

PBSBench: A Multi-Level Vision-Language Framework and Benchmark for Hematopathology Whole Slide Image Interpretation

Supplementary Material

6. Dataset Statistics

Here in Tab. 4, we present more detailed statistics on the number of questions in our datasets, with a breakdown by question type. Note that we remove questions with high similarity or trivial answers, leading to some small subgroups.

7. Code and Data Availability

We publish our code for training **PBS-VL**². We also publish **PBSInstr**, **PBSBench**, and our evaluation toolkit together with the training code.

8. Ethical, Limitation, and Hallucination Statements

8.1. Ethical Statement

Our proposed datasets **PBSInstr** and **PBSBench** are primarily built on a publicly available PBS dataset [11] and several blood cell datasets [31, 39, 50]. They are thoroughly de-identified and reveal no personal information.

PBSInstr, **PBSBench**, and **PBS-VL** are released solely for research and educational purposes. They are not intended for, and must not be used as, a standalone diagnostic device or clinical decision-support system. Any clinical deployment of the would require additional prospective validation, rigorous safety and performance assessment across diverse patient populations.

8.2. Limitation Statement

While our work presents resources for developing multi-modal models for hematopathology, it comes with limitations that need to be considered:

- Our work is built on public datasets for a specific patient cohort, with limited disease coverage and an unknown demographic distribution. It cannot reflect the variability in PBS appearance across institutions, scanners, staining protocols, or more hematologic conditions.
- Our data annotation is noisy despite human evaluation. The noise comes from the subjectivity of cell-type annotation and the potential hallucination from GPT-4o. We only conduct human cross-evaluation on small subsets, and it cannot entirely mitigate human subjectivity.
- We automatically estimate blood cell differentials from slides using cell subtyping models. This introduces addi-

tional model-induced noise. As a result, our reported performance on differential estimation tasks should be interpreted relative to this surrogate reference, not as absolute accuracy against expert quantification.

8.3. Close-loop Effect Discussion

We acknowledge potential bias from GPT-assisted synthesis and mitigate it via (a) different context: synthesis uses extra context (*e.g.*, cell type, condition) while evaluation is image-only, and (b) structured tasks (MCQ, FB) that are less style-sensitive. Empirically, GPT-4o does not trivially solve PBSBench and PBS-VL improves on structured subsets, suggesting gains beyond stylistic alignment.

8.4. Hallucination Statement

Our work involves two sources of potential hallucination: (1) hallucinations in GPT-assisted annotations used to construct **PBSInstr** and **PBSBench**, and (2) hallucinations from PBS-VL.

Our curation of **PBSInstr** and **PBSBench** relies on GPT-4o. It participated in the creation of cell crop captioning, slide description synthesizing, and QA pair creation. Although we tried our best to constrain the model responses by providing detailed instructions (as shown in Sec. 10) and employed multiple human-in-the-loop strategies for error control, these responses may still be hallucinated and erroneous.

We develop **PBS-VL** via adapter fine-tuning based on Qwen2.5-VL [2] with our vision backbone on our dataset. As a result, PBS-VL inherits potential hallucination from Qwen, our curated noisy dataset, and overfitting during fine-tuning. Potential hallucinations include statements that are not fully supported by the visual evidence or misinterpretation of hematologic findings.

9. Additional Data Processing Details

9.1. Patch Quality Control

We apply the tile quality control model from [11] to identify and remove patches with extremely low or high cell concentrations, as well as patches dominated by artifacts. Following [11], we extract patches of size 512×512 at $40\times$ magnification and reuse their released DenseNet121-based quality control model [15] without additional fine-tuning. For each candidate patch, the model outputs a continuous quality score ranging from 0 to 1. We follow the original decision rule from [11] and set the quality score threshold to 0.5.

²<https://github.com/Wang-Yuanlong/PBSBench>

Table 4. Detailed breakdown of the number of questions w.r.t tasks and question types

Image type	Split	Tasks	# QA pairs				
			True/False	MCQ	Fill-in-the-Blank	Open-ended	Subtotal
Cell level (ID)	Train	Morphology	4972	4481	3263	877	13593
		Abnormality	1702	1729	808	3294	7533
		Cell subtyping	134	647	2360	401	3542
		Knowledge	334	1183	397	719	2633
		Subtotal	7142	8040	6828	5291	27301
	Test	Morphology	670	625	454	108	1857
		Abnormality	231	222	103	432	988
		Cell subtyping	22	103	318	50	493
		Knowledge	34	138	41	109	322
		Subtotal	957	1088	916	699	3660
Cell level (OOD)	-	Morphology	720	894	688	134	2436
		Abnormality	7	304	123	824	1258
		Cell subtyping	0	43	85	19	147
		Knowledge	0	252	179	223	654
		Subtotal	727	1493	1075	1200	4495
Slide level (ID)	Train	Morphology	269	157	189	27	642
		Abnormality	16	21	4	6	47
		Cell differential	39	250	16	0	305
		Knowledge	0	1	0	137	138
		Diagnosis	1	78	4	71	154
		Subtotal # QAs	325	507	213	241	1286
	Test	Morphology	31	18	20	1	70
		Abnormality	0	6	1	3	10
		Cell differential	4	26	1	0	31
		Knowledge	0	0	0	13	13
		Diagnosis	0	5	0	9	14
		Subtotal # QAs	35	55	22	26	138

We illustrate typical quality control examples in Fig. 5. The upper panel shows examples of good- and poor-quality patches, highlighting standard failure modes such as overly crowded regions, edge patches without cells, and scanning artifacts. The lower panel exhibits the distribution of predicted quality scores across an entire slide, demonstrating how the model focuses on the monolayer region and the feathered edge area, which aligns with clinical practice.

