Toward Generalizable Whole Brain Representations with High-Resolution Light-Sheet Data

Minyoung E. Kim[*†1] **Dae Hee Yun**[*1] **Aditi V. Patel**[1] **Madeline Hon**[1]
Taegeon Lee[1] **Webster Guan**[1] **Brian Nguyen**[1]

[1]LifeCanvas Technologies, Cambridge, MA 02141, USA
{mykim, daeheeyun, aditi, mhon, t76lee, websterg, brian}@lifecanvastech.com

Abstract

Unprecedented visual details of biological structures are being revealed by subcellular-resolution whole-brain 3D microscopy data, enabled by recent advances in intact tissue processing and light-sheet fluorescence microscopy (LSFM). These volumetric data offer rich morphological and spatial cellular information, however, the lack of scalable data processing and analysis methods tailored to these petabyte-scale data poses a substantial challenge for accurate interpretation. Further, existing models for visual tasks such as object detection and classification struggle to generalize to this type of data. To accelerate the development of suitable methods and foundational models, we present CANVAS, a comprehensive set of high-resolution whole mouse brain LSFM benchmark data, encompassing six neuronal and immune cell-type markers, along with a set of cell annotations and a leaderboard. We also demonstrate challenges in generalization of baseline models built on existing architectures, especially due to the heterogeneity in cellular morphology across phenotypes and anatomical locations in the brain. To the best of our knowledge, CANVAS is the first and largest LSFM benchmark capturing intact mouse brain tissue at subcellular level, and includes extensive annotations of cells throughout the brain.

1 Introduction

The brain is comprised of thousands of distinct cell types with diverse molecular, morphological, and functional properties that exhibit region-specific heterogeneity [2, 6, 16] and reflect local microenvironments. While thin-section imaging can only capture a limited view of cellular diversity[22], recent advances in light-sheet fluorescence microscopy (LSFM), tissue clearing, and labeling techniques now enable high-resolution imaging of intact tissues, including whole mouse brains. Such advances open new opportunities for in-depth phenotyping of individual cells throughout the brain, integrating their molecular, morphological, and microenvironmental features.

However, large-scale, subcellular-resolution 3D data of intact tissues, such as mouse brains, often reach the petabyte scale in real-world applications, and the lack of robust data processing and analysis techniques tailored to these datasets hinders comprehensive interpretation. Scalable end-to-end ETL (Extract, Transform, Load) pipelines and generalizable machine learning methods are essential for extracting meaningful biological insights. While machine learning has made remarkable progress in computer vision tasks including object detection, existing methods in biomedical imaging and related domains still struggle to generalize to these types of data. As new data modalities continue to emerge, the need for foundational models that can capture generalizable features across modalities

[*]Equal contribution
[†]Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

is becoming increasingly important. Although various strategies have been proposed to build such models without large annotated datasets, making both data and annotations publicly available remains critical for accelerating progress.

In this work, we present a new high-resolution light-sheet dataset (CANVAS) and a leaderboard for a cell detection task. The CANVAS data set is composed of 3D images of whole brains labeled with one of six cell-type markers (NeuN, cFos, PV, TH, Iba1, and GFAP), each with cell centroid annotations for three regions of interest (Table 3). The raw volumetric data cover the entire mouse brain except for the cFos marker, which covers a hemisphere, with 1,600-1,850 z-slices at approximately 7,000×10,000 pixels per slice, yielding 44,240 and 45,726 annotated cell centroids for train and test set. Using CANVAS, we demonstrate that baseline models from existing architectures struggle to generalize, underscoring the importance of publicly available datasets like CANVAS and the domain's need for foundational models. To our knowledge, CANVAS represents the first and largest publicly available LSFM dataset with extensive annotations, evaluation metrics, and a benchmark leaderboard, serving as a critical resource for developing robust foundational models for object detection in LSFM and other biological volumetric data. The dataset, annotations, and leaderboard are available at `https://canvas.lct-data.com`.

2 CANVAS: whole-brain volumetric data

CANVAS showcases six distinct cell type markers: Neuron Specific Nuclear Protein (NeuN), Ionized calcium-binding adaptor molecule 1 (IBA1), glial fibrillary acidic protein (GFAP), Tyrosine hydroxylase (TH), cFos (a neural activity marker), and Parvalbumin (PV). Enabling investigation of specific cell populations within complex tissues like the brain, makes molecular cell type markers invaluable tools in neuroscience and biomedical research. The six protein-based markers in this benchmark represent only a small subset of available markers, but each highlights a distinct and functionally important cell class relevant to health and disease. In addition, these markers exhibit distinct morphological characteristics, often varying by brain region.

