MICROCLIP: CONNECTING TEXT AND MICROSCOPY IMAGES VIA CONTRASTIVE LANGUAGE–IMAGE PRE-TRAINING

Anonymous authors
Paper under double-blind review

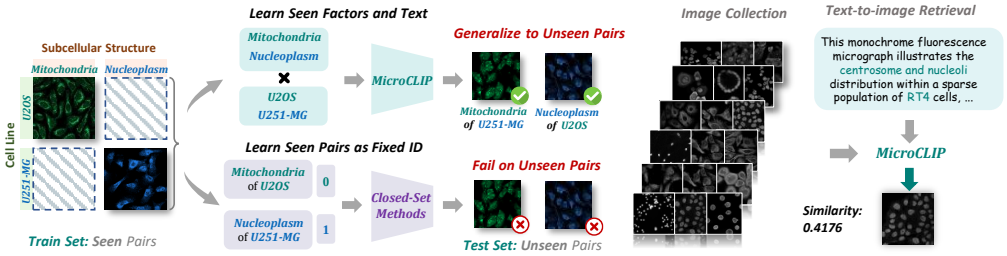


Figure 1: (1) We connect fluorescence microscopy images with text, so that subcellular structures and cell lines can both be represented in a shared semantic space rather than as isolated class IDs. (2) Building on this shared semantic space, MicroCLIP generalizes to unseen structure–cell-line pairs, whereas conventional closed-set methods are restricted to seen pairs. (3) Beyond recognition, this shared representation further enables language-based retrieval, allowing users to search microscopy images through natural language descriptions.

ABSTRACT

Understanding fluorescence microscopy images is fundamental to automating biological research. However, existing methods typically learn predefined concepts from closed-set data, limiting transfer across tasks and biological contexts. In this work, we introduce MicroCLIP, a vision–language model that learns generalizable cellular and subcellular representations from natural language supervision for open-set image classification and retrieval. To pretrain MicroCLIP, we construct MicroPool-2M, a dataset of approximately 2.5 million fluorescence microscopy image–text pairs spanning diverse cellular structures, cell lines, and imaging conditions. Aligning microscopy images with biological concepts in natural language is challenging because most images lack descriptive captions and biological knowledge annotations. We therefore progressively establish image–text associations at three levels to provide MicroCLIP with richer semantic supervision: (1) category-level alignment links images to standardized labels defined by subcellular structure–cell-line combinations; (2) knowledge-level alignment incorporates biological priors derived from these labels to capture relationships across species and cell types; and (3) detail-level alignment uses a multimodal large language model to recaption and verify image–text pairs at scale, emphasizing fine-grained visual characteristics. Extensive experiments demonstrate that MicroCLIP achieves leading performance in zero-shot classification and text–image retrieval on microscopy images. It retrieves images with specific cellular characteristics from thousands of candidates and generalizes to unseen biological concepts without additional supervision. These results highlight MicroCLIP's potential as a foundation model for fluorescence microscopy. Models, training code, and datasets will be publicly released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Large-scale fluorescence microscopy generates massive amounts of biological imaging data, creating a growing need for automated image analysis (Ouyang et al., 2019; Caicedo et al., 2017). A fundamental requirement for such automation is the reliable recognition of biological content in microscopy images. In protein-localization microscopy, images reveal the spatial distribution of proteins across subcellular structures (Thul et al., 2017). The fluorescence pattern associated with a given subcellular structure can vary across cell lines, making both subcellular structure and cell line relevant factors for recognition (Zhang et al., 2025). Yet available datasets cover only a subset of their possible combinations, leaving many biological contexts unobserved during training. Models intended for broad use must therefore generalize beyond the structures, cell lines, and structure–cell-line combinations observed during training. This is difficult for conventional closed-set classifiers, whose predictions are restricted to a predefined label space (Bendale & Boult, 2016).

Language provides a natural interface for moving beyond this fixed label space, allowing recognition targets to be specified through text rather than tied to class IDs alone (Radford et al., 2021). For fluorescence microscopy, however, realizing this potential requires grounding biological semantics in image content. On the semantic side, category annotations identify subcellular structures and cell lines, while biological knowledge describes their properties and functions. On the visual side, microscopy images exhibit instance-specific fluorescence patterns and cellular morphology. Establishing a reliable correspondence between these semantics and visual patterns remains challenging. As illustrated in Section 3, a model can describe relevant biological knowledge and image appearance yet fail to identify the correct biological category. This gap between biological semantics and microscopy images limits reliable language-guided recognition.

Existing approaches use visual information and biological semantics in different ways, but remain limited for language-guided microscopy recognition. Image-only methods such as cytoself (Kobayashi et al., 2022) and Cell-DINO (Moutakanni et al., 2025) learn microscopy representations without language supervision. General-purpose vision–language models such as CLIP (Radford et al., 2021) and SigLIP 2 (Tschannen et al., 2025) align images and text broadly, but lack microscopy-specific biological supervision. Domain-specific multimodal methods move closer, yet target specialized tasks: FluoCLIP (Park et al., 2026) uses stain semantics for focus quality assessment, while CellCLIP (Lu et al., 2025) relates cellular phenotypes to perturbation descriptions.

To capture the correspondence between microscopy images and biological semantics, we curate **MicroPool-2M**, the largest fluorescence microscopy image–text dataset to date, comprising 2.5 million single-channel images paired with textual descriptions. For each image, MicroPool-2M provides natural language descriptions at three levels: category-level captions specify the subcellular structure and cell line, knowledge-level captions describe their biological properties, and detail-level captions capture the fluorescence patterns visible in the image. Together, these image–text pairs cover diverse cell types and subcellular structures, as well as a broad range of biological knowledge.

Building on MicroPool-2M, we introduce **MicroCLIP**, a vision–language model for fluorescence microscopy. Through contrastive learning, MicroCLIP aligns microscopy images with natural language descriptions at three levels, learning representations that generalize to unseen combinations of cell lines and subcellular structures. It enables biological researchers to use natural language to search large-scale microscopy image collections for cellular morphologies and subcellular structures of interest, without task-specific fine-tuning. As shown in Table 1, MicroCLIP outperforms the evaluated general-purpose vision–language models in zero-shot classification and bidirectional image–text retrieval. On unseen structure–cell-line combinations, it achieves 67.70% classification accuracy, compared with 5.12% for the strongest general-purpose baseline.