9.2. Cell Subtyping Model Details

We develop a cell subtyping model based on Vision Transformer (ViT) variants from DinoBloom [21]. In particular, we use the ViT-L/14 (large) variant as the visual encoder. On top of the DinoBloom backbone, we append a feed-forward network (FFN) mapping head, followed by a linear classification layer. The FFN projects the global representation into a feature space tailored for downstream classification, and the final linear layer produces class logits. We freeze the DinoBloom backbone during training and update the FFN mapping head and the classifier. The models are trained with the Adam optimizer and a cosine learning rate schedule, using early stopping based on the area under the

ROC curve (AUC) on a held-out validation set.

We train two separate models using expert-labeled PBS cell crops. The first is a binary classifier that distinguishes white blood cells (WBCs) from non-WBC components (*e.g.*, red blood cells, platelets, and artifacts). The second is a multi-class WBC subtyping model that predicts fine-grained categories according to our predefined label set.

9.3. Cell Type Mapping for OOD Datasets

For out-of-distribution (OOD) evaluations, we normalize the cell type labels from external datasets to match our predefined categories. Specifically, we include three external datasets: AML-Cytomorphology_LMU [31], APL-Kaggle [39], and LISC [50]. The LISC dataset already uses the same WBC categories as our benchmark, so no relabeling is required. For AML-Cytomorphology_LMU and APL-Kaggle, we map their original cell type to either a corresponding class in our label space or to a generic “Others” category when no match exists. The complete mapping between original labels and our standardized categories is shown in Tab. 7.

The “Others” category aggregates several types that

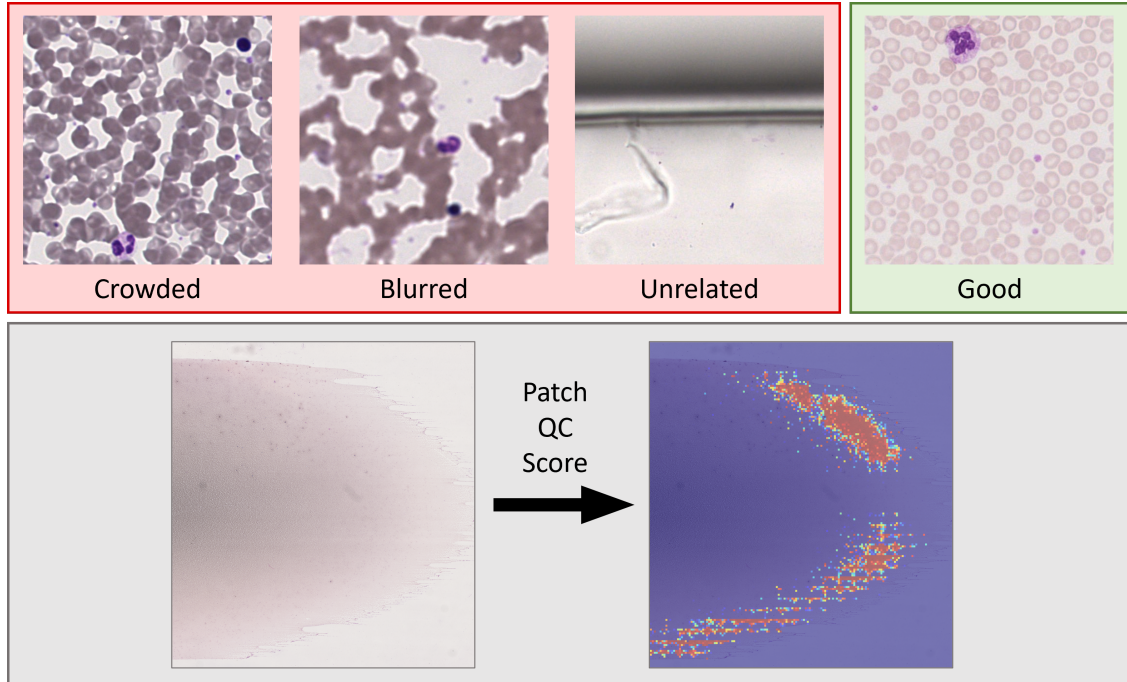


Figure 5. Quality control examples. Poor-quality patches are shown in the red box, and a good-quality patch is shown in the green box. Quality scores are shown in *jet* colormap where red denotes high scores and blue denotes low scores.

would otherwise be too sparse or incompatible with our curated question–answer pairs. These include (1) rare WBC subtypes (*e.g.*, myelocytes, giant thrombocytes), (2) cell types that do not appear in our cohort and thus cannot be consistently represented in our data (*e.g.*, some blast categories), and (3) artifacts or unidentified labels that do not correspond to a well-defined cell category. This normalization step ensures consistent evaluation across datasets while avoiding unstable estimates for extremely rare or dataset-specific labels.

9.4. VQA quality control

We reject questions that are (a) non-PBS questions (*e.g.*, general hematology), (b) factual errors, (c) not answerable from the image, and (d) answer-leaking (*i.e.*, question implies the answer). These are also common error patterns observed from GPT-4o.

For labels mapped to Others, we exclude subtyping questions and ask morphology questions that remain visually grounded.

10. Prompts

As we scale up our annotation using GPT-4o, we present the prompt used for various tasks in Fig. 7, Fig. 8, Fig. 9, and Fig. 10. Note that we remove some prompt sections on input/output formatting and examples for simplicity.

11. Additional Experimental Results

11.1. Breakdown Performance on PBSBench

In this section, we report a detailed breakdown of benchmark performance by task and question types. They can be found in Tab. 8, Tab. 9, Tab. 10, Tab. 11, and Tab. 12. For cell-level questions, we report only one metric per question type for simplicity: accuracy for True-or-False and multiple-choice questions, partial-matching rate for Fill-in-the-blank, and semantic similarity for open-ended questions. Please note that some subgroups have a limited number of questions. As a result, the model’s performance is subject to data randomness, making it unreliable to compare and interpret model performance on those subgroups.

