

BENCHMARKING VISION-LANGUAGE MODELS ON SYNAPSE DETECTION AND PROOFREADING IN CONNECTOMICS

Yicong Li^{1,*}, Junjie Wang², Leander Lauenburg¹, Ella Hugie¹
Alexandra Irger¹, Wanhua Li³, Donglai Wei², Hanspeter Pfister^{1,*}

¹Harvard University, Cambridge, MA, USA ²Boston College, Chestnut Hill, MA, USA

³Nanyang Technological University, Singapore

ABSTRACT

We benchmarked vision-language models (VLMs) on the decisions annotators take when inspecting electron microscopy images in connectomics: synapse detection (presence and polarity) and proofreading (split errors and merge errors). For synapse detection, we evaluated 19 open and 2 closed models across various architectures and sizes under zero-shot, four-shot in-context learning and LoRA settings, against specialist models, on datasets constructed by us using public resources. For proofreading, we evaluated 3 open and 2 closed models on the ConnectomeBench2 dataset, with cross-species transfer from fly and mouse to human and zebrafish. Most models were at chance zero-shot; a few examples helped mainly the closed and largest open ones. LoRA on a few thousand labels brought open models level with specialist models. When evaluated on unseen species, the best adapted VLMs outperformed specialist models trained on the same data in identifying merge errors. The project will be publicly available upon acceptance.

Index Terms— Connectomics, electron microscopy, vision-language models, synapse detection, proofreading

1. INTRODUCTION

Connectomics aims to reconstruct the wiring diagram of the brain: every neuron and every synapse between them. Currently, electron microscopy (EM), while expensive [1, 2], is the primary imaging method that resolves both, at nanometer resolution over volumes that now reach peta-byte scale [3, 4], and the field has drawn considerable attention as these datasets appeared [5]. With the development of computer vision techniques, the reconstruction pipeline is mostly automatic in its two main stages: segmentation traces each neuron through the volume [6, 7], and synapse detection finds the contacts between them and assigns their direction [8, 9]. However, both stages still need human eyes. Synapse detectors are validated and corrected by annotators who judge, section by section, whether a contact is a synapse and which side is presynaptic. Segmentations are proofread, because a split or merge error propagates through every circuit the neuron takes part in; FlyWire took about three million proofreading edits [3, 10], and MICrONS proofreading continues years after imaging [4]. Each of these checks is a small visual decision taken on a few EM sections.

Vision-language models (VLMs) answer visual questions in natural language and can be steered by a few examples or a short fine-tuning run, yet on general microscopy benchmarks they still struggle even with modality identification [11]. Whether they can make

an annotator’s decision from raw EM, and what it takes to get them there, is largely unknown.

Prior work on these two tasks is supervised and task-specific. Synapse detection has been posed as voxel classification [8], as partner assignment from point annotations [9, 12], and as direct mask generation for the pre- and postsynaptic partners of a given cleft [13]. On the proofreading side, guided proofreading trains CNNs to flag split and merge errors and to propose the correction [14]. More recently, ConnectomeBench [15] evaluated several multimodal LLMs on proofreading tasks, but only on mesh renderings rather than EM images. Its successor, ConnectomeBench2 [16] released expert-labelled sites with EM slices from four species and trained a ViT to compare with human performance, without evaluating language models.

To the best of our knowledge, no prior study has evaluated a range of VLMs on synapse detection from EM, and none has benchmarked VLMs on proofreading tasks using ConnectomeBench2 [16], whose only published results are for a supervised ViT. Concretely, our contributions are:

- We built a synapse detection benchmark from public fly and mouse connectomes and evaluated 19 open and two closed VLMs on it zero-shot, with four in-context examples, and with LoRA, against specialist networks trained on the same images.
- We presented the first evaluation of VLMs on ConnectomeBench2: three open models were run on it in full against the dataset’s own ViT, and the same models were compared with two closed VLMs and three specialist networks on mouse and fly subsets.
- We found that most open models were at chance zero-shot, many returning one fixed answer, while the closed models read proofreading above chance. In-context examples helped the closed models and the largest open ones, and LoRA fine-tuning brought open models level with the specialists. Additionally, VLMs seem to be more resilient to cross-species transfer compared with specialist models.