Our contributions are threefold:

- We identify a biological grounding gap in fluorescence microscopy, highlighting the need to explicitly connect microscopy appearance with biological semantics for language-guided image understanding.
- We introduce **MicroPool-2M**, a large-scale multi-source fluorescence image–text dataset that grounds biological semantics in microscopy images through category-level, knowledge-level, and detail-level captions.

- We develop **MicroCLIP**, which learns a shared microscopy image–text representation from this multi-level supervision, substantially improving zero-shot generalization while naturally extending fluorescence microscopy analysis to bidirectional language-based retrieval.

2 Related Work

Representation Learning for Cellular Microscopy. Self-supervised methods such as cytoself (Kobayashi et al., 2022) and Cell-DINO (Moutakanni et al., 2025) learn microscopy representations for protein localization and cellular characterization without language supervision. PUPS (Zhang et al., 2025) incorporates protein sequences and cellular landmarks to predict localization in unseen proteins and cell lines. Cross-modal methods connect cellular images to other biological information: CLOOME (Sanchez-Fernandez et al., 2023) aligns images with chemical structures, while CellCLIP (Lu et al., 2025) aligns Cell Painting phenotypes with textual descriptions of perturbations. For brain histology, Sutton et al. (2026) combine label-linked literature with image-specific properties to supervise cytoarchitectural description generation. These approaches establish complementary links between microscopy and biological information. MicroCLIP focuses on language-guided recognition and retrieval of subcellular structures across cell lines through a shared fluorescence image–text representation.

Vision–Language Pretraining and Supervision Construction. CLIP (Radford et al., 2021) establishes image–text pretraining as a framework for zero-shot recognition and retrieval. SigLIP (Zhai et al., 2023), SigLIP 2 (Tschannen et al., 2025), and PE-Core (Bolya et al., 2025) advance general-purpose representations through improved objectives and training recipes. Data construction is another central direction: MetaCLIP (Xu et al., 2024) balances metadata distributions, and DFN (Fang et al., 2024) learns to filter image–text data. On the text side, LaCLIP (Fan et al., 2023) samples original captions and language rewrites, while VeCLIP (Lai et al., 2024) enriches captions with visual information and mixes them with original text. These studies motivate attention to both supervision content and sampling. MicroCLIP applies this perspective to fluorescence microscopy, pairing images with complementary category identities, biological knowledge, and instance-specific visual descriptions while retaining the standard contrastive objective.

Microscopy Resources and Generalization Evaluation. The Human Protein Atlas (Thul et al., 2017) and OpenCell (Cho et al., 2022) provide annotated protein-localization images, while RxRx1 (Sypetkowski et al., 2023) and Cell Painting resources (Chandrasekaran et al., 2024) capture variation across experimental batches and biological perturbations. MicroPool-2M integrates multiple microscopy sources into a common annotation vocabulary and augments images with three levels of text supervision. Its evaluation distinguishes unseen structures, unseen cell lines, and unseen pairings, complemented by an external source-shift benchmark. The unseen-pair setting relates to compositional zero-shot learning, which evaluates novel combinations of familiar concepts (Nagarajan & Grauman, 2018; Nayak et al., 2023); unseen-factor evaluation is also studied in open-vocabulary compositional learning (Saini et al., 2024). We use these distinctions to assess the transfer of microscopy image–text representations alongside language-based retrieval.

3 Observation: The Biological Grounding Gap

Figure 2 illustrates the biological grounding gap using GPT-6. In response to text queries, the model describes the biological properties of cytosol and SiHa cells. Given a microscopy image, it describes diffuse, granular fluorescence and cellular morphology. Yet when asked to identify the subcellular structure and cell line, it predicts nucleoplasm in U-2 OS cells rather than the annotated cytosol in SiHa cells. This example shows that the ability to describe biological knowledge and visual patterns does not guarantee correct biological identification. We refer to this difficulty in reliably associating microscopy images with their biological semantics as the biological grounding gap.

To examine this difficulty beyond a single example, we select 10 categories from each of the IID and Compositional Shift subsets, with three images per category. Each image is classified among the 10 structure–cell-line candidates selected for its split. As shown in Figure 2(d), GPT-6 achieves joint category accuracies of 40.00% (12/30) on IID and 33.33% (10/30) on Compositional Shift, whereas MicroCLIP achieves 100.00% (30/30) and 90.00% (27/30), respectively. These results show that GPT-6's biological identification remains unreliable in this diagnostic even when candidate identities are supplied, while MicroCLIP achieves higher accuracy on the same subsets. This small

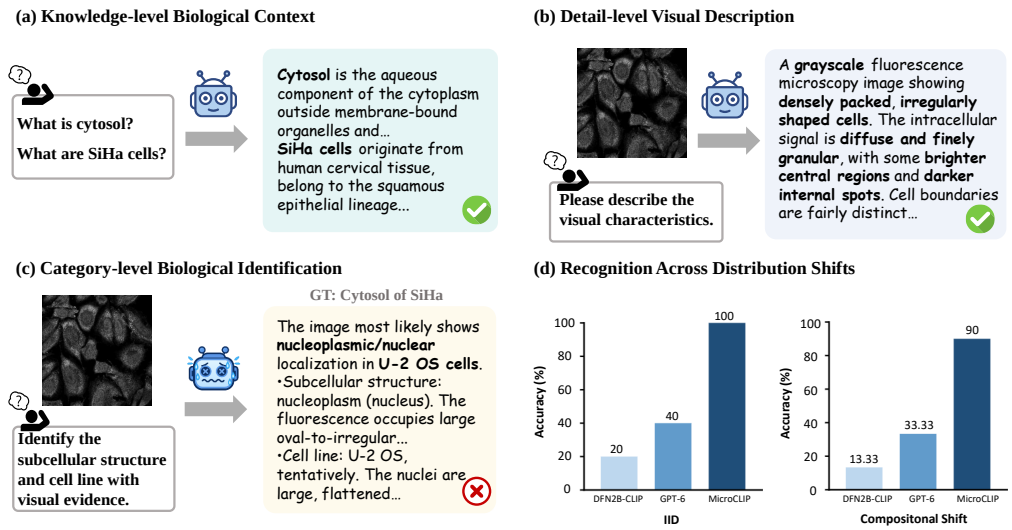


Figure 2: Illustration of the biological grounding gap. (a–c) GPT-6 responses to biological knowledge, visual description, and biological identification queries, respectively. The predicted identity in (c) differs from the reference annotation. (d) Joint structure–cell-line classification accuracy of DFN2B-CLIP, GPT-6, and MicroCLIP on diagnostic subsets of IID and Compositional Shift, with 10 candidate categories and 30 images per split.