NeuN is a marker that localizes neuronal cell bodies across the brain. NeuN expression is ubiquitous throughout the brain with varying regional densities. IBA1 labels microglia, resident immune cells, also with ubiquitous expression, with regional differences in both density and morphology that reflect local immune states. GFAP marks astrocytes, which are highly concentrated along the fiber tracts and surrounding vasculature; astrocytes exhibit significant regional and morphological heterogeneity. TH is a classic marker for dopaminergic neurons, found in high densities within several subcortical nuclei, with extensive long-range axonal projections throughout the brain. PV labels one of the largest classes of interneurons, displaying variable densities and morphologies across the cortical and subcortical structures, as well as the cerebellum. cFOS, an immediate early gene product, serves as a proxy for recent neuronal activation, providing a brain-wide snapshot of activity patterns in response to stimuli or behavior. cFOS expression patterns can vary greatly between individual animals but are typically morphologically consistent. Together, these markers encompass a diverse range of cell types and functions, capturing the spatial complexity and regional specialization of the intact brain.

Data acquisition. The workflow for data acquisition begins with SHIELD preservation[17] of mouse brains to preserve the integrity of biomolecules during extended tissue processing. Next, delipidation removes lipids, the major component of cellular membranes, to improve molecular and optical access for downstream immunolabeling and imaging. Fluorescently labeled molecular probes were uniformly and efficiently delivered throughout intact samples via the SmartBatch+ system, which utilizes probe-binding affinity modulation and electrophoretically enhanced molecular transport[24][12]. Finally, samples are rendered transparent by refractive-index matching with EasyIndex and imaged using SmartSPIM at 3.6X magnification, generating whole-brain datasets with voxel sizes of 1.8 µm × 1.8 µm × 4 µm. Acquiring one channel whole brain dataset takes 40 minutes and generates ~100 GB of data.

Data processing and visualization. A series of post-processing steps are applied after LSFM data acquisition to improve image quality while preserving underlying scientific information[20]. First, destriping removes stripe artifacts introduced by the light paths during SmartSPIM imaging. Second, because multiple z-stack tiles are required for whole mouse brain coverage, each tile stack must be stitched together, producing a series of TIFF images compressed with lossless zlib (level 1). Each

TIFF file represents a complete XY plane at a single z-step, determined during imaging; in our case, this corresponds to a 4 µm z-axis depth. The stitched data are then converted into the Zarr format [25], suitable for efficiently storing and handling large volumetric datasets. Finally, the Zarr-formated data are served via Neuroglancer [15], providing an interactive visualization.

3 Cell detection benchmark

An adult mouse brain contains approximately 70 million neurons and 17 million immune cells. Given the existence of thousands of distinct cell types in the brain, the markers we selected are expressed in hundreds of thousands to millions of cells. For example, microglia, labeled by IBA1, account for 5-12% of the total cellular population. Thus, detecting individual cells labeled by each marker across the brain requires automated and scalable computational approaches. Over the past decades, deep neural networks (DNNs) have achieved state-of-the-art performance in computer vision tasks, including object detection. In principle, these models can also be applied to localize individual cells in LSFM data. Here, we propose a cell detection benchmark using the CANVAS dataset, along with baseline models built on established DNN backbones. We further show that these baseline models struggle to generalize and underperform on subsets CANVAS—particularly when cells exhibit diverse morphological profiles.

3.1 Baseline models

U-Net [18] and ResNet [10] architectures are among the most widely used backbones in computer vision tasks for biomedical image analysis. Numerous variations of these networks have been proposed[13], many incorporating 3D convolution layers to handle volumetric data [5, 3]. More recently, models based on the Vision Transformer (ViT) architecture [7] have also been introduced. However, most of these existing models were developed and trained for other data modalities, such as computed tomography (CT), functional magnetic resonance imaging (fMRI), X-ray, or histopathology [9, 8, 19, 23], with relatively few tailored for microscopy data—and even then, typically for specific subdomains (e.g., retina confocal imaging) [4]. In addition, training ViT models requires extensive computational resources, even for the smallest networks, and they are generally inefficient in terms of inference speed when applied to large-scale datasets such as CANVAS. For our benchmark, we adopted ConvMixerNet [21] as the backbone for baseline models. ConvMixerNet employs a concept similar to patch embeddings in ViT but is simpler and sufficiently effective on our dataset to serve as a baseline. As ConvMixerNet was originally proposed for semantic segmentation, we introduced an additional layer to transform the network's probability heatmap into discrete 3D cell locations using non-maximum suppression, as shown in Appendix A. With this layer, a voxel at position (x, y, z) is identified as a local maximum and defined as a cell centroid if:

$$H(x, y, z) = \max_{\substack{|i| \leq d_{min} \\ |j| \leq d_{min} \\ |k| \leq d_{min}}} H(x + i, y + j, z + k) \tag{1}$$

where H is the 3D probability heatmap derived from the model, d_{min} is the minimum distance between peaks, and a valid detection from Algorithm 2 satisfies:

$$\text{Valid}(x, y, z) = [H(x, y, z) = H_{max}(x, y, z)] \wedge [H(x, y, z) \geq \tau] \tag{2}$$

where H_{max} represents 3D max-pooled output from H for finding local maxima, and τ is a threshold ranged from 0 to 1.