2. BENCHMARK

2.1. Task 1: Synapse detection

We built the synapse detection tasks from two public sources with ground-truth synapse annotations, CREMI [17] for fly and MICrONS [4] for mouse. Every item is a single EM section of 256×256 pixels at 8 nm per pixel, a 2.05 μ m field of view. For the *presence* subtask, a yellow circle of 224 nm radius marks the query point and

*Corresponding authors: Yicong Li (yicong.li.0225@gmail.com) and Hanspeter Pfister (pfister@seas.harvard.edu).

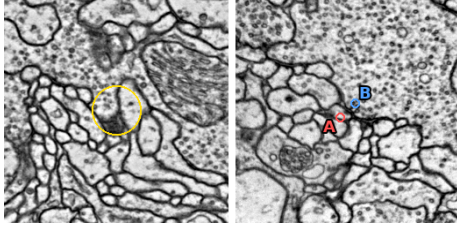


Fig. 1. Synapse detection example inputs of fly. Left: *presence*, the query point marked by the yellow circle (answer: yes). Right: *polarity*, markers A and B inside the two partner processes (answer: B).

the model is asked whether a chemical synapse is present there. For the *polarity* subtask, two markers, A and B, are placed inside the two partner processes of a synapse, with the letters assigned at random, and the model is asked which marker is presynaptic.

Fig. 1 shows one item of each subtask and below is the accompanying prompt:

Preamble (both subtasks): This is a 2 $\mu\text{m} \times 2 \mu\text{m}$ electron micrograph of brain tissue (adult *Drosophila* brain), imaged at 8 nm per pixel. Dark membranes outline neuronal processes; mitochondria appear as dark oval organelles; synaptic vesicles are small round $\sim 40 \text{ nm}$ blobs clustered inside axon terminals. **Presence:** A yellow circle marks the location of interest. Is there a chemical synapse (a synaptic cleft with a presynaptic vesicle cluster on one side and a postsynaptic density on the other) at the marked location? Answer with exactly one word: yes or no. **Polarity:** Two markers, A and B, are placed inside two adjacent neuronal processes that form a chemical synapse at their shared membrane. Which marker is inside the PRESYNAPTIC process, i.e. the side that contains the cluster of synaptic vesicles? Answer with exactly one letter: A or B.

For fly we used the three CREMI volumes [17], adult *Drosophila* brain at $4 \times 4 \times 40 \text{ nm}$ with hand-labelled synaptic clefts and their pre- and postsynaptic partners. We assigned volume A to training, B to validation and C to testing, so the three splits come from different brain regions. Positives samples for the *presence* subtask were created by placing one query at the centroid of every cleft, in the section where the cleft is largest. For each positive we then drew one negative at a random location that lies at least 256 nm from any cleft voxel in its section and 600 nm from every annotated synapse, and we kept the query point at the same offset from the crop centre as in the paired positive, so that the offset alone carries no information. For *polarity* we took the annotated partner pairs and placed the markers 72 to 240 nm from the cleft and at least 112 nm apart. This procedure yielded 246, 262 and 330 *presence* items for training, validation and test, half of them positive, and 170, 275 and 413 *polarity* pairs.

For mouse we used the public MICrONS cubic-millimetre volume of visual cortex (minnie65) [4], available at $8 \times 8 \times 40 \text{ nm}$ in its public mirror, which comes with a cleft segmentation and an automatically generated synapse table. Because the table is automatic, we kept only entries of at least 300 voxels whose cleft is confirmed at the crop centre, and we drew negatives at random points with no cleft within 256 nm. To keep the splits apart, we divided the volume into contiguous blocks along one axis, separated by gaps, so that training, validation and test tissue never overlap. This gave 6000, 1000 and 3000 samples for the *presence* subtask, half positive, and 3000, 500 and 1500 *polarity* pairs.