evaluation does not establish the cause of the errors or imply that every cell-line identity is uniquely recoverable from a single fluorescence image. Sampling, prompts, and full comparison results are provided in Appendix D. These observations motivate learning a more reliable correspondence between microscopy images and biological semantics.

4 MicroPool-2M: A Large-Scale Microscopy Image–Text Dataset

Motivated by the biological grounding gap identified in Section 3, we construct MicroPool-2M to provide explicit supervision linking microscopy images to biological semantics. Each fluorescence microscopy image is paired with three complementary levels of text: category-level captions specify its annotated biological identity, knowledge-level captions provide associated biological context, and detail-level captions describe its visible appearance. Together, these captions provide the supervision used to train MicroCLIP to associate biological meaning with microscopy appearance.

Image collection and annotation harmonization. Covering diverse biological contexts requires varied image sources and consistent category definitions. We aggregate single-channel fluorescence images from the Human Protein Atlas (Thul et al., 2017), Image Data Resource (Williams et al., 2017), RxRx (Sypetkowski et al., 2023), Broad Bioimage Benchmark Collection (Ljosa et al., 2012), and curated literature. Source-specific aliases are mapped to a common vocabulary, assigning each image x_i a structure label s_i and cell-line identity c_i. A structure label can denote a single structure or marker, or a combination of them. Each category is represented as $e_i = (s_i, c_i)$. We remove content-identical images before constructing training and evaluation sets.

Three-level text supervision. Category labels identify biological entities but do not describe their biological context or the visual appearance of individual images. We therefore pair each image x_i with category-level, knowledge-level, and detail-level text:

$$t_i^{\text{cat}} = \mathcal{T}(s_i, c_i), \quad t_i^{\text{know}} = \text{Concat}(\mathcal{K}_s(s_i), \mathcal{K}_c(c_i)), \quad t_i^{\text{det}} = \mathcal{D}(x_i, e_i). \tag{1}$$

Here, $\mathcal{T}$ applies the category template: *a microscopy photo of* $\{structure\}$_of_$\{cell\}$. The mappings $\mathcal{K}_s$ and $\mathcal{K}_c$ retrieve all available biological descriptions generated by GPT-5.6 Sol for the associated structures or markers and cell line, respectively; composite structure labels include descriptions of all constituent structures. For $\mathcal{D}$, we prompt Qwen3.5-397B-A17B with the image and the normalized category name corresponding to e_i to describe fluorescence distribution, morphology, signal intensity, and background. Appendix B provides prompts and caption examples.

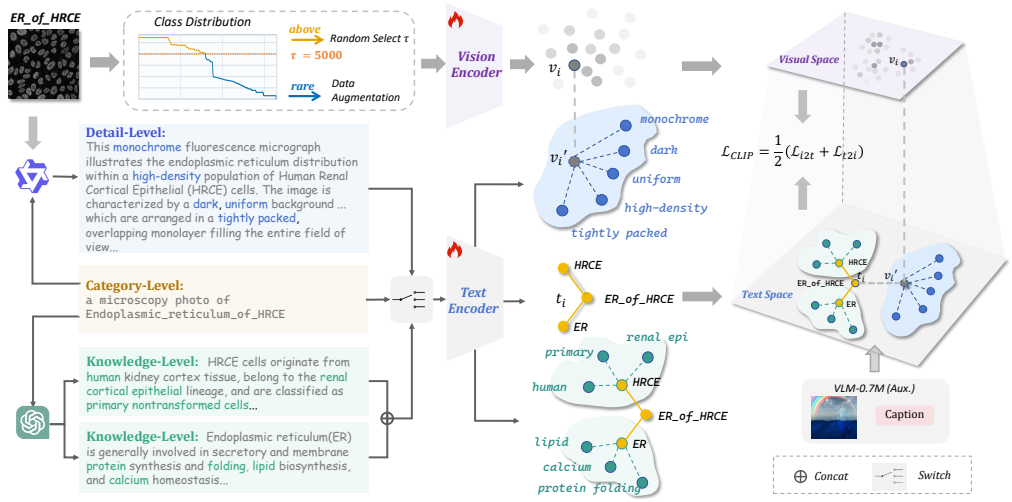


Figure 3: MicroCLIP learns a shared image–text representation from category-level, knowledge-level and detail-level supervision. Each microscopy image is paired with one sampled caption for CLIP training. Composition balancing and rare-class augmentation adjust training exposure, while general-domain image–text pairs provide auxiliary supervision.

Dataset coverage. MicroPool-2M contains 3,301,672 images and 3,391 observed structure–cell-line pairs. After data cleaning and split construction, the training split contains 2,505,017 images spanning 2,032 structure–cell-line categories. The resulting annotations, knowledge tables, and image descriptions define the supervision used by MicroCLIP; the evaluation subsets and protocols are specified in Section 6.1.

5 MicroCLIP: Grounding Biological Semantics in Microscopy

Building on MicroPool-2M, MicroCLIP learns to associate microscopy appearance with biological semantics in a shared image–text space. The three caption levels connect each image to its categorical identity, associated biological knowledge, and instance-specific visual traits. We use the CLIP dual-encoder architecture and symmetric contrastive objective, sampling one caption level for each image occurrence during training so that repeated observations expose the model to complementary textual descriptions. Composition-aware sampling balances training exposure across structure–cell-line pairs, while auxiliary general-domain image–text data provide broader visual and linguistic supervision. Figure 3 summarizes the framework.