11.2. Sensitivity to the Number of Patches

Table 5. Results of sensitivity of numbers of input patches for proprietary models Std in parenthesis

model	# patches	TF-Acc	MCQ-Acc	FB-Pmatch	Open-BLEU1
GPT-4o	15	0.63(0.077)	0.80(0.048)	0.09(0.060)	0.18(0.015)
	30	0.57(0.077)	0.75(0.057)	0.05(0.045)	0.17(0.011)
	50	0.71(0.071)	0.84(0.049)	0.18(0.086)	0.18(0.013)
Gemini-2.5-pro	15	0.51(0.081)	0.67(0.061)	0.09(0.062)	0.10(0.008)
	30	0.49(0.081)	0.73(0.057)	0.09(0.060)	0.09(0.008)
	50	0.57(0.079)	0.76(0.056)	0.09(0.059)	0.10(0.008)
Claude-4.5-Sonnet	15	0.66(0.076)	0.36(0.064)	0.14(0.071)	0.14(0.013)
	30	0.69(0.075)	0.35(0.064)	0.18(0.081)	0.15(0.014)
	50	0.71(0.077)	0.35(0.064)	0.14(0.072)	0.15(0.013)

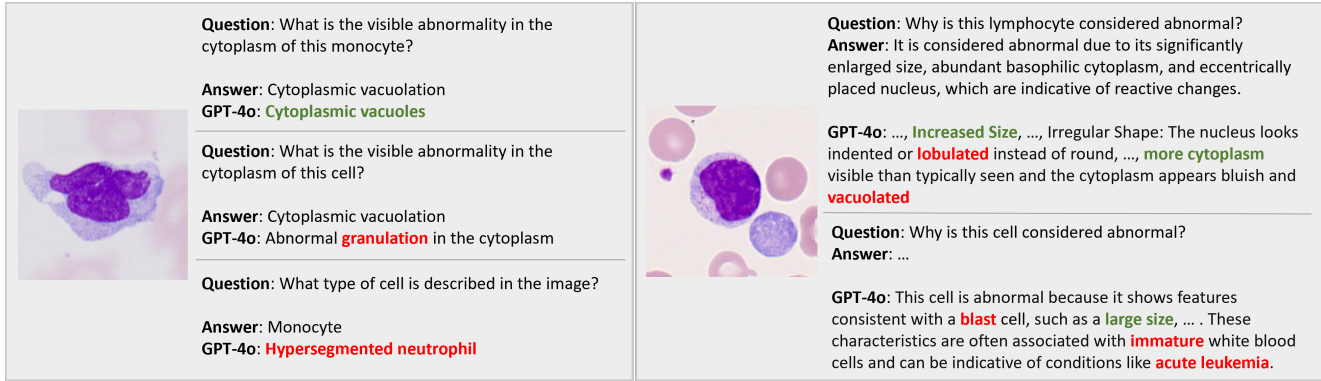


Figure 6. Additional case study on GPT-4o, the model shows performance degradation when no cell types are available as hints in questions.

We conduct sensitivity analysis for the number of patches in Tab. 5. Note that GPT-4o allows up to 50 images. Results show that more patches can improve performance, but gains are modest and remain far from PBS-VL.

11.3. Cell-patch Alignment Training Ablation

Table 6. Results of cell-patch alignment training ablation on slide level QA. Std in parenthesis

model	TF-Acc	MCQ-Acc	FB-Pmatch	Open-BLEU1
PBS-VL w/o align	0.91 (0.047)	0.82(0.051)	0.23(0.088)	0.35(0.017)
PBS-VL	0.86(0.056)	0.85 (0.046)	0.32 (0.096)	0.36 (0.021)

We conduct ablation of cell-patch alignment training by a variation with cell images as input in replacement of patch images for slide-level tasks. The result comparison is in Tab. 6. The original PBS-VL outperforms on MCQ, FB, and open-ended questions. Moreover, it bypasses cell detection at inference, reducing computational overhead.

12. Additional Case Study

We provide an additional case study in Fig. 6 that illustrates how model performance degrades when explicit cell type information is removed from the question. In this case study, we modify the original questions by replacing specific cell types (e.g. monocyte) with a generic reference (“this cell”). Thus, we eliminate hints that may reveal or narrow down the answer without visual understanding.

As shown in Fig. 6, GPT-4o can produce correct answers when the ground-truth cell type is stated in the question. However, it still struggles to correctly identify the cell type in the image on its own once textual hints are removed, leading to erroneous responses. This behavior supports our hypothesis that modern MLLMs possess substantial knowledge of hematopathology but still struggle to connect medical concepts to visual evidence.

13. Large Tables and Figures

See below.

Prompt for producing cell crop captions
<p>You are an expert hematopathologist. You are asked to analyze H&E-stained peripheral blood smear images for blood cells. Every time I send you an image, probably with a comment, I need your help to identify:</p> <ol style="list-style-type: none"> The type of the cell, please choose from this list that best describes the cell: [Basophil, Eosinophils, Lymphocyte, Monocyte, Neutrophil, Artifact, Damaged Cells, Partial Cells] This step can be skipped if the cell type is provided. However, if the cell type is "Others", it means the cell type is rare or Artifact/Damaged/Partial, you still need to identify the cell type based on the image. If the cell does not show normal morphology compared to what is commonly seen in its type, point it out and describe the abnormality, otherwise provide brief morphological description in 1 sentence. Abnormality includes but is NOT limited to: <ul style="list-style-type: none"> - Size (e.g., extremely enlarged, smaller than normal, etc.) - Shape (e.g., irregular, distorted, atypical, damaged, etc.) - Special features (e.g., presence of granules, hypersegmented nucleus, botryoid nucleus, chromatin clumping, etc.) For normal cells, just provide simple morphology description in 1 sentence. <p>For abnormality, please a) briefly describe the abnormal morphology, and b) analyse the possible causes of the abnormality based on your expert knowledge.</p> <ol style="list-style-type: none"> DO NOT include general descriptions (general about the cell type, common sense, etc.) not specific to the cell in the image or unrelated to morphology. For example, <ul style="list-style-type: none"> - "This is a neutrophil, which is a type of white blood cell that helps fight infection." (This is NOT needed) Abnormality only refers to the morphology of the cell, not the number of cells. Answer in the following format: "This cell belongs to <cell_type>. <is_abnormal>. <morphology_description>. <possible_cause_reasoning_if_abnormal_else_ignore>. <p>This is an {{cell_type}}. comments: {{comments}}</p>

Figure 7. Prompts for cell crop captioning

Table 7. The mapping of cell type from out-of-distribution cell image datasets for normalization.

