FluoCLIP: Stain-Aware Focus Quality Assessment in Fluorescence Microscopy

Supplementary Material

8. Spatial Frequency

In Section 5, we used the spatial frequency (SF) [7] metric as a quantitative proxy for image sharpness to analyze stain-dependent focus behavior across datasets. For completeness, we provide the exact formulation of SF below.

Given an image $I \in \mathbb{R}^{M\times N}$, the row and column frequency components are defined as

$$RF = \sqrt{\frac{1}{(M-1)N}\sum_{i=1}^{M-1}\sum_{j=1}^{N}(I(i+1,j)-I(i,j))^2},$$
$$CF = \sqrt{\frac{1}{M(N-1)}\sum_{i=1}^{M}\sum_{j=1}^{N-1}(I(i,j+1)-I(i,j))^2},$$
$$SF = \sqrt{RF^2+CF^2}. \tag{11}$$

Higher SF values indicate sharper, more in-focus images, while lower values correspond to defocus.

9. Dataset Details

9.1. Overview of Datasets

We evaluated FluoCLIP on three datasets used in Section 5 and Section 6: **FocusPath** [13], **BBBC006** [17], and our fluorescence microscopy dataset **FluoMix**. To ensure consistent training across datasets, all z-stacks were reorganized into discrete focus levels, ranging from in-focus to severely defocused, following the ordinal labeling protocol described in the main paper (Section 4). Representative examples from each dataset are shown in Figure 5 and Figure 6. The former illustrates focus-level (rank) variations across datasets, whereas the latter highlights stain-dependent visual differences across the three datasets.

9.2. BBBC006

BBBC006 dataset consists of fluorescence images of Hoechst 33342- and phalloidin-stained U2OS cells. Each field of view contains a 32-plane z-stack (2μm spacing), with the in-focus slice at position 16. Following prior work, we relabeled the stack into 10 ordinal focus levels (0–9). In our experiments, we use 23,341 images for training and 5,843 images for testing.

9.3. FocusPath

FocusPath provides 8,640 bright-field image patches (1024×1024) from nine differently stained slides. Each patch is assigned an absolute z-level from 0 to 13. Due to limited samples in the most defocused slices (levels 12–13), we merge them into level 11, resulting in 12 ordinal levels (0–11). We follow the standard split of 3,876 training and 972 testing samples.

9.4. FluoMix

FluoMix is a fluorescence microscopy dataset designed to capture stain–dependent variability across multiple tissues. The dataset spans three tissue types (**brain**, **lung**, **liver**) and four fluorescent channels (Hoechst 33342/34580 or DAPI, Alexa Fluor 488, Cy3, Alexa Fluor 647). Each stain channel corresponds to distinct biological targets, such as neuronal markers (NeuN, NFM, TH), microglial markers (Iba-1), vascular markers (CD31, Collagen IV), and epithelial markers (CK19, Claudin, ZO-1). Table 8 summarizes the stain–target combinations included in FluoMix.

Each field of view is acquired as a full z-stack (34 planes for FluoMix) covering the progression from in-focus to severely defocused slices. For consistency with FocusPath and BBBC006, all z-stacks are converted into **10 ordinal focus levels** (0 = in-focus, 9 = out-of-focus) using the relative labeling scheme described in Section 4. Representative examples are shown in Figure 5.

For the main experiments, we used brain-tissue subset of FluoMix (D1-D4), consisting of 25,967 training and 6,506 testing images across four stain combinations.

10. Biological Preparation / Imaging Protocols

FluoMix was constructed from adult **rat brain, lung, and liver** tissues, following institutional ethical guidelines approved by the Institutional Review Board (IRB). All samples were processed using standardized fixation, immunostaining, and confocal imaging protocols to ensure cross-tissue consistency while preserving stain-dependent optical characteristics.

10.1. Tissue Preparation

C57BL/6J male rats (8 weeks old) were euthanized using CO_2 anesthesia (30–70% volume/min). Transcardial perfusion was performed with saline followed by fixation with 4% paraformaldehyde (PFA). Brain, lung, and liver tissues were post-fixed for 24 hours, cryoprotected in 30% sucrose, and sectioned into 40 μm slices.

