FlatClip: A Geometry-Aware Surface-Level Baseline for fMRI Representation Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent fMRI foundation models differ substantially in the spatial scale at which they represent brain activity. ROI- and connectivity-based models are efficient but coarse, whereas voxel-level models preserve fine-grained spatial structure but require specialized 3D/4D architectures and costly fMRI-specific pretraining. We ask whether part of this performance gap reflects the importance of preserving cortical geometry. Motivated by evidence that macroscale brain activity is strongly constrained by brain geometry, we introduce **FlatClip**, a frozen-encoder surface-level baseline that renders cortical activity as geometry-aware flatmap sequences and reuses a frozen SigLIP2 image encoder with only a lightweight downstream probe. Across resting-state benchmarks, FlatClip serves as a competitive middle-ground representation, outperforming ROI-level baselines on HCP and ADNI tasks while remaining weaker on PPMI and below the strongest voxel-level models overall. On visual-fMRI decoding, performance improves when the input is restricted from whole cortex to visual or NSD-provided task-active cortex, suggesting that geometry is most useful when it is aligned with the prediction target. We further probe this interpretation with geometry controls: disrupting flatmap spatial organization reduces performance, mapping ROI-level signals back into atlas-defined 4D volumes partially recovers performance, and adaptive patching in highly activated regions improves several voxel-based prediction metrics. Together, these results position surface-level flatmap sequences as a practical middle-ground baseline between ROI and voxel models, and suggest that geometry-preserving spatial organization can be useful when designing fMRI models. Code is available.

1 Introduction

Foundation models have recently become an important direction for fMRI representation learning. Instead of training a separate model for each dataset, task, or cognitive paradigm, fMRI foundation models aim to learn reusable brain representations from large-scale unlabeled neuroimaging data. Such models can reduce task-specific engineering and may improve transfer across heterogeneous downstream analyses [1, 2, 3]. However, fMRI representation learning depends strongly on how brain activity is represented. Unlike natural images, fMRI is a high-dimensional 4D spatiotemporal signal with

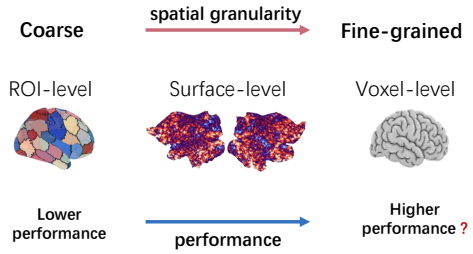

Figure 1: **Motivation.** fMRI representations vary in spatial granularity. FlatClip studies surface-level cortical flatmaps as a geometry-aware middle ground.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

both spatial organization and temporal dynamics. Choosing the right spatial scale is therefore a central modeling decision (Fig. 1).

Most existing fMRI foundation models operate at either ROI level [1, 2, 4, 5, 6] or voxel level [3, 7, 8, 9, 10, 11]. ROI- and connectome-based models summarize voxel signals within atlas-defined regions and then model regional time series or inter-regional relationships. These representations are efficient and scalable, but they are coarse: they depend on parcellation choices and may discard fine-grained cortical geometry [12, 13]. Voxel-based or volumetric models preserve more spatial detail and can capture local structure more directly, but they require specialized 3D/4D architectures, substantial computation, and storage of high-dimensional fMRI volumes. Thus, ROI-level models are computationally attractive but may lose geometry, whereas voxel-level models preserve spatial detail but are costly. This trade-off raises a basic question: whether preserving cortical geometry contributes to the advantage of fine-grained fMRI representations? Pang et al. showed that macroscopic fMRI activity can be parsimoniously explained by geometric eigenmodes derived from brain shape, and that these modes explain spontaneous and task-evoked activity more effectively than connectome-derived modes [14]. This motivates representations that retain the spatial organization of cortical activity rather than reducing all signals to coarse regional averages.

We study surface-level representations as an intermediate point between ROI-level and voxel-level fMRI modeling. Specifically, we render cortical fMRI activity as flatmap sequences: each volumetric fMRI frame is projected onto the cortical surface and represented as a 2D cortical map, while temporal variation is retained across frames. This yields a geometry-aware 2D+time representation. Compared to ROI aggregation, surface flatmaps preserve more of the cortical spatial layout and are better suited for image pre-training models. We introduce **FlatClip**, a frozen-encoder surface-level baseline for fMRI representation learning. FlatClip renders frame-wise cortical activity as flatmap images, extracts frozen features from a SigLIP2 image encoder, and trains only a lightweight downstream probe. The image encoder is not pretrained on fMRI and is not fine-tuned on neuroimaging labels. This design makes FlatClip an out-of-domain reference baseline: it is not affected by the scale, composition, or preprocessing choices of fMRI pretraining corpora, and therefore provides a complementary way to assess how much information can be recovered from cortical geometry without fMRI-specific encoder pretraining.

We first evaluate FlatClip on four resting-state fMRI benchmarks, where subject-level prediction has no explicit stimulus region. Under a frozen-backbone, lightweight-probe protocol, FlatClip exceeds most ROI-level baselines while requiring no fMRI-specific encoder pretraining, although the strongest voxel-level models remain higher overall. We then evaluate visual-fMRI decoding on NSD COCO80, a stimulus-evoked setting in which the relevant cortical region is more explicit; NSD is a large-scale 7T fMRI dataset collected while subjects viewed natural images [15]. Performance improves from whole-cortex input to HCP-MMP visual cortex and NSD-provided subject-level task-active regions, suggesting that fMRI representations should preserve not only cortical geometry, but task-relevant cortical geometry.

Motivated by the resting-state and visual-fMRI observations, we conduct control experiments to test whether the gains are associated with spatial organization and spatial granularity. Geometry-disruption controls show that block-shuffling, vertex permutation, and FC heatmaps reduce performance relative to real flatmaps, indicating that the benefit is not merely due to image-like formatting. Geometry-recovery controls show that mapping ROI-level signals back into atlas-defined 4D volumes improves over ROI-level representations but remains below native voxel inputs, suggesting that coarse spatial layout is useful but cannot fully replace fine-grained voxel geometry. In the 4D voxel-based HCP setting, refining patches in highly activated regions further improves performance, suggesting that finer spatial granularity is most beneficial when allocated to signal-rich regions. Together, these results suggest that preserving cortical spatial organization is a useful design principle for fMRI representation learning, especially when spatial detail is retained in task- or signal-relevant regions. We interpret FlatClip as a practical surface-level reference baseline rather than a fully controlled causal isolation of cortical geometry. Our main findings are summarized as follows:

- We introduce **FlatClip**, a frozen-encoder surface-level fMRI representation baseline that uses geometry-aware flatmap sequences, a frozen SigLIP2 image encoder, and a lightweight downstream probe.
- We show that surface-level flatmaps provide a practical middle ground between ROI-level and voxel-level representations.

- We provide control evidence that spatial organization and spatial granularity matter: disrupting flatmap geometry degrades performance, restoring ROI signals into atlas-defined 4D volumes partially recovers performance, and refining highly activated regions improves 4D voxel-based prediction.

2 Related Work

fMRI foundation models across representation scales. Recent fMRI foundation models aim to reduce task-specific modeling and improve transfer across heterogeneous neuroimaging datasets. A central design choice is the spatial scale at which fMRI is represented. ROI- or atlas-based methods reduce voxel-level signals into region-wise time series or functional-connectivity graphs, and models such as BrainLM, Brain-JEPA, BrainMASS, LCM, and related graph-based approaches learn from ROI tokens, temporal dynamics, or graph-structured connectivity [1, 2, 4, 5, 6, 16]. These representations are efficient and scalable, but they depend on the chosen parcellation and primarily describe interactions among predefined brain regions or networks. At the other end of the scale, voxel- or volume-based models operate directly on 3D or 4D fMRI and preserve more fine-grained spatial structure [7, 3, 8, 9, 10, 11]. These approaches avoid coarse ROI averaging, but they require specialized 3D or 4D architectures and can be computationally expensive. FlatClip studies a surface-level representation between these two regimes: it does not compress fMRI entirely into ROI-level network features, nor does it train a new voxel-level fMRI foundation model.

Geometry-aware cortical surface and flatmap representations. Recent evidence suggests that macroscale brain activity is strongly constrained by brain geometry, with geometric modes explaining spontaneous and task-evoked fMRI more parsimoniously than connectome-derived modes [14, 17]. This motivates fMRI representations that preserve cortical spatial layout rather than reducing all signals to coarse regional averages. Surface-based flatmaps provide a practical way to retain this geometry-aware cortical layout while producing image-compatible inputs. Pycortex is a standard toolbox for projecting volumetric neuroimaging data onto cortical surfaces and generating flattened cortical visualizations [18]. Recent work has explored image-like or flatmap-based interfaces for fMRI, including visual decoding from fMRI signals and spatiotemporal masked-autoencoding on fMRI flatmap videos [19, 20]. In contrast, our goal is not to introduce a new projection method, reconstruct visual stimuli, or pretrain an fMRI-specific flatmap model. Instead, we use standard surface-based flatmaps as a geometry-preserving middle-ground representation between ROI-level summaries and voxel-level volumes, evaluated through a frozen out-of-domain image encoder.