5.1 Learning from Three-Level Text Supervision

Complementary supervision. The three caption levels link microscopy appearance to biological semantics at complementary levels of specificity. Category captions name the annotated structure–cell-line composition, providing a common identity anchor for images in the same category. Knowledge captions describe the biological properties of the annotated entities, providing shared semantic context across compositions that contain the same structure or cell line. Detail captions describe image-specific fluorescence patterns, morphology, signal intensity, and background, capturing visual variation within a category. Together, these text views allow images to be associated with both shared biological meaning and instance-specific appearance.

Text-view sampling. Each time a microscopy image x_i is selected for training, we sample one caption to form the positive image–text pair (x_i, t_i):

$$t_i = \begin{cases} t_i^{\text{cat}}, & \text{with probability } 0.50, \\ t_i^{\text{know}}, & \text{with probability } 0.25, \\ t_i^{\text{det}}, & \text{with probability } 0.25. \end{cases} \tag{2}$$

The higher sampling probability for category captions maintains frequent supervision of target identity, while knowledge and detail captions provide complementary context and visual descriptions.

Captions are resampled on each occurrence throughout training, allowing the same image to receive different text views across repeated observations.

Contrastive learning. Let $\mathbf{v}_i$ and $\mathbf{u}_i$ be the ℓ_2-normalized outputs of the image encoder $f_\theta(x_i)$ and text encoder $g_\phi(t_i)$. For a batch of B pairs, the similarity logits are

$$\ell_{ij} = \alpha\, \mathbf{v}_i^\top \mathbf{u}_j, \tag{3}$$

where $\alpha > 0$ is a learned logit scale. We optimize the standard symmetric CLIP loss (Radford et al., 2021):

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{B} \exp(\ell_{ij})} + \log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{B} \exp(\ell_{ji})} \right]. \tag{4}$$

All text views share the same encoder and objective, bringing the three forms of supervision into a common image–text space.

5.2 Composition-Aware Sampling

Figure 7 shows that the microscopy training set follows a long-tailed distribution over structure–cell-line compositions. To reduce the dominance of frequent compositions, we follow the metadata-balancing strategy of MetaCLIP (Xu et al., 2024). Each candidate occurrence associated with composition e is retained with probability

$$p_{\text{keep}}(e) = \min\left(1, \frac{\tau_a}{n(e)}\right), \tag{5}$$

where $n(e)$ is the number of candidate records for composition e, and $\tau_a = 5{,}000$ controls the effective exposure of frequent compositions. It retains all candidates from compositions below the threshold while progressively reducing the contribution of those with larger image counts.

We additionally apply standard rotation and reflection augmentation to sparsely represented compositions before constructing the training stream. The resulting augmented records are included in $n(e)$, and the augmentation criterion and allocation procedure are provided in Appendix C.2. Retention is applied before text-view sampling, so the exposure of subcellular structure–cell-line compositions and the selection of text levels are controlled independently.

5.3 Auxiliary General-Domain Image–Text Data

MicroPool-2M provides domain-specific microscopy supervision, but its image and text distributions are narrower than those of general-domain vision–language corpora. We therefore mix in VLM-0.7M, a subset of 609,740 image–text records drawn from the VLM-150M dataset released by HQ-CLIP (Wei et al., 2025), to provide broader visual and linguistic exposure during training.

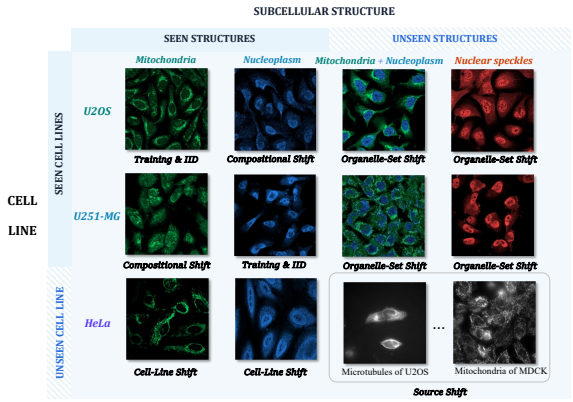
Each auxiliary image is paired with its tag string or VLM-generated description with equal probability. These records are mixed with MicroPool-2M under the same contrastive objective, providing broader visual and linguistic supervision.

6 Experiments

6.1 Experimental Settings

Evaluation data and protocols. We evaluate classification on four MicroPool-2M subsets defined relative to its microscopy training set and summarized in Table 6. *IID* contains new images of observed structure–cell-line pairs. *Compositional Shift* holds out pairs while retaining both constituent labels in training. *Cell-Line Shift* holds out cell lines while retaining the structure labels, and *Organelle-Set Shift* holds out complete structure labels while retaining the cell lines. The latter concerns structure labels, which can themselves be composite. Each internal evaluation category contains 100 images. We additionally evaluate *Source Shift* using 1,158 images from 13 categories drawn from the external Light My Cells dataset (S-BIAD1047) (Kauffmann et al., 2026). Figure 4 shows representative images from these evaluation settings.

For retrieval, we use a held-out pool of 1,000 MicroPool image–text pairs, with each image paired with its detail-level caption. We additionally evaluate zero-shot transfer to projected 3D cell images from the Allen WTC-11 collection in Appendix F.



(a) Evaluation examples.

Subset	Classes	Images
IID	59	5,900
Compositional Shift	64	6,400
Cell-Line Shift	59	5,900
Organelle-Set Shift	54	5,400
Source Shift (external)	13	1,158
Training	2,032	2,505,017

(b) Evaluation statistics.

Figure 4: Classification evaluation subsets: (a) representative images displayed using pseudocolor; (b) category and image counts for the four internal splits, the external Source Shift benchmark, and the training set.