To train the network, we manually generated a small size of cell mask training data for each cell type marker, using a whole-brain dataset not included in this paper. We limited the size of the training set to the minimum sufficient to guide cell centroid annotations due to the impracticality of generating whole-brain cell masks. A binary form of focal loss [14] was used for model training and we trained a separate model for each cell-type marker dataset to further investigate generalizability. An NVIDIA RTX 3090 or 4090 graphics card was used for training, and the model converged within one day.

3.2 Data Annotation

For each cell-type marker dataset, three ROIs were selected for the training set and three for the test set. To generate ground-truth cell annotations, we first performed inference using the baseline models

described in Section 3.1. Based on the predicted cell centroids, we center-cropped volumes around each cell with dimensions of $32 \times 32 \times 8$ pixels (x, y, z), covering the physical area of 57.6 µm $\times$ 57.6 µm $\times$ 32 µm. Each cropped cell volume was then manually annotated as either 0 (non-cell) or 1 (cell). Annotation work was done by seven annotators using the MorPheT annotation tool[11]. 450 ~11,000 cells were annotated per region, with cells labeled as 0 (non-cells) filtered out. False negatives missed by the predictions were subsequently recovered through manual review. In total, more than 130,000 predictions were annotated, yielding approximately 70,000 ground-truth cell centroids. Detailed information on the selected regions for each dataset is shown in Table 3.

3.3 Evaluation method

Model performance was evaluated using accuracy and the F1 score. True positives (TPs) were determined by calculating the Euclidean distance between each predicted cell coordinate and the closest ground-truth cell center using a kd-tree nearest-neighbor search[1]. A prediction was assigned as a TP if the distance fell below a tolerance threshold, defined in voxels and set individually for each cell type and anatomical region, shown in Table B.1. The tolerance threshold was set to the average cell radius, estimated as half the mean diameter of six representative cells sampled from each anatomical region. Each prediction was matched to at most one ground-truth cell center to enforce a one-to-one correspondence. False positives (FPs) and false negatives (FN) were then calculated using the ground-truth and TPs. Accuracy was defined as $\frac{TP}{TP+FP+FN}$, along with a precision-recall curve. F1 score was calculated as the harmonic mean of precision and recall.

3.4 Results

Using the baseline models and annotations described above, we evaluated model performance both within datasets and across datasets, using three different models that were trained on cell type markers with significantly different features. We found that model performance varied across regions and datasets, likely due to the cells' distinct morphological and regional profiles. This is demonstrated by the TH model achieving an F1 score of 0.27 or lower on the cFos dataset, while reaching almost 0.96 in region_3 of the TH dataset, which highlights the challenge model faces generalizing across different cell-type markers and regions. This trend is seen across most marker and cell-type model combinations, as shown in Table B.2. We also evaluated the three models on all six regions of the PV dataset, yielding F1 scores of 0.31 (cFos model), 0.70 (TH model), and 0.64 (IBA1 model). Inter-regional variations in PV+ staining profiles cause large differences in F1 scores, ranging from 0.53 (region_2) to 0.97 (region_6, region_3). Region_2 corresponds to the RT anatomical region, where PV+ cell morphologies significantly differ from those in the other two regions (HIP and mPFC), contributing to the observed generalization issue. Further emphasizing the difficulties in generalizability, PV+ cell morphologies are region dependent, with cells in the RT (Region_2) distinct from those in the other regions (HIP and mPFC). This is also seen in the TH dataset with the TH model where the F1 score ranges from 0.49 to 0.94 between regions due to differences in cell morphologies.

4 Conclusion

In this paper, we present CANVAS, a large-scale and subcellular resolution dataset, labeled with six different markers and accompanied by extensive annotations. A public leaderboard enables benchmarking of baseline models to advance robust foundational models for the LSFM domain. Current limitations include the limited number of manually generated cell centroid labels, which restrict scalability. Approaches such as active learning could help reduce labeling burden andimprove consistency, particularly given the domain expertise required for accurate annotations. The dataset can also be expanded with additional markers, as many brain cell types such as oligodendrocytes, various interneuron, and neurotransmitter-specific populations remain unrepresented.

Nonetheless, we believe the CANVAS dataset holds immediate potential for the broader fields of biomedical and neuroscience research–both as a standalone modality and as a complementary axis for multi-omics analyses, together with other data modalities such as fMRI and mRNA sequencing. As future work, we plan on increasing the number of LSFM brain samples and their cell annotations across additional brain regions, and adding more markers to target a wider range of cell types.