Image-pretrained encoders as out-of-domain frozen baselines. Large-scale image-pretrained models have become reusable feature extractors for downstream vision tasks. Representative approaches include masked image modeling, contrastive image-text pretraining, self-distillation, and sigmoid-loss image–text pretraining, including MAE, CLIP, DINOv2, and SigLIP2 [21, 22, 23, 24]. SigLIP2 extends the SigLIP family with captioning-based pretraining, self-supervised losses, and online data curation, improving visual representation transfer and dense-feature performance [24]. FlatClip uses this frozen-feature reuse paradigm as an implementation strategy rather than as the main scientific claim. Given a cortical flatmap sequence, each frame can be processed by a frozen SigLIP2 image encoder, and only a lightweight downstream probe is trained. Because this encoder is not pretrained on fMRI, FlatClip provides an out-of-domain reference for assessing how much predictive information can be recovered from surface-level cortical geometry without relying on fMRI-specific pretraining corpora.

3 Method

3.1 Overview

FlatClip is designed as a frozen-encoder surface-level baseline for fMRI representation learning. Instead of pretraining a new fMRI encoder on ROI time series or native 4D volumes, FlatClip renders cortical activity as geometry-aware flatmap sequences and evaluates them with a frozen image-pretrained encoder. In this work, we use a frozen SigLIP2 encoder as the default visual backbone and train only a lightweight downstream probe. Because the encoder is pretrained outside the fMRI domain and is kept frozen, FlatClip serves as an out-of-domain reference for assessing how much

predictive information can be recovered from cortical geometry without relying on fMRI-specific pretraining data.

For resting-state fMRI, each subject is represented as a fixed-length sequence of cortical flatmap frames. Each frame is encoded into one global image representation, and the frame-level features are temporally pooled into a subject-level feature. For visual-task fMRI, each sample corresponds to a stimulus-level GLM response map, which is rendered as a flatmap and encoded into one stimulus-level feature. The same lightweight-probe protocol is then used for downstream prediction.

3.2 Geometry-aware Flatmap Sequence Adapter

For each subject i, we construct a sequence of cortical flatmap frames

$$\mathcal{F}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,T}\},$$

where each frame

$$x_{i,t} \in \mathbb{R}^{H \times W \times C}$$

is an image-compatible rendering of cortical fMRI activity. By stacking frames along the temporal dimension, the sequence can be represented as a flatmap tensor

$$\mathbf{X}_i = \mathrm{Stack}_{t=1}^{T}(x_{i,t}) \in \mathbb{R}^{T \times H \times W \times C}.$$

Thus, native volumetric fMRI time series are converted into surface-level flatmap sequences while temporal variation is retained across frames.

For resting-state fMRI, t indexes time points and the sequence length T is fixed within each dataset. For visual-task fMRI, each sample corresponds to a stimulus-level response map estimated by a general linear model (GLM), where each response map is associated with one viewed image.

For NSD visual-fMRI decoding, we evaluate three cortical input regions: whole cortex, HCP-MMP visual cortex, and the NSD-provided subject-level task-active region, denoted as NSDgeneral. The HCP-MMP visual cortex mask is defined from visual ROIs in the HCP-MMP atlas in 32k fsLR space [25]. The NSDgeneral mask is provided by NSD and represents subject-level task-active cortex [26]. These regions are rendered as flatmaps and resized to the input resolution used by the visual encoder.

We render cortical activity as flatmaps using Pycortex [18]. Flatmap rendering is used only as a deterministic surface-level adapter; representation extraction is performed by the frozen SigLIP2 encoder and prediction is performed by the lightweight downstream probe.

3.3 Frozen SigLIP2 Encoder and Feature Construction

Let f_θ denote the frozen SigLIP2 image encoder. For each flatmap frame, we extract one global image representation:

$$z_{i,t} = f_\theta^{\mathrm{global}}(x_{i,t}) \in \mathbb{R}^d,$$

where $f_\theta^{\mathrm{global}}(\cdot)$ denotes the global representation returned by the encoder. The encoder parameters θ are fixed throughout all experiments, and no gradient is propagated into the image encoder.

For resting-state fMRI, each subject has T flatmap frames. We aggregate the frame-level global representations by average temporal pooling:

$$h_i = \frac{1}{T} \sum_{t=1}^{T} z_{i,t} \in \mathbb{R}^d.$$

In our main resting-state experiments, we use $T = 40$ frames for each subject and $d = 768$. Thus, each subject is represented by one pooled 768-dimensional feature derived from 40 frame-level global representations.

For visual-task fMRI, each sample corresponds to one stimulus-level GLM response map rather than a time sequence. For subject s and stimulus image j, the response flatmap is encoded as

$$h_{s,j} = f_\theta^{\mathrm{global}}(x_{s,j}) \in \mathbb{R}^d.$$

Therefore, both resting-state and visual-task experiments use the same global-representation extraction rule; the difference is that resting-state prediction pools features over time, whereas visual-task prediction uses one stimulus-level feature per GLM response map.

3.4 Downstream Prediction

For each downstream task, we train a lightweight probe g_ϕ on top of the frozen SigLIP2 feature. Only the probe parameters ϕ are optimized.

For single-label resting-state classification, let

$$\mathcal{D}_{\text{rest}} = \{(h_i, y_i)\}_{i=1}^{N_{\text{sub}}}$$

denote the subject-level training set. The probe predicts

$$\hat{y}_i = g_\phi(h_i),$$

and is trained with cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N_{\text{sub}}} \sum_{i=1}^{N_{\text{sub}}} \sum_{c=1}^{C} \mathbf{1}[y_i = c] \log p_\phi(y_i = c \mid h_i),$$

where C is the number of classes and N_{sub} is the number of training subjects.

For visual-fMRI recognition, let $\mathcal{D}_{\text{vis}}$ denote the set of subject–stimulus pairs (s, j), and let

$$N_{\text{vis}} = |\mathcal{D}_{\text{vis}}|.$$

Each pair has a feature $h_{s,j}$ and a COCO80 multi-hot target

$$y_j \in \{0, 1\}^C, \quad C = 80,$$

where y_j depends only on the stimulus image and is shared across subjects. The lightweight probe outputs

$$o_{s,j} = g_\phi(h_{s,j}) \in \mathbb{R}^C,$$

and is trained with class-weighted binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N_{\text{vis}}} \sum_{(s,j)\in\mathcal{D}_{\text{vis}}} \sum_{c=1}^{C} \left[w_c y_{j,c} \log \sigma(o_{s,j,c}) + (1 - y_{j,c}) \log(1 - \sigma(o_{s,j,c})) \right],$$

where $\sigma(\cdot)$ is the sigmoid function and w_c is the positive-class weight computed from the training split.

For all FlatClip variants, the SigLIP2 encoder is frozen and only the lightweight probe is trained. This protocol controls downstream readout capacity while allowing us to evaluate the utility of surface-level flatmap representations.

4 Evaluating Geometry-Aware Surface Representations

4.1 Resting-state fMRI Benchmarks

We first evaluate FlatClip on four subject-level resting-state fMRI benchmarks: HCP sex classification, PPMI diagnosis prediction, ADNI MCI vs. CN classification, and ADNI AD vs. CN classification [27, 28, 29]. These benchmarks cover demographic prediction in healthy adults and diagnostic-label prediction across neurodegenerative cohorts. All methods are evaluated on the same subject-level splits with frozen representations and lightweight downstream heads. Because the models have different native input requirements, the input construction follows each model's interface, while the readout protocol and splits are matched. This provides a matched readout protocol for comparing frozen representations. Voxel-level models were evaluated using 40 input frames, following their model-specific requirements; FlatClip used the same temporal window for consistency. In contrast, ROI/FC-based models used approximately 200 frames for feature computation, according to the requirements of their respective pipelines. We report Accuracy and weighted F1 against general-purpose fMRI foundation-model baselines, including ROI-level models (Brain-LM [2], Brain-MASS [4], LCM [6], Brain-Harmony [30]) and voxel-level models (SwiFT [7], NeuroSTORM [9], Omni-fMRI [3]). Details about datasets, task design, ablation of time frame, and model-specific settings are provided in Appendix A.

Table 1: Performance on HCP, ADNI (MCI), ADNI (AD), and PPMI with Accuracy/weighted F1. Darker cell color indicates stronger performance after column-wise min–max normalization.

Model \ Dataset	HCP *Sex Classif.*		ADNI (MCI) *Diagnosis*		ADNI (AD) *Diagnosis*		PPMI *PD Diagnosis*	
	ACC ↑	**wF1 ↑**	**ACC ↑**	**wF1 ↑**	**ACC ↑**	**wF1 ↑**	**ACC ↑**	**wF1 ↑**
Brain-LM	64.12±2.89	64.14±2.84	45.86±5.18	45.78±5.12	66.09±2.22	66.19±2.11	62.50±4.89	57.15±6.62
BrainMASS	68.24±1.29	68.18±1.36	53.79±5.27	53.89±5.18	60.87±5.32	60.91±5.44	52.50±3.22	52.95±2.89
LCM	67.94±3.23	67.93±3.26	53.10±3.99	50.89±4.39	64.35±3.53	64.05±3.82	58.89±3.75	49.65±4.92
Brain Harmony	67.84±2.56	67.73±2.57	58.33±3.86	58.28±4.01	70.90±5.28	71.04±5.35	60.55±5.17	57.51±4.98
FlatClip	82.06±2.05	82.04±2.05	61.11±1.48	60.75±1.16	75.46±2.12	75.41±2.11	54.86±1.96	55.72±1.63
SwiFT	86.03±1.39	84.91±1.76	69.78±4.63	69.31±5.23	74.69±2.86	74.70±2.63	61.55±2.52	58.23±4.20
NeuroSTORM	85.72±1.44	84.36±2.00	57.25±3.18	56.76±3.21	71.64±0.66	65.35±2.06	66.78±0.51	57.37±0.86
Omni-fMRI	92.86±0.63	92.08±0.70	63.40±2.55	56.67±9.60	84.26±0.87	83.65±1.37	68.13±0.60	63.65±2.04

