Natural Language-guided Neural Encoding Benchmark for Vision

Taha Razzaq **Hisan Naeem** **Asim Iqbal***
Tibbling Technologies
asim@tibbtech.com

Abstract

Understanding the link between visual stimuli and their neural representations is key to advancing Human-Computer Interaction, particularly for therapeutic and assistive technologies. Developing language-guided visual response systems could significantly enhance support for individuals with visual impairments, providing personalized assistance through descriptive language for daily tasks. Advancements in generative multimodal networks highlight the promise of image captioning models for such systems. However, evaluating their biological plausibility requires a rigorous benchmark that assesses how well these models produce captions that align with neural encoding in the visual cortex. In this paper, we present a novel benchmarking approach to evaluate the alignment of image captioning models with neural activity patterns, using a dataset of visual exposures and neural recordings from primates and mice. This method allows for a comparison of various models based on their congruence with biological neural responses, aiding in the development of assistive technologies for visually impaired individuals. Our work extends beyond computational vision, providing valuable insights for designing neuro-inspired generative multimodal networks. These advancements hold transformative potential for health-related applications, including natural language-driven visual aids and therapeutic interventions for individuals with visual impairments.

1 Introduction

A key question in neuroscience is how neuronal activity patterns correspond to visual stimuli. Neurons, each tuned to specific inputs, generate distinct activity patterns in response to visual cues [3]. For instance, expert pianists can identify music by observing key-touch movements, suggesting that visual input activates brain neurons to "hear" music [9]. Research [25, 14] indicates that visual neurons encode specific stimuli, linking visual input to neural activity in the visual cortex, which informs studies on deep learning models and their relationship to the visual cortex [19, 41, 31, 12].

One promising application of modeling neural activity is language-guided vision, which could help visually impaired individuals by describing visual stimuli through text-to-speech systems, mimicking the human visual system. This requires advanced image captioning models that accurately encode images and detect subtle visual changes.

To identify the best models, we introduce a benchmarking technique that evaluates image2text models using a text2neural mapping mechanism. This method maps image captions to visual cortex neurons, as shown in **Figure 1**. Validated with data from primates and mice, our method ensures cross-species consistency. The mouse visual cortex, similar to the primate cortex, also serves as a model for studying neural responses [27, 37, 20, 34]. Through this comprehensive evaluation, GIT and ClipCap emerged as top performers, demonstrating the strongest alignment with visual neuron responses across species.

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

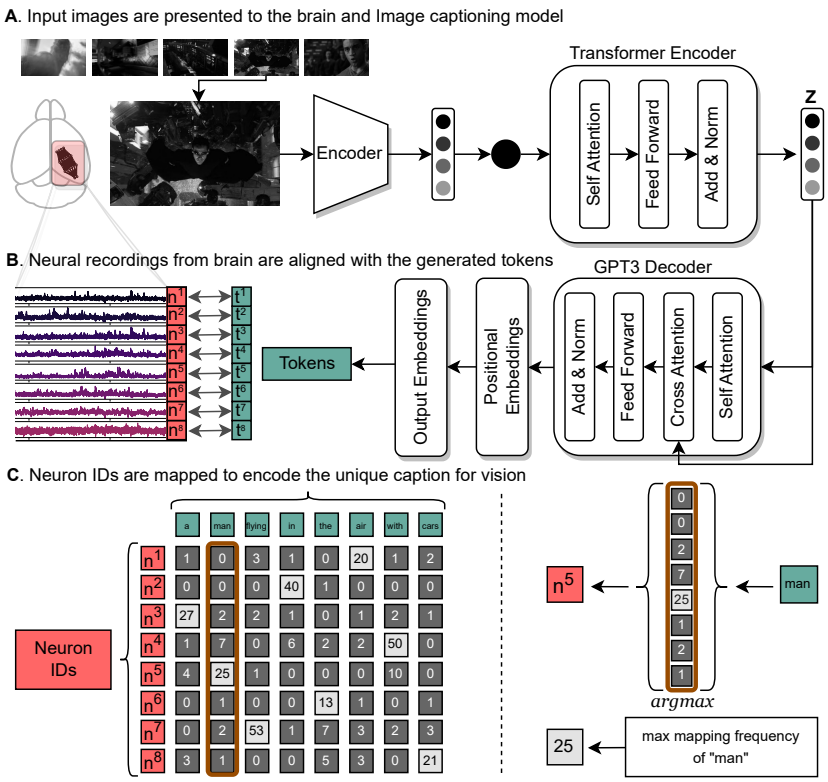

Figure 1: Block diagram framework of our technique: **A.** Frames are processed by the image captioning model and shown to the animal. **B.** Neural responses are mapped to text tokens (captions). **C.** Dataframe section shows mapping frequencies between neuron IDs and text tokens; the most frequent neuron is assigned to each token.

2 Datasets

Mouse visual cortex dataset: We use MICrONS [4], the largest open-source multi-modal connectomics dataset from the mouse visual cortex. This dataset offers diverse visual stimuli including natural scene images, and corresponding functional traces from $70,000$+ neurons across four cortex regions (V1, LM, AL, RL). It also includes 3D EM reconstructions and two-photon calcium imaging, covering over $200,000$ neurons and 523 million synapses, enabling robust neural response modeling.

Primate visual cortex dataset: To ensure generalizability to the human visual cortex, we use a primate dataset [2] with $1,960$ synthetic images presented to adult monkeys with neural responses recorded from V4 and IT regions. This dataset is crucial because the monkey visual cortex closely resembles the human's in structure and cell types [24], helps bridge the gap between primate and human visual processing, providing insights applicable to computational vision and aiding comparisons between biological networks and CNNs [7].

Curating Natural Movie Clips dataset from mouse and primate for caption diversity: The MICrONS dataset includes 19 sessions with over 90 minutes of visual stimuli, including natural movie clips and directional visuals. We focused on Natural Movie clips, initially filtering to $24,000$ frames per session. Sampling every $6th$ frame, we retained $82,243$ frames. Additionally, we included the entire primate dataset to create our final *Natural Movie Clips* dataset which was used for experimentation.

3 Methods

3.1 Image2text module

As a first step to our benchmarking framework, we use image2text models to convert visual stimuli to text captions. The text captions are then mapped to the neural activity of the mouse visual cortex (via

the text2neural framework) and the ability of each image2text model to create a text2neural mapping is evaluated. The entire pipeline of our framework is summarized in **Figure 1**.

Shortlisted image captioning models: To evaluate our benchmarking technique, we analyzed ten state-of-the-art image2text models, including ViT-GPT2-IC [28], BLIP [18], ExpansionNet-V2 [10], show-attend-tell [40], ClipCap [23], ViT-GPT2 [1], BLIP2 [17], PromptCap [11], mPLUG [16] and GIT [39]. We examined how their text representations correlate with neural activity in the mouse and primate visual cortices using frames from the Natural Movie Clips dataset. Our model selection spans various architectures and attention mechanisms. All models were tested on Google Colab's T4 GPU.

Visualizing unique caption distributions through embedding projections: To visualize the distribution of unique captions generated by each image2text model, we applied the UMAP dimensionality reduction algorithm to all the captions. Using word2vec model [21], a consistent embedding vector for each caption (S) was created, resulting in a 1×100 vector for each token (w). These token vectors were then stacked to form a matrix ($\mathbf{V}$) of shape $n \times 100$, where n is the number of tokens in the caption. The final embedding vector ($\vec{s}$) was obtained by averaging across the first axis, yielding a 1×100 vector, calculated as $\vec{s_i} = \frac{\sum_j^n \mathbf{V}_{ji}}{n}$, where $S = (w_1, w_2, \ldots, w_n)$ and $\vec{x_i} = word2vec(w_i)$.

$$\mathbf{V}_{n\times 100} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,100} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,100} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,100} \end{bmatrix} \tag{1}$$

For K unique captions, we compute a matrix of size $K \times 100$, with each row representing a caption's embedding vector ($\vec{s}$), which is used to generate UMAP embeddings. The UMAP plots for all models, shown in **Figure 2A**, serve as a metric to evaluate the models' efficiency in encoding visual features.

3.2 Text2neural module

After generating image2text representations, we mapped the captions to neural activity in the visual cortex. Each neuron was treated as an independent unit with a preferred token, assuming a direct mapping between neurons and tokens, justified by their response to specific visual features. This creates a correlation between the captions and the most active neurons for a given visual stimulus. For a given image frame, its generated caption ($\vec{s}$) is converted into a text embedding vector ($T = (t_1, t_2, \ldots, t_k)$) using a pre-trained CLIP model [29], ensuring standardized text token generation within its vocabulary (size C). Simultaneously, we consider all the visual cortex neurons (X) and extract the top k active neurons to form an activation vector, $A = (a_1, a_2, \ldots, a_k)$ and their corresponding neuron IDs, $N = (n_1, n_2, \ldots, n_k)$. The activation and neuron ID of the i^{th} top active neuron for the given input stimulus are represented by a_i and n_i, while $k = length(T) = length(A)$.

To achieve a direct mapping between T, A, and N, we use rank-based sorting. We apply the $argsort$ function to T to obtain its rank vector, then sort A and N accordingly, which aligns them with the order of T, allowing for a direct index-based mapping, where the i^{th} text token in T corresponds to the i^{th} neuron ID in N. We repeat this process for all captions generated by an image2text model across the entire Natural Movie Clips dataset, storing the resulting mappings in a text2neural dataframe (NT) of size (X, C), where X is the total number of neurons in the mouse visual cortex. Each row of NT represents a neuron ID as a bag-of-words, and the frequency of a text token's mapping to that neuron is represented by the column. Using NT, we extract the neuron ID which best represents a specific token along with the number of times the text token was mapped to that neuron (mapping frequency). This process is shown in **Figure 1B-C**. For all (10) shortlisted image2text models, we repeat this process and eventually are left with the highest mapping frequency for each token corresponding to each model which is stored in a vector P where $P = (p_1, p_2, \ldots, p_{10})$. where p_i is the mapping frequency of the i^{th} image2text model for a specific text token. After repeating this for all the text tokens, we create a Token Distribution Dataframe (DT) of shape $(C, 10)$.

$$\text{DT} = \begin{bmatrix} P_{1,1} & P_{1,2} & \ldots & P_{1,10} \\ P_{2,1} & P_{2,2} & \ldots & P_{2,10} \\ \vdots & \vdots & \ldots & \vdots \\ P_{C,1} & P_{C,2} & \ldots & P_{C,10} \end{bmatrix} \tag{2}$$

where $p_{i,j}$ represents the mapping frequency of the i^{th} text token for the j^{th} image2text model.