References

[1] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[2] Annie Bryant, Zhaozhi Li, Rojashree Jayakumar, Alberto Serrano-Pozo, Benjamin Woost, Miwei Hu, Maya E Woodbury, Astrid Wachter, Gen Lin, Taekyung Kwon, Robert V Talanian, Knut Biber, Eric H Karran, Bradley T Hyman, Sudeshna Das, and Rachel E Bennett. Endothelial cells are heterogeneous in different brain regions and are dramatically altered in alzheimer's disease. *J. Neurosci.*, 43(24):4541–4557, June 2023.

[3] Yimin Cai, Yuqing Long, Zhenggong Han, Mingkun Liu, Yuchen Zheng, Wei Yang, and Liming Chen. Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution. *BMC Med. Inform. Decis. Mak.*, 23(1):33, February 2023.

[4] Danny Chen, Wenzhong Yang, Liejun Wang, Sixiang Tan, Jiangzhaung Lin, and Wenxiu Bu. PCAT-UNet: UNet-like network fused convolution and transformer for retinal vessel segmentation. *PLoS One*, 17(1):e0262689, January 2022.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.

[6] Soumen Das and Narendrakumar Ramanan. Region-specific heterogeneity in neuronal nuclear morphology in young, aged and in alzheimer's disease mouse brains. *Front. Cell Dev. Biol.*, 11:1032504, February 2023.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[8] Rim El Badaoui, Ester Bonmati Coll, Aleka Psarrou, and Barbara Villarini. 3d catbrats: Channel attention transformer for brain tumour semantic segmentation. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 489–494, 2023.

[9] Yong Hao, Chengxiang Zhang, and Xiyan Li. Dbm-vit: A multiscale features fusion algorithm for health status detection in cxr / ct lungs images. *Biomedical Signal Processing and Control*, 87:105365, 2024.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] Minyoung E. Kim. Mapping the cellular landscape of the brain: A scalable approach to comprehensive microscopy data analysis. [Online; accessed 2025-08-28].

[12] Sung-Yon Kim, Jae Hun Cho, Evan Murray, Naveed Bakh, Heejin Choi, Kimberly Ohn, Luzdary Ruelas, Austin Hubbert, Meg McCue, Sara L Vassallo, et al. Stochastic electrotransport selectively enhances the transport of highly electromobile molecules. *Proceedings of the National Academy of Sciences*, 112(46):E6274–E6283, 2015.

[13] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *CoRR*, abs/1806.05034, 2018.

[14] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[15] Jeremy Maitin-Shepard, Alex Baden, William Silversmith, Eric Perlman, Forrest Collman, Tim Blakely, Jan Funke, et al. Neuroglancer: Webgl-based viewer for volumetric data, 2021.

[16] Sean J Miller. Astrocyte heterogeneity in the adult central nervous system. *Frontiers in cellular neuroscience*, Nov 2018.

[17] Young-Gyun Park, Chang Ho Sohn, Ritchie Chen, Margaret McCue, Dae Hee Yun, Gabrielle T Drummond, Taeyun Ku, Nicholas B Evans, Hayeon Caitlyn Oak, Wendy Trieu, et al. Protection of tissue physicochemical properties using polyfunctional crosslinkers. *Nature biotechnology*, 37(1):73–83, 2019.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[19] Chen Sheng, Lin Wang, Zhenhuan Huang, Tian Wang, Yalin Guo, Wenjie Hou, Laiqing Xu, Jiazhu Wang, and Xue Yan. Transformer-based deep learning network for tooth segmentation on panoramic radiographs. *J. Syst. Sci. Complex.*, 36(1):257–272, 2023.

[20] Justin Swaney, Lee Kamentsky, Nicholas B Evans, Katherine Xie, Young-Gyun Park, Gabrielle Drummond, Dae Hee Yun, and Kwanghun Chung. Scalable image processing techniques for quantitative analysis of volumetric biological images from light-sheet microscopy. March 2019.

[21] Asher Trockman and J. Zico Kolter. Patches are all you need? *CoRR*, abs/2201.09792, 2022.

[22] Hiroki R Ueda, Ali Ertürk, Kwanghun Chung, Viviana Gradinaru, Alain Chédotal, Pavel Tomancak, and Philipp J Keller. Tissue clearing and its applications in neuroscience. *Nat. Rev. Neurosci.*, 21(2):61–79, February 2020.

[23] Lian Wang, Liangrui Pan, Hetian Wang, Mingting Liu, Zhichao Feng, Pengfei Rong, Zuo Chen, and Shaoliang Peng. Dhunet: Dual-branch hierarchical global–local fusion network for whole slide image segmentation. *Biomedical Signal Processing and Control*, 85:104976, 2023.