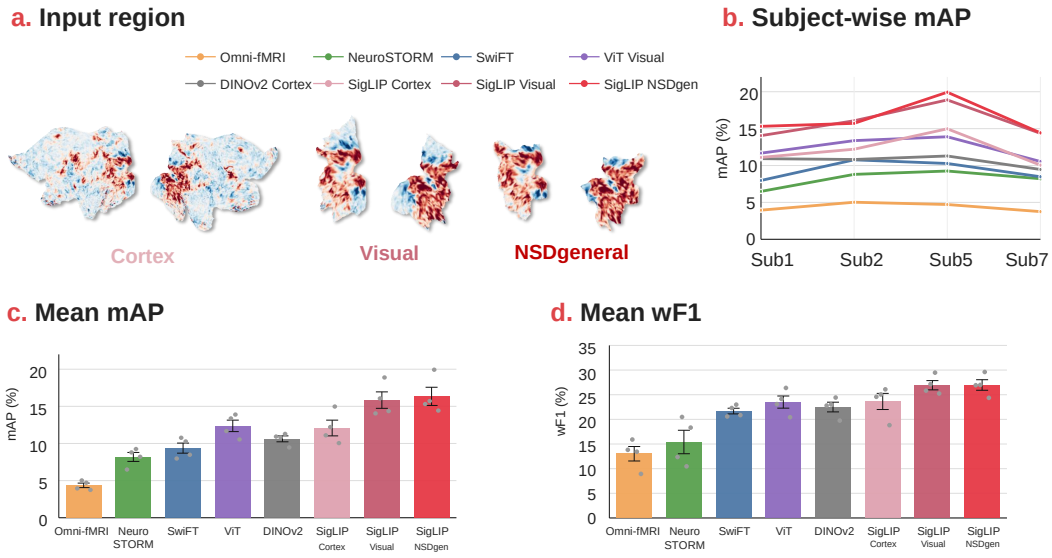

Figure 2: Visual-fMRI COCO80 multi-label recognition on NSD. (a) Illustration of the cortical input regions used by FlatClip: whole cortex, HCP-MMP visual cortex, and the NSD-provided subject-level task-active region (NSDgeneral). (b) Subject-wise mAP across four NSD subjects for Omni-fMRI, NeuroSTORM, SwiFT and FlatClip variants. (c,d) Mean mAP and weighted F1 across subjects. All FlatClip variants use one global representation per GLM response map and the same downstream MLP classifier. Dots denote individual subjects, and error bars denote SEM across subjects. Stimulus-level alignment statistics are reported in Appendix B.2.

As shown in Table 1, FlatClip achieves competitive performance on HCP and ADNI without fMRI-specific encoder pretraining or end-to-end fine-tuning. It outperforms most ROI-level baselines while remaining below the strongest voxel-level models, consistent with its role as a surface-level middle-ground representation. This pattern suggests that cortical flatmaps recover part of the information lost by ROI aggregation, while native voxel inputs still retain additional fine-grained spatial detail. FlatClip is relatively weaker on PPMI, which may reflect a limitation of the current cortical-only input. Parkinson's disease involves subcortical and brainstem systems that are not fully captured by cortical surface flatmaps [31, 32].

4.2 Visual-fMRI COCO80 Multi-label Recognition

We next evaluate FlatClip on stimulus-level visual-fMRI decoding using NSD COCO80 multi-label recognition. Each sample corresponds to a GLM-estimated cortical response map evoked by one NSD stimulus image, and the target is an 80-dimensional COCO multi-hot label vector. Unlike resting-state

benchmarks, this setting is stimulus-evoked and has a more explicit task-relevant cortical substrate. It therefore allows us to test whether preserving cortical geometry is most useful when the retained geometry is relevant to the prediction target. We report within-subject decoding using mAP and weighted F1, and compare FlatClip with general-purpose fMRI foundation-model baselines [7, 3, 9].

For FlatClip variants, each GLM response map is encoded by the frozen SigLIP2 encoder into one global representation, followed by the same lightweight probe for COCO80 prediction. We vary the cortical input region across whole cortex, HCP-MMP visual cortex, and the NSD-provided subject-level task-active region, denoted NSDgeneral. The NSDgeneral setting is treated as a task-informed input-region setting. For voxel-level baselines on NSD, we follow each model's closest available visual-stimulus evaluation protocol. NeuroSTORM includes an NSD visual-stimulus evaluation, and Omni-fMRI provides an NSD-style protocol using repeated GLM response maps. For SwiFT, which supports variable-length inputs, each GLM response map is treated as a single-volume sequence.

As shown in Fig. 2, FlatClip variants achieve stronger COCO80 recognition performance than the evaluated general-purpose fMRI foundation-model baselines. More importantly, performance improves as the input region becomes more specific to the visual task: whole-cortex input is weaker, HCP-MMP visual cortex performs better, and the NSDgeneral subject-level task-active region achieves the best mean mAP. For weighted F1, both visual-cortex and NSDgeneral inputs improve over whole-cortex input, while their difference is smaller. This pattern suggests that visual-fMRI decoding benefits not only from preserving cortical geometry, but also from selecting geometry that is relevant to the evoked visual task. Additional image–fMRI feature-alignment analyses in Appendix B.2 further show that SigLIP2-NSDgeneral features are more closely aligned with the corresponding image-feature geometry than both fMRI foundation-model baselines and broader FlatClip region choices. More details are provided in Appendix B.

4.3 Geometry Controls and Representation Scale

We next examine whether the benefit of FlatClip is associated with preserving cortical geometry rather than merely converting fMRI into an image-like input. We evaluate two complementary directions: geometry reduction and geometry recovery. Detailed control construction is provided in Appendix C.

Geometry reduction. We perturb cortical flatmaps while keeping the frozen SigLIP2 encoder and MLP probe unchanged. Spatial block-shuffling permutes image blocks, vertex permutation randomly reassigns cortical activation values, and the three stronger controls further test whether the effect can be explained by low-level image statistics: left-right hemisphere swap preserves most pixel statistics but disrupts laterality, within-hemisphere region shuffle disrupts cortical layout within each hemisphere, and phase/frequency-matched randomization preserves the Fourier amplitude spectrum and intensity histogram while randomizing spatial phase. We also include FC heatmaps, random images, and a randomly initialized encoder as additional controls.

As shown in Table 2, real flatmaps achieve the best HCP sex-classification performance. The performance drops under spatial perturbations, especially within-hemisphere shuffling and phase/frequency-matched randomization, indicating that FlatClip benefits from meaningful cortical geometry rather than only rendering-induced image statistics. FC heatmaps and random images perform much worse, and the randomly initialized encoder also degrades performance, supporting the importance of both geometry-preserving flatmap inputs and pretrained visual representations. DINOv2 controls show the same trend (Appendix C).

Geometry recovery. We then ask the complementary question: if coarse ROI signals are placed back into a geometry-aware 4D structure, can performance be partially recovered? To test this, we compare ROI-, 4D ROI-, and voxel-based representations (Table 3). For a controlled comparison between 4D ROI and voxel inputs, both are evaluated with the same ViT-Small backbone and an input budget of approximately 1000 patches; only the spatial granularity of the input signal is changed. The 4D ROI input is constructed by mapping ROI-level signals back into an atlas-defined 4D volume, where all voxels within the same ROI share the ROI-mean time series [33, 34]. Thus, 4D ROI preserves the 4D input format and coarse atlas geometry, but removes fine-grained within-ROI spatial variation.

Compared with ROI-level time-series representations, 4D ROI inputs achieve substantially higher performance, suggesting that restoring spatial layout and 4D structure is beneficial. Native voxel

Table 2: **Geometry-control ablation on HCP sex classification.** All results use the frozen SigLIP2 encoder with one global representation per frame, followed by the same MLP probe. Results are reported as mean±std over three seeds. Significance markers indicate paired t-tests comparing Real flatmap against each control after Holm correction. **Red** indicates the best performance.

Control input	ACC ↑	wF1 ↑	Balanced ACC ↑	Macro F1 ↑
Real flatmap	**82.06 ± 2.05**	**82.04 ± 2.05**	**81.81 ± 2.03**	**81.88 ± 2.06**
Spatial block-shuffle	$76.14 \pm 1.83^{*}$	$76.18 \pm 1.83^{*}$	$76.21 \pm 1.85^{*}$	$76.06 \pm 1.84^{*}$
Left-right hemisphere swap	$74.96 \pm 1.17^{*}$	$74.89 \pm 1.23^{*}$	$74.56 \pm 1.31^{*}$	$74.64 \pm 1.28^{*}$
Within-hemisphere region shuffle	$70.90 \pm 1.63^{*}$	$70.87 \pm 1.62^{*}$	$70.62 \pm 1.61^{*}$	$70.64 \pm 1.62^{*}$
Vertex-permutation	$72.93 \pm 1.92^{*}$	$72.87 \pm 1.97^{*}$	$72.58 \pm 2.05^{*}$	$72.62 \pm 2.02^{*}$
Real flatmap with random initialized	$70.56 \pm 1.34^{*}$	$70.56 \pm 1.36^{*}$	$70.37 \pm 1.41^{*}$	$70.35 \pm 1.38^{*}$
Phase/frequency-matched randomization	$64.64 \pm 2.29^{*}$	$64.49 \pm 2.14^{*}$	$64.12 \pm 2.03^{*}$	$64.14 \pm 2.07^{*}$
FC heatmap	$57.87 \pm 2.69^{*}$	$57.93 \pm 2.68^{*}$	$57.98 \pm 2.52^{*}$	$57.80 \pm 2.62^{*}$
Random image	$48.90 \pm 3.26^{*}$	$48.89 \pm 3.42^{*}$	$48.69 \pm 3.69^{*}$	$48.59 \pm 3.65^{*}$

Markers denote Holm-corrected paired tests against Real flatmap within each metric: $^{*}p < 0.05$.