10.2. Immunostaining

Free-floating sections from all tissues were blocked in PBS containing 3% bovine serum albumin (BSA) and 0.1% Triton X-100 for 1 hour at room temperature. Slices were in-

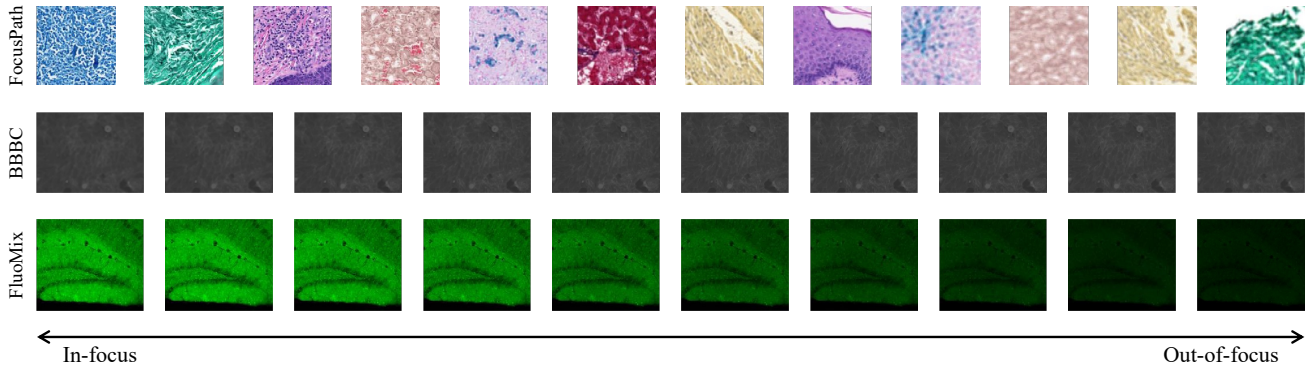

Figure 5. Examples of dataset classes: The figure displays samples of images from three datasets across different focus levels. (Top) FocusPath dataset images numbered from 0 to 11, each showing different staining techniques. (Middle) BBBC006 dataset images numbered from 0 to 9. (Bottom) FluoMix dataset images numbered from 0 to 9, representing different focus levels.

Table 8. Overview of the FluoMix dataset, summarizing tissue types, fluorescent stains, and associated biological targets. Each stain is paired with distinct protein markers (e.g., neuronal, microglial, vascular, epithelial), reflecting the heterogeneity of fluorescence signals across tissues. The rightmost column reports the number of fields of view collected for each stain–target combination.

Brain Tissue					
Dataset	**Hoechst 34580**	**Alexa 488**	**Cy3**	**Alexa 647**	**# Sets**
D1	nucleus	Iba-1	Tuj-1	Collagen IV	504
D2	nucleus	Neurofilament-M (NFM)	Tyrosine Hydroxylase (TH)	Collagen IV	152
D3	nucleus	NeuN	Tyrosine Hydroxylase (TH)	Collagen IV	554
D4	nucleus	GFAP	Tuj-1	CD31	623
Lung Tissue					
Dataset	**Hoechst 33342**	**Alexa 488**	**Cy3**	**Alexa 647**	**# Sets**
D5	nucleus	CD31	Vimentin	Collagen IV	634
D6	nucleus	CD31	Vimentin	Collagen IV	596
Liver Tissue					
Dataset	**DAPI**	**Alexa 488**	**Cy3**	**Alexa 647**	**# Sets**
D7	nucleus	CK19	Claudin	ZO-1	96