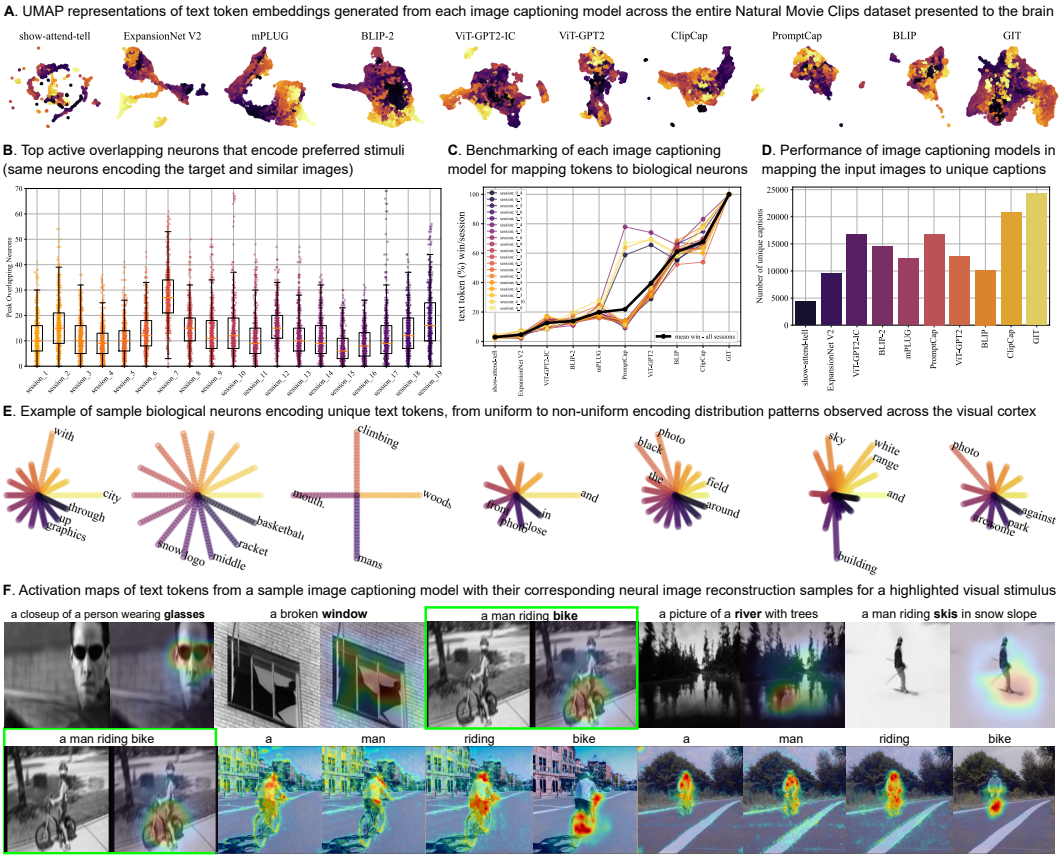

Figure 2: **A.** UMAP of ten image captioning models based on unique captions. **B.** Alignment of active neurons in response to visual stimuli from the mouse dataset. **C.** Benchmarking models by their alignment with active neurons. **D.** Performance of models in mapping images to captions. **E.** Neurons and their mapping to text tokens (colors for tokens, length for alignment). **F.** Top: Activation maps for captions on images. Bottom: Image reconstruction from text.

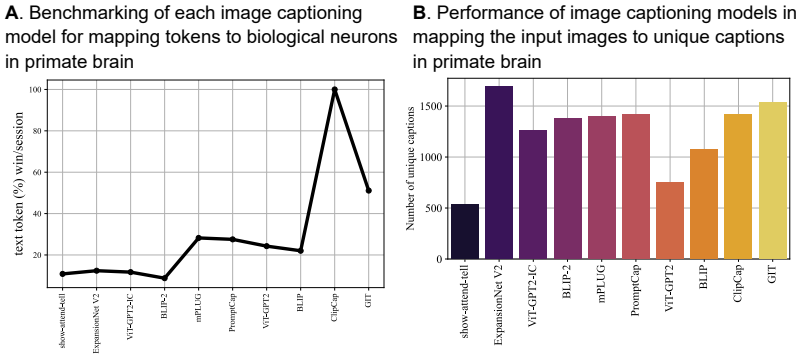

Figure 3: **A.** Model alignment with active neurons in the primate brain. **B.** Performance in mapping images to captions for the primate visual cortex.

To determine the best-performing image2text model for each text token, we use the DT dataframe, applying a row-wise $argmax$ to identify the winning model, and store these results in a vector W. We then calculate the frequency of wins for each model by counting the occurrences of its index in W, resulting in a vector Z. Each element in Z corresponds to the performance of a specific image2text model across sessions, representing its encoding capability, where $Z = (z_1, z_2, \ldots, z_{10})$ and z_i denotes the performance of the i^{th} model. We repeat the process for the entire Natural Movie Clip dataset and evaluate the performance of the image2text models.

4 Results

Effective image2text encoding in latent space: The UMAP projections in **Figure 2A** illustrate how image2text models encode visual features through spread, density, and color variation. Minimal color variation indicates caption similarity, while varying captions produced for similar input frames show models' sensitivity to visual changes. Ideal models tend to have wide, dense latent spaces with small clusters, indicating caption diversity and noise resistance. GIT and ClipCap perform best, while show-attend-tell shows limited, highly clustered space, indicating weaker performance.

Encoding similarity in visual cortex and image2text models: The encoding pattern of image2text models, which shows high diversity for significant changes and low sensitivity to minor stimuli variations, mirrors neural encodings as depicted in **Figure 2B**. The figure illustrates that while the mean number of overlapping neurons encoding preferred stimuli ranges from 5 to 15, the data exhibits significant spread and variability. This indicates that the visual cortex clusters similar stimuli, with certain neurons being sensitive to subtle visual changes. For language-guided visual therapy, it's crucial that the image2text model is both diverse and noise-resistant, generating natural captions that effectively stimulate corresponding neurons in the visual cortex. Our text2neural method successfully captures and validates this dynamic correlation between brain encoding and image2text models.

Image2text alignment with neural encoding: Our findings on the text2neural technique support our hypothesis that the information encoded by image2text models resembles that of the visual cortex. As shown in **Figure 2C** and **Figure 3A**, models with a broader latent space demonstrate stronger text2neural mapping. The black curve in **Figure 2C** represents the mean number of wins across all sessions, with GIT and ClipCap consistently outperforming other models in encoding neural representations across all text tokens. Conversely, show-attend-tell exhibits the fewest wins for both datasets. The strength of a model's text2neural mapping is closely linked to its ability to generalize across diverse image stimuli, aligning with our UMAP experiment findings where GIT and ClipCap excelled in representing the visual encoding capacity of the visual cortex (**Figure 2A**).

Quantitative analysis of image caption diversity validates text2neural mapping: We quantified the sensitivity of image2text models by analyzing the unique captions generated, as shown in **Figure 2D** and **Figure 3B**. Both datasets show that GIT and ClipCap produce the most unique captions, supporting our text2neural mapping results. These models effectively capture subtle features in images, unlike show-attend-tell, which assigns similar captions to different images. ExpansionNet V2 performs best on the primate dataset but less so on the mouse dataset, likely due to differences in dataset size and image resolution. Our text2neural mappings also reflect the neural profiles in the visual cortex, as shown in **Figure 2E**. We observe diverse spiking profiles among neurons, with each encoding a varying number of tokens, indicating sensitivity to a range of contrasting images [6].

Multi-modal evaluation via attention fields and image reconstruction: As a qualitative measure of performance, we visualized the attention maps from our best-performing image2text model for various images, as shown in **Figure 2F**. Each image is paired with its caption and the attention map generated for a specific token. Notably, these attention maps resemble Gaussian fields rather than point clouds, closely mirroring the receptive fields of visual cortex neurons [13, 38].

5 Conclusion

Our approach offers a novel framework for benchmarking state-of-the-art image captioning models based on their ability to encode neural activity from the visual cortex. By systematically mapping neural responses to text tokens generated by image2text models, we present a unique evaluation criterion that aligns with the visual cortex's functionality. This has significant implications for applications such as natural language-guided visual therapy, providing both qualitative and quantitative metrics to assess image2text models using mammalian visual cortex data. To our knowledge, this is the first attempt to benchmark image2text models through neural responses.

While the results of our technique are positive, and corroborated by the standard benchmarking results **Supplementary Table 1**, our method is limited to mapping neurons to text tokens in isolation. A promising area for future research is the mapping of entire captions across populations of neurons. This approach is based on the concept of neuronal ensembles, which suggests that neurons respond collectively to their preferred stimuli. Such studies could provide valuable insights into the relationship between individual neurons and the encoding capacity of interconnected populations.

References

[1] bipin. image-caption-generator, 2021.

[2] Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):1–18, 12 2014.

[3] Luis Carrillo-Reid and Rafael Yuste. Playing the piano with the cortex: role of neuronal ensembles and pattern completion in perception and behavior. *Current Opinion in Neurobiology*, 64:89–95, 2020. Systems Neuroscience.