Table 3: **Comparison of ROI-, 4D ROI-, and voxel-based representations.** 4D ROI denotes atlas-defined 4D inputs where all voxels within each ROI are assigned the ROI-mean time series. The 4D ROI and voxel models use the same ViT-Small backbone with approximately 1000 input patches.

Representation	Resolution	HCP Sex		AD Diagnosis	
		ACC ↑	wF1 ↑	ACC ↑	wF1 ↑
ROI time series	450 ROIs	65.68±4.68	65.51±4.47	60.65±5.61	60.75±3.90
4D ROI volume	100 ROIs	78.75±4.09	78.72±4.20	70.69±5.11	69.89±7.90
4D ROI volume	450 ROIs	80.47±6.44	80.42±6.46	69.10±4.49	68.15±4.20
Voxel model	Native 4D volume	80.74±2.41	80.32±3.21	72.22±10.2	**71.66±9.54**
Voxel model	Adaptive patches	**82.81±0.57**	**82.88±0.55**	**72.99±1.03**	70.76±3.16

inputs are comparable to or slightly stronger than 4D ROI inputs in several metrics, and adaptive patching further improves performance, indicating that fine-grained voxel-level geometry may provide additional information beyond coarse ROI-mean structure. However, under the current three-fold evaluation, the difference between 4D ROI and native voxel inputs is not statistically significant. We further test an adaptive patch strategy in the 4D voxel setting, where highly activated regions are represented with finer patches [3]. This strategy further improves performance, consistent with the visual-fMRI finding that task- or signal-relevant regions benefit from finer spatial granularity.

Together, the geometry-reduction and geometry-recovery controls suggest that fMRI representations benefit from preserving spatial geometry at an appropriate granularity. They also support the view that surface flatmaps provide a middle ground between coarse ROI representations and full voxel-level models.

4.4 Implementation Ablations

We perform implementation ablations to identify stable design choices for FlatClip. For resting-state benchmarks, Table 4 evaluates three factors: frozen image backbone, feature representation, and flatmap-value normalization. For the backbone comparison, all methods use the same flatmap input, one global representation, and the same lightweight probe. This avoids giving any backbone an advantage from higher-dimensional token concatenation and focuses the comparison on the frozen encoder. Under this controlled setting, SigLIP2 achieves the strongest overall performance among the evaluated image backbones and is therefore used as the default FlatClip encoder. We also examine whether using more token information improves performance. Aggregating patch-level or frame-level token information can improve some tasks, but the one-global-representation setting remains compact, stable, and directly comparable with foundation-model baselines. We therefore use one global representation for the main comparison. For visual-fMRI COCO80, we use the same principle of controlled comparison: all FlatClip variants use one global representation and the same lightweight probe. Additional visual-fMRI ablations are provided in Appendix B.

Table 4: Implementation ablations for frozen image backbone, feature representation, and flatmap normalization. Results are reported as Accuracy / weighted F1. (%)

Setting \ Dataset	HCP *Sex Classif.*		ADNI (MCI) *Diagnosis*		ADNI (AD) *Diagnosis*		PPMI *PD Diagnosis*	
	ACC ↑	**wF1** ↑	**ACC** ↑	**wF1** ↑	**ACC** ↑	**wF1** ↑	**ACC** ↑	**wF1** ↑
DINOv2	80.37±2.55	80.33±2.53	60.68±2.67	60.44±1.98	75.46±4.01	75.20±3.69	51.39±3.18	51.58±2.25
ViT-B	80.10±1.71	80.12±1.68	54.48±2.80	54.82±3.18	76.23±4.06	**76.05±3.88**	56.67±4.51	54.03±5.19
SigLIP2	**82.06±2.05**	**82.04±2.05**	**61.11±1.48**	**60.75±1.16**	**75.46±2.12**	75.41±2.11	**54.86±1.96**	**55.72±1.63**
40 Patch-Mean	**84.94±1.06**	**84.91±1.06**	58.12±1.96	57.82±1.87	**76.39±4.17**	**76.36±3.93**	**56.94±4.21**	**56.48±4.27**
40 Global tokens	83.08±0.59	83.02±0.66	60.68±1.96	60.02±2.07	74.07±4.88	74.13±4.98	53.47±7.54	53.93±6.14
One Global token	82.06±2.05	82.04±2.05	**61.11±1.48**	**60.75±1.16**	75.46±2.12	75.41±2.11	54.86±1.96	55.72±1.63
Frame Z-Score	**82.23±1.34**	**82.23±1.28**	58.12±3.23	58.28±3.27	73.15±0.80	73.52±0.91	49.65±3.35	50.68±3.10
Voxel Z-Score	74.49±2.20	74.52±2.22	58.55±0.60	56.22±1.28	**80.56±1.96**	**79.68±2.33**	50.35±2.60	51.17±0.76
Global Z-Score	82.06±2.05	82.04±2.05	**61.11±1.48**	**60.75±1.16**	75.46±2.12	75.41±2.11	**54.86±1.96**	**55.72±1.63**

5 Discussion

This work studies fMRI representation learning through spatial scale and task-relevant geometry. In resting-state benchmarks, FlatClip acts as a surface-level middle ground: it outperforms most ROI/FC-level baselines under a frozen-backbone, lightweight-probe protocol, while voxel-level models remain strong upper comparisons. This pattern suggests that cortical flatmaps recover part of the information lost by ROI aggregation, but that native voxel inputs still preserve additional fine-grained spatial detail. Thus, FlatClip should be interpreted as a geometry-aware surface-level reference baseline rather than a replacement for voxel-level fMRI foundation models.

The visual-fMRI results extend this observation by showing that the useful geometry should also be task-relevant. In NSD COCO80 decoding, performance improves when the input is restricted from whole cortex to HCP-MMP visual cortex and NSD-provided task-active cortex. This suggests that preserving all cortical regions indiscriminately is not always optimal; for stimulus-evoked decoding, task-responsive regions provide a cleaner signal than broad cortex-wide inputs. Together, the resting-state and visual-fMRI results indicate that fMRI representations benefit from preserving cortical geometry, especially when the preserved geometry is aligned with the prediction target.

The control experiments support this interpretation. Disrupting flatmap geometry through block-shuffling, vertex permutation, or FC heatmap conversion reduces performance, indicating that the gain is not explained by image-like formatting alone. Conversely, mapping ROI-level signals back into atlas-defined 4D volumes improves over ROI time-series representations, showing that coarse spatial layout and 4D structure are useful. Adaptive patching in highly activated regions further improves voxel-based prediction, suggesting that finer spatial granularity is most useful when allocated to signal-rich regions. These controls support the role of spatial organization and spatial granularity, while not claiming to fully isolate cortical topology from all other image statistics.

The role of the frozen SigLIP2 encoder should be understood in this context. FlatClip does not assume that fMRI is equivalent to natural images. Instead, SigLIP2 provides an out-of-domain frozen feature extractor for evaluating image-compatible surface representations without fMRI-specific encoder pretraining. This makes FlatClip a useful reference baseline for assessing how much predictive information can be recovered from surface-level cortical geometry with only a lightweight downstream probe.

Limitations. Cortical flatmaps are not lossless representations of 4D fMRI. They preserve surface-level cortical layout, but may introduce flattening distortions, discard volumetric relationships, and omit subcortical or brainstem structures that can be important for some tasks. FlatClip also uses simple average pooling over frame-level representations, so temporal dynamics are only coarsely summarized. Finally, although our controls reduce several image-format and low-level-statistic confounds, they do not fully isolate cortical geometry from effects of rendering, cropping, encoder inductive bias, or adaptive patch tokenization. Future work should extend FlatClip to subcortical and whole-brain geometry-aware representations with stronger temporal modeling.

References

[1] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

[2] Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pages 2023–09, 2023.

[3] Mo Wang, Wenhao Ye, Junfeng Xia, Junxiang Zhang, Xuanye Pan, Minghao Xu, Haotian Deng, Hongkai Wen, and Quanying Liu. Omni-fmri: A universal atlas-free fmri foundation model. *arXiv preprint arXiv:2601.23090*, 2026.

[4] Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 2024.

[5] Xiangmin Han, Rundong Xue, Jingxi Feng, Yifan Feng, Shaoyi Du, Jun Shi, and Yue Gao. Hypergraph foundation model for brain disease diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[6] Ziquan Wei, Tingting Dan, and Guorong Wu. Large connectome model: An fmri foundation model of brain connectomes empowered by brain-environment interaction in multitask learning landscape. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2155–2163, 2026.

[7] Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. *Advances in Neural Information Processing Systems*, 36:42015–42037, 2023.