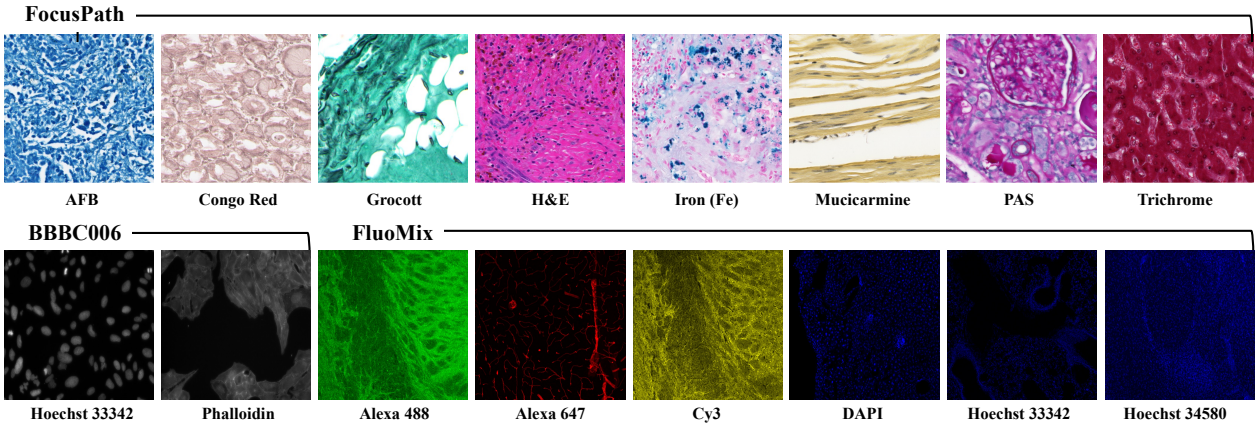

Figure 6. Sample images illustrating stain diversity in the three datasets. (Top) **FocusPath**: bright-field microscopy images with eight different histological dyes. (Bottom left) **BBBC006**: fluorescence microscopy images of cell lines labeled with Hoechst 33342 and Phalloidin. (Bottom right) **FluoMix**: tissue-level fluorescence microscopy images covering six distinct fluorescent staining protocols.

cubated overnight at 4°C with primary antibodies targeting diverse cellular structures, including:

- **Brain**: Iba-1, NeuN, NFM, Tuj-1, GFAP, TH, CD31, Collagen IV
- **Lung**: CD31, Vimentin, Collagen IV
- **Liver**: CK19, Claudin, ZO-1

After washing, sections were incubated with Alexa Fluor–conjugated secondary antibodies (Alexa Fluor 488, Cy3, Alexa Fluor 647). Hoechst 33342, Hoechst 34580 or DAPI (1:2000) was used to counterstain nuclei. Slides were washed, mounted with anti-fade medium, and allowed to dry under light-protected conditions.

10.3. Imaging Conditions

All fluorescence images were acquired using a Leica MICA confocal microscope with a 10× objective (NA 0.32) at a resolution of 1600 × 1352 pixels (0.586 μm/pixel). Excitation wavelengths were selected based on fluorophore spectra: 405 nm (Hoechst, DAPI), 488 nm (Alexa Fluor 488), 561 nm (Cy3), and 647 nm (Alexa Fluor 647), with matched bandpass emission filters to minimize spectral bleed-through.

To capture focus variability across staining conditions, each field of view (FOV) was acquired as a full z-stack:

- **Brain**: 34-plane stacks (1-2 μm spacing)
- **Lung**: 34-plane stacks (2 μm spacing)
- **Liver**: 34-plane stacks (2 μm spacing)

Laser power and exposure settings were adjusted per channel to optimize signal-to-noise ratio while maintaining consistency across tissue types. No deconvolution or contrast enhancement was applied to preserve the native optical properties of each stain.

11. Few-Shot Validations

Fluorescence microscopy exhibits substantial stain- and tissue-dependent variability, making it unrealistic to obtain dense focus annotations across all stain–tissue combinations encountered in practice. As a result, few-shot adaptation is a key requirement for stain-aware FQA models deployed in real imaging workflows, where supervision is scarce and domain shifts naturally arise across laboratories.

To evaluate FluoCLIP under these realistic constraints, we consider three complementary few-shot settings: (1) label-efficient learning with limited training samples, (2) generalization to unseen tissues under known stains, and (3) generalization to unseen stain–tissue pairs. As summarized in Table 9, these settings vary in whether the stain, the tissue, or both differ from the base training domain, thereby separating the effect of supervision scarcity from the effects of stain- and tissue-driven domain shifts. This structured evaluation allows us to systematically assess how well FluoCLIP transfers its stain-grounded representations under increasingly challenging and practically relevant scenarios.

11.1. Few-Shot Learning on FluoMix

To evaluate whether FluoCLIP can model Stain-Aware FQA with only a small number of labeled training samples, we first perform few-shot fine-tuning on the FluoMix brain-tissue subset (D1-D4). This restricted setup allows us to assess how effectively the model learns stain-dependent focus behavior when supervision is severely limited. All methods were initialized from the same pretrained CLIP image encoder (ResNet50) and text encoders. For baselines, we finetuned CoOp [26] following the protocol from the original paper, where the encoders are frozen and only the prompt embeddings are learned. Because both OrdinalCLIP [15] and FluoCLIP update both the image encoder and text prompts, we additionally include CoOp+E_{img}, which extends CoOp by also fine-tuning the image encoder to ensure a fair architectural comparison.