[24] Dae Hee Yun, Young-Gyun Park, Jae Hun Cho, Lee Kamentsky, Nicholas B Evans, Nicholas DiNapoli, Katherine Xie, Seo Woo Choi, Alexandre Albanese, Yuxuan Tian, et al. Uniform volumetric single-cell processing for organ-scale molecular phenotyping. *Nature Biotechnology*, pages 1–12, 2025.

[25] Zarr Developers. Zarr: Chunked, compressed, n-dimensional arrays for python, 2025. Released August 25, 2025.

A Algorithms

Algorithm 1: Model with Location Prediction

Input : Detection Model M, threshold $\tau \in [0, 1]$
Output : Location prediction model M_{loc}
Parameters : $d_{min} = 3$ (minimum distance between peaks)

$I_{img} \leftarrow$ Volumetric Image Input from Model M
$I_{idx} \leftarrow$ Batch index
$H \leftarrow$ Model output; 3D probability heatmap
$k \leftarrow 2 \times d_{min} + 1$
$H_{max} \leftarrow \texttt{MaxPool3D}(H, \text{kernel} = (k, k, k), \text{stride} = (1, 1, 1), \text{padding} = \text{"valid"})$
$H_{max} \leftarrow \texttt{ZeroPad3D}(H_{max}, \text{padding} = d_{min})$
$L \leftarrow \texttt{FindMaxima}(H, H_{max}, I_{idx}, \tau)$
$M_{loc} \leftarrow \texttt{Model}(I_{img}, I_{idx}, L)$
return M_{loc}

Algorithm 2: FindMaxima Layer

Input : Heatmap H (shape (B, X, Y, Z)), max-pooled heatmap H_{max} (shape (B, X, Y, Z)), batch indices I (shape $(B,)$)
Output : Cell locations L (shape $(N, 4)$) with columns $[b, x, y, z]$
Parameters : Threshold $\tau \in [0, 1]$

$M_{local} \leftarrow (H = H_{max})$
$M_{thresh} \leftarrow (H \geq \tau)$
$M_{valid} \leftarrow M_{local} \wedge M_{thresh}$
$L \leftarrow \texttt{where}(M_{valid})$
$i_{min} \leftarrow \texttt{cast}(\texttt{min}(I), \text{int64})$
$L_{batch} \leftarrow L[:, 0] + i_{min}$
$L_{batch} \leftarrow \texttt{expand_dims}(L_{batch}, \text{axis} = -1)$
$L \leftarrow \texttt{concatenate}([L_{batch}, L[\cdot, 1:]], \text{axis} = 1)$
return L

B Evaluation

B.1 Evaluation parameters

Table 1: Tolerance thresholds (per region) for each dataset

Brain ID	Marker	Region_1	Region_2	Region_3
brain_1	GFAP	4.1	5.0	6.0
brain_3	IBA1	5.6	5.2	4.7
brain_5	NeuN	6.0	5.9	5.9
brain_7	TH	6.3	5.6	6.1
brain_8	cFos	7.8	8.4	7.9
brain_9	PV	9.7	5.8	7.8

B.2 Baseline models evaluation

Table 2: Performance metrics (Accuracy and F1 Score) for three different cell-type models across regions.