[4] MICrONS Consortium, J. Alexander Bae, Mahaly Baptiste, Agnes L. Bodor, Derrick Brittain, JoAnn Buchanan, Daniel J. Bumbarger, Manuel A. Castro, Brendan Celii, Erick Cobos, Forrest Collman, Nuno Maçarico da Costa, Sven Dorkenwald, Leila Elabbady, Paul G. Fahey, Tim Fliss, Emmanouil Froudakis, Jay Gager, Clare Gamlin, Akhilesh Halageri, James Hebditch, Zhen Jia, Chris Jordan, Daniel Kapner, Nico Kemnitz, Sam Kinn, Selden Koolman, Kai Kuehner, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Gayathri Mahalingam, Sarah McReynolds, Elanine Miranda, Eric Mitchell, Shanka Subhra Mondal, Merlin Moore, Shang Mu, Taliah Muhammad, Barak Nehoran, Oluwaseun Ogedengbe, Christos Papadopoulos, Stelios Papadopoulos, Saumil Patel, Xaq Pitkow, Sergiy Popovych, Anthony Ramos, R. Clay Reid, Jacob Reimer, Casey M. Schneider-Mizell, H. Sebastian Seung, Ben Silverman, William Silversmith, Amy Sterling, Fabian H. Sinz, Cameron L. Smith, Shelby Suckow, Zheng H. Tan, Andreas S. Tolias, Russel Torres, Nicholas L. Turner, Edgar Y. Walker, Tianyu Wang, Grace Williams, Sarah Williams, Kyle Willie, Ryan Willie, William Wong, Jingpeng Wu, Chris Xu, Runzhe Yang, Dimitri Yatsenko, Fei Ye, Wenjing Yin, and Szi chieh Yu. Functional connectomics spanning multiple areas of mouse visual cortex. *bioRxiv*, 2021.

[5] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khc, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 7 2021.

[6] Justin L. Gardner, Pei Sun, R. Allen Waggoner, Kenichi Ueno, Keiji Tanaka, and Kang Cheng. Contrast adaptation and representation in human early visual cortex. *Neuron*, 47(4):607–620, 2005.

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022.

[8] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2022.

[9] Takehiro Hasegawa, Ken-Ichi Matsuki, Takashi Ueno, Yasuhiro Maeda, Yoshihiko Matsue, Yukuo Konishi, and Norihiro Sadato. Learned audio-visual cross-modal associations in observed piano playing activate the left planum temporale. an fmri study. *Cognitive Brain Research*, 20(3):510–518, 2004.

[10] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. Expansionnet v2: Block static expansion in fast end to end training for image captioning, 2022.

[11] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning, 2023.

[12] Asim Iqbal, Phil Dong, Christopher M Kim, and Heeun Jang. Decoding neural responses in mouse visual cortex through a deep neural network. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[13] Andreas J Keller, Morgane M Roth, and Massimo Scanziani. Feedback generates a second receptive field in neurons of the visual cortex. *Nature*, 582(7813):545–549, 2020.

[14] Roxanne Khamsi. Jennifer aniston strikes a nerve. *Brain*, 2004.

[15] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization, 2022.

[16] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections, 2022.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[19] Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[20] James H Marshel, Yoon Seok Kim, Timothy A Machado, Sean Quirin, Brandon Benson, Jonathan Kadmon, Cephra Raja, Adelaida Chibukhchyan, Charu Ramakrishnan, Masatoshi Inoue, et al. Cortical layer–specific critical dynamics triggering perception. *Science*, 365(6453):eaaw5202, 2019.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[22] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[23] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021.

[24] Felipe Mora-Bermúdez, Farhath Badsha, Sabina Kanton, J Gray Camp, Benjamin Vernot, Kathrin Köhler, Birger Voigt, Keisuke Okita, Tomislav Maricic, Zhisong He, et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife*, 5:e18683, 2016.

[25] Yunjun Nam, Takayuki Sato, Go Uchida, Ekaterina Malakhova, Shimon Ullman, and Manabu Tanifuji. View-tuned and view-invariant face encoding in it cortex is explained by selected natural image fragments. *Scientific reports*, 11(1):7827, 2021.

[26] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[27] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.

[28] NLP Connect. vit-gpt2-image-captioning (revision 0e334c7), 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[31] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[34] Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

[35] stabilityai. sd-vae-ft-ema-original, 2022.

[36] stabilityai. sd-vae-ft-mse-original, 2022.

[37] Anne E Urai, Brent Doiron, Andrew M Leifer, and Anne K Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature neuroscience*, 25(1):11–19, 2022.

[38] Pooja Viswanathan and Andreas Nieder. Comparison of visual receptive fields in the dorsolateral prefrontal cortex and ventral intraparietal area in macaques. *European Journal of Neuroscience*, 46(11):2702–2712, 2017.

[39] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022.

[40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[41] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

A Supplementary Material

A.1 Mouse visual cortex dataset

As discussed in **section 2** of the primary manuscript, we use MICrONS dataset for our experiments which is an open-source anatomical and functional data collected from mouse visual cortex. **Figure 4** provides an overview of the MICrONS dataset, summarising its different aspects quantitatively. **Figure 4 (a)** shows the overall connectivity of the set of neurons that were part of both the anatomical and functional data, whereas **Figure 4 (b-e)** shows the sample connections from unique sub-regions inside the mouse visual cortex. These chord plots demonstrate the anatomy (i.e. neuronal locations and their connections) as well as the functional profile of the entire data with respect to the presented input stimuli. Each bar on the circumference of the chord plot represents a unique neuron in mouse visual cortex whereas the height of each bar represents the highest activity of that neuron on its preferred stimulus. As an example, sample preferred stimuli are also shown around the chord plots for the top 10 highly active neurons, depicting the diversity of the stimuli encoded by the neural population.

The colors of the connections are an indicator of the strength of the synapses, with darker colors representing a stronger connection. **Figure 4 (b-e)** shows the connections of the neurons in individual regions, as well as the outgoing connections from that region to other regions. As a general trend, this can be seen that neurons in a particular region are more densely connected with each other, and have relatively less connections with neurons present in other regions. **Figure 4 (h)** is a quantitative summary of the incoming and outgoing synapses of the 4 regions shown previously. **Figure 4 (i)** shows the total neurons present in the data from each sub-region in mouse visual cortex and **Figure 4 (j)** shows sample functional responses of neurons captured using Calcium imaging (two-photon) with their corresponding presented stimuli on top. The diversity in connectivity between different regions is observed to be non-uniform and different regions have different connection density profiles. This data serves ideal for our purpose to map the input images to neuro-aligned captions by introducing a benchmark of the image captioning models.

We utilize the MICrONS and primate datasets to conduct our experiments. A subset of their unique and distinctive input stimuli is also shown in **Figure 7**.

A.2 Text2neural module

We primarily discussed the text2neural mapping of our schema in **subsection 3.2** and further expand in this subsection. We expect that if the visual encoding of image features by text tokens is functionally similar to the neural encoding of the same visual features in the visual cortex, then the encoded preferred stimuli space of a population of neurons should have a similar text representation and the corresponding text2neural representation. **Figure 2E** already depicts an aspect of this observation where a population of neurons encodes unique text tokens. After further analysis, we report the pool of such neural encoding patterns into uniform and non-uniform clusters as shown in **Figure 5** and **Figure 6**, respectively. Each plot in these figures represents a unique neuron's token distribution. The subplots in **Figure 5** (uniform clusters) demonstrate a few such examples, where the number of tokens mapped to a single neuron varies, their mapping frequency remains uniform. **Figure 6** (non-uniform clusters), on the other hand, shows a non-uniform distribution of encoding tokens. The overall distribution of the cumulative clusters as text tokens, across neurons for each session and for all image2text models is shown in **Figure 8**. It can be observed that the number of unique text tokens mapped to unique neurons varies across different image2text models. For instance, some models tend to have only $2-3$ neurons with meaningful text tokens mapping whereas other models have a larger number of unique text tokens being mapped to unique neurons. This is in line with our previous analysis performed to quantify model performance using text2neural mapping strength, as shown in **Figure 2C** and **Figure 3A**. To further demonstrate this similarity between the encoding of visual information in the visual cortex and image2text models, we employ different techniques which are detailed ahead.

A.2.1 Neural population encoding for similar visual stimuli

We systematically compare the stimuli (image) frames, generated captions (text tokens) and corresponding highly active neurons with their mapped text2neural representations. For the entire Natural

Movie Clips dataset, we compute the similarity between all the input frames presented to the animals. We first extract the CLIP embeddings [29] of the frames and compute their pair-wise MSE. Repeating this step for each input, we find its top k most similar image frames.

In theory, similar visual features should be encoded by a common set of neurons (with similar preferred input stimuli encoded in their receptive field) and hence we expect that for a given input frame and its corresponding similar frames, some overlap should be observed between the set of top activated neurons for each frame. To quantify this, we start by extracting top P firing neurons against each image, resulting in a set of neurons s_i, where i corresponds to the i^{th} similar image. Here, s_x represents the set of top firing neurons for our target image. For the purpose of our analysis, we define *match* as the cardinality of the set $s_x \cap s_i$. Subsequently, for each target frame, its final *match* score is obtained by $\sum_i^k |s_x \cap s_i|$. The results of this experiment on the MICrONS dataset are shown in Figure 2B.

It is interesting to observe that sessions 7 and 19 have a higher median score, compared to the other sessions. This can be attributed to the internal distribution of the stimuli frames in these sessions, where a high similarity is present. In addition, this can also be referred to the internal connectivity of the neurons observed in these sessions, which impacted this score. We also observe a similar trend for the primate visual cortex as shown in **Figure 11**.