For each method, we sample k labeled samples per focus rank per stain ($k = \{1, 8, 16, 32, 64, 128\}$). Since D1-D4 contain four stain combinations and 10 ordinal focus ranks, each few-shot training set consists of $4 \times 10 \times k$ training images. Although all models start from the generic CLIP checkpoint, this experiment evaluates whether they can adapt to the fluorescence FQA domain using only this small subset. In particular, it measures how effectively each method acquires stain-aware focus behavior when fine-tuning from a natural-image pretraining domain to the fluorescence microscopy domain with extremely limited supervision. This setting allows us to directly test whether FluoCLIP's stain-aware design provides a stronger and more data-efficient adaptation capability than stain-agnostic baselines.

Adapting from CLIP to the fluorescence FQA domain with only a handful of labeled examples is inherently challenging, particularly because stain-dependent appearance variations amplify the effect of limited supervision. Nevertheless, the few-shot results in Table 10 reveal consistent behavioral patterns across methods. CoOp performs the worst, and CoOp+E_{img} often degrades further, consistent with observations from [26] that naive image-encoder fine-tuning can be harmful. OrdinalCLIP slightly outperforms FluoCLIP at very low shot counts (1–32), but its gains saturate quickly. In contrast, FluoCLIP exhibits the steepest improvement and ultimately achieves the best performance in the 64–128 shot range. Taken together, these results demonstrate that FluoCLIP's stain-aware design provides a more data-efficient adaptation mechanism, enabling the model to acquire stain-dependent focus behavior more rapidly than stain-agnostic baselines once modest supervision is available.

11.2. Few-Shot Generalization to Unseen Tissues

Next, we evaluated FluoCLIP along the *tissue generalization* axis by testing its ability to adapt to previously unseen

Table 9. Overview of the three few-shot evaluation settings used to assess label efficiency, cross-stain transfer, and cross-tissue generalization.

Generalization Type	Model Initialization	Adaptation Stain	Adaptation Tissue	Description
Few-Shot Learning	CLIP-pretrained on Natural Images	Alexa 488, Cy3, Alexa 647, Hoechst 34580	Brain	Test whether the model can learn stain-aware FQA with only k-shot samples
Cross-Tissue	FluoCLIP-pretrained on Brain (D1-D4)	Alexa 488, Cy3, Alexa 647 (Seen Stains)	Lung (Unseen Tissue)	Adaptation to new tissue with distinct morphology while keeping stains unchanged
Cross-Stain & Cross-Tissue	FluoCLIP-pretrained on Brain (D1-D4)	Hoechst 33342, DAPI (Unseen Stains)	Lung, Liver (Unseen Tissues)	Most challenging: test adaptation to entirely new stain-tissue pairs

Table 10. Comparison of accuracy (%) under few-shot learning settings on the FluoMix mouse brain dataset. CoOp follows the experimental protocol from the original paper [26], using prompt tuning only. CoOp+E_{img} extends CoOp by additionally fine-tuning the image encoder together with the prompts. OrdinalCLIP and FluoCLIP follow the main ResNet-50 experiment configuration described in Section 6. FluoCLIP exhibits the **steepest** improvement curve as supervision increases, demonstrating a strong stain-aware prior that enables rapid adaptation in low-data regimes. Results are reported as mean ± standard deviation over five runs with different random seeds.

#-Shots	1	8	16	32	64	128
CoOp [26]	11.72±1.10	16.32±0.91	17.88±0.44	18.36±0.61	18.38±0.44	18.50±0.50
CoOp+E_{img} [26]	13.58±1.88	14.92±1.10	14.98±0.88	15.82±1.17	17.06±0.78	19.22±1.17
OrdinalCLIP [15]	15.51±1.15	19.43±0.98	19.55±0.29	20.60±0.91	21.18±0.50	19.99±2.06
FluoCLIP	11.69±1.15	15.92±0.89	18.90±1.32	20.45±0.87	24.41±0.52	**29.11±2.13**

Table 11. Comparison of accuracy (%) results under few-shot settings on a held-out fluorescence microscopy dataset acquired from **mouse lung tissue** with a **known staining combination** (Alexa 488, Cy3, and Alexa 647). We observe that from the 32-shot regime, FluoCLIP exhibits a clear advantage, suggesting that its stain-aware modeling begins to effectively leverage **stain-specific signal variation** when sufficient examples are available. The results are mean ± standard deviation over five random seeds.