AML-Cytomorphology_LMU		APL-kaggle	
fine-coarsed types	normalized types	fine-coarsed types	normalized types
BAS	Basophil	Artifact	Others
EBO	Others	Band neutrophils	Neutrophil
EOS	Eosinophil	Basophil	Basophil
KSC	Others	Blast (no lineage spec)	Others
LYA	Lymphocyte	Eosinophils	Eosinophil
LYT	Lymphocyte	Erythroblast	Others
MMZ	Others	Giant thrombocyte	Others
MOB	Others	Lymphocyte	Lymphocyte
MON	Monocyte	Lymphocyte (variant)	Lymphocyte
MYB	Others	Metamyelocyte	Others
MYO	Others	Monocyte	Monocyte
NGB	Neutrophil	Myelocyte	Others
NGS	Neutrophil	Plasma cells	Others
PMB	Others	Prolymphocyte	Others
PMO	Others	Promonocyte	Others
		Promyelocyte	Others
		Segmented neutrophils	Neutrophil
		Smudge cells	Others
		Thrombocyte aggregation	Others
		Unidentified	Others
		Young Unidentified	Others

Table 8. The accuracy of cell-level **True/False** questions for benchmarked models breaks down by task types.

	In-domain				Out of Domain	
	Morphology	Abnormality	Cell subtyping	Knowledge	Morphology	Abnormality
	Proprietary VL Models					
GPT-4o	0.85(0.013)	0.50(0.032)	0.77(0.090)	0.91(0.048)	0.86(0.013)	0.86(0.129)
Gemini-2.5-pro	0.77(0.017)	0.46(0.032)	0.82(0.081)	0.94(0.040)	0.65(0.018)	0.29(0.170)
Claude-4.5-Sonnet	0.57(0.019)	0.59(0.032)	0.77(0.089)	0.89(0.053)	0.54(0.019)	0.29(0.178)
	Open-source General-purpose VL Model					
InternVL-8B	0.75(0.016)	0.65(0.031)	0.91(0.059)	0.80(0.066)	0.64(0.018)	0.86(0.136)
Idefics3-8B	0.57(0.019)	0.62(0.032)	0.59(0.106)	0.77(0.069)	0.56(0.019)	0.29(0.178)
Qwen2-VL-7B	0.71(0.017)	0.53(0.034)	0.68(0.098)	0.80(0.068)	0.69(0.016)	0.57(0.190)
Qwen2.5-VL-7B	0.71(0.017)	0.40(0.031)	0.68(0.097)	0.80(0.069)	0.78(0.015)	0.57(0.184)
DeepSeek-VL-7B	0.77(0.016)	0.24(0.028)	0.45(0.105)	0.66(0.081)	0.94(0.008)	1.00(0.000)
LLaVA-1.6-Mistral-7B	0.66(0.018)	0.38(0.033)	0.45(0.104)	0.80(0.067)	0.75(0.016)	0.29(0.172)
LLaVA-1.6-Vicuna-7B	0.64(0.018)	0.39(0.032)	0.68(0.101)	0.57(0.084)	0.70(0.017)	0.43(0.185)
BLIP2-OPT-6.7B	0.49(0.019)	0.57(0.032)	0.50(0.112)	0.54(0.083)	0.43(0.018)	0.29(0.166)
	Domain-specific - Medical VL Model					
LLaVA-Med	0.60(0.018)	0.35(0.031)	0.55(0.105)	0.66(0.082)	0.70(0.017)	0.71(0.170)
Med-flamingo	0.38(0.019)	0.65(0.031)	0.32(0.096)	0.46(0.085)	0.25(0.016)	0.43(0.185)
	Domain-specific - Pathology VL Model					
PA-LLaVA	0.26(0.017)	0.79(0.027)	0.50(0.103)	0.29(0.077)	0.02(0.005)	0.14(0.133)
Quilt-LLaVA	0.73(0.018)	0.23(0.029)	0.55(0.105)	0.71(0.074)	0.97(0.006)	0.86(0.137)
PathGen-LLaVA	0.78(0.016)	0.25(0.027)	0.73(0.094)	0.80(0.068)	0.97(0.007)	0.71(0.170)
Ours - PBS-VL	0.87(0.013)	0.30(0.029)	1.00(0.000)	0.83(0.063)	0.97(0.007)	0.86(0.124)

Table 9. The accuracy of cell-level **MCQ** questions for benchmarked models breaks down by task types.