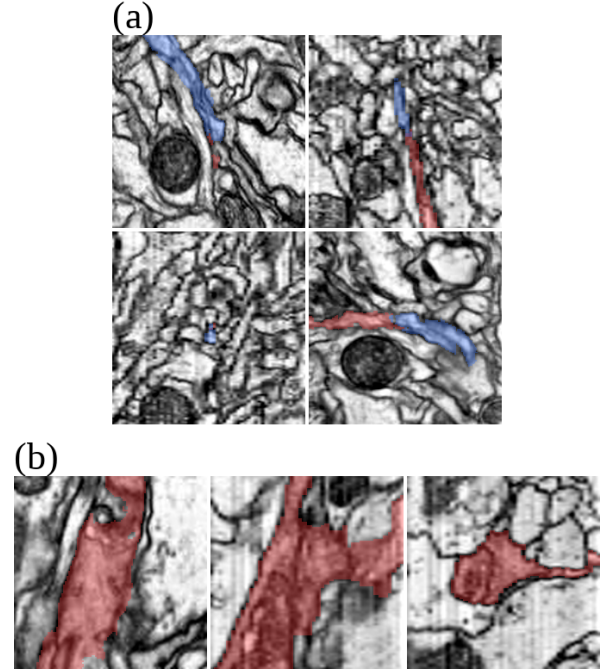


Fig. 2. Proofreading example inputs of fly. (a) A *split-error* site: the two segments are tinted red and blue; the proofreaders merged them, so the answer is yes. (b) A *merge-error* site: the segment is tinted red; the proofreaders split it, so the answer is no.

2.2. Task 2: Proofreading

For proofreading we did not build a new dataset but adopted ConnectomeBench2 (CB2) [16], which released proofreading sites from the edit histories of expert proofreaders, together with matched controls, in mouse, fly, human and zebrafish. Each site comes as four EM views taken before the edit, the xy , xz and yz planes and an oblique plane, at 224×224 pixels, together with the masks of the segments involved; the cutouts span $2.5 \mu\text{m}$ in mouse and human and $1.5 \mu\text{m}$ in fly and zebrafish. We kept the released images unchanged and only tinted the masks, so that the model knows which segments the question is about. In the *split-error* subtask, the two segments are tinted red and blue in all four planes, and the answer is yes when the proofreaders merged them, that is, at a split error. In the *merge-error* subtask, the union of the segments is tinted red in the three planes, and the answer is no when the proofreaders split it, that is, at a merge error. We dropped the oblique plane for this task because its orientation is chosen from the post-split masks and would leak the answer, as the dataset card itself warns.

Fig. 2 shows one item of each subtask and below is the accompanying prompt:

Split error: The four panels are electron-microscopy slices through the same location in adult *Drosophila* brain, each about $1.5 \mu\text{m}$ across: top-left xy , top-right xz , bottom-left yz , bottom-right an oblique plane. Two segments from an automatic neuron segmentation are tinted RED and BLUE and meet near the centre. Automatic segmentations sometimes wrongly split one neuron into two segments, whereas two different cells are separated by a continuous dark membrane. Do the red and blue segments belong to the SAME neuron, so that they should be merged? Answer with exactly

Table 1. Synapse detection, balanced accuracy (%) on the test splits. P: *presence*; Pol: *polarity*. Bold: best VLM per column; underlined: second best; *Italic rows*: specialist models, not ranked; –: not applicable.