Cell Type	Region	cFOS Model		TH Model		Iba1 Model	
		Accuracy	F1_Score	Accuracy	F1_Score	Accuracy	F1_Score
cFos	**Total**	**0.64**	**0.78**	**0.15 -0.28**	**0.26 -0.34**	**0.58 -0.15**	**0.74 -0.08**
	Train Set	**0.64**	**0.78**	**0.14 -0.31**	**0.25 -0.37**	**0.58 -0.06**	**0.73 -0.05**
	region_1	0.67	0.80	0.13 -0.27	0.23 -0.34	0.56 -0.16	0.72 -0.12
	region_2	0.72	0.84	0.14 -0.23	0.25 -0.30	0.58 +0.01	0.73 +0.01
	region_3	0.53	0.70	0.16 -0.72	0.27 -0.67	0.61 -0.05	0.75 -0.04
	Test Set	**0.65**	**0.78**	**0.16 -0.25**	**0.27 -0.31**	**0.59 -0.15**	**0.74 -0.11**
	region_4	0.67	0.80	0.15 -0.21	0.25 -0.27	0.57 -0.17	0.72 -0.12
	region_5	0.73	0.85	0.15 -0.17	0.26 -0.22	0.62 -0.03	0.76 -0.03
	region_6	0.55	0.71	0.19 -0.62	0.31 -0.58	0.59 -0.24	0.74 -0.16
NeuN	**Total**	**0.01 -0.63**	**0.02 -0.76**	**0.12 -0.31**	**0.21 -0.39**	**0.40 -0.29**	**0.57 -0.25**
	Train Set	**0.01 -0.62**	**0.02 -0.75**	**0.11 -0.34**	**0.20 -0.42**	**0.38 -0.26**	**0.55 -0.23**
	region_1	0.00 -0.67	0.00 -0.80	0.13 -0.27	0.23 -0.35	0.41 -0.31	0.58 -0.26
	region_2	0.07 -0.65	0.13 -0.71	0.13 -0.25	0.23 -0.32	0.31 -0.25	0.48 -0.25
	region_3	0.01 -0.53	0.02 -0.68	0.09 -0.79	0.17 -0.77	0.37 -0.28	0.54 -0.25
	Test Set	**0.01 -0.63**	**0.02 -0.76**	**0.12 -0.28**	**0.22 -0.36**	**0.41 -0.32**	**0.59 -0.26**
	region_4	0.00 -0.67	0.00 -0.80	0.13 -0.22	0.23 -0.29	0.38 -0.36	0.55 -0.30
	region_5	0.09 -0.65	0.16 -0.69	0.13 -0.20	0.23 -0.26	0.32 -0.33	0.48 -0.30
	region_6	0.01 -0.54	0.02 -0.69	0.11 -0.69	0.21 -0.69	0.47 -0.36	0.64 -0.26
TH	**Total**	**0.02 -0.62**	**0.04 -0.74**	**0.43**	**0.60**	**0.16 -0.53**	**0.27 -0.55**
	Train Set	**0.01 -0.62**	**0.02 -0.75**	**0.45**	**0.62**	**0.16 -0.48**	**0.27 -0.51**
	region_1	0.01 -0.65	0.02 -0.79	0.40	0.57	0.13 -0.59	0.23 -0.61
	region_2	0.00 -0.72	0.01 -0.83	0.38	0.55	0.07 -0.49	0.14 -0.58
	region_3	0.04 -0.49	0.08 -0.61	0.88	0.94	0.29 -0.36	0.45 -0.34
	Test Set	**0.03 -0.61**	**0.06 -0.72**	**0.41**	**0.58**	**0.15 -0.58**	**0.26 -0.58**
	region_4	0.01 -0.66	0.02 -0.78	0.35	0.52	0.1 -0.64	0.17 -0.68
	region_5	0.01 -0.72	0.03 -0.82	0.32	0.49	0.03 -0.62	0.06 -0.73
	region_6	0.09 -0.45	0.17 -0.53	0.80	0.89	0.36 -0.47	0.53 -0.38
PV	**Total**	**0.19 -0.46**	**0.31 -0.47**	**0.54 +0.11**	**0.70 +0.10**	**0.47 -0.21**	**0.64 -0.17**
	Train Set	**0.19 -0.45**	**0.32 -0.46**	**0.54 +0.09**	**0.70 +0.08**	**0.50 -0.14**	**0.67 -0.12**
	region_1	0.21 -0.46	0.35 -0.45	0.67 +0.26	0.80 +0.22	0.30 -0.42	0.46 -0.38
	region_2	0.00 -0.72	0.01 -0.83	0.36 -0.02	0.53 -0.02	0.47 -0.09	0.64 -0.08
	region_3	0.39 -0.14	0.56 -0.13	0.93 +0.05	0.96 +0.03	0.84 +0.18	0.91 +0.12
	Test Set	**0.18 -0.46**	**0.31 -0.48**	**0.54 +0.14**	**0.70 +0.13**	**0.45 -0.29**	**0.62 -0.23**
	region_4	0.19 -0.48	0.31 -0.49	0.61 +0.26	0.76 +0.24	0.26 -0.47	0.41 -0.43
	region_5	0.01 -0.73	0.01 -0.84	0.38 +0.06	0.56 +0.07	0.42 -0.23	0.59 -0.19
	region_6	0.38 -0.16	0.56 -0.15	0.94 +0.14	0.97 +0.08	0.86 +0.03	0.92 +0.02
GFAP	**Total**	**0.00 -0.64**	**0.00 -0.78**	**0.02 -0.41**	**0.03 -0.57**	**0.23 -0.46**	**0.37 -0.45**
	Train Set	**0.00 -0.64**	**0.00 -0.77**	**0.01 -0.44**	**0.02 -0.60**	**0.20 -0.44**	**0.33 -0.45**
	region_1	0.00 -0.67	0.00 -0.80	0.00 -0.40	0.00 -0.57	0.20 -0.52	0.34 -0.50
	region_2	0.00 -0.72	0.00 -0.84	0.02 -0.36	0.04 -0.51	0.19 -0.38	0.32 -0.41
	region_3	0.00 -0.53	0.00 -0.70	0.10 -0.78	0.18 -0.76	0.20 -0.46	0.33 -0.46
	Test Set	**0.00 -0.65**	**0.00 -0.78**	**0.03 -0.38**	**0.05 -0.53**	**0.29 -0.45**	**0.45 -0.40**
	region_4	0.00 -0.67	0.00 -0.80	0.01 -0.34	0.02 -0.50	0.22 -0.52	0.36 -0.49
	region_5	0.00 -0.73	0.00 -0.85	0.05 -0.27	0.09 -0.40	0.30 -0.35	0.46 -0.33
	region_6	0.00 -0.54	0.01 -0.70	0.05 -0.76	0.09 -0.80	0.45 -0.38	0.62 -0.29
IBA1	**Total**	**0.01 -0.63**	**0.03 -0.75**	**0.40 -0.03**	**0.57 -0.03**	**0.69**	**0.82**
	Train Set	**0.01 -0.63**	**0.02 -0.76**	**0.35 -0.10**	**0.52 -0.10**	**0.64**	**0.78**
	region_1	0.02 -0.65	0.03 -0.77	0.48 +0.07	0.65 +0.07	0.72	0.84
	region_2	0.01 -0.72	0.01 -0.83	0.38 0.00	0.55 0.00	0.57	0.72
	region_3	0.00 -0.53	0.00 -0.70	0.21 -0.67	0.34 -0.59	0.65	0.79
	Test Set	**0.02 -0.63**	**0.04 -0.75**	**0.43 0.03**	**0.61 0.03**	**0.74**	**0.85**
	region_4	0.02 -0.65	0.03 -0.77	0.47 +0.12	0.64 +0.12	0.74	0.85
	region_5	0.01 -0.72	0.03 -0.82	0.48 +0.15	0.64 +0.16	0.65	0.79
	region_6	0.03 -0.52	0.06 -0.65	0.37 -0.44	0.54 -0.35	0.83	0.91