	In-domain				Out of Domain			
	Morphology	Abnormality	Cell subtyping	Knowledge	Morphology	Abnormality	Cell subtyping	Knowledge
Proprietary VL Models								
GPT-4o	0.88(0.012)	0.87(0.022)	0.94(0.023)	0.99(0.007)	0.91(0.010)	0.95(0.013)	0.95(0.031)	0.96(0.011)
Gemini-2.5-pro	0.85(0.015)	0.87(0.023)	0.94(0.023)	0.99(0.007)	0.89(0.010)	0.93(0.015)	0.95(0.032)	0.99(0.007)
Claude-4.5-Sonnet	0.73(0.018)	0.66(0.033)	0.79(0.042)	0.74(0.038)	0.87(0.011)	0.91(0.016)	0.98(0.023)	0.97(0.011)
Open-source General-purpose VL Model								
InternVL-8B	0.89(0.013)	0.93(0.015)	0.93(0.025)	0.98(0.012)	0.88(0.011)	0.94(0.014)	0.95(0.032)	0.97(0.010)
Idefics3-8B	0.30(0.018)	0.34(0.032)	0.32(0.047)	0.62(0.041)	0.34(0.016)	0.54(0.028)	0.26(0.068)	0.68(0.029)
Qwen2-VL-7B	0.74(0.017)	0.92(0.018)	0.81(0.040)	0.97(0.014)	0.82(0.012)	0.93(0.015)	0.84(0.056)	0.97(0.011)
Qwen2.5-VL-7B	0.70(0.018)	0.81(0.026)	0.79(0.041)	0.93(0.020)	0.80(0.014)	0.88(0.019)	0.91(0.045)	0.94(0.014)
DeepSeek-VL-7B	0.46(0.019)	0.60(0.033)	0.42(0.047)	0.64(0.040)	0.52(0.017)	0.74(0.026)	0.44(0.073)	0.66(0.030)
LLaVA-1.6-Mistral-7B	0.46(0.020)	0.70(0.031)	0.39(0.048)	0.63(0.039)	0.53(0.017)	0.74(0.024)	0.42(0.077)	0.60(0.030)
LLaVA-1.6-Vicuna-7B	0.41(0.020)	0.57(0.034)	0.39(0.047)	0.59(0.042)	0.49(0.016)	0.71(0.027)	0.44(0.074)	0.56(0.031)
BLIP2-OPT-6.7B	0.19(0.016)	0.22(0.028)	0.20(0.040)	0.20(0.034)	0.22(0.014)	0.23(0.023)	0.14(0.051)	0.24(0.027)
Domain-specific - Medical VL Model								
LLaVA-Med	0.29(0.018)	0.30(0.031)	0.25(0.042)	0.30(0.039)	0.26(0.015)	0.31(0.027)	0.14(0.053)	0.21(0.026)
Med-flamingo	0.27(0.017)	0.27(0.029)	0.25(0.042)	0.25(0.037)	0.26(0.014)	0.27(0.025)	0.33(0.072)	0.25(0.027)
Domain-specific - Pathology VL Model								
PA-LLaVA	0.23(0.016)	0.25(0.027)	0.28(0.044)	0.25(0.037)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)
Quilt-LLaVA	0.34(0.018)	0.41(0.032)	0.36(0.047)	0.39(0.042)	0.39(0.016)	0.38(0.027)	0.49(0.075)	0.41(0.031)
PathGen-LLaVA	0.76(0.016)	0.85(0.024)	0.75(0.041)	0.78(0.036)	0.81(0.013)	0.88(0.018)	0.74(0.066)	0.60(0.032)
Ours - PBS-VL	0.94(0.009)	0.96(0.012)	0.98(0.013)	0.99(0.007)	0.94(0.008)	0.97(0.010)	0.93(0.037)	0.99(0.007)

Table 10. The partial match rate of cell-level **Fill-in-the-blank** questions for benchmarked models breaks down by task types.

	In-domain				Out of Domain			
	Morphology	Abnormality	Cell subtyping	Knowledge	Morphology	Abnormality	Cell subtyping	Knowledge
Proprietary VL Models								
GPT-4o	0.40(0.023)	0.62(0.048)	0.57(0.028)	0.34(0.074)	0.23(0.016)	0.24(0.038)	0.51(0.055)	0.40(0.037)
Gemini-2.5-pro	0.36(0.022)	0.50(0.049)	0.47(0.027)	0.34(0.072)	0.17(0.015)	0.22(0.038)	0.36(0.052)	0.44(0.037)
Claude-4.5-Sonnet	0.50(0.024)	0.62(0.047)	0.58(0.027)	0.54(0.078)	0.22(0.016)	0.23(0.040)	0.49(0.053)	0.45(0.037)
Open-source General-purpose VL Model								
InternVL-8B	0.34(0.022)	0.46(0.050)	0.44(0.029)	0.24(0.065)	0.19(0.015)	0.16(0.033)	0.27(0.047)	0.17(0.028)
Idefics3-8B	0.15(0.017)	0.36(0.047)	0.16(0.021)	0.10(0.047)	0.11(0.012)	0.07(0.023)	0.06(0.026)	0.18(0.029)
Qwen2-VL-7B	0.24(0.020)	0.47(0.049)	0.24(0.024)	0.20(0.061)	0.12(0.013)	0.20(0.036)	0.09(0.031)	0.09(0.021)
Qwen2.5-VL-7B	0.14(0.016)	0.39(0.048)	0.16(0.020)	0.07(0.040)	0.10(0.011)	0.15(0.033)	0.15(0.038)	0.06(0.017)
DeepSeek-VL-7B	0.12(0.015)	0.40(0.048)	0.12(0.019)	0.02(0.024)	0.06(0.009)	0.13(0.031)	0.12(0.035)	0.02(0.011)
LLaVA-1.6-Mistral-7B	0.12(0.015)	0.27(0.045)	0.03(0.009)	0.10(0.048)	0.08(0.010)	0.05(0.020)	0.08(0.029)	0.18(0.029)
LLaVA-1.6-Vicuna-7B	0.14(0.016)	0.23(0.040)	0.02(0.008)	0.02(0.025)	0.08(0.010)	0.04(0.018)	0.05(0.023)	0.15(0.027)
BLIP2-OPT-6.7B	0.18(0.018)	0.39(0.048)	0.21(0.022)	0.17(0.057)	0.21(0.015)	0.27(0.039)	0.20(0.042)	0.17(0.027)
Domain-specific - Medical VL Model								
LLaVA-Med	0.07(0.012)	0.42(0.047)	0.04(0.011)	0.00(0.000)	0.06(0.009)	0.15(0.032)	0.06(0.026)	0.03(0.013)
Med-flamingo	0.03(0.009)	0.26(0.044)	0.04(0.011)	0.00(0.000)	0.03(0.006)	0.13(0.030)	0.09(0.033)	0.01(0.006)
Domain-specific - Pathology VL Model								
PA-LLaVA	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)
Quilt-LLaVA	0.15(0.016)	0.37(0.047)	0.11(0.017)	0.15(0.054)	0.10(0.011)	0.16(0.034)	0.13(0.036)	0.08(0.020)
PathGen-LLaVA	0.34(0.023)	0.41(0.047)	0.25(0.024)	0.12(0.049)	0.21(0.016)	0.19(0.033)	0.26(0.046)	0.11(0.024)
Ours - PBS-VL	0.41(0.023)	0.59(0.048)	0.74(0.024)	0.41(0.078)	0.32(0.017)	0.24(0.036)	0.45(0.053)	0.47(0.036)

Table 11. The semantic similarity of cell-level **Open-ended** questions for benchmarked models breaks down by task types.

