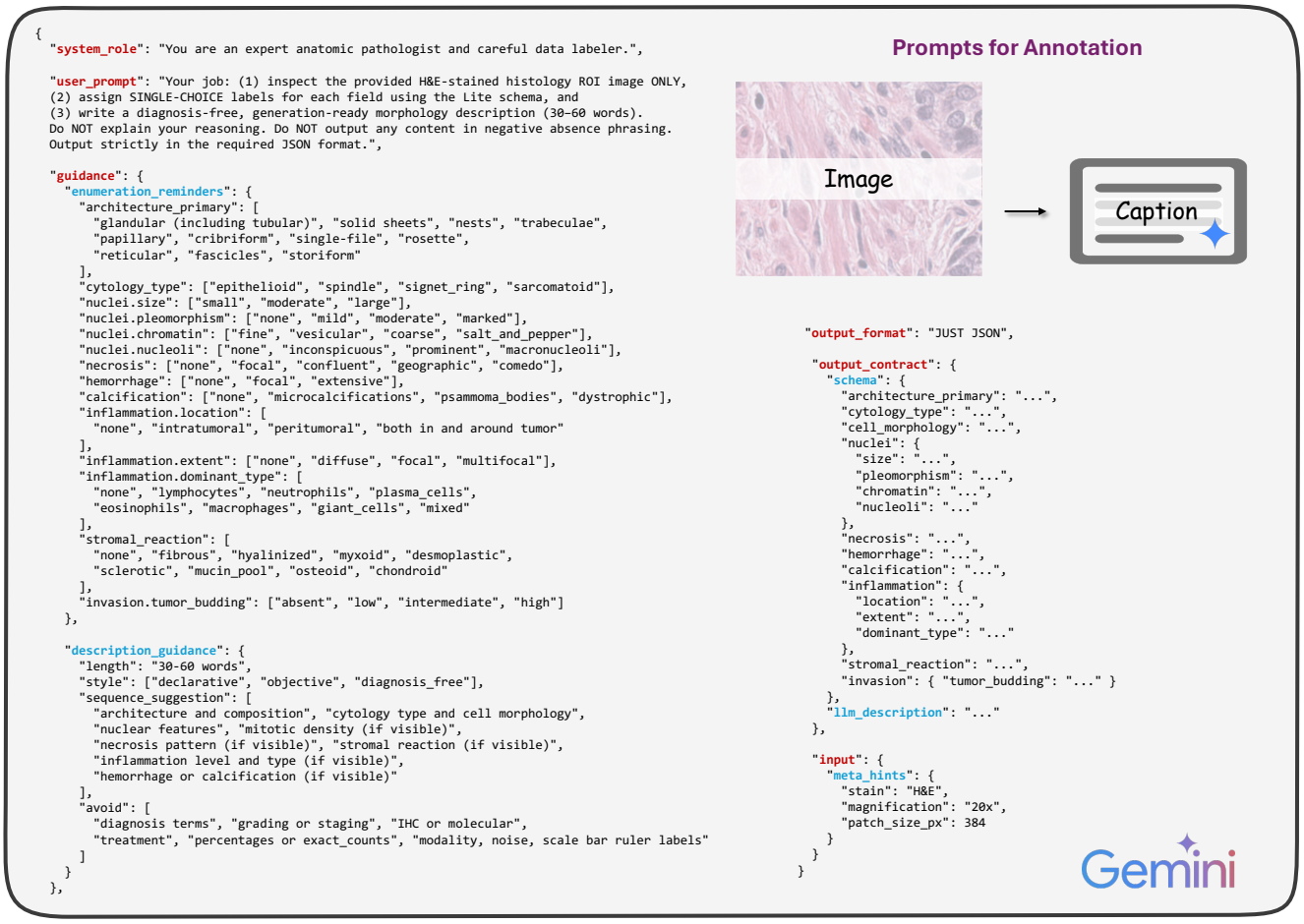


Figure S8. Prompts used for the preliminary annotation of pathology images with Gemini,-2,5 Pro formatted in JSON. Distinct colors are applied to differentiate hierarchical levels within the JSON structure:**top-level keys** and **secondary levels**. The overall workflow is illustrated in the upper-right panel.

Figure S9. Prompts used to perform cross-validation of Gemini-2.5 Pro's preliminary pathology annotations via GPT-5, encoded in JSON. Distinct colors are applied to differentiate hierarchical levels within the JSON structure: **top-level keys** and **secondary levels**. The overall workflow is illustrated in the upper-right panel.