Model	zero-shot				4-shot				LoRA			
	fly P	fly Pol	mouse P	mouse Pol	fly P	fly Pol	mouse P	mouse Pol	fly P	fly Pol	mouse P	mouse Pol
Qwen2.5-VL-3B	49.1	50.0	49.5	50.0	50.9	50.0	53.9	50.0	50.3	57.8	74.4	76.4
Qwen2.5-VL-7B	50.0	49.0	49.6	49.2	48.8	49.1	48.2	49.3	71.8	70.0	86.3	81.4
Qwen2.5-VL-32B	48.2	49.1	46.2	49.8	55.2	49.4	52.2	50.3	68.2	61.7	83.5	77.4
Qwen3-VL-2B	51.2	49.1	48.3	50.9	50.0	48.7	54.1	50.0	69.7	56.9	81.0	79.9
Qwen3-VL-4B	53.3	48.2	47.5	51.9	59.1	51.7	48.8	49.8	72.4	56.3	83.7	82.1
Qwen3-VL-8B	52.4	56.8	47.1	49.1	57.6	54.7	56.5	49.9	73.0	66.0	83.1	87.8
Qwen3-VL-32B	50.0	54.9	49.9	51.3	71.2	58.4	56.1	50.7	71.5	<u>76.2</u>	82.9	89.2
Qwen3.5-2B	53.0	56.9	51.7	49.7	57.0	52.1	55.9	48.7	75.5	65.1	87.1	87.6
Qwen3.5-4B	50.0	50.5	50.0	50.1	52.1	64.2	51.0	50.3	80.0	74.6	87.9	<u>90.1</u>
Qwen3.5-9B	50.0	71.9	50.0	52.5	75.8	72.4	56.5	49.4	<u>82.7</u>	75.4	89.0	89.9
Qwen3.5-27B	<u>68.5</u>	58.5	55.4	51.8	73.0	69.0	<u>63.9</u>	52.4	<u>82.7</u>	72.0	87.5	89.3
InternVL3-2B	49.7	50.2	49.3	50.1	50.0	48.0	49.9	46.0	77.3	61.2	88.8	77.2
InternVL3-8B	51.5	45.2	48.4	48.4	67.6	46.1	47.1	46.3	79.1	73.4	89.8	87.0
InternVL3-14B	51.5	47.1	50.4	48.4	60.3	48.4	52.5	48.2	80.0	72.3	89.8	86.9
InternVL3-38B	50.6	47.2	56.1	49.5	65.2	56.3	62.2	51.2	80.6	75.6	<u>91.6</u>	88.6
Gemma 4 E2B	52.1	48.4	47.3	49.8	51.2	49.6	49.8	52.7	60.9	53.6	79.4	76.2
Gemma 4 E4B	50.0	52.1	49.5	50.2	65.5	59.0	48.1	48.2	72.1	59.0	83.2	75.5
Gemma 4 12B	60.6	<u>75.0</u>	51.8	<u>61.5</u>	72.1	83.2	56.4	<u>61.5</u>	81.5	91.1	93.5	94.5
Gemma 4 31B	73.6	70.1	60.9	66.6	81.8	76.0	64.4	66.9	85.5	75.7	88.0	88.1
GPT-5.6 Sol	59.7	77.2	45.0	56.8	79.4	<u>83.1</u>	56.3	60.9	–	–	–	–
Claude Opus 5	59.7	63.4	<u>57.2</u>	52.0	68.2	<u>65.7</u>	62.2	53.7	–	–	–	–
<i>ResNet-50</i>	–	–	–	–	–	–	–	–	83.8	75.9	95.6	95.3
<i>ConvNeXt-T</i>	–	–	–	–	–	–	–	–	85.6	82.6	96.4	95.6
<i>ViT-S/14 (DINOv2)</i>	–	–	–	–	–	–	–	–	79.7	60.5	95.3	92.2

one word: yes or no. **Merge error:** The three panels are electron-microscopy slices (xy, xz, yz) through the same location in adult *Drosophila* brain, each about 1.5 μ m across. One segment from an automatic neuron segmentation is tinted RED. Automatic segmentations sometimes wrongly merge two different cells across a membrane near the centre of the view. Is the red segment a SINGLE neuron here, with no merge error at the centre? Answer with exactly one word: yes or no.

We evaluated proofreading on three different setups. The first is the complete test split of the dataset’s second release, 101 929 operations with their natural class mix, namely 20 879 split-error and 22 094 merge-error operations in mouse, 24 065 and 24 203 in fly, 5 351 and 60 in human, and 5 223 and 54 in zebrafish; models were trained on its complete training and validation split of 531 734 and 82 822, respectively. The second is a balanced subset of mouse and fly with 3000, 600 and 1000 sites per species and task for training, validation and test, on which the models can be compared with closed-source VLMs and specialist networks under our budget. The third measures transfer to the species that no model developed using the second setup ever saw in training: namely, human and zebrafish.

2.3. Evaluation protocol

We report balanced accuracy throughout and each model saw one image and one prompt per item; the prompt states the image scale, the marker convention and the definition of the target, and asks for exactly one word or letter. We decoded greedily with at most 64 new tokens and took the last *yes/no* or *A/B* in the reply as the decision; any other reply counts as wrong. All three regimes shared the same images and prompts. In the zero-shot regime the model received only the test sample. In the 4-shot regime we prepended four training

examples of the same species and subtask as prior turns, two for each answer. In the LoRA regime the model was fine-tuned with the recipe given in Sec. 2.4. As references, we trained specialist networks on exactly the same rendered images and the same label pools as the adapters.

2.4. Models

For synapse detection, we evaluated several open families of instruction-tuned VLMs across their size ladders: Qwen2.5-VL (3B, 7B, 32B) [18], Qwen3-VL (2B, 4B, 8B, 32B) [19], Qwen3.5 (2B, 4B, 9B, 27B) [20], InternVL3 (2B, 8B, 14B, 38B) [21] and Gemma 4 (E2B, E4B, 12B, 31B) [22]. For proofreading, due to budget constraints, we kept three mid-sized models, one each from the Qwen, InternVL, and Gemma lines: Qwen3.5-9B, InternVL3-8B and Gemma 4 12B. We also evaluated two closed models, GPT-5.6 Sol and Claude Opus 5, as zero-shot and in-context reference rows.