Evaluation metrics. For classification, image–text models rank all candidate categories within each evaluated subset using the category template in Section 4. We report Top-1 accuracy without adaptation to the evaluation subsets. For retrieval, we report Recall@1 for image-to-text (I2T) and text-to-image (T2I), ranking the full corresponding pool for each query.

Baselines. The fully supervised classifier is trained on the full microscopy training set and evaluated on the 59-category IID subset. For Cell-DINO (Moutakanni et al., 2025), we self-supervise the backbone on the same 2.5M microscopy training images and fit separate linear probes for structure and cell-line recognition; their predictions are combined for IID and Compositional Shift evaluation. CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), SigLIP 2 (Tschannen et al., 2025), PE-Core (Bolya et al., 2025), and DFN2B-CLIP (Fang et al., 2024) use released weights and native configurations, without microscopy adaptation or prompt ensembling. These comparisons measure direct transfer from general-purpose pretraining.

Implementation details. MicroCLIP initializes both encoders from DFN2B-CLIP ViT-B/16 and fine-tunes all parameters with 224×224 image inputs and a 256-token text context. The training stream combines 2,505,017 microscopy images, 609,740 natural-image records, and 175,516 transformed copies of rare-class microscopy images. We train for 12,000 updates using AdamW, a global batch size of 8,192, a peak learning rate of 4×10^{-5}, 2,500 warmup steps, cosine decay, and weight decay 0.2. Training uses bf16 mixed precision on eight NVIDIA RTX PRO 5000 GPUs with 72 GB of memory each and takes approximately 55 hours.

6.2 Classification and Cross-Modal Retrieval

Table 1 shows that MicroCLIP generalizes to unseen structure–cell-line pairings while maintaining high IID accuracy. It achieves 67.70% on Compositional Shift, compared with 5.12% for the strongest general-purpose VLM and 0.98% for Cell-DINO under factor-wise linear probing. On IID images, it reaches 98.46%, compared with 97.74% for the fully supervised classifier.

Transfer also extends to held-out cell lines, structure labels, and an external source. MicroCLIP achieves 34.07% on Cell-Line Shift and 46.13% on Organelle-Set Shift, versus 7.63% and 2.54% for the strongest general-purpose baselines. Organelle-Set Shift concerns unseen complete labels, which may be composite, rather than necessarily unseen atomic structures. On Light My Cells (Kauffmann et al., 2026), MicroCLIP reaches 36.53%, versus 22.54% for PE-Core, although accuracy under distribution shift remains well below IID performance.

The same representation supports retrieval with image-specific descriptions. On 1,000 held-out image–caption pairs with detail-level captions, MicroCLIP achieves 61.5% I2T and 63.3% T2I Recall@1, compared with at most 0.7% and 0.3% for the evaluated general-purpose VLMs.

Table 1: Comparison of classification and cross-modal retrieval performance. Classification results are Top-1 accuracy (%), retrieval results are Recall@1 (%). Best and second-best results are shown in bold and underlined, respectively.

Method	Params.	In-domain classification	Out-of-domain classification				Retrieval	
		IID	Compositional Shift	Cell-Line Shift	Organelle-Set Shift	Source Shift	I2T	T2I
Closed-set methods								
Fully Supervised Classifier	–	<u>97.74</u>	–	–	–	–	–	–
Cell-DINO	304M	36.06	0.98	–	–	–	–	–
Open-set methods								
CLIP (OpenAI)	150M	2.97	<u>5.12</u>	1.80	<u>2.54</u>	8.38	<u>0.7</u>	0.2
SigLIP	203M	3.05	1.88	1.07	1.26	9.33	0.2	0.2
SigLIP 2	375M	3.08	2.28	3.59	1.11	10.79	0.2	<u>0.3</u>
PE-Core	448M	3.34	3.78	<u>7.63</u>	2.19	<u>22.54</u>	0.4	0.0
DFN2B-CLIP	150M	1.44	1.41	4.32	0.81	16.49	0.5	0.1
MicroCLIP (Ours)	150M	**98.46**	**67.70**	**34.07**	**46.13**	**36.53**	**61.5**	**63.3**

Table 2: MicroCLIP component ablations. Cat./Det./Know.: category/detail/knowledge captions. Bal.: composition balancing. Aux.: general-domain image–text pairs from VLM-0.7M.

Method	Cat.	Det.	Know.	Bal.	Aux.	In-domain classification	Out-of-domain classification				Retrieval	
						IID	Compositional Shift	Cell-Line Shift	Organelle-Set Shift	Source Shift	I2T	T2I
Category-level only	✓					97.37	56.47	23.25	40.74	<u>37.31</u>	15.8	12.6
+ Detail-level captions	✓	✓				**98.90**	63.94	28.69	41.63	**40.59**	28.4	27.8
+ Knowledge-level captions	✓	✓	✓			<u>98.86</u>	<u>66.36</u>	25.41	<u>48.31</u>	36.36	37.1	37.4
+ Data Balancing	✓	✓	✓	✓		98.24	64.95	<u>31.53</u>	**49.91**	36.44	<u>39.4</u>	<u>41.5</u>
+ VLM-0.7M	✓	✓	✓	✓	✓	98.46	**67.70**	**34.07**	46.13	36.53	**61.5**	**63.3**

6.3 Ablation Studies

Ablation on Main Components. Table 2 compares components using the same DFN2B-initialized ViT-B/16 backbone and contrastive objective. Category-level supervision supports compositional recognition but provides limited alignment with image-specific descriptions. Adding detail captions raises retrieval to 28.4%/27.8%, and adding knowledge captions further raises it to 37.1%/37.4%. Compositional Shift accuracy increases from 56.47% to 66.36%.

Composition balancing then improves Cell-Line Shift accuracy from 25.41% to 31.53%. With the auxiliary general-domain image–text pairs from VLM-0.7M, the full recipe reaches 61.5%/63.3% retrieval and the highest Compositional and Cell-Line Shift accuracy in this ablation. Gains vary across shifts: the balanced microscopy-only variant performs best on Organelle-Set Shift, while category-plus-detail supervision performs best on Source Shift.