A.2.2 Text2Neural mapping across image frames

While the previous experiment reveals novel information, we conduct further analysis to gain deeper insights. For each image2text model, we extract the captions for the target image, c_x, and the corresponding k similar images, denoted as c_i, computing a *caption-matching* score. By creating a set V containing all the unique text tokens from c_1 to c_k, and treating c_x as a set containing tokens for the target image, we use $|c_x \cap V|$ to compute the *caption-matching* score for a single target frame. This score is a representative of the encoding capability of an image2text model, where a lower value indicates drastic variations across the generated captions of similar images. On one hand, variations in generated captions represent the model's ability to detect minor changes in visual features, while on the other hand, such variability for highly similar image frames is also a depiction of the model's inconsistent mapping between image frames and the corresponding text captions. However, in our case, the input stimuli contain considerable variation and hence a lower *caption-matching* score is preferred. **Figure 9** summarizes these results across all the image2text models for 19 sessions of the MICrONS dataset while the result of this experiment on the primate visual cortex dataset is summarized in **Figure 10**. ClipCap and GIT have a lower median *caption-matching* score compared to other models, which further supports our findings from the previous experiments (Figures 2C and 3A. Similarly, show-attend-tell is appears to be a poor performing model, with a large variance in its *caption-matching* scores. Plots in **Figure 12** (except the top-left one) represent the number of unique text tokens for each session of the MICrONS dataset, by sweeping a threshold for the minimum mapping frequency of each token, from 1 (dark purple) to 22 (bright yellow) for each model per session. The plot on the top-left of **Figure 12** corresponds to Figure 2C that shows the average of all the sessions.

A.2.3 Mapping between visual and text responsive neurons

For evaluation from a different perspective, we compute which model best depicts the mapping between biological and artificial neurons, that are encoding the visual features and mapping to their text captions. Here, we refer to biological neurons as the top active neurons for input stimuli, whereas artificial neurons represent the neural IDs obtained by mapping most frequent text tokens to biological neurons. In order to find the text token that encodes a specific neuron response, we extend the text2neural mapping as described in subsection 2.5. For a given session and neuron, the mapping frequency of all the text tokens for the image2text models is stored. We consider the text token with the highest mapping frequency across each image2text model and find the overall highly representative text tokens among them. Since commonly occurring text tokens such as "a", "and", "the", "of" tend to get mapped to multiple neurons repeatedly, considering the text tokens only with their highest mapping frequency will be insufficient to capture the broad variability in text tokens. Hence, in order to ensure that a wide variety of tokens are mapped to unique neurons, we ensure that a single text token is not mapped to multiple neurons. In case the highest mapped text token for a

neuron already has a corresponding neuron mapped to it, we consider the next highly mapped token, and so on.

In order to capture the model's text2neural encoding capabilities, we compute the overlap between the artificial and biological neurons. Set B, representing biological neurons, is created by $(\bigcup_i^k s_i) \cup s_x$, whereas set A containing artificial neurons is created by $\bigcup_i^k m_i$, where m_i is the set produced by mapping each token in c_i to its corresponding neural ID. A final *overlap score* is computed by $|A \cap B|$, which is a representative of the model's ability to emulate the mouse brain, successfully mapping biological neurons to artificial ones. Each plot in **Figure 13** represents an image2text model and its overlap between biological and artificial neurons per session in the MICrONS dataset. A higher biological to artificial neuron overlap indicates stronger text2neural encoding capabilities of the image2text model. We observe that our best performing models, ClipCap and GIT, have significant overlap which correlates with our text2neural encoding results (**Supplementary Table 1**), whereas models like show-attend-tell have a fairly low overlap between artificial and biological neurons. We also observe a direct correlation between the image2text capability of a model and its text2neural mapping as observed by its *overlap score* between biological and artificial neurons.

A.3 Multi-modal evaluation via attention fields

Another approach to quantify the performance of image2text models is to reconstruct images from their generated captions and compare them with the original visual stimuli presented to the animal subject. In order to perform this comparison, we feed the generated captions from image captioning models to pre-trained state-of-the-art text2image models and compare their generated outputs with the input frames.

We include the state-of-the-art Stable Diffusion [32] model and its variants (stable-diffusion-EMA [35] and stable-diffusion-MSE [36]), VQ-Diffusion([8], Latent Diffusion Model (LDM) [32], DeepFloyd-IF [33], unCLIP [30] and DALL·E-Mini [5]. In addition, GLIDE [26] and RQ-Transformer [15] are also used to cover different model architectures. Although, majority of these models are diffusion-based models, they vary in terms of their architectures and training.

We compare the original input frames with the generated outputs of text2image models and compute the SSIM score across them, as summarized in **Supplementary Table 1**. In order to utilize the most unique and feature-rich image frames from the MICrONS dataset, we perform preliminary filtering over the stimuli. We capture image frames across the dataset, and filter them using BRISQUE [22] - an Image Quality Assessment (IQA) algorithm. BRISQUE generates a score for each frame independently, which captures its perceptual quality. A lower score means the image is not blurred, and contains less noise. The filtered image frames through BRISQUE score are the best ones from each session. For a balanced comparison, top 15 image frames, based on BRISQUE scores from each session are considered for text2image reconstruction.

We demonstrate some examples of the generated images from the text2image models for GIT (best performing image2text model) in **Figure 14**. Finally, its corresponding attention maps, shown in **Figure 15**, capture how each text token in the caption is mapped to its visual feature in the generated image. Although, we use a range of different state-of-the-art text2image models, we provide a generalized code snippet detailing the layout of a forward pass for a typical text2image model as *Forward pass through a Text2Image model* below.

Furthermore, we also provide a code snippet for the generic layout of a forward pass of an image2text model as *Forward pass through an Image2Text model*. **Figure 16** provides a holistic view of our experimental pipeline, showcasing all the modules, namely image2text, text2neural, and text2image.

Supplementary Table 1: The table's top section summarizes SSIM scores between reconstructed and original stimuli. The middle section is benchmarking results of image-to-text models on real-world datasets, and the bottom section shows the number of unique captions generated by all models for the mouse visual cortex dataset.

Text2Image \ Image2Text	show-attend-tell	Expansion-Net V2	ViT-GPT2-IC	BLIP2	mPLUG	Prompt Cap	ViT-GPT2	BLIP	ClipCap	GIT
GLIDE	**0.147**	**0.151**	0.159	0.150	**0.160**	**0.153**	**0.153**	**0.138**	**0.153**	**0.161**
DeepFloyd-IF	0.135	0.120	0.141	0.128	0.119	0.134	0.087	0.126	0.099	0.149
RQ-Transformer	0.132	0.129	0.122	0.137	0.080	0.133	0.106	0.133	0.115	0.129
stable-diffusion	0.131	0.124	0.132	0.116	0.103	0.111	0.095	0.116	0.085	0.113
VQ-Diffusion	0.120	0.137	0.150	0.135	0.112	0.135	0.122	0.120	0.112	0.146
LDM	0.095	0.094	0.099	0.095	0.095	0.101	0.093	0.098	0.081	0.094
DALL·E-Mini	0.134	0.137	0.139	0.127	0.170	0.108	0.111	0.112	0.089	0.130
stable-diffusion-EMA	0.130	0.125	0.134	0.123	0.105	0.107	0.100	0.118	0.091	0.122
stable-diffusion-MSE	0.123	0.125	0.121	0.107	0.106	0.114	0.099	0.112	0.089	0.116
unCLIP	0.135	0.145	**0.169**	**0.150**	0.112	0.131	0.105	0.128	0.119	0.138
CIDEr-D (MS COCO)	-	138.5	-	145.8	155.1	150.1	-	136.7	108.35	144.6
BLEU@4 (MS COCO)	20.3	42.1	-	43.7	46.5	45.4	-	40.4	32.15	42.3
Unique Captions	4384	9587	16749	14482	12301	16836	12711	10206	20799	24275

Forward pass through an **Image2Text** model

```
def generate_caption(image):
    # extract features
    features = encoder(image)
    # find latent vector
    latents = ViT_enc(features)
    # find embeddings for caption
    word_embed = gpt_dec(latents)
    # embeddings to words
    cap_list = []
    for token in word_embed:
        word = tokenizer(token)
        cap_list.append(word)
    caption = "␣".join(cap_list)
    return caption
```

Forward pass through a **Text2Image** model

```
def generate_images(caption):
    # extract caption embedding
    word_embeddings = text_encoder(caption)
    # embeddings to latents
    latents = []
    for token in word_embeddings:
        embedd = UNET(token)
        latent = model(embedd)
        latents.append(latent)
    # convert latents to image
    image = visual_decoder(latents)
    return image
```