[8] Mo Wang, Junfeng Xia, Wenhao Ye, Enyu Liu, Kaining Peng, Jianfeng Feng, Quanying Liu, and Hongkai Wen. Slim-brain: A data-and training-efficient foundation model for fmri data analysis. *arXiv preprint arXiv:2512.21881*, 2025.

[9] Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Danie Barron, Quanzheng Li, Randy Hirschtick, Byung-Hoon Kim, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for fmri analysis. *ArXiv*, abs/2506.11167, 2025.

[10] Yifei Sun, Daniel Chahine, Qinghao Wen, Tianming Liu, Xiang Li, Yixuan Yuan, Fernando Calamante, and Jinglei Lv. Voxel-level brain states prediction using swin transformer. *IEEE Journal of Biomedical and Health Informatics*, 29:8719–8726, 2025.

[11] Junbeom Kwon, Jungwoo Seo, Heehwan Wang, Taesup Moon, Shinjae Yoo, and Jiook Cha. Predicting task-related brain activity from resting-state brain dynamics with fmri transformer. *Imaging Neuroscience*, 3, 2024.

[12] Mo Wang, Kaining Peng, Jingsheng Tang, Hongkai Wen, and Quanying Liu. DCA: Graph-guided deep embedding clustering for brain atlases. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Mehraveh Salehi, Abigail S Greene, Amin Karbasi, Xilin Shen, Dustin Scheinost, and R Todd Constable. There is no single functional atlas even for a single individual: Functional parcel definitions change with task. *NeuroImage*, 208:116366, 2020.

[14] James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. Geometric constraints on human brain function. *Nature*, 618(7965):566–574, 2023.

[15] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse Breedlove, Jacob S. Prince, Logan T Dowdle, Matthias Nau, Bradley Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Norris Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116 – 126, 2021.

[16] Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, and Quanying Liu. Brain-dit: A universal multi-state fmri foundation model with metadata-conditioned pretraining. *arXiv preprint arXiv:2604.12683*, 2026.

[17] Hamid Behjat and Martin Larsson. Spectral characterization of functional mri data on voxel-resolution cortical graphs. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 558–562. IEEE, 2020.

[18] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, 9:23, 2015.

[19] Jianxiong Gao, Yu Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:11802–11816, 2024.

[20] Connor Lane, Daniel Z Kaplan, Tanishq Mathew Abraham, and Paul S Scotti. Scaling vision transformers for functional mri with flat maps. *arXiv preprint arXiv:2510.13768*, 2025.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[22] 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.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[24] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[25] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[26] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[27] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[28] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

[29] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[30] Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, et al. Brain harmony: A multimodal foundation model unifying morphology and function into 1d tokens. *arXiv preprint arXiv:2509.24693*, 2025.

[31] Heiko Braak, Kelly Del Tredici, Udo Rüb, Rob AI De Vos, Ernst NH Jansen Steur, and Eva Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of aging*, 24(2):197–211, 2003.

[32] Ronald B Postuma, Daniela Berg, Matthew Stern, Werner Poewe, C Warren Olanow, Wolfgang Oertel, José Obeso, Kenneth Marek, Irene Litvan, Anthony E Lang, et al. Mds clinical diagnostic criteria for parkinson's disease. *Movement disorders*, 30(12):1591–1601, 2015.

[33] Ye Tian, Daniel S Margulies, Michael Breakspear, and Andrew Zalesky. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature neuroscience*, 23(11):1421–1432, 2020.

[34] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

A Details of Rest fMRI experiments

A.1 Details of Datasets and Task

Human Connectome Project (HCP). We utilize the HCP dataset as our primary benchmark for evaluation, comprising 1010 subjects. They are split to 606/202/202 for train/evaluation/test set respectively. The dataset provides high-quality, standardized data ideal for establishing baseline model performance.

Parkinson's Progression Markers Initiative (PPMI). PPMI is a global, longitudinal, observational study aimed at identifying biomarkers of Parkinson's disease (PD) progression [28]. For our disease classification task, we utilized a curated subset of the PPMI cohort; after quality control and task-specific filtering, 331 subjects with complete resting-state fMRI and diagnostic annotations were retained (age range: 35–84 years), covering prodromal PD, clinically diagnosed PD, and healthy control groups.

Alzheimer's Disease Neuroimaging Initiative (ADNI). The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a large-scale, longitudinal, observational study designed to characterize the progression of Alzheimer's disease [27]. For our disease classification task, we utilized a curated subset of the ADNI cohort; after quality control and task-specific filtering, 497 subjects with complete resting-state fMRI and diagnostic annotations were retained (age range: 55–90 years), spanning cognitively normal, mild cognitive impairment (MCI), and Alzheimer's disease (AD) groups.

A.2 Evaluation Protocol

All resting-state evaluations are performed at the subject level. For FlatClip and all compared fMRI foundation-model baselines, the pretrained backbone is kept frozen, and only a lightweight downstream classification head is trained. All models are evaluated on the same subject-level train/validation/test splits. This matched readout protocol controls downstream probe capacity and focuses the comparison on frozen representation quality.

For each subject, FlatClip uses a fixed sequence of 40 cortical flatmap frames. Frame-wise global representations are mean-pooled into one subject-level feature before downstream prediction. For classification tasks, the probe is trained with cross-entropy loss. Model selection is performed on the validation split, and final performance is reported on the held-out test split. To reduce sensitivity to a single validation/test partition, we repeat the evaluation across five subject-level random splits and report mean and standard deviation.

Because the compared models differ in native input representation, architecture, and pretraining data, we interpret the results as a matched-probe representation benchmark rather than a fully controlled architectural comparison.

A.3 Details of Baseline Models

In this section, we introduce our baseline models.

BrainLM is the first fMRI foundation model specifically designed to capture the spatiotemporal dynamics of brain activity through a Transformer-based masked autoencoder architecture. It employs the AAL-424 atlas for the parcellation of brain regions, transforming fMRI recordings into a 424-dimensional set of ROIs. The model is trained on an extensive dataset consisting of 6,700 hours of fMRI data derived from 77,298 recordings across the UK Biobank and the Human Connectome Project.

BrainMASS utilizes the Schaefer 100-region atlas for the parcellation of brain networks. It was trained on 26 datasets, encompassing a total of 64,584 subjects, including data from the UK Biobank (UKB), Human Connectome Project (HCP), and ADHD-200, etc. The model is designed to learn representations of functional brain networks by integrating both masked ROI modeling and latent representation alignment techniques, thereby enhancing its diagnostic performance.

Brain-Harmony is a multimodal brain foundation model designed to jointly represent brain morphology and function within a unified one-dimensional token space. The model first encodes functional MRI ROI time-series and structural T1-weighted MRI into modality-specific token embeddings, and then uses a Transformer-based harmonizer to fuse these tokens into a shared subject-level representation. In our baseline setting, we use the official pretrained Brain-Harmony backbone with a ViT-Base architecture and 128 latent tokens.

LCM is an fMRI foundation model for connectome-based brain representation learning, which formulates pretraining as a multitask learning problem by leveraging brain–environment interaction variables, including demographic and behavioral targets, together with large-scale functional neuroimaging data. Its architecture represents brain connectomes as node-level embeddings and uses a decoder-style prediction head with learnable task queries to support multiple downstream objectives such as sex prediction, behavioral recognition, and disease classification.

SwiFT (Swin 4D fMRI Transformer) is an end-to-end framework for modeling high-dimensional spatiotemporal fMRI dynamics directly from raw 4D volumes, avoiding the information loss that can arise from hand-crafted features or atlas-level summaries. It adopts a computationally efficient 4D shifted-window attention design to capture local and long-range spatiotemporal dependencies in brain activity. SwiFT also supports contrastive self-supervised pretraining, which further improves downstream transfer performance on large-scale fMRI benchmarks.

NeuroSTORM is a general-purpose neuroimaging foundation model designed to learn representations directly from raw 4D fMRI volumes. Its backbone is based on Shifted-Window Mamba (SWM), which is used to model long-range spatiotemporal dependencies while maintaining computational efficiency. NeuroSTORM is pre-trained on large-scale fMRI data from more than 50,000 subjects across UK Biobank, ABCD, and HCP, and introduces a Spatiotemporal Redundancy Dropout (STRD) module to reduce redundant temporal information during representation learning. This design enables NeuroSTORM to serve as a transferable foundation model for downstream fMRI prediction tasks.

Omni-fMRI is an atlas-free fMRI foundation model designed to learn directly from voxel-level signals, avoiding the information loss and atlas-dependent biases introduced by predefined region-level parcellations. To scale pretraining to 49,497 fMRI sessions from nine datasets, Omni-fMRI introduces a dynamic patching mechanism that reduces computational cost while preserving informative spatial structure. The model is evaluated through a comprehensive benchmark suite covering 11 datasets and diverse resting-state and task-based fMRI tasks, where it consistently outperforms existing fMRI foundation models. In our experiments, we use Omni-fMRI as a strong voxel-level foundation-model baseline for assessing whether FlatClip can achieve competitive transfer performance without fMRI-specific encoder pretraining.

A.4 Ablation of Architectures

We further compare different frozen image backbones and feature extraction choices for resting-state prediction (Table 5). For the main comparisons, all methods use one global representation and the same downstream MLP probe to ensure a fair evaluation of the frozen backbone. This setting follows the standard use of global or `[CLS]` representations in ViT-style encoders for image-level prediction, where the global token summarizes the input image into a single feature vector. The ViT baseline uses a standard ImageNet-pretrained ViT-B/16 visual encoder from torchvision. Across backbones, SigLIP2 achieves the strongest performance on HCP and is used as the default backbone in subsequent geometry-control analyses. The additional 40-token and patch-mean variants are included as sensitivity analyses rather than main comparison settings.