Composition-balancing threshold. As shown in Table 3, lowering τ_a reduces the sampling dominance of frequent structure–cell-line pairs. Our default, $\tau_a = 5{,}000$, improves Cell-Line Shift accuracy from 26.61% without balancing to 34.07% and gives the best Compositional Shift and bidirectional retrieval results among the tested thresholds, with little change in IID accuracy. The gains are not monotonic: Organelle-Set Shift favors $\tau_a = 20{,}000$, while Source Shift favors $\tau_a = 50{,}000$.

Table 3: Effect of the composition-balancing threshold τ_a (Equation 5). Retained (%) denotes the online sample acceptance fraction. Metrics and highlighting follow Table 1.

Balancing threshold τ_a	Retained (%)	In-domain classification	Out-of-domain classification				Retrieval	
		IID	Compositional Shift	Cell-Line Shift	Organelle-Set Shift	Source Shift	I2T	T2I
off (no balancing)	100.0	<u>98.76</u>	<u>66.64</u>	26.61	46.78	30.05	32.4	35.7
50,000	49.8	**98.83**	64.73	<u>28.44</u>	<u>47.35</u>	**39.29**	<u>47.0</u>	<u>47.6</u>
20,000	30.7	98.24	65.52	27.88	**47.80**	34.54	44.5	45.5
5,000	10.5	98.46	**67.70**	**34.07**	46.13	<u>36.53</u>	**61.5**	**63.3**

Training-data size. Table 4 shows that increasing microscopy data from 25% to 100% under an equal-epoch protocol improves Compositional Shift accuracy from 60.39% to 66.97% and strengthens retrieval, but reduces Cell-Line and Source Shift accuracy. Thus, more data benefits compositional recognition and retrieval without uniformly improving transfer. The final row reports the full recipe with VLM-0.7M (Wei et al., 2025); it also changes balancing and rare-class augmentation, so it does not isolate the effect of additional data.

Table 4: Training-data scale and the full recipe. Samples (M) counts images before augmentation and balancing. Metrics and highlighting follow Table 1.

Training data	Samples (M)	In-domain classification	Out-of-domain classification				Retrieval	
		IID	Compositional Shift	Cell-Line Shift	Organelle-Set Shift	Source Shift	I2T	T2I
25%	0.63	**98.76**	60.39	33.05	44.56	**40.93**	32.1	32.0
50%	1.25	98.49	65.05	30.44	45.19	39.81	32.4	30.8
100%	2.51	98.54	66.97	30.05	45.54	37.31	60.1	60.7
100% + VLM-0.7M	3.11	98.46	**67.70**	**34.07**	**46.13**	36.53	**61.5**	**63.3**

6.4 Qualitative Visualization

Retrieval. To illustrate cross-modal retrieval, Figure 5 presents both retrieval directions. MicroCLIP retrieves descriptions and images matching the queried structure–cell-line identities, whereas DFN2B-CLIP and CLIP return mismatched annotations in these examples.

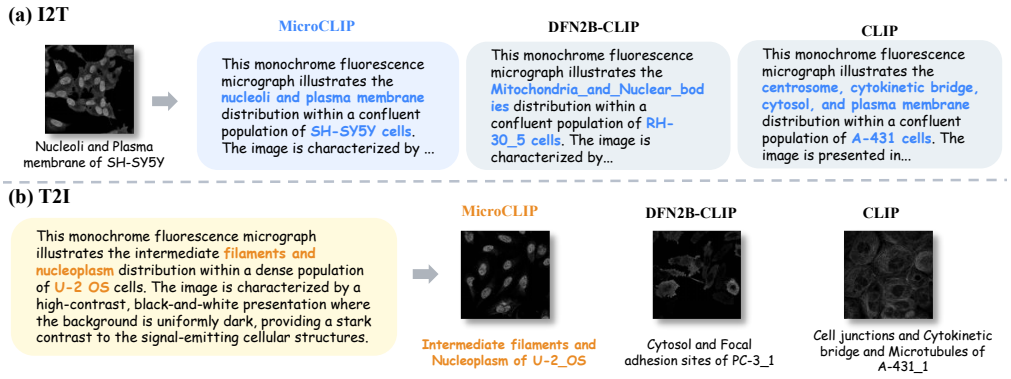


Figure 5: Qualitative comparison of (a) image-to-text and (b) text-to-image retrieval.

Joint image–text embedding space. Figure 6 shows that DFN2B-CLIP's category-text embeddings are separated from image embeddings, whereas MicroCLIP exhibits more distinct image clusters and greater overlap between modalities. This projection qualitatively illustrates the learned representation, while classification and retrieval results quantify alignment.

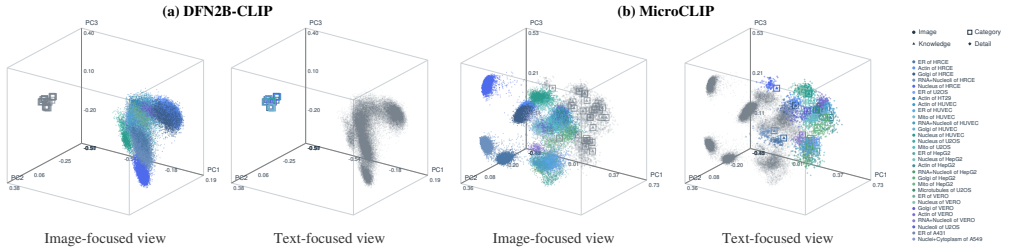


Figure 6: **Joint image–text embeddings visualized with 3D PCA.** (a) DFN2B-CLIP and (b) MicroCLIP. Each pair of views uses the same projection, highlighting images or text by category and showing the other modality in gray. Colors identify categories consistently across panels; marker shapes identify images and text levels. DFN2B-CLIP shows category text only, while MicroCLIP includes all three text levels.

7 Limitations and Conclusion

MicroPool-2M provides three-level text supervision for MicroCLIP to connect fluorescence microscopy images with biological semantics. MicroCLIP outperforms the evaluated general-purpose vision–language models in zero-shot structure–cell-line recognition and bidirectional retrieval, supporting language-guided analysis beyond training categories. Our evaluation is limited to these tasks in fluorescence microscopy, while broader biological targets and research tasks remain outside the scope of this study. Extending language-guided microscopy analysis to these settings is a promising direction toward a more general-purpose microscopy research assistant.