To fine-tune the open VLMs we used LoRA [23] with one recipe throughout: rank-16 adapters with $\alpha = 32$ and dropout 0.05 on the attention and MLP projections of the language model, the vision encoder and the bfloat16 base frozen, AdamW at a learning rate of 10^{-4} , a short linear warm-up followed by a cosine schedule, and only the answer token and the end-of-turn token supervised.

As references we fine-tuned three specialist networks: ResNet-50 [24] and ConvNeXt-Tiny [25], initialized from ImageNet, and a ViT-S/14 initialized from DINOv2 [26]. Each was trained end to end on exactly the rendered images and the label pool of the corresponding adapters, one model per subtask covering both species.

Table 2. Proofreading on the complete ConnectomeBench2, balanced accuracy (%) on the test split: all 101,929 samples across 4 species. S: *split-error* subtask, M: *merge-error* subtask. Bold: best VLM per column; *Italic row*: the dataset’s own reported EM-only ViT-B; –: not applicable.

Model	zero-shot		4-shot		LoRA	
	S	M	S	M	S	M
Qwen3.5-9B	50.0	52.9	58.7	57.3	94.6	90.2
InternVL3-8B	51.0	50.0	49.0	57.3	93.9	90.1
Gemma 4 12B	50.1	50.0	62.7	58.0	95.0	90.9
<i>CB2 ViT-B, EM only</i>	–	–	–	–	<i>94.2</i>	<i>91.1</i>

Table 3. Proofreading on the subset of ConnectomeBench2 (Sec. 2.2), balanced accuracy (%). S: *split-error* subtask, M: *merge-error* subtask. Bold: best VLM per row; *italic columns*: specialist models, not ranked; –: not applicable.

	Qwen3.5-9B	InternVL3-8B	Gemma 4 12B	GPT-5.6 Sol	Claude Opus 5	<i>ResNet-50</i>	<i>ConvNeXt-T</i>	<i>ViT-S/14</i>
<i>zero-shot</i>								
mouse S	50.0	53.4	50.4	67.2	73.1	–	–	–
mouse M	59.3	50.0	50.0	67.6	67.3	–	–	–
fly S	50.0	52.0	50.0	63.0	64.7	–	–	–
fly M	51.0	50.0	50.0	57.0	57.0	–	–	–
<i>4-shot</i>								
mouse S	60.2	49.7	64.5	74.1	76.4	–	–	–
mouse M	69.1	54.1	57.7	65.7	68.0	–	–	–
fly S	60.4	48.9	64.7	74.4	66.4	–	–	–
fly M	61.7	50.0	60.0	67.7	61.9	–	–	–
<i>LoRA, trained on mouse + fly</i>								
mouse S	93.3	89.3	92.5	–	–	94.9	94.9	85.3
mouse M	84.6	84.4	83.4	–	–	86.1	85.9	82.0
fly S	86.6	83.1	85.2	–	–	89.2	90.1	80.8
fly M	80.1	78.8	74.5	–	–	76.7	79.7	74.2
<i>transfer to unseen species</i>								
human S	83.1	72.2	79.4	–	–	81.9	85.4	65.7
human M	71.9	67.6	77.3	–	–	59.8	66.9	57.9
zebrafish S	88.0	79.0	85.4	–	–	87.3	89.3	78.8
zebrafish M	77.3	73.5	82.7	–	–	70.5	74.7	65.6

3. RESULTS

3.1. Synapse detection

Table 1 shows the synapse detection results. Zero-shot, most open models couldn’t interpret the images: across the 76 cells of the 19 open models the median score was 50.0 and 62 cells fell below 55. The exceptions were the two largest Gemma 4 and Qwen3.5 models, mostly on fly, where Gemma 4 31B reached 73.6 on *presence* and Gemma 4 12B 75.0 on *polarity*; on mouse no open model passed 67. The closed models stayed at or near chance on mouse, but GPT-5.6 Sol read fly *polarity* at 77.2, the highest zero-shot value in the table, while staying at 59.7 on fly *presence*.

Four in-context examples raised the performance: Gemma 4 12B reached 83.2 on fly *polarity*, above the best specialist at 82.6, Gemma 4 31B 81.8 on fly *presence*, and Qwen3.5-9B 75.8 on fly *presence*, whereas the best mouse cells stayed at 64.4 and 66.9. GPT-5.6 Sol reached 79.4 and 83.1 on the two fly subtasks, and Claude

Opus 5 68.2 and 65.7.