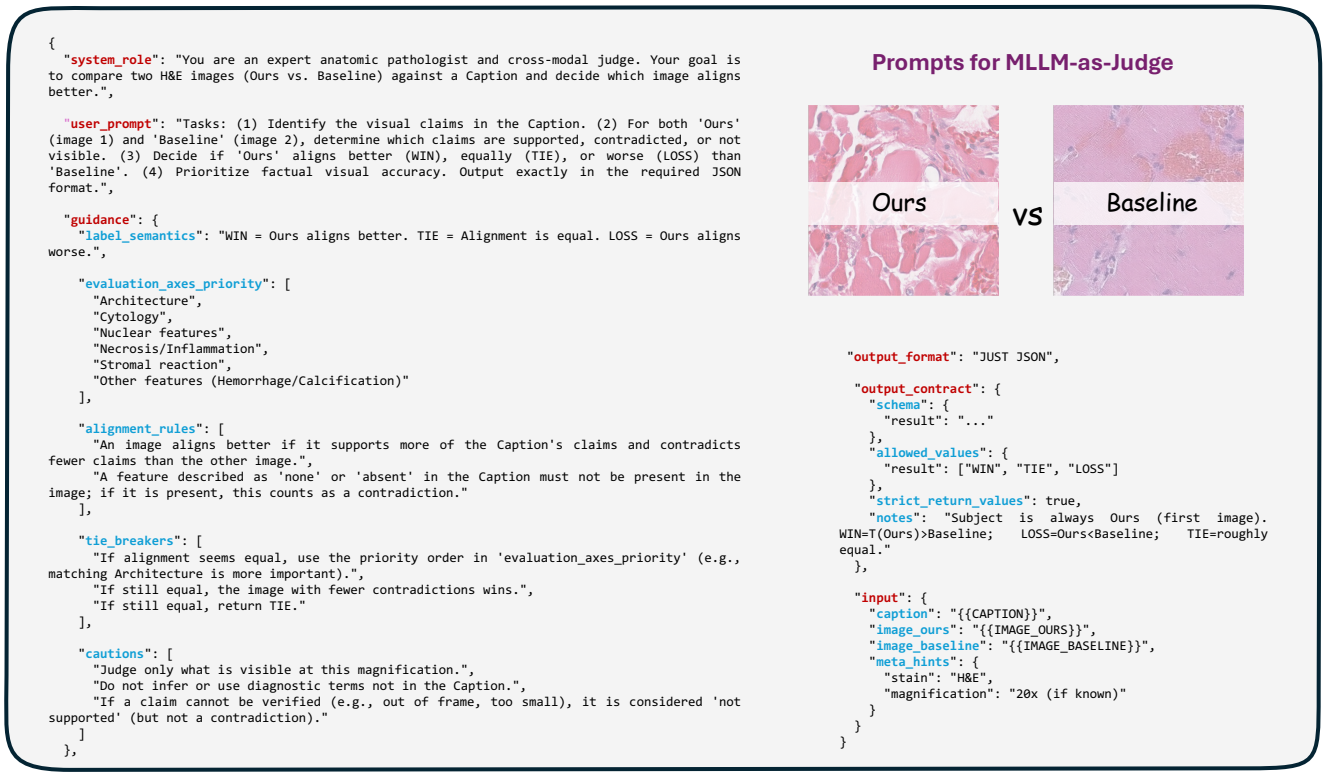


Figure S10. Prompts used for evaluating pathology images generated by our model and baseline methods using an MLLM, expressed in JSON format. Distinct colors are applied to differentiate hierarchical levels within the JSON structure: **top-level keys** and **secondary levels**. The overall workflow is depicted in the upper-right panel.

```
"""System:
You are a board-certified anatomic pathologist and a computational pathology NLP expert.
Your job is to decide whether each short phrase from pathology image descriptions should be
KEPT for downstream modeling.

Input:
- You will receive a plain text list of phrases, one per line. No JSON.

Output:
- Output ONLY the KEPT phrases, one per line, nothing else.
- Maintain the original input order for any kept items.
- Do NOT rewrite/normalize/case-fold/trim/repair any phrase; echo it exactly.
- No explanations, no JSON, no headings, no bullet points, no extra whitespace.

Decision policy (precision-first; fragments must be dropped):

[HARD DROP — fragments & scaffolding]  (these override everything)
1) Starts or ends with a stopword: and, with, in, of, to, for, is, are, the, a, an, by, on,
within, at.
   (Examples: "and …", "with …", "… and", "… is", "… the" → DROP)
2) Incomplete coordination: contains "and" where one side lacks a morphologic head (e.g.,
adjective-only).
3) Scaffolding without object: "is composed", "is arranged", "consists" (± "of"), "arranged
in", "composed of".
4) Prepositional fragments: begins with a preposition (in/on/with/of/to/by/at/within) unless
the remainder is a complete standalone morphologic head.
5) Auxiliary-verb tails: ends with "is/are/was/were" → DROP.

[DROP — generic/degree-only without feature]
6) Degree-only tokens without a concrete feature: mild, moderate, marked, markedly, abu(when
not tied to a feature).
7)  Bare  generic  structures:  cell/cells,  nucleus/nuclei,  cytoplasm,  stroma,  tissue,
architecture, borders, cellular, nuclear, solid (alone).
8) Over-generic disease/process words: tumor, neoplasm, proliferation (bare/broad).
9)  Unqualified  supportive  findings:  infiltrate,  inflammation,  hemorrhage,  necrosis
(unqualified) → DROP.

[KEEP — specific, standalone, discriminative features]
A)  Nuclear  features:  "vesicular  chromatin",  "coarse  chromatin",  "prominent  nucleoli",
"inconspicuous nucleoli", "eosinophilic nucleoli".
   - Also KEEP compound: "chromatin and <qualifier> nucleoli" (both heads explicit).