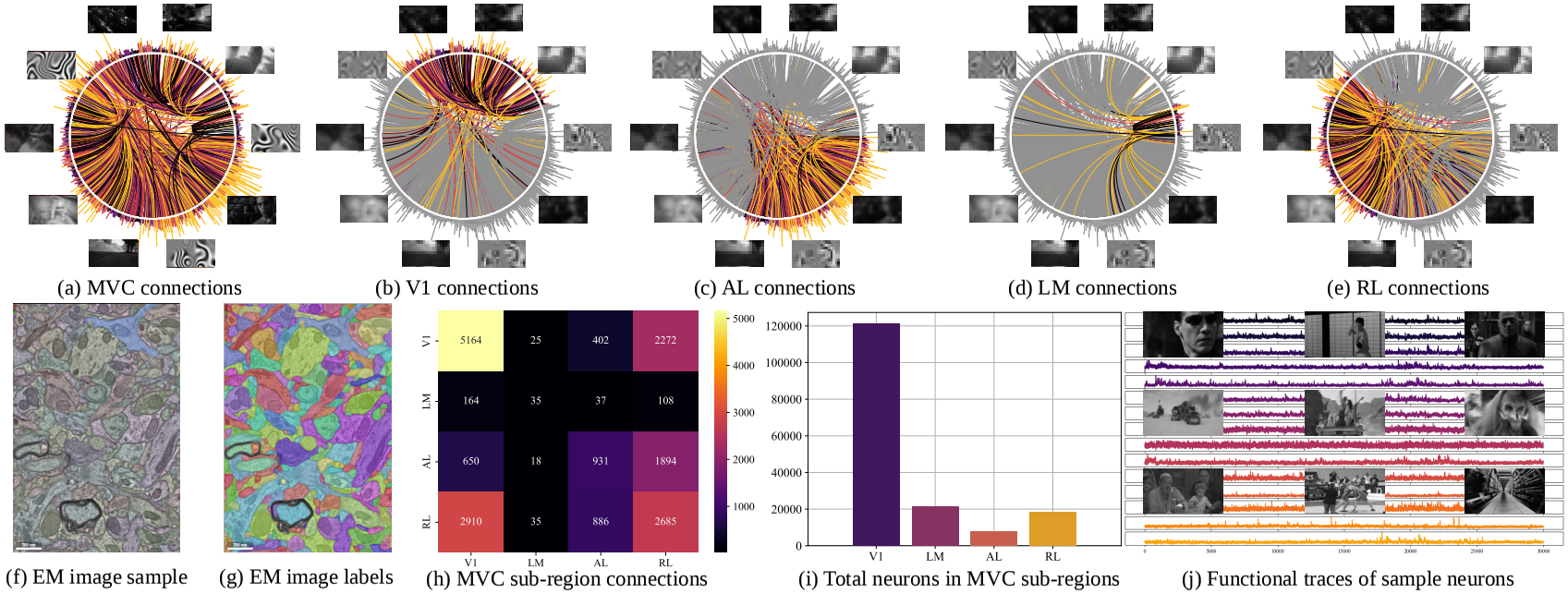

Figure 4: The complete profile of the MICrONS Explorer data is shown with overall connectivity of neurons in (a) as well as mouse visual cortex shortlisted sub-regions (b-e). The spikes on the chord plot shows the highest active neurons for 10 sample stimuli. A snippet of the anatomical EM (Electromagnetic) data is shown in (f) with the corresponding segmentations of the neuron and connections in (g). The sub-region based connectivity across mouse visual cortex is plotted in (h) and the total number of neurons captured in the data are plotted in (i). The functional profiles (neural representations or neurons' response profile) of 15 sample neurons for the input natural movie stimuli clips are plotted in (j).

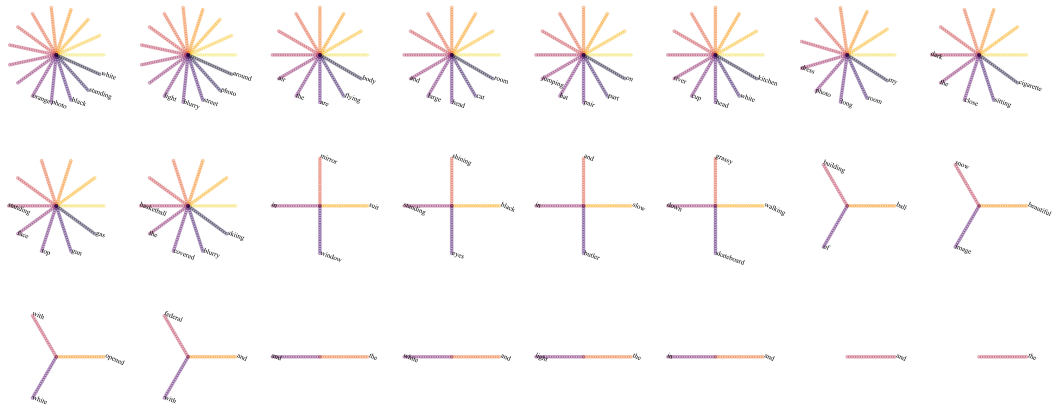

Figure 5: Neurons encoding text tokens with uniform distributions. Each sub-plot represents a neuron along with its overall uniform text token distribution.

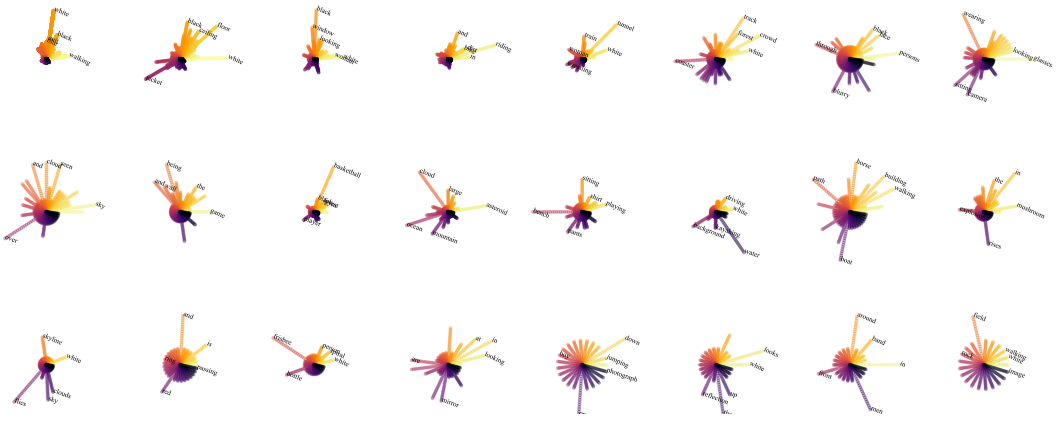

Figure 6: Neurons encoding text tokens with non-uniform distributions. Each sub-plot represents a neuron along with its varying token distribution.

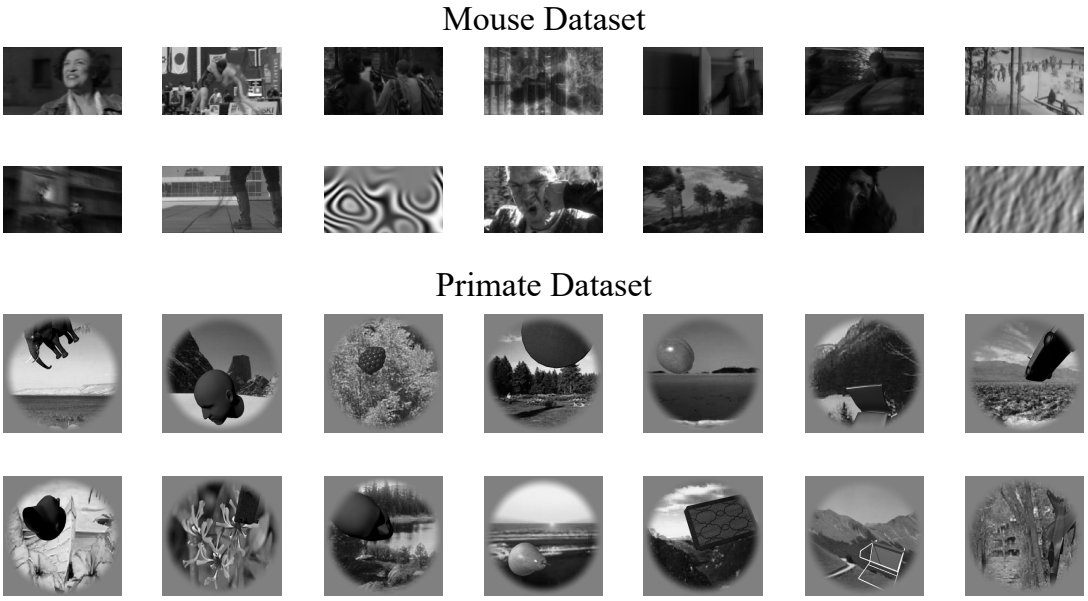

Figure 7: A few examples of input stimuli from the mouse and primate datasets are visualised. The mouse dataset contained 2 different types of stimuli, natural images and paramteric stimuli (named monet and trippy), whereas the primate dataset only contains synthetic natural images, generated using specialized software.