#-Shots	1	8	16	32	64	128
OrdinalCLIP [15]	20.09±0.42	21.87±1.23	23.99±0.50	29.14±0.66	41.26±1.01	61.73±0.62
FluoCLIP	20.35±0.79	20.95±1.05	23.51±0.62	29.43±0.63	43.27±1.19	67.97±1.31

Table 12. Comparison of accuracy (%) under few-shot settings on held-out fluorescence microscopy datasets of unseen **mouse lung, liver** tissues with an unseen staining combination (Hoechst 33342 and DAPI). Because both stains target the nucleus and are visually similar to Hoechst 34580 used during base training, OrdinalCLIP performs better in the extremely low-shot regime. However, from the 32-shot setting onward, FluoCLIP shows a clear advantage, demonstrating its ability to leverage **stain-aware adaptation** once sufficient examples become available. The results are reported as mean ± standard deviation over five random seeds.

#-Shots	1	8	16	32	64	128
OrdinalCLIP [15]	19.82±1.85	25.36±1.23	29.37±0.68	34.46±1.25	42.82±0.74	54.35±1.17
FluoCLIP	18.75±2.04	24.54±0.93	29.20±1.00	34.90±1.45	44.57±1.18	57.71±0.46

tissues while keeping the stain conditions fixed. The model was pretrained on brain tissue (D1–D4) and then adapted using k labeled samples per focus rank per stain from lung-tissue images (D5) stained with Alexa 488, Cy3, and Alexa 647. For this setting, the few-shot training set is constructed using the same sampling protocol–k labeled samples per focus rank for each stain ($k = \{1, 8, 16, 32, 64, 128\}$)–resulting in $3 \times 10 \times k$ training samples.

Tissue-dependent variations introduce substantial domain shifts in cellular morphology, density, and spatial texture, making this setting more challenging. As shown in Table 11, OrdinalCLIP [15] performs slightly better in the lowest-shot regime due to its stain-agnostic design, which avoids committing to stain priors that may behave differently across tissues. However, FluoCLIP again shows a sharp performance increase as k grows, surpassing OrdinalCLIP from the 32-shot regime onward. This indicates that FluoCLIP's stain-grounded modeling generalizes effectively across tissue-dependent fluorescence variations once modest supervision becomes available.

11.3. Few-Shot Generalization to Unseen Stain & Tissue Pairs

We now evaluate the most challenging generalization setting, where both the stain and the tissue differ from those seen during base training. Starting from a checkpoint trained on the FluoMix brain-tissue subset (D1–D4), we adapt the model to two entirely unseen nuclear stains–Hoechst 33342 (D5–D6) and DAPI (D7)–each acquired from different tissues than the training domain. Adaptation uses only k labeled samples per focus rank for each stain ($k = \{1, 8, 16, 32, 64, 128\}$), resulting in $2 \times 10 \times k$ training

images. To ensure comparability with earlier experiments, we restrict evaluation to nucleus-targeting stains (DAPI, Hoechst 33342). Since these stains highlights the same biological structure, morphological differences caused by tissue variability are reduced, allowing stain-dependent optical differences to be examined more clearly (Figure 6)

As shown in Table 12, OrdinalCLIP [15] slightly outperforms FluoCLIP at 1–16 shots. Because all unseen stains in this setting target the nucleus and therefore appear visually similar to Hoechst 34580 stained images from the base training domain, OrdinalCLIP's stain-agnostic prompts generalize reasonably well in this extremely low-shot regime. In contrast, FluoCLIP has insufficient samples to learn the subtle stain-specific variations. However, as more labeled samples become available, FluoCLIP quickly surpasses all baselines and widens the performance gap at 32–128 shot range. This demonstrates that explicit stain grounding provides superior adaptability when both the stain and tissue shift simultaneously, becoming increasingly advantageous once even modest supervision is available.