	In-domain				Out of Domain			
	Morphology	Abnormality	Cell subtyping	Knowledge	Morphology	Abnormality	Cell subtyping	Knowledge
Proprietary VL Models								
GPT-4o	0.70(0.013)	0.62(0.006)	0.77(0.010)	0.63(0.015)	0.74(0.010)	0.62(0.004)	0.78(0.020)	0.68(0.009)
Gemini-2.5-pro	0.66(0.012)	0.58(0.007)	0.74(0.012)	0.54(0.015)	0.68(0.011)	0.54(0.005)	0.73(0.019)	0.58(0.010)
Claude-4.5-Sonnet	0.58(0.022)	0.47(0.012)	0.69(0.023)	0.45(0.029)	0.66(0.009)	0.58(0.004)	0.70(0.016)	0.62(0.008)
Open-source General-purpose VL Model								
InternVL-8B	0.70(0.013)	0.59(0.008)	0.80(0.010)	0.62(0.017)	0.73(0.012)	0.59(0.005)	0.78(0.021)	0.67(0.010)
Idefics3-8B	0.64(0.015)	0.52(0.009)	0.69(0.018)	0.56(0.018)	0.67(0.016)	0.52(0.006)	0.68(0.028)	0.61(0.012)
Qwen2-VL-7B	0.69(0.012)	0.60(0.006)	0.75(0.010)	0.58(0.016)	0.74(0.011)	0.57(0.006)	0.77(0.019)	0.63(0.011)
Qwen2.5-VL-7B	0.62(0.012)	0.57(0.006)	0.72(0.009)	0.57(0.014)	0.66(0.010)	0.55(0.005)	0.72(0.016)	0.60(0.010)
DeepSeek-VL-7B	0.58(0.017)	0.50(0.009)	0.67(0.018)	0.50(0.017)	0.61(0.017)	0.49(0.006)	0.68(0.013)	0.51(0.013)
LLaVA-1.6-Mistral-7B	0.61(0.012)	0.53(0.007)	0.68(0.012)	0.56(0.015)	0.64(0.011)	0.53(0.005)	0.71(0.024)	0.58(0.010)
LLaVA-1.6-Vicuna-7B	0.60(0.013)	0.51(0.006)	0.66(0.010)	0.54(0.013)	0.64(0.010)	0.51(0.005)	0.67(0.024)	0.56(0.010)
BLIP2-OPT-6.7B	0.45(0.019)	0.38(0.008)	0.52(0.027)	0.39(0.017)	0.46(0.020)	0.34(0.007)	0.45(0.043)	0.35(0.016)
Domain-specific - Medical VL Model								
LLaVA-Med	0.62(0.013)	0.53(0.007)	0.69(0.012)	0.54(0.016)	0.67(0.012)	0.52(0.006)	0.72(0.025)	0.59(0.012)
Med-flamingo	0.44(0.017)	0.43(0.008)	0.54(0.024)	0.36(0.016)	0.45(0.017)	0.38(0.006)	0.42(0.042)	0.36(0.011)
Domain-specific - Pathology VL Model								
PA-LLaVA	-0.04(0.004)	-0.06(0.002)	-0.05(0.007)	-0.07(0.005)	-0.00(0.003)	0.01(0.001)	-0.03(0.004)	0.00(0.003)
Quilt-LLaVA	0.68(0.014)	0.57(0.008)	0.74(0.011)	0.55(0.017)	0.71(0.012)	0.55(0.006)	0.75(0.025)	0.59(0.011)
PathGen-LLaVA	0.69(0.013)	0.61(0.007)	0.80(0.010)	0.58(0.017)	0.70(0.012)	0.56(0.006)	0.78(0.020)	0.61(0.012)
Ours - PBS-VL	0.75(0.013)	0.70(0.007)	0.82(0.014)	0.72(0.016)	0.79(0.012)	0.68(0.005)	0.80(0.022)	0.72(0.011)

Table 12. The performance of slide-level QAs for benchmarked models breaks down by task types.