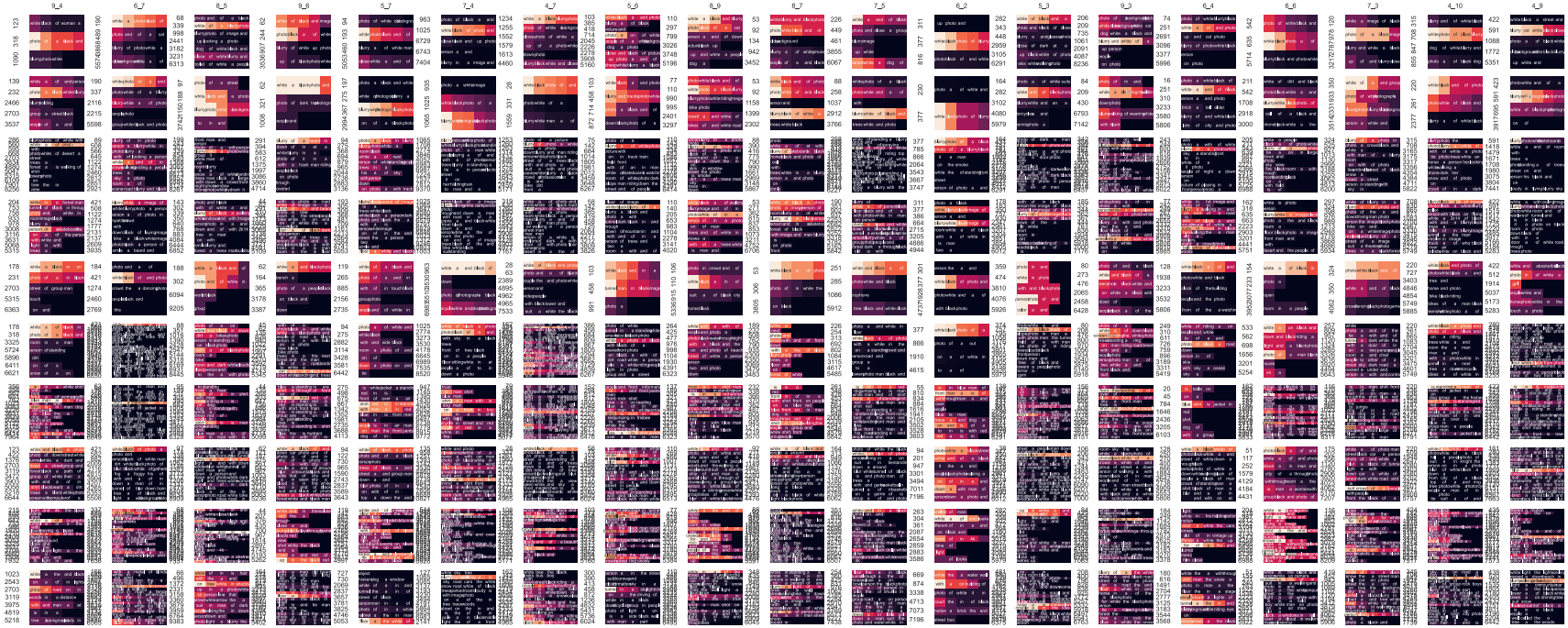

Figure 8: Confusion Matrix - based sub-plots representing the unique text2neural mapping for each image2text model/session of the MICrONS dataset. Each row represents an image2text model while the columns correspond to a session. Each confusion matrix shows the top text tokens mapped to unique neurons for the given image2text model/session. While some models have text tokens mapped to a larger number of neurons, other models only have 2-3 neurons being encoded by text tokens.

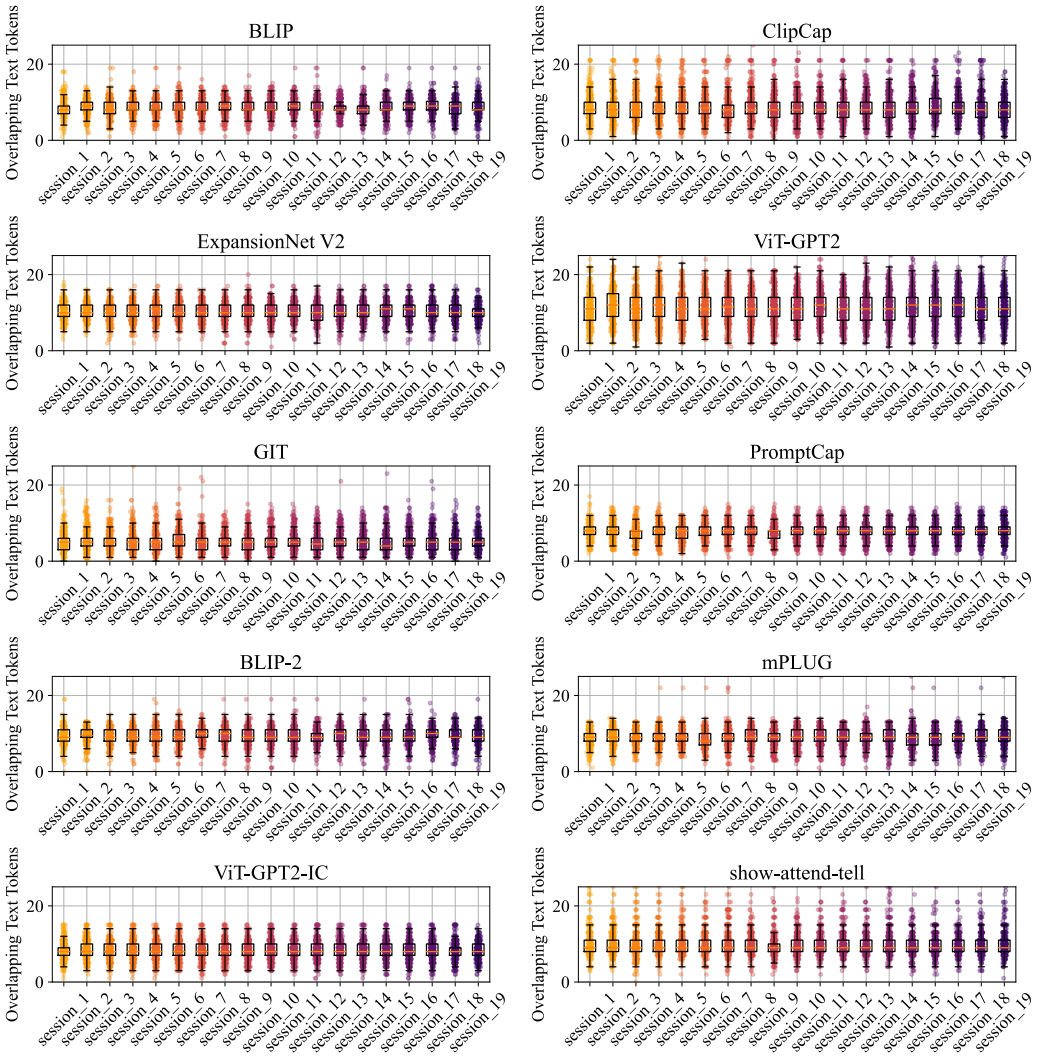

Figure 9: Overlap between the caption of the target image and the captions for the k similar images for the mouse data, computed for each image2text model, across each session. It is a representation of the diversity of the captions generated by each image2text model.

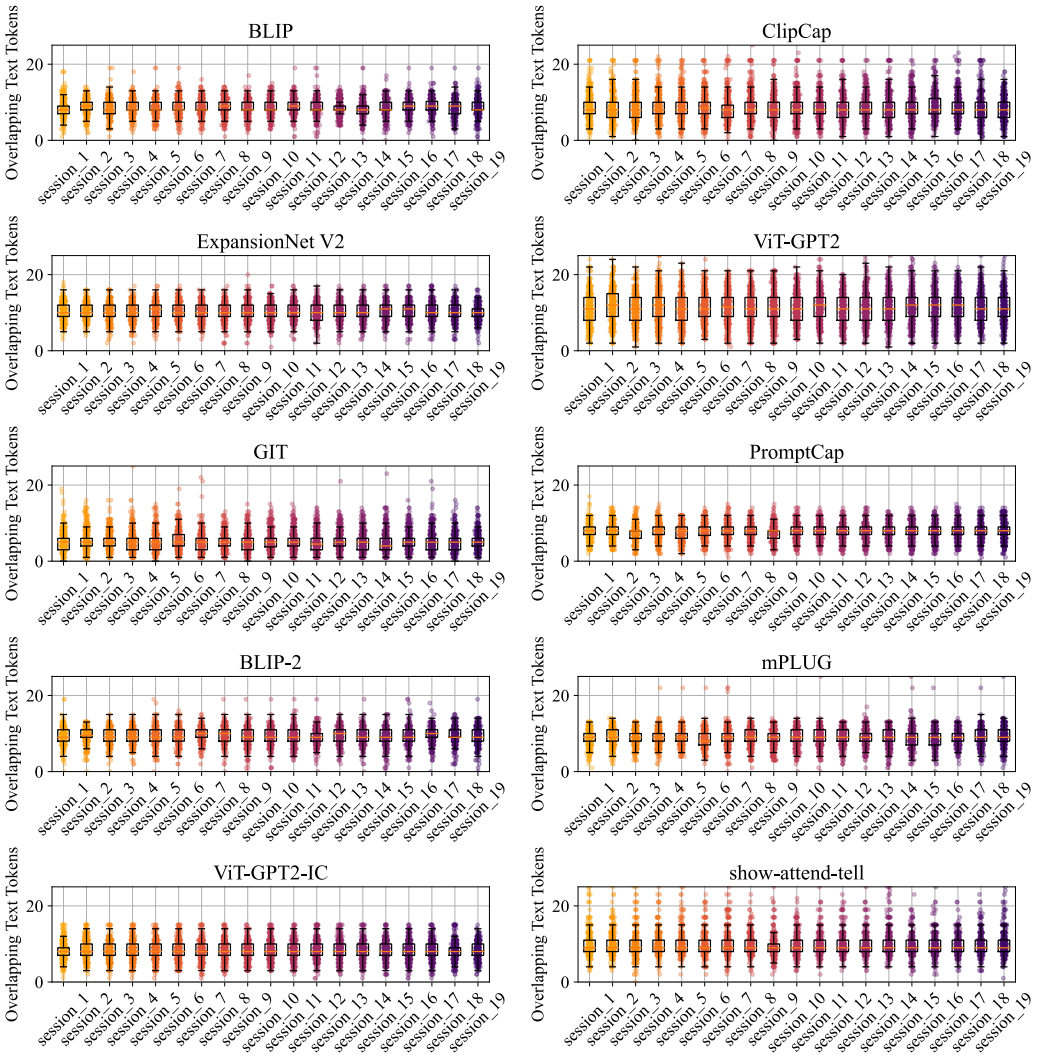

Figure 10: Overlap between the caption of the target image and the captions for the k similar images for the primate data, computed for each image2text model. It is a representation of the diversity of the captions generated by each image2text model.