LoRA changed the picture entirely. 63 of 76 cells reached 70, and the best adapter per cell reached 85.5 on fly *presence*, 91.1 on fly *polarity*, 93.5 on mouse *presence* and 94.5 on mouse *polarity*, ahead of or level with the specialists on fly and within three points of them on mouse. Size mattered less once adapted: Gemma 4 12B beat Gemma 4 31B on three of the four cells, and Qwen3.5-9B matched or beat Qwen3.5-27B on all four.

3.2. Proofreading

Full ConnectomeBench2 evaluation. Table 2 reports the three open models evaluated on the complete ConnectomeBench2 (CB2) under the dataset’s own protocol. Zero-shot, all three sat at or near chance. Four in-context examples lifted the *split-error* subtask scores of Qwen3.5-9B and Gemma 4 12B to 58.7 and 62.7 and the *merge-error* scores of all three to 57 to 58. LoRA brought all three level with the dataset’s own EM-only ViT-B, at 94.2 and 91.1: Gemma 4 12B reached 95.0 and 90.9, and the other two were within about one point on both tasks.

Balanced subsets, closed models, and transfer. Table 3 compares the same three models with the two closed models and the specialist models on the balanced mouse and fly subset of CB2, and shows how their adapters transfer. Zero-shot, the open models were again at or near chance, apart from Qwen3.5-9B at 59.3 on mouse merge errors, whereas the closed models performed better: Claude Opus 5 scored 73.1 and 64.7 on split errors in mouse and fly, GPT-5.6 Sol 67.2 and 63.0. With four examples the closed models rose to 76.4 and 74.4 at best, Gemma 4 12B and Qwen3.5-9B reached 58 to 69, and InternVL3-8B barely moved. LoRA took the open models to 83.1 to 93.3 on split errors and 74.5 to 84.6 on merge errors, within 1.6 points of the references on mouse, 3.5 points behind on fly split errors, and ahead on fly merge errors, 80.1 against 79.7. Applied unchanged to human and zebrafish, the adapters kept most of this accuracy. On split errors the best adapter transferred about as well as the best specialist network, Qwen3.5-9B at 83.1 and 88.0 against 85.4 and 89.3 for ConvNeXt-Tiny. On merge errors the best adapter beat the best network on both species, Gemma 4 12B reaching 77.3 and 82.7 against 66.9 and 74.7, and on human all three adapters did.

4. DISCUSSION AND CONCLUSION

We benchmarked vision-language models for synapse detection and proofreading tasks on EM images in connectomics. Without adaptation, most open models are at or near chance and cannot be dropped into an annotation pipeline; the closed models performed better under zero-shot and gained from in-context examples, but stay well below the adapted open models. A few thousand labels with LoRA bring an open model level with a purpose-trained network, and the adapters carry to species never seen in training, the best of them beating the specialist networks on merge errors. What makes the VLM worth its cost is not accuracy in the trained setting, where a small network is about as good and cheaper per image, but what happens away from it. One adapter per task answered both questions in both species, where the specialist models needed one network per question, and the adapters generally lost less accuracy on unseen species, which suggests that large-scale pretraining on images and text supplies context the small models lack. A VLM is therefore the right tool when the decisions are many or when a volume comes from a species with no training data; a purpose-trained network remains the right tool for one fixed decision on one well-labelled dataset.

5. ACKNOWLEDGEMENTS

This research is supported by NSF grant NCS-FO-2124179 and NIH grant R01HD104969. The authors have no relevant financial or non-financial interests to disclose. The use of AI systems is only for editing and grammar enhancement.

6. COMPLIANCE WITH ETHICAL STANDARDS

This study used only publicly released, de-identified datasets, CREMI, MICrONS and ConnectomeBench2, under their open licenses, including the human tissue sites in ConnectomeBench2's public release. No new human or animal data was collected and ethical approval was not required.