AI USE STATEMENT

In this work, we used generative AI to generate synthetic text annotations for MicroPool-2M. We also evaluated GPT-6 as a subject of study in the qualitative examples and classification diagnostic presented in Section 3 and Appendix D. We take full responsibility for the content of this work.

ETHICS STATEMENT

MicroPool-2M is constructed from existing microscopy resources and published data. We do not collect new human-subject data or personally identifiable information. Data provenance, preprocessing, and source-specific usage information are documented in the appendix and dataset documentation.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the construction and evaluation splits of MicroPool-2M, as well as the MicroCLIP architecture, training configurations, and evaluation protocols, in the main text and appendix. We will publicly release MicroPool-2M, the MicroCLIP model weights, and training code to facilitate reproduction of the reported results.

REFERENCES

Eva Bártová, Gabriela Šustáčková, Lenka Stixová, Stanislav Kozubek, Soňa Legartová, and Veronika Foltankova. Recruitment of oct4 protein to uv-damaged chromatin in embryonic stem cells. *PLoS One*, 6(12):e27281, 2011.

Soufiane Belharbi, Mara KM Whitford, Phuong Hoang, Shakeeb Murtaza, Luke McCaffrey, and Eric Granger. SR-CACO-2: A dataset for confocal fluorescence microscopy image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59948–59983, 2024. doi: 10.52202/079017-1914.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016. doi: 10.1109/CVPR.2016.173.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2036. URL https://papers.neurips.cc/paper_files/paper/2025/hash/57bc0a850255e2041341bf74c7e2b9fa-Abstract-Conference.html.

Michal Breker, Melissa Gymrek, and Maya Schuldiner. A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *Journal of Cell Biology*, 200(6):839–850, 2013.

Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D. Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A. Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G. Linington, and Anne E. Carpenter. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, 2017. doi: 10.1038/nmeth.4397.

Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular Cancer Therapeutics*, 9(6):1913–1926, 2010.

Srinivas Niranj Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G. Doench, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114–1121, 2024. doi: 10.1038/s41592-024-02241-6.

Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, et al. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, 2022. doi: 10.1126/science.abi6983.

Michael J Cox, Steffen Jaensch, Jelle Van de Waeter, Laure Cougnaud, Daan Seynaeve, Soulaiman Benalla, Seong Joo Koo, Ilse Van Den Wyngaert, Jean-Marc Neefs, Dmitry Malkov, et al. Tales of 1,008 small molecules: phenomic profiling through live-cell imaging in a panel of reporter cell lines. *Scientific Reports*, 10(1):13262, 2020.

Jayme L. Dahlin, Bruce K. Hua, Beth E. Zucconi, Jr. Nelson, Shawn D., Shantanu Singh, Anne E. Carpenter, Jonathan H. Shrimp, Evelyne Lima-Fernandes, Mathias J. Wawer, Lawrence P. W. Chung, et al. Reference compounds for characterizing cellular injury in high-content cellular morphology assays. *Nature Communications*, 14(1):1364, 2023.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6fa4d985e7c434002fb6289ab9b2d654-Abstract-Conference.html.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9bf0810a4a1597a36d27ceea58667d92-Abstract-Conference.html.

Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman A Carrel, Todd R Golub, Stuart L Schreiber, et al. Multiplex cytological profiling assay to measure diverse cellular states. *PloS one*, 8(12):e80999, 2013.

Katie Heiser, Peter F. McLean, Chadwick T. Davis, Ben Fogelson, Hannah B. Gordon, Pamela Jacobson, Brett Hurst, Ben Miller, Ronald W. Alfa, Berton A. Earnshaw, et al. Identification of potential treatments for covid-19 through artificial intelligence-enabled phenomic analysis of human cells infected with sars-cov-2. *bioRxiv*, 2020. doi: 10.1101/2020.04.21.054387.

Ivaylo I Ivanov and Kenya Honda. Intestinal commensal microbes as immune modulators. *Cell Host & Microbe*, 12(4):496–508, 2012.

Dorian Kauffmann, Guillaume Gay, Julio Mateos-Langerak, et al. 2d multimodal image collection for fluorescence prediction from transmitted light microscopy. *Scientific Data*, 13:743, 2026. doi: 10.1038/s41597-026-07004-w.

Hirofumi Kobayashi, Keith C Cheveralls, Manuel D Leonetti, and Loic A Royer. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature methods*, 19(8):995–1003, 2022. doi: 10.1038/s41592-022-01541-z.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. VeCLIP: Improving CLIP training via visual-enriched captions. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/5836_ECCV_2024_paper.php.

Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637, 2012. doi: 10.1038/nmeth.2083.

David J Logan, Jing Shan, Sangeeta N Bhatia, and Anne E Carpenter. Quantifying co-cultured cell phenotypes in high-throughput using pixel-based classification. *Methods*, 96:6–11, 2016.

MingYu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP – learning perturbation effects in cell painting via text-guided contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jason Moffat, Dorre A Grueneberg, Xiaoping Yang, So Young Kim, Angela M Kloepfer, Gregory Hinkle, Bruno Piqani, Thomas M Eisenhaure, Biao Luo, Jennifer K Grenier, et al. A lentiviral rnai library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 124 (6):1283–1298, 2006.

Théo Moutakanni, Camille Couprie, Seungeun Yi, Michael Doron, Zitong S Chen, Nikita Moshkov, Elouan Gardes, Mathilde Caron, Hugo Touvron, Armand Joulin, et al. Cell-DINO: Self-supervised image-based embeddings for cell fluorescent microscopy. *PLOS Computational Biology*, 21(12):e1013828, 2025. doi: 10.1371/journal.pcbi.1013828.

Micha Müller, Merve Avar, Daniel Heinzer, Marc Emmenegger, Adriano Aguzzi, Lucas Pelkmans, and Scott Berry. High content genome-wide sirna screen to investigate the coordination of cell size and rna production. *Scientific Data*, 8(1):162, 2021.

