NEURONS: Emulating the Human Visual Cortex Improves Fidelity and Interpretability in fMRI-to-Video Reconstruction

Supplementary Material

A. Details of Instruction Prompts

Our NEURONS consists of four decoupled tasks, *i.e.*, scene description generation, concept name generation, segmentation mask generation, and rule-based key object discovery. However, the cc2017 dataset [45] only contains paired fMRI and visual stimuli. To generate high-quality labels to enable the decoupled tasks, we have designed a series of detailed instruction prompts for Qwen and Grounded-SAM. The complete instruction prompts are shown in Fig. a.

B. More Details about NEURONS

B.1. More Details about Brain Model

The Brain Model employs the MindEye2 [34] as the backbone, which consists of a ridge regression module, a Residual MLP module, and a diffusion prior network. The ridge regression module maps x_c to a lower dimensional for easier follow-up, the Residual MLP module further learns the representation in a deeper hidden space, and the diffusion prior network transforms the fMRI hidden features to image embeddings.

The training of the Brain Model mainly consists of three losses: contrastive learning loss $\mathcal{L}_{\text{CLIP}_\text{t}}$ between CLIP text embedding $\hat{e}^t$ and e^t, contrastive learning loss $\mathcal{L}_{\text{CLIP}_\text{v}}$ between CLIP video embedding $\hat{e}^v \in \mathbb{R}^{BF\times N\times C}$ and e^v, and prior loss $\mathcal{L}_{\text{prior}}$. $\mathcal{L}_{\text{CLIP}_\text{t}}$ and $\mathcal{L}_{\text{CLIP}_\text{v}}$ are the implementation of BiMixCo loss which aligns all the frames of a video y_c and its corresponding fMRI signal x_c using bidirectional contrastive loss and MixCo data augmentation. The MixCo needs to mix two independent fMRI signals. For each x_c, we random sample another fMRI x_{m_c}, which is the keyframe of the clip index by m_c. Then, we mix x_c and x_{m_c} using a linear combination:

$$x_c^* = mix(x_c, x_{m_c}) = \lambda_c \cdot x_c + (1-\lambda_c)x_{m_c}, \quad (5)$$

where x_c^* denotes mixed fMRI signal and λ_c is a hyperparameter sampled from Beta distribution. Then, we adapt the ridge regression to map x_c^* to a lower-dimensional $x_c^{*'}$ and obtain the embedding $e_{x_c^*}$ via the MLP, *i.e.*, $e_{x_c^*} = \mathcal{E}(x_c^{*'})$. Based on this, the BiMixCo loss can be formed as:

$$\begin{aligned}\mathcal{L}_{\text{BiMixCo}} = &-\frac{1}{2F}\sum_{i=1}^{F}\lambda_i \cdot \log\frac{\exp(sim(e_{x_i^*}, e_{y_i})/\tau)}{\sum_{k=1}^{F}\exp\left(sim(e_{x_i^*}, e_{y_k})/\tau\right)}\\ &-\frac{1}{2F}\sum_{i=1}^{F}(1-\lambda_i)\cdot\log\frac{\exp(sim(e_{x_i^*}, e_{y_{m_i}})/\tau)}{\sum_{k=1}^{F}\exp(sim(e_{x_i^*}, e_{y_k})/\tau)}\\ &-\frac{1}{2F}\sum_{j=1}^{F}\lambda_j\cdot\log\frac{\exp(sim(e_{x_j^*}, e_{y_j})/\tau)}{\sum_{k=1}^{F}\exp(sim(e_{x_k^*}, e_{y_j})/\tau)}\\ &-\frac{1}{2F}\sum_{j=1}^{F}\sum_{\{l|m_l=j\}}(1-\lambda_j)\\ &\cdot\log\frac{\exp(sim(e_{x_l^*}, e_{y_j})/\tau)}{\sum_{k=1}^{F}\exp(sim(e_{x_k^*}, e_{y_j})/\tau)},\end{aligned} \quad (6)$$

where $\hat{e}^t \in \mathbb{R}^{BF\times N\times C}$ denotes the OpenCLIP embeddings for video y_c.