			GPT-4o	Gemini-2.5-pro	Claude-4.5	HistoGPT	SlideChat	PBS-VL-Slide
Morphology	T/F	Accuracy	0.61(0.087)	0.52(0.089)	0.74(0.081)	0.19(0.069)	0.77(0.074)	0.94(0.045)
	MCQ	Accuracy	0.78(0.094)	0.67(0.109)	0.28(0.104)	0.28(0.102)	0.83(0.088)	0.94(0.056)
	FB	EMatch	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.25(0.098)
		PMatch	0.00(0.000)	0.05(0.049)	0.15(0.078)	0.00(0.000)	0.00(0.000)	0.30(0.101)
	Open	BLEU-1	0.10(0.000)	0.02(0.000)	0.08(0.000)	0.01(0.000)	0.05(0.000)	0.28(0.000)
		ROGUE-L	0.16(0.000)	0.08(0.000)	0.13(0.000)	0.02(0.000)	0.11(0.000)	0.40(0.000)
		Sim	0.82(0.000)	0.74(0.000)	0.76(0.000)	0.20(0.000)	0.75(0.000)	0.74(0.000)
Abnormality	MCQ	Accuracy	0.83(0.149)	0.67(0.193)	0.50(0.193)	0.00(0.000)	0.67(0.198)	0.83(0.150)
	FB	EMatch	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)
		PMatch	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)
	Open	BLEU-1	0.12(0.025)	0.07(0.019)	0.12(0.052)	0.02(0.008)	0.24(0.042)	0.39(0.029)
		ROGUE-L	0.22(0.027)	0.12(0.029)	0.18(0.050)	0.03(0.012)	0.23(0.032)	0.44(0.044)
		Sim	0.76(0.027)	0.70(0.007)	0.68(0.028)	0.13(0.046)	0.62(0.074)	0.76(0.021)
Cell differential	T/F	Accuracy	0.25(0.214)	0.25(0.209)	0.25(0.218)	0.25(0.215)	0.75(0.217)	0.25(0.208)
	MCQ	Accuracy	0.69(0.090)	0.85(0.071)	0.38(0.097)	0.23(0.081)	0.19(0.080)	0.77(0.081)
	FB	EMatch	1.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	0.00(0.000)	1.00(0.000)
		PMatch	1.00(0.000)	1.00(0.000)	1.00(0.000)	0.00(0.000)	0.00(0.000)	1.00(0.000)
	Open	BLEU-1	0.18(0.013)	0.09(0.009)	0.17(0.021)	0.03(0.004)	0.16(0.022)	0.37(0.032)
		ROGUE-L	0.19(0.015)	0.11(0.009)	0.17(0.020)	0.04(0.006)	0.19(0.019)	0.35(0.045)
Knowledge	Open	Sim	0.73(0.041)	0.65(0.042)	0.68(0.042)	0.17(0.020)	0.63(0.048)	0.77(0.034)
	MCQ	Accuracy	0.80(0.186)	0.40(0.225)	0.20(0.180)	0.40(0.210)	0.80(0.175)	1.00(0.000)
	Open	BLEU-1	0.19(0.019)	0.11(0.015)	0.15(0.015)	0.04(0.005)	0.15(0.033)	0.34(0.036)
		ROGUE-L	0.21(0.024)	0.13(0.013)	0.19(0.020)	0.05(0.005)	0.23(0.019)	0.43(0.037)
Diagnosis		Sim	0.72(0.054)	0.71(0.035)	0.65(0.044)	0.14(0.026)	0.58(0.060)	0.76(0.040)

Prompt for producing whole slide captions
<p>You are a hematology and pathology report assistant. Your task is to generate a concise and clinically realistic <i>*Microscopic Description*</i> section for a blood smear slide, based on the aggregated cell-level observations provided below. Each slide contains multiple cells, each described by its morphology and possible abnormalities.</p> <p>### Guidelines</p> <p>1. Primary Focus</p> <ul style="list-style-type: none">- Emphasize <i>abnormal cells</i> and their key morphological characteristics.- Summarize <i>normal cells</i> only briefly (e.g., “Most neutrophils appear normal” or “Lymphocytes largely display typical morphology.”)- Highlight distinct abnormal findings such as:<ul style="list-style-type: none">+ Nuclear abnormalities (e.g., bilobed, hypersegmented, irregular, or elongated nuclei)+ Cytoplasmic irregularities (e.g., vacuolation, basophilia, projections)+ Cell type-specific atypia (e.g., atypical lymphocytes, immature neutrophils)- Focus on synthesizing the abnormal findings clearly and concisely.- Mention normal cell populations only briefly, without repeating details. <p>2. Integration</p> <ul style="list-style-type: none">- Integrate all descriptions for the same slide into one cohesive paragraph.- Avoid listing each cell individually; instead, merge overlapping findings and capture overall trends. <p>3. Language and Style</p> <ul style="list-style-type: none">- Write in the formal tone of a pathology report.- Start with: “Microscopic description: Examination of the slide reveals...”- Maintain clarity, precision, and avoid unnecessary repetition.- Do not mention that the information came from AI, GPT, or separate cells.- Avoid vague or generic opening sentences such as “Examination of the slide reveals a mixture of normal and abnormal hematologic cells.”- Each sentence must describe specific observable findings, not general statements. <p>4. Length</p> <ul style="list-style-type: none">- Typically, 3–5 sentences. Include enough detail to capture all major abnormalities without redundancy. <p>### Example Output: {{examples}}</p> <p>Now generate the final microscopic description for the following input:{{cell_descriptions}}</p>