C CANVAS Dataset Overview

C.1 LSFM data

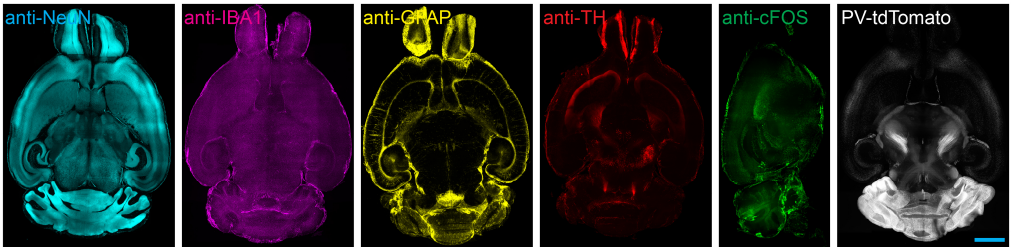

Figure 1: Overview of datasets in CANVAS showing six cell type markers imaged using light sheet fluorescence microscopy. Markers include NeuN (cyan), IBA1 (magenta), GFAP (yellow), TH (red), cFOS (green), and PV (grey). All markers except PV are based on immunolabeling; PV is transgenically labeled with fluorescent proteins. All datasets represent whole-brain imaging, except cFOS, which is hemisphere-only. Images are 500 µm maximum intensity projections. Scale bar: 2 mm.