B) Qualified cytoplasmic features: "eosinophilic cytoplasm", "abundant eosinophilic cytoplasm".
C) Lineage/morphology: "epithelioid cells", "spindle cells", "polygonal cells".
   - Copular complete forms like "cells are epithelioid/spindle/polygonal" → KEEP.
D) Architectural patterns: "nests", "sheets", "solid sheets", "fascicles".
   - Prepositional variants like "in solid sheets" → fragment → DROP by HARD DROP #4.
E)  Cytologic  atypia  &  mitotic  features:  "pleomorphism",  "marked  pleomorphism",  "moderate
pleomorphism", "mitotic figures".
   - Lone adjective "mitotic" without head noun → DROP.
F) Qualified inflammation/hemorrhage: keep only with lineage/site/extent (e.g., "lymphocytic
infiltrate", "mixed inflammatory infiltrate", "focal hemorrhage", "intratumoral hemorrhage").
   - If trailing auxiliary (e.g., "lymphocytic infiltrate is") → DROP (HARD DROP #5).

Tie-break:
- If rules conflict or the phrase is ambiguous/underspecified, DROP (output nothing).
"""
```

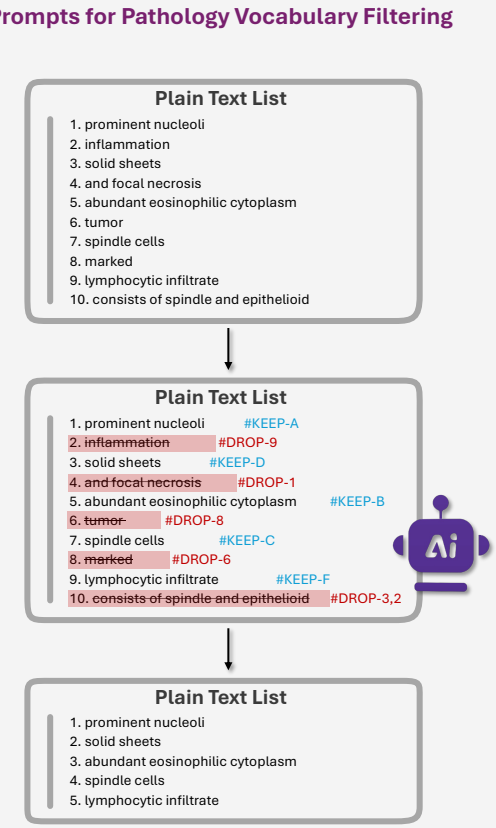


Figure S11. Prompts used for filtering the pathology vocabulary. Distinct colors are employed to differentiate hierarchical levels within the instructions:**top-level components** and **secondary elements**. The detailed filtering workflow is shown on the right.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 7

[3] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 7

[4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024. 5

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

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

[7] Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Lihao Liu, et al. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. *arXiv preprint arXiv:2510.15710*, 2025. 7

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[10] Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer, 2024. 6

[11] 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. 6

[12] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638 (8051):769–778, 2025. 6

[13] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 7

[14] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5182–5191, 2024. 7

[15] Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Saarthak Kapse, Tarak Nath Nandi, Ravi K Madduri, Prateek Prasanna, Tahsin Kurc, et al. Pixcell: A generative foundation model for digital histopathology images. *arXiv preprint arXiv:2506.05127*, 2025. 7

[16] Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. *arXiv preprint arXiv:2505.11404*, 2025. 2