We use the Diffusion Prior to transform fMRI embedding e_{x_c} into the reconstructed OpenCLIP embeddings of video e^v. Similar to DALLE·2, Diffusion Prior predicts the target embeddings with mean-squared error (MSE) as the supervised objective:

$$\mathcal{L}_{\text{Prior}} = \mathbb{E}_{e_{y_c}, e_{x_c}, \epsilon\sim\mathcal{N}(0,1)}||\epsilon(e_{x_c}) - e_{y_c}||. \quad (7)$$

B.2. Key Object Segmentation Objective

$$\begin{aligned}\mathcal{L}_{seg}(y^{seg}, \hat{y}^{seg}) = -\frac{1}{B\times F}\sum_{i=1}^{B\times F}&[y_i^{seg}\log(\hat{y}_i^{seg})\\ &+(1-y_i^{seg})\log(1-\hat{y}_i^{seg})]\end{aligned} \quad (8)$$

where $\hat{y}^{seg}$ is the ground truth masks of the key objects.

C. Details of Data Pre-processing

We utilized fMRI data from the cc2017 dataset, preprocessed by [10] using the minimal preprocessing pipeline [6]. The preprocessing steps included artifact removal, motion correction (6 degrees of freedom), registration to standard space (MNI space), and transformation onto cortical surfaces, which were coregistered to a cortical surface template [7]. To identify stimulus-activated voxels, we computed the voxel-wise correlation between fMRI signals for each repetition of the training movie across subjects. The correlation coefficients for each voxel were Fisher z-transformed, and the average z-scores across 18 training movie segments were evaluated using a one-sample t-test. Voxels with significant activation (Bonferroni-corrected, P < 0.05) were selected for further analysis. This process

Scene Description Generation

```
System Prompts: You will assist with image analysis tasks. Your role is to analyze the given image and generate a descriptive caption for the provided image. Please follow the rules:

- The output format should be: "Caption: *****.".
- Cover all primary objects in the given image.
- Focus on the relationship between the objects.
- Strictly keep the caption in one phrase.
- Do not use a celebrities name such as "harry potter" in the caption, use "boy", "girl", "man", "woman" instead.
- Do not include any unnecessary commentary or explanations.
- Do not say anything else from the output format.

For example:
[Input]: Image size of 224 * 224
[Output]: Caption: a boy sitting on a chair holding a cup.
[Input]: Image size of 224 * 224
[Output]: Caption: an airplane flying to an airport.
[Input]: Image size of 224 * 224
[Output]: caption: two girls sitting on a couch.
[Input]: Image size of 224 * 224
[Output]: Caption: a woman with long black hair brushing teeth.
[Input]: Image size of 224 * 224
[Output]: Caption: a close up of a yellow eel in the ocean.
[Input]: Image size of 224 * 224
[Output]: Caption: a fish swimming in the ocean.
```

(a)

Primary Object Detection

```
System Prompts: You will assist with image analysis tasks. Your role is to analyze the given image and identify the primary objects and list their names. Please follow the rules:

- Strictly follow the output format: "Objs: [***, ***, …]".
- Cover all primary objects in the given image.
- Only provide the object names of the primary objects.
- Make sure the object names are in one-word noun format, do not add any descriptive adjectives or descriptive nouns.
- Keep only the object name in its singular form.
- Do not use a persons' name as object name such as "harry potter", use "boy", "girl", "man", "woman" instead.
- Separate each object name with a comma.
- Do not repeat any same word.
- Do not include any unnecessary commentary or explanations.
- Do not say anything else from the output format.

For example:
[Input]: Image size of 224 * 224
[Output]: Objs: [boy, chair, cup]
[Input]: Image size of 224 * 224
[Output]: Objs: [airplane, airport]
[Input]: Image size of 224 * 224
[Output]: Objs: [girl, couch]
[Input]: Image size of 224 * 224
[Output]: Objs: [woman]
```

(b)

Concept Generation

```
System Prompts: You will assist with text-related tasks. I will give you a list of object names in the following form: Objs: [***, ***, ***, …]. Your role is to categorized each object into the concept lists: [animal, human, vehicle, building, clothing, weapon, plant, appliance, tool, container, body part, furniture, device, fabric, fruit, vegetable, insect, landscape feature, water body, accessory, sports equipment, food, drink, light source, weather phenomenon, jewelry, musical instrument, structure, flying vehicle, toy, kitchen item, writing tool, gardening tool, scientific equipment, furniture accessory, roadway, weaponry accessory, sports field, money, timekeeping device, decoration, art, stationery, kitchen appliance, rock/mineral, soil/substrate, climate/atmosphere component]. Please follow the rules:

- Strictly follow the output format: "Concept: [***, ***, …]".
- Only output the concept name in the given list.
- You can categorize the object into more than one concepts.
- If two objects belong to one concept, only list concept once.

For example:
[Input]: Objs: [boy, chair, cup]
[Output]: Concept: [human, appliance, furniture]
[Input]: Objs: [airplane, airport]
[Output]: Objs: [building, flying vehicle]
[Input]: Objs: [girl, couch]
[Output]: Objs: [human, furniture]
[Input]: Objs: [woman]
[Output]: Objs: [human]
```