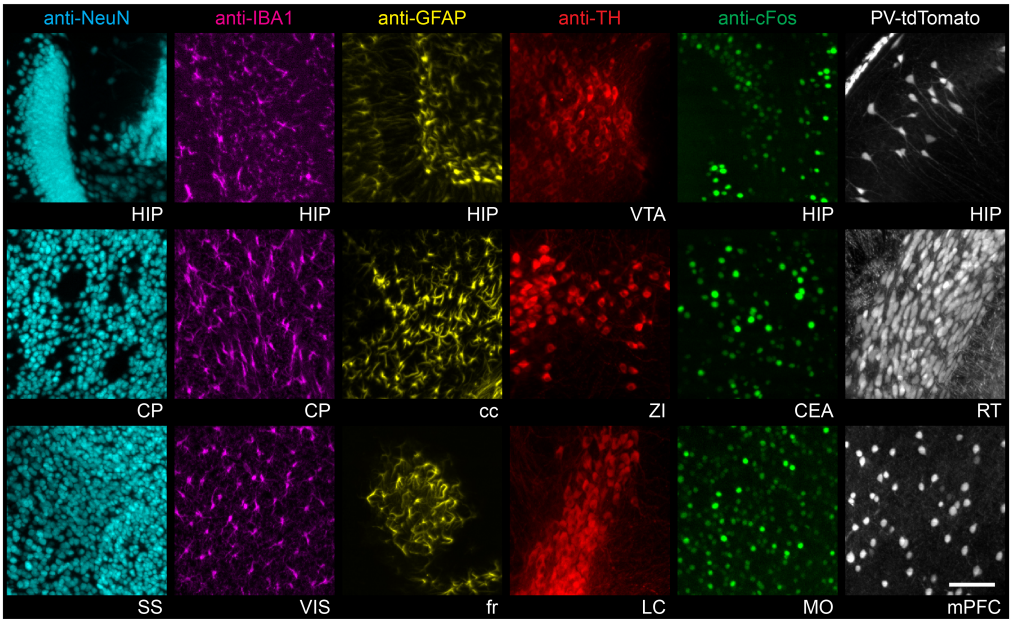

Figure 2: Zoomed-in views from the CANVAS dataset showing six cell type markers acros various brain regions: NeuN (cyan), IBA1 (magenta), GFAP (yellow), TH (red), cFOS (green), and PV (grey), presented as 80 µm maximum intensity projections. Images include the following brain regions: HIP (hippocampus), VTA (ventral tegmental area), CP (caudate putamen), ZI (zona incerta), CEA (central amygdalar nucleus), RT (reticular nucleus of the thalamus), SS (somatosensory areas), VIS (visual areas), LC (locus coeruleus), MO (somatomotor areas), and mPFC (medial prefrontal cortex); and fiber tracts: cc (corpus callosum) and fr (fasciculus retroflexus). Scale bar: 100 µm.

C.2 Annotation data

The table 3 shows the details of ROI selection for each cell type marker dataset for the train set, along with the number of cells annotated. The test set also contains three regions per each marker dataset, while not reported here.

Table 3: Region selections for each dataset

Brain ID	Marker	Region ID	ROI (x, y, z)	Size (px)	# annotated cells
brain_5	NeuN	region_1	(2316, 4627, 137)	$300 \times 300 \times 150$	7,233
		region_2	(2236, 5212, 582)	$300 \times 300 \times 150$	1,766
		region_3	(2442, 3001, 779)	$300 \times 300 \times 150$	8,321
brain_8	cFos	region_1	(4076, 4422, 899)	$500 \times 500 \times 250$	3,782
		region_2	(4402, 3198, 1132)	$300 \times 300 \times 150$	2,827
		region_3	(2481, 5237, 484)	$400 \times 400 \times 200$	2,781
brain_7	TH	region_1	(2990, 5337, 1143)	$300 \times 300 \times 150$	1,185
		region_2	(2698, 6705, 941)	$300 \times 300 \times 150$	1,118
		region_3	(3204, 4272, 1126)	$300 \times 300 \times 150$	400
brain_9	PV	region_1	(1380, 5294, 697)	$600 \times 600 \times 300$	715
		region_2	(2048, 4494, 871)	$400 \times 400 \times 200$	3,915
		region_3	(3025, 2114, 623)	$500 \times 500 \times 250$	1,497
brain_12	GFAP	region_1	(2666, 4850, 443)	$300 \times 300 \times 150$	1,944
		region_2	(2688, 2876, 707)	$200 \times 200 \times 100$	477
		region_3	(3267, 5465, 1008)	$300 \times 300 \times 150$	109
brain_11	IBA1	region_1	(2638, 4805, 422)	$400 \times 400 \times 200$	2,076
		region_2	(1866, 6035, 217)	$400 \times 400 \times 200$	2,021
		region_3	(2240, 3556, 737)	$400 \times 400 \times 200$	2,073

C.3 Data structure

```
[Marker]_[Brain_ID] (cFos_brain_8)
├── image
│   ├── 0000_431340_465490_032660_Ch2.tif
│   ├── 0001_431340_465490_032640_Ch2.tif
│   ├── 0002_431340_465490_032620_Ch2.tif
│   ├── 0003_431340_465490_032600_Ch2.tif
│   ├── ..
└── label
    ├── TRAIN
    │   ├── cFos_brain8_region1_gt.csv
    │   ├── cFos_brain8_region2_gt.csv
    │   └── cFos_brain8_region3_gt.csv
    └── TEST (not publicly available)
        ├── cFos_brain8_region4_gt.csv
        ├── cFos_brain8_region5_gt.csv
        └── cFos_brain8_region6_gt.csv
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main purpose of the paper is to introduce a new LSFM dataset, along with annotations and a benchmark leaderboard, as described in the Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the presented dataset are discussed in the Conclusion (Section 4).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: As this is a dataset benchmark paper, the raw data, corresponding annotations, and evaluation metrics are provided. For replicating the baseline model results, we described the novel layer introduced in Section 3.1 and provided further details in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: As a benchmark dataset paper, the leaderboard will be made publicly available with open access to the data and annotations, although the code used to generate the dataset will not be released.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The main focus of the paper is about releasing a new public dataset, along with a benchmark leaderboard.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources and training time for the baseline models are described in Section 3.1.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Authors have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the open-source code and models that were used in the paper are properly referenced.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The data will be publicly shared on the leaderboard website with descriptions and instructions upon the paper publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.