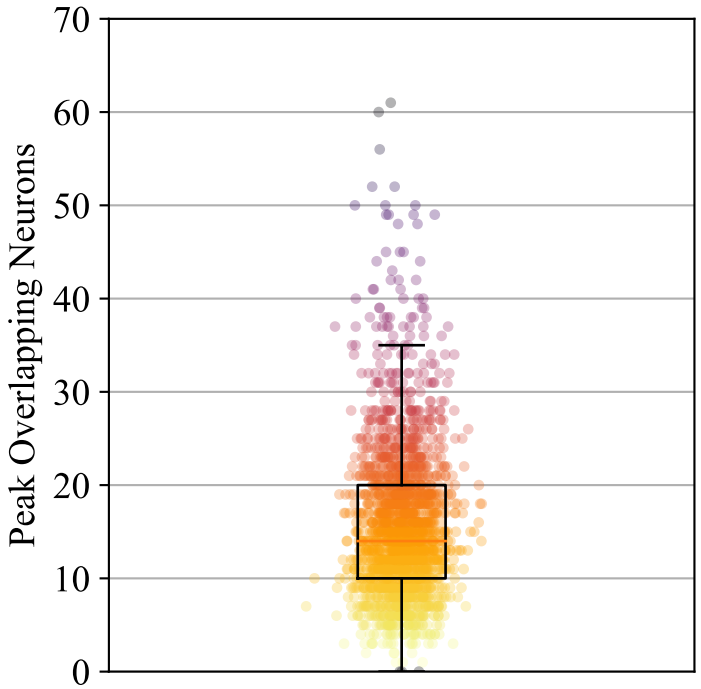

Figure 11: The distribution of overlapping neurons for the primate dataset. For an image, and its k similar images, it measures the overlap between top active neurons that fire for the target image and its similar images, where significant overlap implies that neurons encode preferred stimuli.

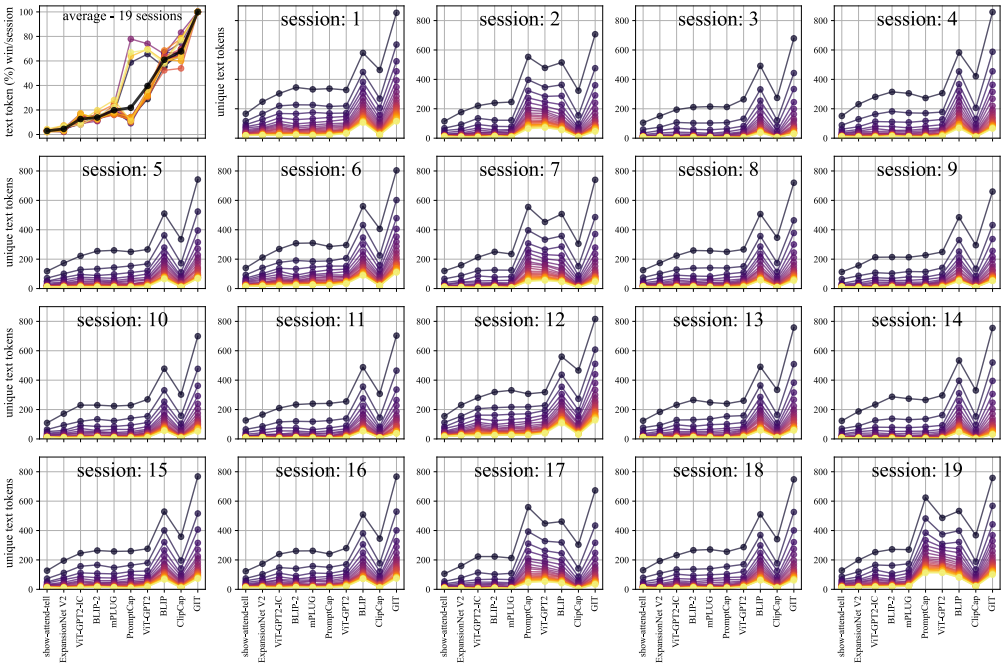

Figure 12: The top-left plot shows the normalized wins of each image2text model across all sessions of the MICrONS dataset. The black curve represents the average wins across all session. The rest of the 19 plots represent the results of each session when a threshold sweep is applied on the minimum mapping frequency of each text token. The sweep ranges from 1 to 22, where lower values are represented by darker colors (1-black), and lighter colors (22-yellow) represent higher and more stringent values.

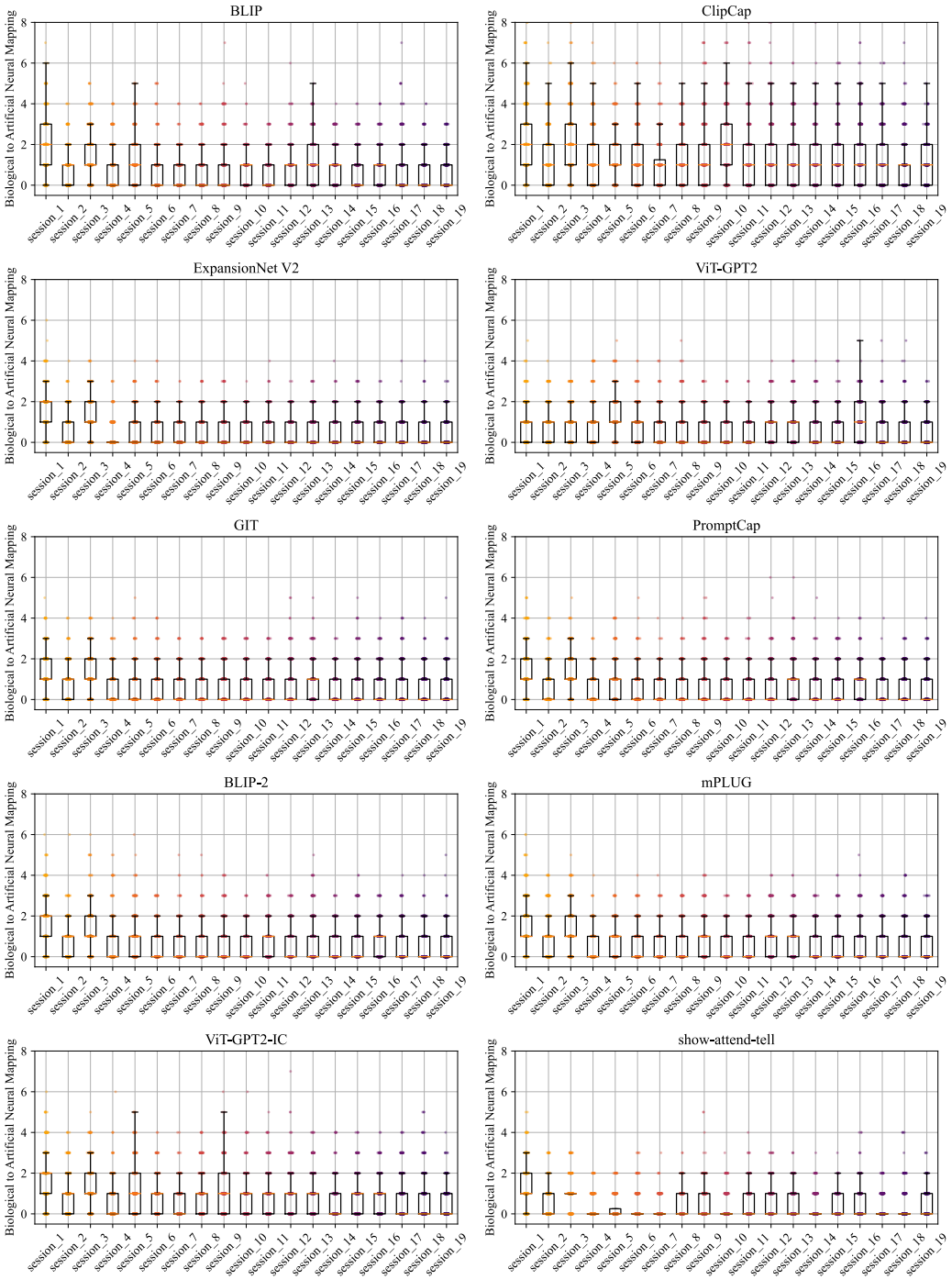

Figure 13: Biological to artificial neuron mapping for each image2text model, across all sessions of the MICrONS dataset. Each sub-plot represents an image2text model's session-wise mapping between the biologically active neurons and the neurons extracted using the text2neural representations. It is a measure of a ranked overlap between biological and artificial neurons.

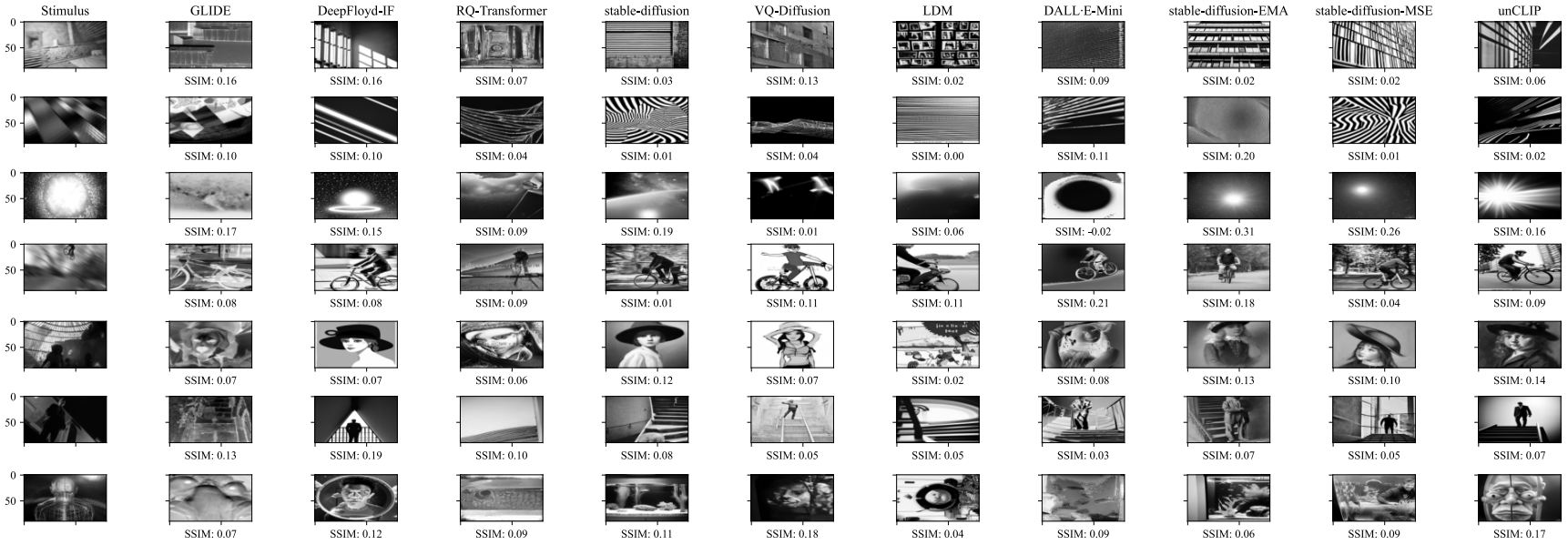

Figure 14: Generated images by text2image models for a few sample stimuli. GIT's generated captions for the stimuli were used as inputs to the text2image models. The generated images are used to assess the decoding capabilities of each image2text model.