A.5 Configuration

All FlatClip encoders are frozen during downstream evaluation. We first render scalar cortical fMRI activity into RGB cortical flatmaps and then extract image features with a frozen SigLIP2-NaFlex encoder. Only the downstream MLP probe is trained for task prediction. The main settings are summarized in Table 6. We choose $T = 40$, following the setting with prior works [9, 3].

Table 5: Ablation of frozen image backbone and feature extraction strategy. Results are reported as Accuracy / weighted F1.

Backbone	Feature	HCP *Sex Classif.*		PPMI *PD Diagnosis*		ADNI (MCI) *Diagnosis*		ADNI (AD) *Diagnosis*	
		ACC ↑	wF1 ↑	ACC ↑	wF1 ↑	ACC ↑	wF1 ↑	ACC ↑	wF1 ↑
DINOv2	One global token 80.37±2.55 / 80.33±2.53	51.39±3.18	51.58±2.25	60.68±2.67	60.44±1.98	75.46±4.01	75.20±3.69		
DINOv2	40 Global tokens	79.53±1.17	79.50±1.15	47.57±5.14	48.82±2.67	55.98±2.67	55.58±2.09	75.00±2.41	74.61±2.34
ViT-B	40 Global tokens	70.00±1.13	70.02±1.15	60.00±1.36	55.45±1.72	52.76±4.57	52.19±4.71	78.11±2.56	76.84±2.71
SigLIP2	One Global token	82.06±2.05	82.04±2.05	54.86±1.96	55.72±1.63	61.11±1.48	60.75±1.16	75.46±2.12	75.41±2.11
SigLIP2	40 Global tokens	83.08±0.59	83.02±0.66	53.47±7.54	53.93±6.14	60.68±1.96	60.02±2.07	74.07±4.88	74.13±4.98
SigLIP2	40 patch_mean	84.94±1.06	84.91±1.06	56.94±4.21	56.48±4.27	58.12±1.96	57.82±1.87	76.39±4.17	76.36±3.93

Table 6: Hyperparameter settings of FlatClip.

Configuration	*Value*
Flatmap rendering	
surface space	fsLR cortical surface
renderer	PyCortex / cached flatmap projection
Normalization	global z-score with bounded color scale
colormap	diverging red–blue colormap
background pixels	transparent outside cortex/ROI, composited on white RGB
flatmap crop	cropped to non-background cortical pixels
Resting-state configs	
image encoder	frozen SigLIP2-base-patch16-NaFlex
input format	sequence of cortical flatmap frames
input region	whole cortex
frames per subject	first 40 resting-state frames
encoder input	RGB flatmap, padded/processed by SigLIP2-NaFlex processor
frame feature	global image embedding
frame feature dimension	768
subject feature	mean pooling over 40 frame features
probe architecture	2-layer MLP, hidden dim 256, dropout 0.2
optimizer	AdamW
learning rate	1×10^{-3}
weight decay	1×10^{-4}
batch size	64
training epochs	up to 250, early stopping patience 40
validation metric	weighted-F1
loss function	class-weighted cross-entropy
reported metrics	accuracy, weighted-F1
NSD visual-fMRI configs	
image encoder	frozen SigLIP2-base-patch16-NaFlex
input format	stimulus-level GLM cortical flatmap
input regions	whole cortex / HCP-MMP visual / NSDgeneral
encoder output	mean-pooled global representation over patch tokens
main feature dimension	768
probe architecture	MLP with LayerNorm, hidden dim 1024, dropout 0.1
optimizer	AdamW
learning rate	3×10^{-4}
weight decay	1×10^{-4}
batch size	128
training epochs	10
validation metric	mAP
loss function	class-weighted BCE with logits
reported metrics	mAP, weighted-F1

Flatmap construction and inference cost. For resting-state fMRI, each subject-level dtseries file is converted to cortical vertex time series and the first 40 frames are rendered. The selected

Table 7: Within-subject COCO80 multi-label classification on NSD. **Red** indicates the best performance. wF1 denotes weighted-F1. Unmarked FlatClip rows use the main setting: one 768-dimensional global representation obtained by mean-pooling the 16×16 patch-token grid. Rows marked with * use the token-rich ablation with all 256 patch tokens.

Model \ Subject	Sub1		Sub2		Sub5		Sub7		Mean	
	mAP ↑	wF1 ↑	mAP ↑	wF1 ↑	mAP ↑	wF1 ↑	mAP ↑	wF1 ↑	mAP ↑	wF1 ↑
SigLIP2-NSDgeneral	**15.311**	**26.922**	15.714	26.986	**19.933**	**29.611**	**14.428**	24.366	**16.347**	**26.971**
SigLIP2-Visual	14.050	25.829	**16.065**	**27.151**	18.889	29.476	14.361	**25.238**	15.841	26.924
DINOv2-Visual	14.765	26.338	14.780	25.105	17.213	27.304	12.605	22.547	14.841	25.323
ViT-Cortex	11.674	23.092	13.373	24.128	13.895	26.376	10.544	20.432	12.371	23.507
SigLIP2-Cortex	11.100	24.576	12.211	25.048	14.974	26.090	10.053	18.822	12.085	23.634
DINOv2-Cortex	10.930	22.800	10.815	23.022	11.290	24.427	9.467	19.751	10.625	22.500
SwiFT	7.974	20.582	10.769	22.303	10.274	22.989	8.478	20.858	9.374	21.683
NeuroSTORM	6.491	12.358	8.798	20.484	9.253	10.471	8.209	18.326	8.188	15.410
Omni-fMRI	3.949	13.818	5.024	15.910	4.715	13.441	3.751	8.934	4.360	13.026
SigLIP2-Cortex*	16.604	29.776	15.771	28.262	19.218	31.379	14.129	27.836	16.430	29.313
SigLIP2-Visual*	20.192	**33.255**	19.623	29.727	23.323	**33.801**	19.490	30.457	20.657	31.810
SigLIP2-NSDgeneral*	**21.565**	32.330	**22.218**	**32.084**	**25.300**	32.744	**20.282**	**31.123**	**22.341**	**32.070**

vertex-by-time matrix is globally z-scored before rendering, and the resulting cortical values are projected to a 2D flatmap. For NSD, each sample corresponds to a stimulus-level GLM response map. The response values are z-scored within each map, mapped to RGB using a diverging colormap, and pixels outside the selected cortex/ROI are treated as transparent background. Before feature extraction, transparent pixels are alpha-composited onto a white RGB background, matching the three-channel input expected by the image encoder.

The whole-cortex input uses a 943×384 flatmap canvas. For NSD visual-fMRI experiments, the rendered flatmaps are cropped to the valid cortical extent of each input region before feature extraction. The HCP-MMP visual-cortex input uses a 273×264 cropped canvas. Because NSDgeneral is a subject-specific task-active cortical mask, its cropped canvas size varies slightly across subjects: 293×208 for Sub1, 293×209 for Sub2 and Sub5, and 292×209 for Sub7. All images are then processed by SigLIP2-NaFlex, which supports variable-size inputs and converts each flatmap into a fixed-budget patch sequence before downstream feature pooling.

Feature extraction is performed offline with the frozen SigLIP2-NaFlex encoder. From the logged extraction jobs, HCP required 3.96 hours on four A100 GPUs for 40,440 flatmap images, corresponding to 2.84 images/s in wall-clock throughput. PPMI required 39.5 minutes on four A100 GPUs for 18,960 images, and ADNI required 61.6 minutes on two A100 GPUs for 19,840 images. After features are cached, downstream MLP training is lightweight and does not require backpropagation through the image encoder.

A.6 The influence of frame number

We further evaluate the effect of temporal sampling length in resting-state FlatClip. For each subject, we uniformly sample 10, 20, or 40 frames from the available 40-frame sequence. Each frame is encoded independently by the frozen SigLIP2 encoder, and the selected frame-level global representations are mean-pooled into a single 768-dimensional subject-level feature. As shown in Table 8, increasing the number of frames generally improves or stabilizes performance, but the gain is modest under the one-token pooling strategy. HCP and ADNI show the clearest benefit from using more temporal samples, with 40 frames giving the best or tied-best performance in most metrics. ADNI (AD) already saturates around 20 frames, while PPMI does not show a consistent improvement with additional frames, suggesting that temporal coverage alone is not the main limiting factor for this task. Overall, these results indicate that FlatClip can extract useful subject-level information even from sparse temporal samples, while the 40-frame setting provides a more stable default for the main experiments.

Table 8: Ablation on the number of resting-state frames used by FlatClip. All results use frozen SigLIP2 with one global representation per frame. Selected frame features are mean-pooled into a 768-dimensional subject-level token. Results are reported as Accuracy / weighted F1 over three seeds.