Figure 8. Prompts for whole slide captioning

Prompt for producing cell crop QAs
<p>You are a biomedical expert specializing in cellular morphology and vision-language reasoning. Your task is to generate diverse and high-quality Question–Answer (Q&A) pairs for each given cell entry. Each data entry describes a single cell, including its type, abnormality status, description, and reasoning.</p> <p>### OBJECTIVE: Based only on the provided information, generate meaningful, accurate, and image-grounded questions and answers about the cell. Questions should focus primarily on what can be inferred or observed from the cell’s visible features (shape, nucleus, granules, cytoplasm, abnormalities). Avoid using metadata or external knowledge. All questions must be non-trivial and require reasoning or observation beyond simple recall.</p> <p>### QUESTION GENERATION PRINCIPLES</p> <ul style="list-style-type: none">- Generate 4–6 Q&A pairs for each cell.- At least 70% of questions must depend on visible morphological features.- Use a diverse mix of question types: “True/False”, “Multi-choice”, “Close-ended”, “Open-ended”.- Limit “True/False” to at most one per cell.- Avoid repetitive, trivial, or metadata-based questions.- Each question must have distinct focus and clear wording.- Every question must be grammatically complete and use correct punctuation.<ul style="list-style-type: none">+ If it is phrased as a question, end with a question mark (?).+ If it is phrased as a factual statement (especially for True/False), end with a period (.). <p>### DIFFICULTY AND LEAKAGE RULES</p> <ul style="list-style-type: none">- All question types must demonstrate high difficulty. Simple recall of definitions, names, or colors is not acceptable.- At least half of the questions should require interpretive or comparative reasoning, not just direct observation. <p>### QUANTITATIVE DETAIL RULE</p> <ul style="list-style-type: none">- Avoid including explicit numeric thresholds or quantitative values (e.g., “5 lobes”, “three segments”, “two nuclei”) that appear in the description or reasoning. Instead, use comparative or qualitative phrasing such as “greater than usual segmentation,” “unusually divided nucleus,” or “excessive lobulation.” <p>### QUESTION TYPE RULES</p> <ol style="list-style-type: none">1. True/False<ul style="list-style-type: none">- All True/False questions must start with “(True/False)”. - If the statement is phrased as a question, add a question mark (?) at the end.- If the statement is declarative (a fact to judge true or false), do not add a question mark.- Ensure every True/False question is explicitly a factual statement that can be judged true or false.- Avoid obvious or definitional statements. Focus on subtle morphological or diagnostic distinctions.2. Multi-choice<ul style="list-style-type: none">- Do not include options (A/B/C/D) in the question text.- Use an “options” field containing exactly 3–5 clear and distinct choices.- “answer” must exactly match one of the “options”.- Never include options inline in the question sentence.- Distractor options should be plausible but slightly incorrect, requiring careful observation to distinguish the correct one.3. Fill-in-the-blank<ul style="list-style-type: none">- Short factual or fill-in-the-blank questions that require concise, direct answers. These questions can test recognition, identification, or specific morphological details4. Open-ended Reasoning or explanatory questions requiring short justification. <p>### TOPIC FIELD: Each question must include “topic” describing its conceptual focus. Only use the following four predefined topics unless a truly new and justified category is required:</p> <ol style="list-style-type: none">1. “morphology” — questions about the visible structural characteristics of the cell, such as nuclear shape, chromatin texture, or cytoplasmic granules.2. “abnormality” — questions describing abnormal or atypical cell features compared to normal morphology.3. “cell_type” — questions focused on identifying or classifying the type of cell based on structural features.4. “knowledge” — questions related to biological reasoning or hematologic concepts that extend beyond direct observation. <p>### EXAMPLE OUTPUT: {{examples}}</p> <p>Now generate 4–6 (or more if justified) high-quality Q&A pairs for the following cell, following all rules strictly.</p>

Figure 9. Prompts for cell crop QA

Prompt for producing whole slide QAs
<p>You are a hematology and pathology question generation assistant. Your task is to generate diverse, clinically realistic, and morphologically grounded Question–Answer (Q&A) pairs for each given blood smear slide entry. Each data entry describes a blood smear slide at the report level, including the morphological description, manual differential count (CBC), and diagnosis context.</p> <p>### OBJECTIVE: Based only on the provided information, generate meaningful, exam-quality, and visually grounded Q&A pairs that reflect what a hematology student or pathologist might be asked based on this slide. Questions should test understanding of microscopic findings, abnormalities, and diagnostic reasoning that can be inferred from the morphological description and CBC data. Avoid introducing external textbook content or general biomedical facts unless they are explicitly supported by the provided data.</p> <p>### QUESTION GENERATION PRINCIPLES</p> <ul style="list-style-type: none"> - Generate 3–6 Q&A pairs for each slide (more if justified by rich content). - Ensure a balanced mix of question types: <ul style="list-style-type: none"> + True/False + Multi-choice + Close-ended + Open-ended - Include one True/False and at least one Multi-choice question per slide. - Questions must be answerable directly from the provided data, especially from: <ul style="list-style-type: none"> + the microscopic description + the CBC comment + the diagnosis information - Avoid overly general or knowledge-based questions that could be answered without the provided slide information. - Each question must be grounded in specific morphological, quantitative, or diagnostic features explicitly described in the input. - Do NOT ask general textbook questions such as: <ul style="list-style-type: none"> + “What is MDS?” + “What does SF3B1 mutation mean?” + “What are the normal functions of neutrophils?” - Focus instead on questions that test visual interpretation, morphological reasoning, or recognition of abnormalities mentioned in the description. - Every question should reference at least one observable or measurable feature (e.g., nuclear irregularity, granulation, cytoplasmic color, or cell proportions). - Emphasize morphology, nuclear/cytoplasmic features, and abnormal findings. - For CBC-related questions, do not directly ask for exact or approximate percentages. Instead, Phrase such questions as True/False or Multi-choice to test approximate interpretation rather than numeric recall. - If diagnostic information is present, include at least one reasoning-style question (e.g., “Which morphological feature supports the diagnosis of MDS?”). - Avoid generic, filler, or repetitive questions. <p>### QUESTION TYPE RULES</p> <ol style="list-style-type: none"> 1. True/False- Always start with ““(True/False)””. <ul style="list-style-type: none"> + If phrased as a question, end with a question mark. + If phrased as a statement, omit the question mark. + Must refer to observable or directly stated findings. 2. Multi-choice <ul style="list-style-type: none"> + Always use “Multi-choice” exactly. + Include “options” with 3–5 distinct and plausible choices. + “answer” must exactly match one of the “options”. + Do not include “A)”, “B)”, etc. 3. Close-ended <ul style="list-style-type: none"> + Short factual or descriptive questions requiring concise answers (text only). + Avoid any numeric or percentage-based answers. 4. Open-ended <ul style="list-style-type: none"> + Short reasoning or interpretation questions requiring justification. <p>### TOPIC FIELD: Each question must include a “topic” field describing its conceptual focus. Use one of the following categories:</p> <ul style="list-style-type: none"> - “morphology” — nuclear or cytoplasmic shape, granules, color - “abnormality” — atypical, dysplastic, or hypersegmented features - “cbc” — quantitative or proportional findings from CBC - “diagnosis” — morphological features supporting diagnosis - “reasoning” — interpretive or integrative reasoning <p>### EXAMPLE OUTPUT: {{examples}}</p> <p>Now generate between 3 and 6 Q&A pairs following the above structure for the slide below.</p>

Figure 10. Prompts for whole slide QA