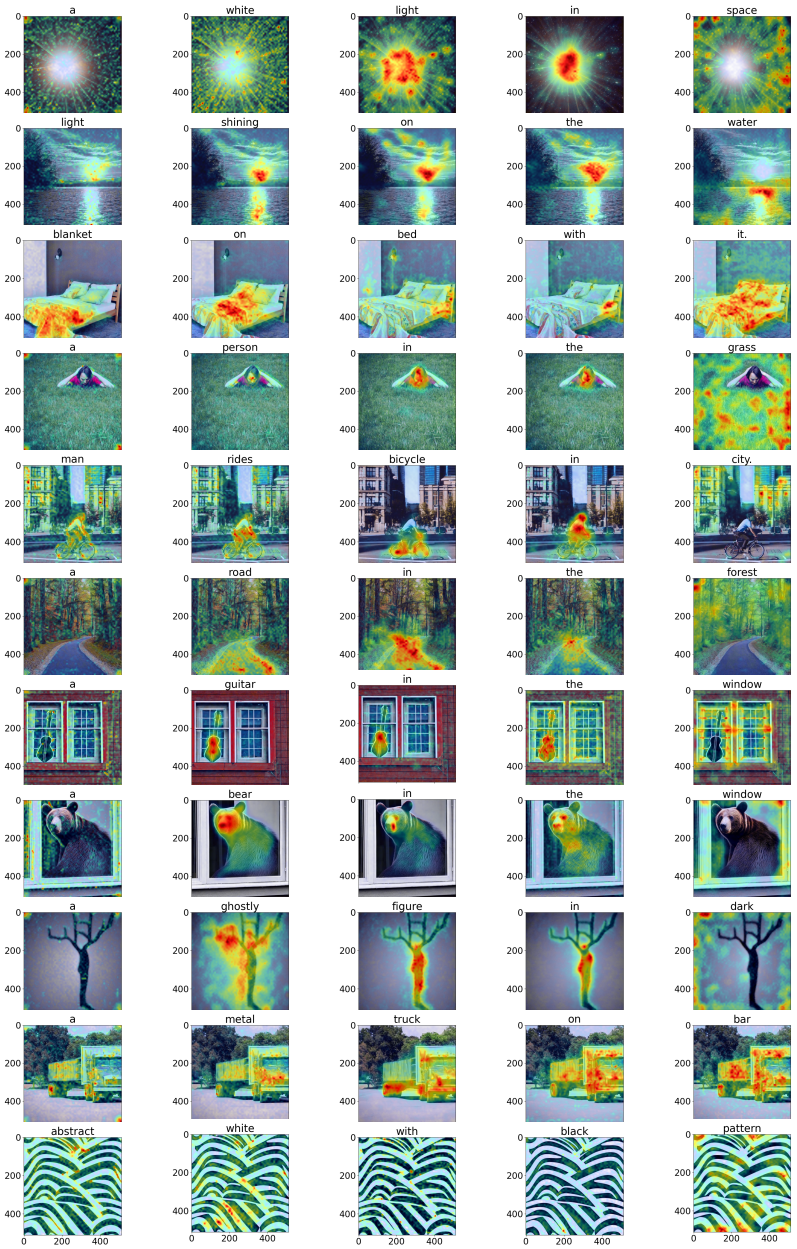

Figure 15: Attention maps of a text2image model for a sample of text captions generated by GIT (image2text model). Each row corresponds to a single caption, while each column shows the mapping between the text tokens and visual features in the generated image.

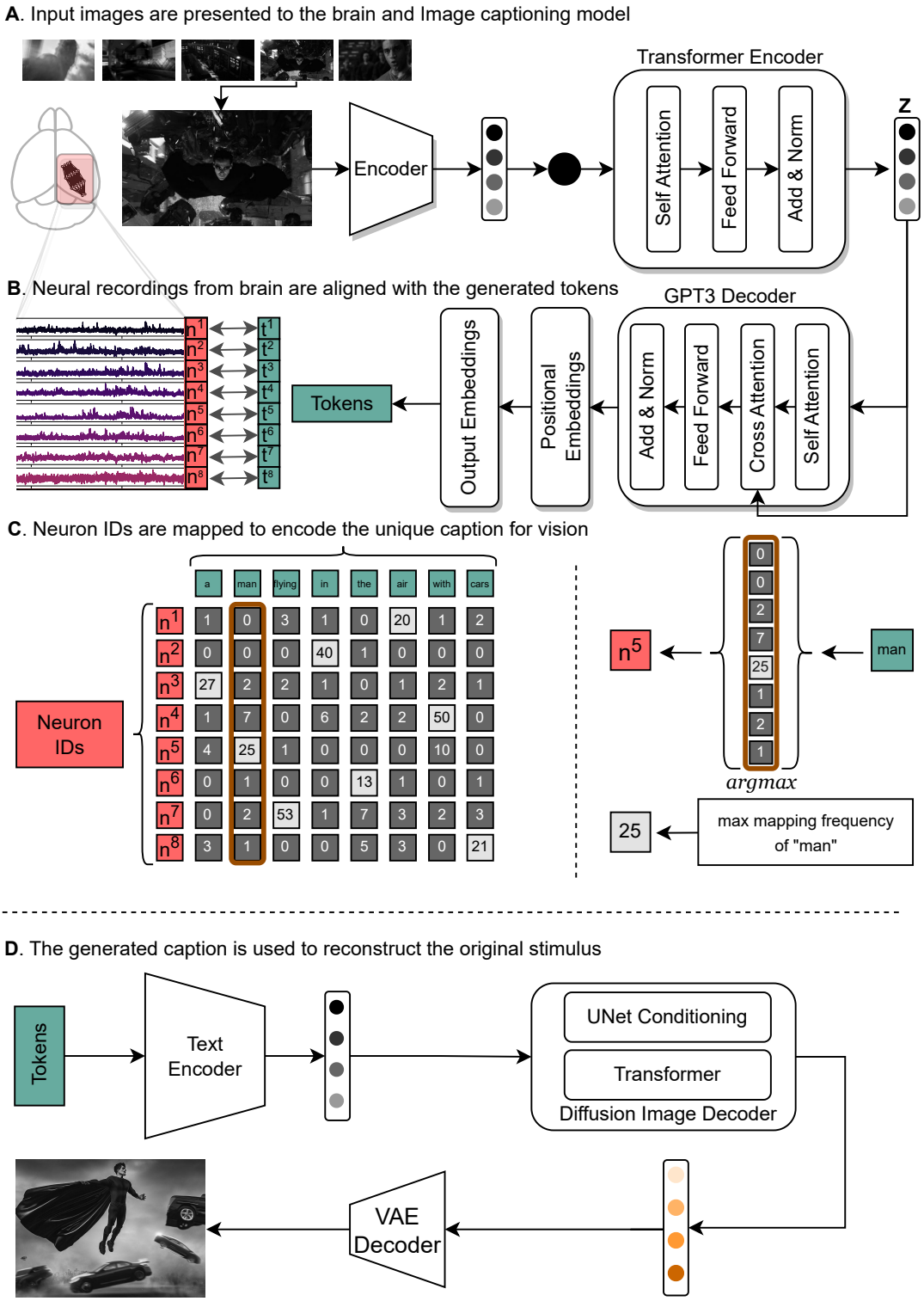

Figure 16: An overview of the entire experimental pipeline. subsections A, B, and C correspond to Figure 1, whereas subsection D represents the text2image module. The caption generated in subsection B is fed to a text2image model which generates an image based on the caption.

B 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: As mentioned in the abstract and the introduction, we propose a novel natural language-guided benchmarking technique which is explained in detail throughout the manuscript. In section 3, we give a step-by-step explanation of the individual modules of our pipeline.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations are clearly mentioned in the Conclusion section.

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: We do not present any theoretical results and show the performance of our techniques through qualitative and quantitative figures. We explain each module in section 3.

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: For our paper, we used publicly available datasets which are detailed in section 2. In section 3, we provide the details of the individual modules of our pipeline along with necessary equations, which can be used to reproduce the technique and hence the results shown in the paper.

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: We used publicly available open-sourced datasets for the paper and provide a pseudo code for the image2text and text2image module of our pipeline however the code will be made open-source upon the acceptance of the paper.

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: [Yes]

   Justification: In section 3, we detail the values of different parameters involved in our text2neural module.

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: We report qualitative and quantitative results to demonstrate the performance of our technique.

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 relevant information regarding the computer resources is mentioned section 3.

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: NeurIPS Code of Ethics was not violated during this study.

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: The impact of our work is mentioned in Section 1 of the paper.

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:

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: The original owner of all the assets used in the paper are properly cited.

13. **New Assets**

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

    Answer: [Yes]

    Justification: We introduce a novel benchmarking technique which utilizes existing image2text models and combines their output with our text2neural module. We provide concise documentation regarding the technique and break it down into multiple steps in section 3.1.

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: No crowdsourcing or research with human subjects was involved in our paper.

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: No crowdsourcing or research with human subjects was involved in our paper.