(c)

Figure a. The overall detailed instruction prompts for scene description generation (a) and key concept generation (b-c).

identified 13,447, 14,828, and 9,114 activated voxels in the visual cortex for the three subjects, respectively. Consistent with prior studies [13, 26, 41], we incorporated a 4-second delay in the BOLD signals to account for hemodynamic response latency when mapping movie stimulus responses.

D. Implementation Details

In this paper, videos from the cc2017 dataset were downsampled from 30FPS to 3FPS to make a fair comparison with the previous methods, and the blurred video was interpolated to 8FPS to generate the final 8FPS video during inference. The training of the Brain Model and Decoupler was performed with 30 and 50 epochs, respectively, and the batch size of training the Brain Model was set to 120, while 10 for the Decoupler. We use the AdamW [22] for optimization, with a learning rate set to 5e-5, to which the OneCircle learning rate schedule [35] was set. Theoretically, our approach can be used in any text-to-video diffusion model, and we choose the open-source available AnimateDiff [12] as our inference model following [10]. The inference is performed with 25 DDIM [36] steps. All experiments were conducted using a single A100 GPU.

E. More Experimental Results

E.1. Frame-based comparison with SOTAs.

We compare the frame-based evaluation metrics in Table a with previous SOTA methods. NEURONS consistently excels in semantic-level frame understanding while maintaining competitive pixel-level performance. As shown in Table 1, our method achieves the highest Semantic-level 2-way accuracy (0.811), outperforming NeuroClips by 0.6% and MinD-Video by 1.9%. For 50-way semantic classification,

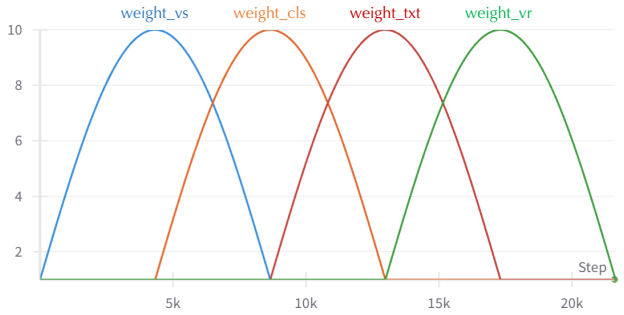

Figure b. Illustration of the progressive learning strategy.

Table a. Quantitative comparison of NEURONS reconstruction performance against other methods (Frame-based). Bold font signifies the best performance, while underlined text indicates the second-best performance. MinD-Video and *NeuroClips* are both results averaged across all three subjects, and the other methods are results from subject 1. Results of baselines are quoted from [10].

Method	Semantic-level		Pixel-level	
	2-way↑	50-way↑	SSIM↑	PSNR↑
Nishimoto [26]	0.727±0.04	-	0.116±0.09	8.012±2.31
Wen [45]	0.758±0.03	0.070±0.01	0.114±0.15	7.646±3.48
Wang [42]	0.713±0.04	-	0.118±0.08	**11.432**±2.42
Kupershmidt [19]	0.764±0.03	0.179±0.02	0.135±0.08	8.761±2.22
MinD-Video [4]	0.796±0.03	0.174±0.03	0.171±0.08	8.662±1.52
NeuroClips [10]	<u>0.806</u>±0.03	<u>0.203</u>±0.01	**0.390**±0.08	9.211±1.46
NEURONS (ours)	**0.811**±0.03	**0.210**±0.01	<u>0.365</u>±0.11	<u>9.527</u>±2.26
subject 1	0.810±0.03	0.206±0.01	0.373±0.14	9.591±2.24
subject 2	0.810±0.03	0.214±0.01	0.353±0.08	9.502±2.40
subject 3	0.817±0.03	0.210±0.01	0.369±0.13	9.488±2.16

NEURONS attains 0.210, a 3.4% improvement over Neuro-

Clips, highlighting its capability to discern fine-grained semantic patterns. In pixel-level metrics, NEURONS achieves the second-best PSNR (9.527), surpassing NeuroClips by 3.4%, while its SSIM (0.365) remains competitive. The observed trade-off between Wang's high PSNR (11.432), and its poor semantic performance (0.713 in 2-way) underscores the challenge of balancing reconstruction fidelity with semantic alignment—a challenge NEURONS addresses effectively through its unified architecture. These results validate that our approach advances the state of the art in frame-based video understanding by harmonizing semantic and low-level feature learning.