Robert F Murphy, Michael V Boland, and Meel Velliste. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. In *ISMB*, volume 8, pp. 251–259, 2000.

Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *Computer Vision – ECCV 2018*, pp. 172–190. Springer International Publishing, 2018. doi: 10.1007/978-3-030-01246-5_11.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations*, 2023.

Beate Neumann, Thomas Walter, Jean-Karim Hériché, Jutta Bulkescher, Holger Erfle, Christian Conrad, Phill Rogers, Ina Poser, Michael Held, Urban Liebel, et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727, 2010.

Wei Ouyang, Casper F Winsnes, Martin Hjelmare, Anthony J Cesnik, Lovisa Åkesson, Hao Xu, Devin P Sullivan, Shubin Dai, Jun Lan, Park Jinmo, et al. Analysis of the human protein atlas image classification competition. *Nature methods*, 16(12):1254–1261, 2019. doi: 10.1038/s41592-019-0658-6.

Hyejin Park, Jiwon Yoon, Sumin Park, Suree Kim, Sinae Jang, Eunsoo Lee, Dongmin Kang, and Dongbo Min. FluoCLIP: Stain-aware focus quality assessment in fluorescence microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28288–28297, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Park_FluoCLIP_Stain-Aware_Focus_Quality_Assessment_in_Fluorescence_Microscopy_CVPR_2026_paper.html.

Chang Qiao, Di Li, Yuting Guo, Chong Liu, Tao Jiang, Qionghai Dai, and Dong Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods*, 18(2):194–202, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Beyond seen primitive concepts and attribute-object compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14466–14476, 2024.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. CLOOME: Contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14:7339, 2023. doi: 10.1038/s41467-023-42328-w.

Morgan Sarah Schwartz, Erick Moen, Geneva Miller, Tom Dougherty, Enrico Borba, Rachel Ding, William Graf, Edward Pao, and David Van Valen. Caliban: Accurate cell tracking and lineage construction in live-cell imaging experiments with deep learning. *bioRxiv*, 2024. doi: 10.1101/803205. URL https://www.biorxiv.org/content/early/2024/08/07/803205.

Christoph Spahn, Estibaliz Gómez-de Mariscal, Romain F Laine, Pedro M Pereira, Lucas von Chamier, Mia Conduit, Mariana G Pinho, Guillaume Jacquemet, Séamus Holden, Mike Heilemann, et al. Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches. *Communications Biology*, 5(1):688, 2022.

Ronald Springer, Alexander Zielinski, Catharina Pleschka, Bernd Hoffmann, and Rudolf Merkel. Unbiased pattern analysis reveals highly diverse responses of cytoskeletal systems to cyclic straining. *PLoS One*, 14(3):e0210570, 2019.

Matthew Sutton, Katrin Amunts, Timo Dickscheid, and Christian Schiffer. Cytoarchitecture in words: Weakly supervised vision-language modeling for human brain microscopy. *arXiv preprint arXiv:2602.23088*, 2026. doi: 10.48550/arXiv.2602.23088. URL https://arxiv.org/abs/2602.23088.

David Svoboda and Vladimir Ulman. Mitogen: a framework for generating 3d synthetic time-lapse sequences of cell populations in fluorescence microscopy. *IEEE Transactions on Medical Imaging*, 36(1):310–321, 2017.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4285–4294, 2023. doi: 10.1109/CVPRW59228.2023.00451.

Peter J. Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, et al. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, 2017. doi: 10.1126/science.aal3321.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. doi: 10.48550/arXiv.2502.14786. URL https://arxiv.org/abs/2502.14786.

Matheus P. Viana, Jianxu Chen, Theo A. Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E. Arakaki, Matte Bailey, Ben Berry, Antoine Borensztejn, Eva M. Brown, et al. Integrated intracellular organization and its variations in human iPS cells. *Nature*, 613(7943):345–354, 2023. doi: 10.1038/s41586-022-05563-7.

Alessandra Vigilante, Anna Laddach, Nathalie Moens, Ruta Meleckyte, Andreas Leha, Arsham Ghahramani, Oliver J Culley, Annie Kathuria, Chloe Hurling, Alice Vickers, et al. Identifying extrinsic versus intrinsic drivers of variation in cell behavior in human ipsc lines from healthy donors. *Cell Reports*, 26(8):2078–2087.e3, 2019.

Zhixiang Wei, Guangting Wang, Xiaoxiao Ma, Ke Mei, Huaian Chen, Yi Jin, and Fengyun Rao. Hq-clip: Leveraging large vision-language models to create high-quality image-text datasets and clip models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22447–22456, October 2025.

Jens Wenzel, Christian Held, Ralf Palmisano, Stefan Teufel, Jean-Pierre David, Thomas Wittenberg, and Roland Lang. Measurement of tlr-induced macrophage spreading by automated image analysis: differential role of myd88 and mapk in early and late responses. *Frontiers in Physiology*, 2:71, 2011.

Eleanor Williams, Josh Moore, Simon W. Li, Gabriella Rustici, Aleksandra Tarkowska, Anatole Chessel, Simone Leo, Bálint Antal, Richard K. Ferguson, Ugis Sarkans, Alvis Brazma, Rafael E. Carazo Salas, and Jason R. Swedlow. Image Data Resource: a bioimage data integration and publication platform. *Nature Methods*, 14(8):775–781, 2017. doi: 10.1038/nmeth.4326.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5BCFlnfE1g.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Chi Zhang, Hao Jiang, Weihuang Liu, Junyi Li, Shiming Tang, Mario Juhas, and Yang Zhang. Correction of out-of-focus microscopic images by deep learning. *Computational and Structural Biotechnology Journal*, 20:1957–1966, 2022.

Xinyi Zhang, Yitong Tseo, Yunhao Bai, Fei Chen, and Caroline Uhler. Prediction of protein subcellular localization in single cells. *Nature Methods*, 22:1265–1275, 2025. doi: 10.1038/s41592-025-02696-1.

Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11710–11718, 2019.