Frames \ Dataset	HCP *Sex Classif.*		ADNI (MCI) *Diagnosis*		ADNI (AD) *Diagnosis*		PPMI *PD Diagnosis*	
	ACC ↑	**wF1** ↑	**ACC** ↑	**wF1** ↑	**ACC** ↑	**wF1** ↑	**ACC** ↑	**wF1** ↑
10 frames	76.31±4.61	76.27±4.64	59.40±3.23	58.52±3.13	76.39±1.39	76.39±1.43	54.17±2.08	54.12±2.11
20 frames	77.83±1.92	77.79±1.86	60.68±1.48	59.90±1.77	**78.24±0.80**	**78.29±1.26**	52.43±1.59	53.39±2.27
40 frames	**82.06±2.05**	**82.04±2.05**	**61.11±1.48**	**60.75±1.16**	75.46±2.12	75.41±2.11	**54.86±1.96**	**55.72±1.63**

B Details of NSD Visual-fMRI Experiments

Natural Scenes Dataset (NSD). We evaluate visual-fMRI decoding on the Natural Scenes Dataset (NSD), a large-scale 7T fMRI dataset in which subjects viewed natural images from COCO. We use stimulus-level GLM beta estimates, specifically the `betas_fithrf_GLMdenoise_RR` version, as neural response inputs. The decoding target is an 80-dimensional COCO multi-hot label vector.

Cortical input regions. We evaluate three cortical input regions: whole cortex, HCP-MMP visual cortex, and NSDgeneral. NSDgeneral denotes the NSD-provided subject-level task-active region and is used as a stimulus-responsive cortical mask. All regions are rendered as flatmaps and encoded using the same frozen SigLIP2 encoder. Cortex-wide flatmaps are used to test whether broad cortical coverage helps or hurts visual decoding.

B.1 GLM for Baselines models

SwiFT extracted features in a volume-wise manner: each NSD beta volume is processed independently, and the resulting representation is used as the feature for that volume.

NeuroSTORM follows a similar strategy. Its NSD extraction pipeline also processes each beta volume independently with a 3D backbone and saves the resulting backbone feature for each volume.

Therefore, in the current setup, both SwiFT and NeuroSTORM operate on native single-volume inputs and do not require an additional repeat-based adaptation step at feature extraction time.

Omni-fMRI, however, is less naturally matched to this setting. Because the pretrained encoder expects 40 input channels, each single 3D beta volume is duplicated 40 times along the channel dimension and then fed into the encoder. We then keep the CLS token as the feature for that volume and concatenate the volume-wise CLS features across the sequence. This difference in input construction is important when interpreting performance. In particular, Omni-fMRI is not applied to the original single-volume signal directly, but to a repeated version of the same volume. As a result, its extracted representation is influenced not only by the backbone itself, but also by the repeat-based adaptation strategy, which may weaken the effectiveness of the model in the current NSD setting.

B.2 Supplementary Analysis of Image–fMRI Feature Alignment

We further quantified image–fMRI feature alignment in SigLIP2 space using a randomly sampled 100-image subset from NSD shared1000 (Fig. 3, Table 9). For each shared image, we computed the alignment score for each subject and then averaged across the four subjects, yielding one value per image. For Local-RDM, each image was represented by its distance profile to the other shared images, and we correlated the fMRI-derived distance profile with the corresponding SigLIP2 image-feature distance profile. Paired cosine measured the similarity between the aligned fMRI-derived feature and its matched image feature, while the paired-random margin subtracted a randomly mismatched image cosine.

Across all three stimulus-level metrics, SigLIP2-NSDgeneral achieved the strongest alignment with the corresponding image features. Compared with the strongest non-FlatClip baseline, NeuroSTORM, SigLIP2-NSDgeneral improved Local-RDM by 0.083, paired cosine by 0.088, and paired-random

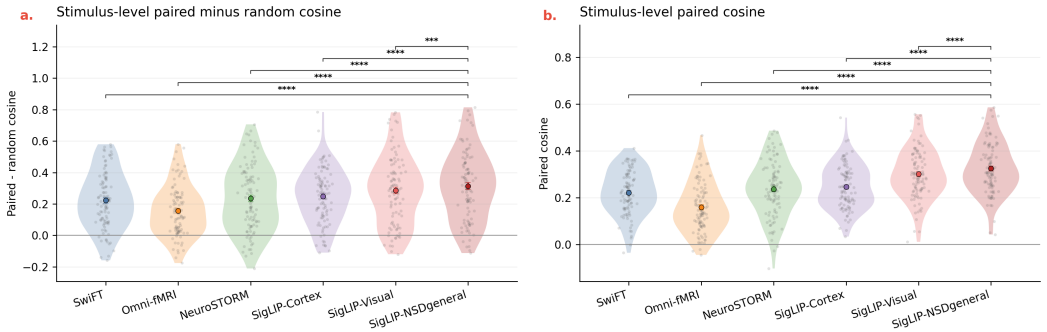

Figure 3: Stimulus-level image–fMRI feature alignment analysis on a randomly sampled subset of NSD shared1000 in SigLIP2 space. (a) Paired-random cosine margin, defined as the cosine similarity between the aligned fMRI-derived feature and the matched image feature minus a randomly mismatched image cosine. (b) Paired cosine similarity between the aligned fMRI-derived feature and its matched image feature. For each shared image, scores are first computed for each subject and then averaged across the four subjects. Dots denote individual shared stimuli, and large markers denote the mean. Significance brackets compare SigLIP2-NSDgeneral against each other method using paired Wilcoxon signed-rank tests over stimuli.

Table 9: Stimulus-level image–fMRI feature alignment analysis on a randomly sampled subset of NSD shared1000 in SigLIP2 space. Values are averaged over four subjects, leaving one value per image. Δ denotes SigLIP2-NSDgeneral minus the comparison method. p_t and p_w denote Holm-corrected paired t-test and Wilcoxon signed-rank p-values, respectively.

Metric	Comparison	Baseline	SigLIP2-NSDgeneral	Δ	95% CI	p_t	p_w	Improved
Local-RDM	vs SwiFT	0.026	0.134	0.108	[0.086, 0.129]	9.01×10^{-16}	3.18×10^{-12}	83.0%
Local-RDM	vs Omni-fMRI	0.009	0.134	0.125	[0.102, 0.147]	2.70×10^{-17}	5.27×10^{-13}	86.0%
Local-RDM	vs NeuroSTORM	0.052	0.134	0.083	[0.064, 0.101]	5.78×10^{-13}	1.34×10^{-10}	79.0%
Local-RDM	vs SigLIP2-Cortex	0.032	0.134	0.102	[0.084, 0.121]	2.09×10^{-17}	2.87×10^{-13}	89.0%
Local-RDM	vs SigLIP2-Visual	0.091	0.134	0.043	[0.026, 0.061]	5.41×10^{-6}	5.73×10^{-6}	74.0%
Paired cosine	vs SwiFT	0.222	0.325	0.103	[0.087, 0.119]	3.82×10^{-21}	7.31×10^{-15}	90.0%
Paired cosine	vs Omni-fMRI	0.159	0.325	0.166	[0.150, 0.182]	1.50×10^{-35}	3.61×10^{-17}	99.0%
Paired cosine	vs NeuroSTORM	0.237	0.325	0.088	[0.072, 0.104]	5.07×10^{-17}	4.31×10^{-14}	89.0%
Paired cosine	vs SigLIP2-Cortex	0.246	0.325	0.079	[0.064, 0.094]	3.70×10^{-16}	1.17×10^{-13}	89.0%
Paired cosine	vs SigLIP2-Visual	0.301	0.325	0.024	[0.015, 0.033]	3.83×10^{-6}	7.96×10^{-6}	69.0%
Margin	vs SwiFT	0.223	0.314	0.091	[0.066, 0.116]	7.19×10^{-10}	8.30×10^{-9}	76.0%
Margin	vs Omni-fMRI	0.156	0.314	0.157	[0.128, 0.186]	5.07×10^{-17}	2.87×10^{-13}	87.0%
Margin	vs NeuroSTORM	0.235	0.314	0.079	[0.053, 0.106]	1.81×10^{-7}	4.49×10^{-7}	72.0%
Margin	vs SigLIP2-Cortex	0.248	0.314	0.066	[0.042, 0.090]	1.38×10^{-6}	2.49×10^{-6}	69.0%
Margin	vs SigLIP2-Visual	0.286	0.314	0.028	[0.012, 0.044]	4.23×10^{-4}	4.06×10^{-4}	68.0%

margin by 0.079. The advantage also remained significant when compared with SigLIP2-Visual, indicating that restricting the flatmap input to NSD task-active cortex yields representations that are more tightly aligned with image-feature geometry than using a broader visual mask. These results support the interpretation that FlatClip's advantage is not only reflected in downstream COCO80 classification, but also in a closer stimulus-level correspondence between cortical flatmap representations and image encoder representations. Because the statistical unit in this analysis is the stimulus after averaging over subjects, we treat it as supplementary mechanistic evidence rather than as a replacement for subject-level evaluation.

B.3 Ablation and Detailed Results

We report detailed within-subject NSD COCO80 results in Table 7. All FlatClip variants use the same frozen SigLIP2 encoder, one global image representation per GLM response map, and the same lightweight MLP probe. This setting keeps the downstream feature dimensionality fixed and makes the comparison focus on the cortical input region.

Across the evaluated general-purpose fMRI baselines, FlatClip variants obtain substantially higher mAP and weighted F1. Among FlatClip variants, performance improves when the input region is restricted from whole cortex to visual or NSD-provided task-active cortex. These results support the main finding that visual-fMRI decoding benefits from task-relevant cortical geometry rather than indiscriminate whole-cortex coverage.

C Details of Geometry perturbation Experiments

C.1 Geometry-control perturbations.

To test whether FlatClip benefits from the spatial organization of cortical flatmaps rather than only from the marginal distribution of fMRI values, we constructed several geometry-control inputs for HCP sex classification. All controls used the same subjects, train/validation/test split, number of frames, frozen encoder, and downstream MLP probe. For each subject, we used a fixed sequence of 40 frames. All rendered inputs were converted to the same image canvas before feature extraction, so that performance differences primarily reflect changes in spatial organization under matched downstream settings, while keeping image canvas handling and feature extraction fixed within each control family.