E.2. More Qualitative Comparison Results

To further highlight the superior performance of NEURONS, we provide additional qualitative comparisons between our method and the previous SOTA approach, NeuroClips [10], as a supplement to Fig. 5. As we can see in Fig. d, NeuroClips produces many semantic errors (*e.g.*, confusing a "boat" with a "highway road"), whereas NEURONS produces more accurate and visually coherent results.

E.3. Qualitative Analysis of Ablation Study

We present qualitative results from the ablation study, as illustrated in Fig. c. From the top case, the results demonstrate that the Brain Model utilizing only $\mathcal{L}_{rec}$ exhibits limited semantic information. The inclusion of $\mathcal{L}_{seg}$ could capture more motion information. $\mathcal{L}_{cls}$ improves the accuracy of concept recognition, such as distinguishing between humans and vehicles. Further incorporating $\mathcal{L}_{txt}$ enables the model to perceive broader scenes and environments, such as bodies of water. Finally, the combination of progressive learning and the aggregated video reconstruction pipeline ensures that the video output is both semantically and spatially accurate. In the more complex scene from the bottom case, where the main elements (humans, streets) are easier to identify, the reconstructed videos show no significant visual differences. However, we observe that $\mathcal{L}_{seg}$ tends to make the model concentrate on key objects (*e.g.*, only one person generated in the video). In contrast, the loss functions $\mathcal{L}_{cls}$ and $\mathcal{L}_{txt}$ contribute to enriching the richness of concepts and details within the scene. We also provide the corresponding descriptions for the videos. It is evident that the descriptions generated by our trained GPT-2 model are more accurate. For example, in the bottom case, it successfully generates terms like "people" and "busy street" which are consistent with the ground truths.

E.4. Concept Recognition Accuracy.

For one of our decoupled tasks, *i.e.*, concept recognition, we provide the accuracy score for each concept as a supplement to Table. 3. The results are shown in Table. b.