12. Ablation Study

To understand the contribution of each component in FluoCLIP, we conduct a systematic ablation by varying the stain-token type, the use of the two-stage grounding pipeline, and the inclusion of the text-side adapter.

We compare three types of stain tokens: S^{plain} is a non-learnable text token that simply inserts the stain name, S^{train} is a learnable token updated during training without explicit grounding, and S is our grounded stain token obtained through the two-stage procedure. Table 13 varies two factors that jointly define the stain-token type—Fine-tune (S) (trainable vs. frozen tokens) and Two Stage (with vs. without stain grounding)—along with an additional orthogonal choice of whether to include the text-side Adapter after CLIP text encoder.

With this ablation, we evaluate how different learning strategies for stain tokens-from frozen tokens to naïve trainable tokens to grounded tokens-and the presence of adapter affect the model's ability to acquire robust stain-aware focus representations.

12.1. Effects of Stain-Grounding

Configuration (A) uses non-learnable stain tokens (S^{plain}), while (C) replaces it with learnable stain tokens (S^{train}) but without the grounding stage (Two Stages). Interestingly, (C) performs worse than (A), indicating that simply making stain tokens trainable does not help. Without the grounding phase, the model receives no visual constraint linking stain-specific text features to fluorescence appearance, and learning tends to inject noise rather than meaningful stain semantics.

Table 13. Ablation of FluoCLIP's learning strategy. We evaluate how stain tokens should be trained and whether the text-side adapter provides additional gains. Configuration (A-F) combines three stain-token variants (S^{plain}: frozen, S^{train}: naïvely trained, S: grounded via two-stage learning) with optional use of the grounding pipeline (Two Stages) and the adapter (Adapter). Results (mean ± std over five seeds) isolate the contribution of stain grounding, trainability, and adapter design.

Type	Configuration	Fine-tune (S)	Two Stage	Adapter	Accuracy (%) ↑
(A)	S^{plain}	×	×	×	81.04±0.81
(B)	S^{plain} + Adapter	×	×	o	83.21±3.93
(C)	S^{train}	o	×	×	80.50±0.67
(D)	S^{train} + Adapter	o	×	o	81.38±0.63
(E)	S	o	o	×	81.83±0.95
(F)	S + Adapter	o	o	o	84.28±0.88

When the grounding phase is enabled (E), the performance increases over both (A) and (C) (by +0.79% and +1.33%, respectively), demonstrating that explicit grounding is essential for producing stain tokens that actually reflect stain-dependent optical cues. Importantly, these gains appear even before introducing the adapter, confirming that grounding–not trainability alone–is the key factor.

12.2. Effects of Adapter

Configuration (B) and (D) evaluate the adapter with non-grounded stain tokens (S^{plain} in (B), S^{train} in (D)). The adapter yields moderate improvements in both cases, but the gains remain limited because the underlying stain tokens are not aligned with visual features. Thus, architectural augmentation alone cannot produce meaningful stain-aware embeddings—the adapter benefits only materialize when the stain tokens themselves are grounded.

12.3. Combined Effects

Configuration (E) and (F) corresponds to the two-stage FluoCLIP pipeline, with (F) additionally using the text-side adapter. Grounded stain tokens alone (E) already outperform all non-grounded alternatives (A, C), verifying that the grounding stage is the primary source of performance improvement. Adding the adapter (F) achieves the best results overall (84.28%), confirming that grounded stain semantics and the adapter contribute complementary benefits. Overall, these ablations demonstrate that (i) stain-grounding is necessary for producing stable and informative stain-aware embeddings, and (ii) the adapter further refines these representations, but works best when grounding is present. This explains why the full FluoCLIP pipeline achieves the strongest performance.