7. REFERENCES

- [1] Yicong Li, Yaron Meirovitch, Aaron T. Kuan, Jasper S. Phelps, Alexandra Pacureanu, Wei-Chung Allen Lee, Nir Shavit, and Lu Mi, "X-Ray2EM: Uncertainty-aware cross-modality image reconstruction from X-ray to electron microscopy in connectomics," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023.
- [2] Aaron T Kuan, Jasper S Phelps, Logan A Thomas, Tri M Nguyen, Julie Han, Chiao-Lin Chen, Anthony W Azevedo, John C Tuthill, Jan Funke, Peter Cloetens, et al., "Dense neuronal reconstruction through x-ray holographic nanotomography," *Nature neuroscience*, vol. 23, no. 12, pp. 1637–1643, 2020.
- [3] Sven Dorkenwald, Arie Matsliah, Amy R. Sterling, et al., "Neuronal wiring diagram of an adult brain," *Nature*, vol. 634, pp. 124–138, 2024.
- [4] The MICrONS Consortium, "Functional connectomics spanning multiple areas of mouse visual cortex," *Nature*, vol. 640, pp. 435–447, 2025.
- [5] Nature Methods, "Method of the Year 2025: electron microscopy-based connectomics," *Nature Methods*, vol. 22, no. 12, 2025.
- [6] Jan Funke, Fabian Tschoop, William Grisaitis, et al., "Large scale image segmentation with structured loss based deep learning for connectome reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1669–1680, 2019.
- [7] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H. Sebastian Seung, "Superhuman accuracy on the SNEMI3D connectomics challenge," *arXiv preprint arXiv:1706.00120*, 2017.
- [8] Larissa Heinrich, Jan Funke, Constantin Pape, Juan Nunez-Iglesias, and Stephan Saalfeld, "Synaptic cleft segmentation in non-isotropic volume electron microscopy of the complete *Drosophila* brain," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. 2018, vol. 11071 of *Lecture Notes in Computer Science*, pp. 317–325, Springer.
- [9] Julia Buhmann, Arlo Sheridan, Caroline Malin-Mayor, et al., "Automatic detection of synaptic partners in a whole-brain *Drosophila* electron microscopy data set," *Nature Methods*, vol. 18, pp. 771–774, 2021.
- [10] Philipp Schlegel, Yijie Yin, Alexander S. Bates, Sven Dorkenwald, Katharina Eichler, Paul Brooks, et al., "Whole-brain annotation and multi-connectome cell typing of *Drosophila*," *Nature*, vol. 634, pp. 139–152, 2024.
- [11] Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy, "Micro-bench: A microscopy benchmark for vision-language understanding," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.
- [12] Yicong Li, Wanhua Li, Qi Chen, Wei Huang, Yuda Zou, Xin Xiao, Kazunori Shinomiya, Pat Gunn, Nishika Gupta, Alexey Polilov, et al., "Waspsyn: A challenge for domain adaptive synapse detection in microwasp brain connectomes," *IEEE transactions on medical imaging*, vol. 43, no. 11, pp. 3719–3730, 2024.
- [13] Nicholas Turner, Kisuk Lee, Ran Lu, Jingpeng Wu, Dodam Ih, and H. Sebastian Seung, "Synaptic partner assignment using attentional voxel association networks," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2020.
- [14] Daniel Haehn, Verena Kaynig, James Tompkin, Jeff W. Lichtman, and Hanspeter Pfister, "Guided proofreading of automatic segmentations for connectomics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9319–9328.
- [15] Jeff Brown, Andrew Kirjner, Annika Vivekananthan, and Ed Boyden, "ConnectomeBench: Can LLMs proofread the connectome?," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.
- [16] Jeff Brown, Tim Farkas, Gleb Razgar, and Edward S. Boyden, "ConnectomeBench2: A unified benchmark for automated connectomic proofreading," *arXiv preprint arXiv:2606.21116*, 2026.
- [17] Jan Funke, Stephan Saalfeld, Davi D. Bock, Srinivas C. Turaga, and Eric Perlman, "CREMI: MICCAI challenge on circuit reconstruction from electron microscopy images," 2016.
- [18] Shuai Bai, Keqin Chen, Xuejing Liu, et al., "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025.
- [19] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al., "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.
- [20] Qwen Team, "Qwen3.5: Accelerating productivity with native multimodal agents," Qwen blog, <https://qwen.ai/blog?id=qwen3.5>, 2026.
- [21] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al., "InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.
- [22] Gemma Team, "Gemma 4 technical report," *arXiv preprint arXiv:2607.02770*, 2026.
- [23] Edward J. Hu, Yelong Shen, Phillip Wallis, et al., "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [25] Zhuang Liu, Hanzhi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al., "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.