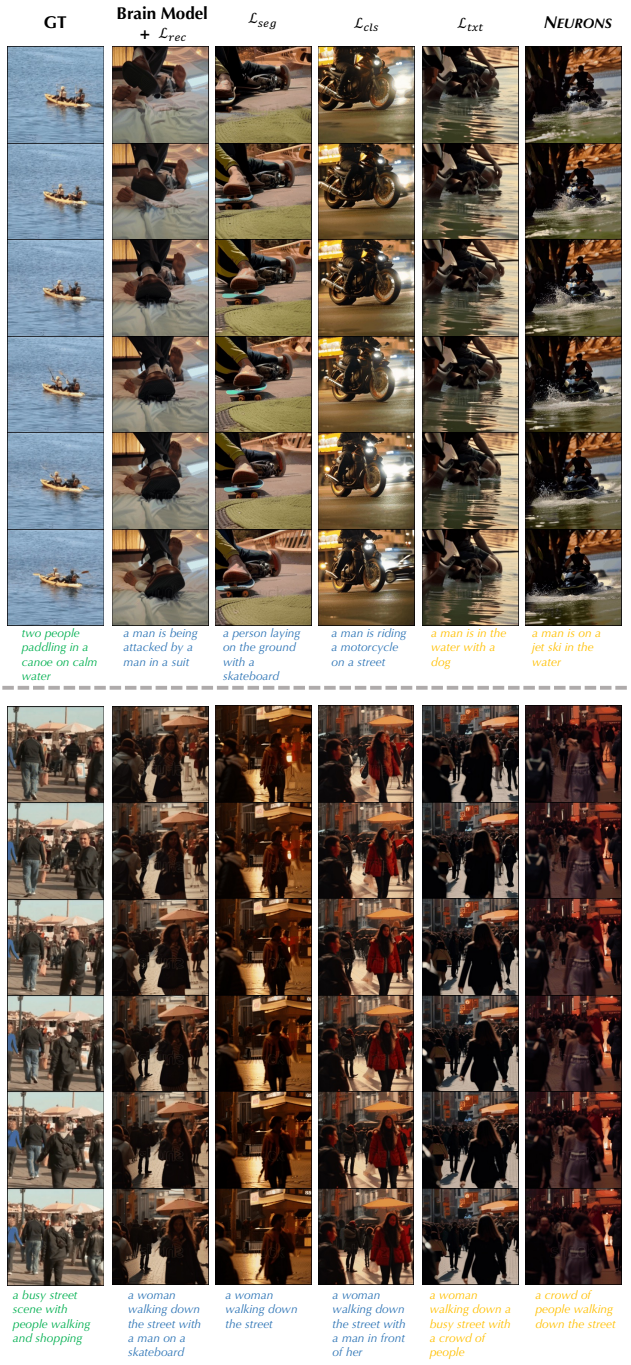

Figure c. We present the generated videos alongside their corresponding text descriptions. Note that the descriptions in the last two columns are generated using our model $\mathcal{D}_{txt}$ (highlighted in yellow), while the other descriptions are produced by captioning the middle frame using BLIP-2.

E.5. Details of Caption and Verb Evaluation

For caption evaluation, we follow traditional metrics for image captioning [3], and report BLEU and CIDEr scores.

BLEU analyzes the co-occurrences of n-grams between the candidate and reference sentences, and CIDEr employs TF-IDF weighting to focus more on semantically informative words that capture image-specific contents.

To better analyze the model's motion understanding capabilities, we specifically evaluate verb accuracy within the generated captions, providing a deeper insight into how effectively the model identifies dynamic actions during video reconstruction. We first extract verbs from both generated and ground truth captions using the part-of-speech (POS) tagging model and then adopt Word2Vec embedding to calculate semantic similarity between the verb pairs. The generated verb is considered correct if the similarity score exceeds the pre-defined threshold of 0.8.

F. Details of Brain Decoding Interpretation

To validate the inspiration behind our NEURONS, which is drawn from the human visual cortex (see Fig.1 in the Introduction section), we employ a visualization tool, *i.e.*, BrainDecodesDeepNets [47], to project the embeddings of each decoupled task onto a brain map. Specifically, we use fMRI data as input and extract four projected embeddings corresponding to the four decoupled tasks of NEURONS. Next, we train a BrainNet [47] to reconstruct the original fMRI image. Finally, following the approach of BrainDecodesDeepNets, we visualize the individual layer weights for each decoupled task (see Fig. 6), further confirming our initial insights. For the training of the brainNet, we utilize the Algonauts 2023 challenge dataset, the same as BrainDecodesDeepNets [47]. We train it for 50 epochs using AdamW optimizer on 1 RTX3090 GPU card. The learning rate is set to 1e-5.

G. Limitations and Future Work

While NEURONS demonstrates strong performance in fMRI-to-video reconstruction, several limitations remain. First, the model is evaluated on a single dataset with limited subject diversity, restricting generalizability. Second, the reconstructed videos, though temporally smooth, remain low in resolution and frame rate due to the coarse temporal granularity of fMRI data. Additionally, current evaluation metrics may not fully capture perceptual realism or narrative coherence. Lastly, although the model shows functional alignment with visual cortex regions, the neurobiological interpretation remains correlational rather than causal.

Table b. Classification accuracy of all the concepts. "-" denotes no such concept in the test set.

Index	Class Name	Accuracy
0	animal	0.450
1	human	0.735
2	vehicle	0.228
3	building	0.070
4	clothing	0.184
5	weapon	-
6	plant	0.196
7	appliance	-
8	tool	0.0625
9	container	0.057
10	body part	0.163
11	furniture	0.155
12	device	0.033
13	fabric	-
14	fruit	0.0
15	vegetable	0.0
16	insect	-
17	landscape feature	0.144
18	water body	0.310
19	organism	0.212
20	fish	0.292
21	reptile	0.0623
22	mammal	-
23	accessory	0.067
24	sports equipment	0.182
25	food	0.6
26	drink	0.0
27	light source	0.0
28	weather phenomenon	0.091
29	jewelry	-
30	musical instrument	-
31	structure	0.209
32	flying vehicle	0.283
33	toy	-
34	kitchen item	0.214
35	writing tool	-
36	gardening tool	-
37	scientific equipment	0.0
38	furniture accessory	0.0
39	roadway	0.147
40	weaponry accessory	-
41	sports field	0.042
42	money	-
43	timekeeping device	-
44	decoration	-
45	art	0.0
46	stationery	0.111
47	kitchen appliance	-
48	rock/mineral	0.0
49	soil/substrate	0.0
50	climate/atmosphere component	0.262