Algorithm 1 Overall Workflow of FluoCLIP

Stage 1: Stain-Grounding
Input (Stage 1):
Training set $\mathcal{D}_{\text{stain}} = \{(x_i, s_i)\}$
Stain tokens $\mathbf{S}$, adapter $\mathcal{A}$
Context tokens $C = \{[\textit{stain-context}]\}$
Image encoder E_{img}
Text encoder E_{text}
Output (Stage 1):
Updated $\mathbf{S}, \mathcal{A}$
for epoch $= 1$ to stage1_max_epoch **do**
  Generate Prompt $\mathcal{P}^{\text{stain}}$ (Eq. 3)
  for $(x_i, s_i) \in \mathcal{D}_{\text{stain}}$ **do**
    Encode Image $v_i \leftarrow \mathrm{E}_{\text{img}}(x_i)$
    Encode Text $t_l^{stain} \leftarrow \mathcal{A}\big(\mathrm{E}_{\text{text}}(\mathcal{P}^{\text{stain}})\big)$
    Calculate Similarity $z_{i,l}^{stain} = \langle v_i, t_l^{stain} \rangle$
    Update $\mathbf{S}, \mathcal{A}$
  end for
end for

Stage 2: Stain-Guided Ranking
Input (Stage 2):
Training set $\mathcal{D}_{\text{rank}} = \{(x_i, s_i, r_i)\}$
Grounded stain tokens $\mathbf{S}$, adapter $\mathcal{A}$ from Stage 1
Base ranks $\mathbf{R}^{\text{base}}$
Conditioning network f_θ
Context tokens $C = \{[\textit{stain-context}], [\textit{rank-context}]\}$
Output (Stage 2):
Updated $\mathbf{R}^{\text{base}}$, f_θ, $\mathrm{E}_{\text{img}}, C$
for epoch $= 1$ to stage2_max_epoch **do**
  Generate Stain-Guided Rank $\tilde{\mathbf{R}}_k^l$ (Eq. 5, 6)
  for $(x_i, s_i, r_i) \in \mathcal{D}_{\text{rank}}$ **do**
    Encode Image $v_i \leftarrow \mathrm{E}_{\text{img}}(x_i)$
    Generate Prompt $\mathcal{P}^{\text{rank},i}$ (Eq. 7)
    Encode Text $t^{\text{rank},i} \leftarrow \mathcal{A}\big(\mathrm{E}_{\text{text}}(\mathcal{P}^{\text{rank},i})\big)$
    Calculate Similarity $z_k^{rank,i} = \langle v_i, t_k^{\text{rank},i} \rangle$
    Update $\mathbf{R}^{\text{base}}$, $f_\theta, \mathrm{E}_{\text{img}}, C$
  end for
end for

13. Algorithm

Algorithm 1 outlines the overall two-stage pipeline of FluoCLIP. **Stage 1** performs *stain-grounding*, in which the stain tokens and adapter are optimized with stain-class supervision to align vision–language representations with fluorescence-specific stain semantics. **Stage 2** conducts *stain-guided ranking*. The grounded stain embeddings modulate the base rank embeddings through a conditioning network, and interpolation yields continuous stain-conditioned rank tokens. These rank tokens, together with learnable context tokens and the fixed grounded stain tokens, are assembled into sample-specific sentences and encoded by the frozen CLIP text encoder. Image–text similarities are then used to predict the focus rank, while the base rank tokens, context tokens, conditioning network, and image encoder are updated with ranking supervision.

Table 14. Comparison with biomedical CLIP variants on FluoMix. Each method uses its own pretrained ResNet50-based vision encoder corresponding to its original pretraining. Results are reported as mean ± standard deviation over five runs.

Method	PubMedCLIP	PMC-CLIP	FluoCLIP
Accuracy (%)	81.48±0.34	84.06±0.23	85.21±0.88

14. Biomedical CLIP Variants

To evaluate whether domain-specific pretraining alone suffices for fluorescence FQA, we compare FluoCLIP with biomedical CLIP variants, including PubMedCLIP [8] and PMC-CLIP [16]. These models are pretrained on large-scale biomedical image–text corpora and provide stronger domain alignment than the original CLIP.

As shown in Table 14, both PubMedCLIP and PMC-CLIP achieve competitive performance on FluoMix, outperforming standard image-only baselines. In particular, PMC-CLIP benefits from broader biomedical coverage and achieves higher accuracy than PubMedCLIP.

However, FluoCLIP consistently outperforms both models, achieving the highest accuracy. This result indicates that domain-specific pretraining alone is insufficient for FQA, as it does not explicitly capture stain-dependent optical variations or ordinal focus structure. By incorporating stain-aware grounding and stain-conditioned ordinal modeling, FluoCLIP effectively resolves these limitations and yields more consistent and accurate focus predictions across diverse fluorescence conditions.