Real flatmap. The real-flatmap condition uses the original cortical flatmap rendering without spatial perturbation. Each frame preserves the native 2D cortical surface layout produced by the flatmap projection, including both local activity patterns and their global anatomical arrangement.

Left-right hemisphere swap. For the left-right swap control, we exchanged the left and right hemispheric halves of each rendered flatmap image. The same transformation was applied to all subjects and frames. This control preserves the pixel intensity distribution, local texture, and most low-level image statistics, but disrupts hemispheric laterality and the correspondence between cortical location and visual image position. It therefore tests whether FlatClip depends on the anatomically meaningful placement of the two hemispheres rather than only on the presence of cortical-shaped image content.

Within-hemisphere region shuffle. For the within-hemisphere shuffle control, we first separated the rendered flatmap into left- and right-hemisphere image regions. Within each hemisphere, the valid cortical area was divided into local image blocks, and these blocks were randomly permuted only within the same hemisphere. The same block permutation was applied consistently across subjects and frames. This preserves hemispheric identity and local block-level activation patterns, while disrupting the finer global layout of cortical regions within each hemisphere. Compared with left-right swapping, this provides a stronger test of whether within-hemisphere cortical organization contributes to downstream prediction.

Phase/frequency-matched randomization. To further control for rendering-induced low-level image statistics, we constructed a phase-randomized control for each flatmap frame. For each RGB channel, we preserved the Fourier amplitude spectrum of the original flatmap but randomized the phase, and then matched the resulting image back to the original channel-wise intensity histogram. This keeps the global frequency content and marginal color/intensity distribution similar to the real flatmap, while destroying spatial phase and therefore cortical geometry. A performance drop under this control indicates that the model is not driven only by low-level image frequency or intensity statistics.

Random initialized encoder. In addition to input perturbations, we included a randomly initialized SigLIP2 encoder control. This condition uses the real flatmap input and the same downstream MLP probe, but removes pretrained visual representations by replacing the frozen pretrained encoder with the same architecture initialized from random weights. This control tests whether the observed performance comes only from the model architecture and image-format input, or whether pretrained visual features contribute to the FlatClip representation.

Spatial block-shuffle. For the block-shuffle control, each flatmap image was divided into non-overlapping 16×16 image blocks. A single random block permutation was generated and applied consistently to all subjects and all frames. This preserves the local texture and activation pattern inside each block, but disrupts the global arrangement of cortical regions. Thus, this control tests whether local patch-level fMRI patterns alone are sufficient when large-scale cortical topology is scrambled.

Table 10: Geometry-control ablation on HCP sex classification. All results use the frozen DINOv2 encoder with one global token per frame, followed by the same MLP probe. Results are reported as mean±std over three seeds. Significance markers indicate paired t-tests comparing Real flatmap against each control after Holm correction. **Red** indicates the best performance.

Control input	ACC ↑	wF1 ↑	Balanced ACC ↑	Macro F1 ↑
Real flatmap	**80.37 ± 2.55**	**80.33 ± 2.53**	**80.08 ± 2.47**	**80.15 ± 2.53**
Spatial block-shuffle	$75.47 \pm 0.78^{*}$	$75.49 \pm 0.76^{*}$	$75.44 \pm 0.70^{*}$	$75.34 \pm 0.75^{*}$
Vertex-permutation	$73.77 \pm 2.80^{*}$	$73.80 \pm 2.81^{*}$	$74.06 \pm 2.61^{*}$	$73.73 \pm 2.76^{*}$
FC heatmap	$54.99 \pm 5.08^{**}$	$55.04 \pm 5.07^{**}$	$54.80 \pm 5.08^{**}$	$54.77 \pm 5.07^{**}$
Random image	$49.24 \pm 1.52^{**}$	$49.33 \pm 1.52^{**}$	$49.24 \pm 1.54^{**}$	$49.14 \pm 1.53^{**}$

Markers denote Holm-corrected paired t-tests against Real flatmap: $^{*}p < 0.05$, $^{**}p < 0.01$.

Vertex-permutation. For the vertex-permutation control, we identified valid cortical pixels using the flatmap alpha mask and randomly permuted the values within this cortical mask. The same permutation was applied consistently across all subjects and frames. This preserves the distribution of activation values within each frame, and also avoids changing the background region, but destroys both local neighborhood structure and global cortical topology. Therefore, this is a stronger geometry disruption than block-shuffling.

FC heatmap. For the FC heatmap control, we replaced frame-wise flatmaps with a subject-level functional connectivity image. For each subject, one 200-frame Schaefer100 ROI time series was used to compute a 100×100 Pearson correlation matrix after ROI-wise normalization. The resulting FC matrix was clipped to $[-1, 1]$, rendered as a red–blue heatmap, resized to the same image canvas, and repeated as 40 identical frames. This control keeps subject-level functional information but removes cortical surface geometry and frame-wise flatmap dynamics.

Random image. For the random-image control, each frame was replaced by a random RGB image with the same spatial size as the flatmap input. Random images were generated with deterministic subject- and frame-specific seeds. This control preserves only the image-format interface to the frozen visual encoder, while removing both fMRI signal and cortical geometry.

As an additional robustness check, we repeat the geometry-control ablation with a frozen DINOv2 encoder. Table 10 shows the same qualitative trend as the main SigLIP2 geometry-control experiment.

C.2 Details of geometry recovery experiments

To examine whether the spatial organization of fMRI activity contributes to downstream performance, we compare ROI time-series, 4D ROI-volume, native voxel-volume, and adaptive-patch voxel-volume representations under controlled model capacity. The purpose of this experiment is to separate the effect of signal content from the effect of geometry: ROI time-series preserve regional temporal signals but remove spatial layout, 4D ROI volumes restore coarse atlas geometry, and native voxel volumes preserve fine-grained spatial variation.

4D ROI-volume processing. For volumetric fMRI inputs, we construct ROI-restricted 4D volumes from the preprocessed resting-state fMRI scans. Each scan is aligned to the common template space used by the dataset, and the target ROI mask is resampled to the same spatial resolution using nearest-neighbor interpolation. Given a 4D fMRI volume $X_i \in \mathbb{R}^{T \times H \times W \times D}$ for subject i, voxel signals within each ROI are first averaged to obtain an ROI-level time series. We then map the ROI-level signal back to the atlas-defined 4D volume, so that all voxels belonging to the same ROI share the corresponding ROI-mean time series. Voxels outside the selected ROI mask are set to zero. This produces a 4D ROI-volume representation that preserves the 4D input format and coarse anatomical geometry, while removing fine-grained within-ROI spatial variation.

To ensure a consistent input length, we sample a fixed sequence of 40 frames from each scan and use the resulting ROI-volume input

$$X_i^{ROI} \in \mathbb{R}^{40 \times H \times W \times D}$$

as the subject-level representation.

Native voxel-volume processing. For the voxel-volume setting, we use the same preprocessed 4D fMRI scans and the same 40-frame sampling strategy, but preserve the native voxel-wise signals instead of replacing them with ROI means. Voxels outside the selected brain or ROI mask are set to zero, and temporal z-score normalization is applied voxel-wise. This setting keeps fine-grained voxel-level spatial variation and serves as a stronger geometry-preserving volumetric reference.

Patch budget and adaptive patching. All 4D ROI-volume and voxel-volume models use the same ViT-Small backbone. For the standard 4D ROI and voxel settings, we use a fixed patch size of 8 and adjust the effective input crop so that each sample contains approximately 1,000 spatiotemporal patches. This keeps the token budget comparable across ROI-volume and voxel-volume inputs despite differences in spatial extent.

We also evaluate a patch-adaptive voxel setting following the adaptive patching strategy of Omni-fMRI [3]. In this variant, the base patch size is set to 12, and the patch allocation is adapted according to local signal properties so that informative regions can be represented at finer effective resolution while maintaining an overall token budget close to 1,000 patches. This experiment tests whether allocating more spatial detail to signal-relevant regions further improves geometry-preserving volumetric modeling.

40-frame ROI time-series processing. For the ROI time-series setting, we convert each 4D resting-state scan into a region-level temporal representation using the same ROI definition. At each time point, voxel signals within each ROI are averaged to obtain a regional activity vector. For subject i, this produces a time series

$$R_i = [r_{i,1}, r_{i,2}, \ldots, r_{i,T}], \quad r_{i,t} \in \mathbb{R}^K,$$

where K is the number of ROIs. The regional signals are temporally z-scored within each ROI, and the same 40-frame sampling strategy is applied to obtain

$$R_i^{40} \in \mathbb{R}^{40 \times K}.$$

This representation removes within-ROI spatial geometry and keeps only region-level temporal dynamics. We feed the 40-frame ROI sequence to a lightweight temporal model with the same train/validation/test splits as the flatmap and ROI-volume settings.

All experiments are performed at the subject level. Frames from the same subject are never split across train, validation, and test sets, preventing frame-level leakage.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations have been discussed in the paper.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The details of experiments are provided.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Code is provided in supplementary.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Details are provided.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Statistical significance is calculated and provided.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Details of compute resources are provided.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: yes.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: It's a general model research.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets used in this project is licensed.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Code and documents are provided.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification:

    Guidelines: The paper does not involve crowdsourcing.

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLM is used only for writing, editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.