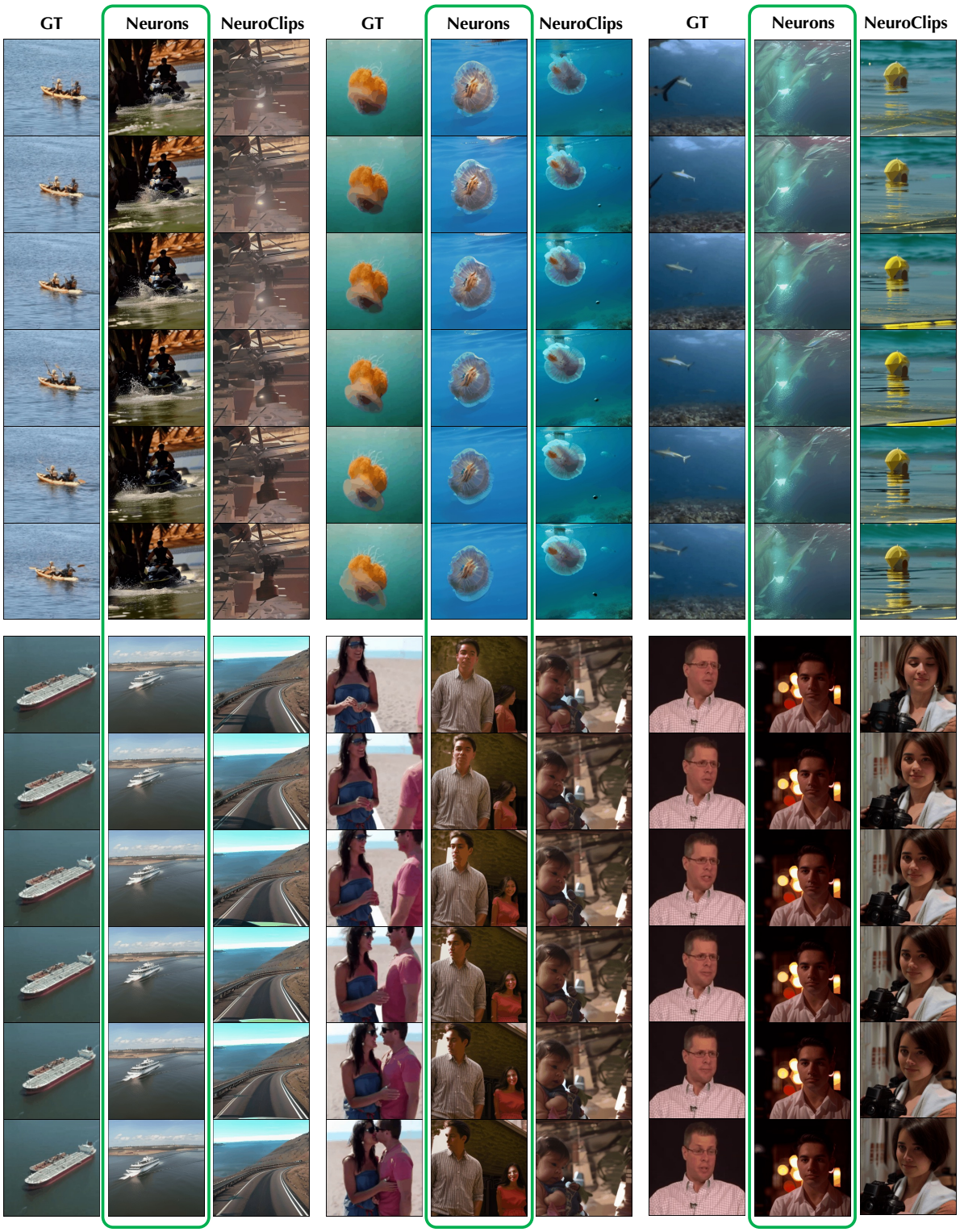

Figure d. More qualitative comparison between NEURONS and the previous SOTA, NeuroClips.

H. Overview of AI Methodologies for Neuroscience Readers

This study leverages several state-of-the-art AI models to decode and reconstruct visual experiences from brain activity. At the core is a neural network-based brain model trained using *contrastive learning*, which aligns fMRI signals with visual and textual embeddings from *CLIP*, a widely used vision-language model. The decoding process is decomposed into four explicit sub-tasks—segmentation, classification, captioning, and reconstruction—each modeled by deep learning modules optimized with task-specific loss functions. For video synthesis, we use *diffusion models*, a class of generative models that produce high-quality video frames conditioned on the outputs of the brain model. These AI components are organized in a biologically inspired, hierarchical manner to simulate the functional specialization of the human visual cortex.