
A Hierarchical Tokenization Framework for Brain fMRI Representation Learning

Karanpartap Singh* Yue Zhao Mohammad H. Abbasi Ehsan Adeli
Stanford University

Abstract

We introduce **HBAR** (Hierarchical Brain Activity Representation), a hierarchical tokenization framework for learning compact, multi-scale representations of neuroimaging data. HBAR organizes voxel-level resting-state fMRI into a hierarchy spanning sub-parcel voxel groups, parcels, and large-scale functional networks, enabling approximately $415\times$ compression while preserving fine-grained spatial information beyond conventional parcellation. The framework supports fixed atlas-based, fully learned, and softly prior-guided hierarchies, allowing us to systematically study the role of structured priors in brain representation learning. Evaluated on large-scale fMRI data, HBAR substantially improves reconstruction over parcellation-based and single-scale tokenization baselines, while supporting downstream modeling of brain dynamics and demographic prediction tasks such as sex classification. Across reconstruction and downstream evaluation, atlas-guided hierarchies provide the strongest overall performance, suggesting that established functional atlases encode organizational structure that remains difficult to recover from reconstruction alone. By combining voxel-level fidelity, compact discrete tokens, and multiscale brain structure, HBAR offers a practical framework for representation learning in high-dimensional neuroimaging data.

1 Introduction

Many modern machine learning systems operate on high-dimensional structured data, where observations are defined over sets of interrelated entities rather than independent features. Examples include images defined over spatial grids (e.g. CNNs) [18], graphs defined over relational neighborhoods (e.g. Graph-CNNs) [16], and biological signals defined over anatomically organized domains such as brain networks [28]. A central challenge in these settings is to construct representations that are both compact enough to enable efficient learning and inference, yet expressive enough to preserve the dependencies and organization inherent to the data.

A common approach is *tokenization*, where high-dimensional inputs are mapped to discrete or continuous tokens used for downstream modeling. Existing tokenization strategies, however, typically rely on fixed partitions or uniform downsampling, imposing strong assumptions about the structure of the data. While effective for regular domains, such approaches can be suboptimal for structured signals whose organization is not aligned with simple spatial or topological partitions. An alternative perspective is to view tokenization as a learned aggregation process, in which tokens are formed by adaptively grouping elements according to their statistical dependencies [29]. This naturally leads to *hierarchical tokenization*, where representations are organized across multiple scales from local structure to global organization. However, prior approaches are often tied to fixed spatial hierarchies (e.g., image pyramids), limiting their flexibility and applicability to more general structured domains.

*Correspondance to: karanps@stanford.edu

We make our anonymized code available at <https://anonymous.4open.science/r/HBAR-anon-CC10/>.

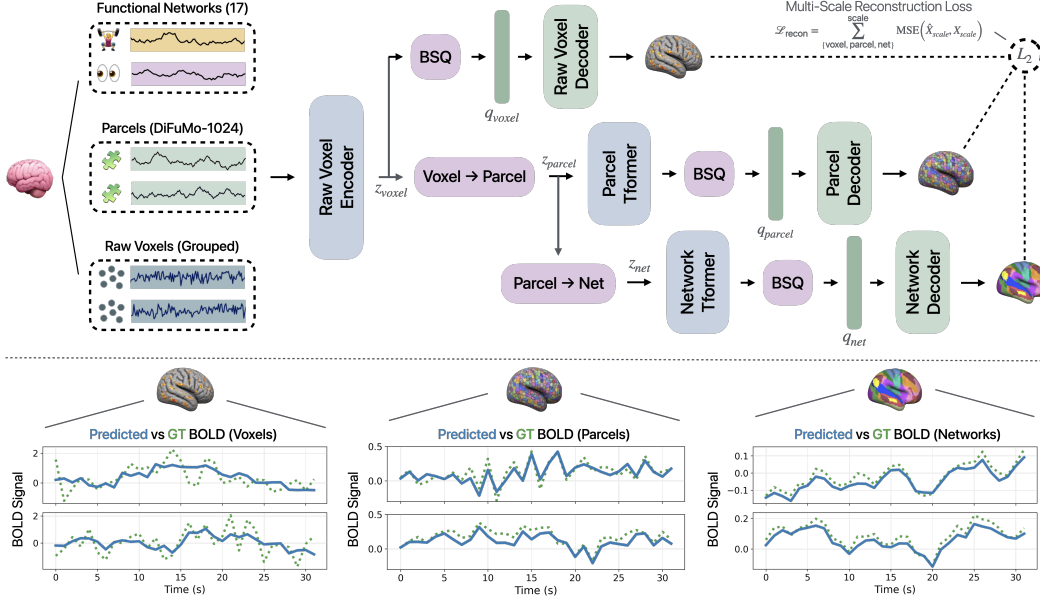


Figure 1: **Hierarchical tokenization and reconstruction for resting-state fMRI.** Voxel-level BOLD signals are organized into a three-level hierarchy spanning sub-parcel voxel groups, parcels (DiFuMo-1024), and large functional networks (Yeo-17). A shared encoder produces voxel-level representations, which are progressively aggregated to parcel- and network-level tokens via hierarchical mappings. At each scale, tokens are discretized using binary spherical quantization (BSQ) and decoded to reconstruct the corresponding signals. The model is trained using a multi-scale reconstruction objective, encouraging representations to preserve information across voxel, parcel, and network levels. Example reconstructions (bottom) demonstrate recovery of BOLD dynamics at all scales.

Resting-state functional magnetic resonance imaging (rs-fMRI) provides a compelling testbed for hierarchical tokenization. Brain activity is inherently structured across multiple scales, from voxel-level signals to parcels and functional networks, and exhibits rich dependencies that are not strictly localized in space. At the same time, existing representation learning approaches for rs-fMRI typically operate at one of two extremes: either directly modeling voxel-level data at high computational cost [31], or aggregating signals into predefined atlases or parcellations [6, 9, 34], which can discard fine-grained spatial information. This tension makes rs-fMRI an ideal setting in which to study learned, multiscale tokenization. To study this setting, we introduce **HBAR (Hierarchical Brain Activity Representation)**, a multiscale tokenization framework for structured neuroimaging data.

Our contributions are summarized as follows: (1) we introduce a general framework for hierarchical tokenization of structured data based on hierarchical aggregation rather than fixed spatial downsampling; (2) we instantiate HBAR for rs-fMRI using a voxel-to-parcel-to-network hierarchy, achieving approximately $415\times$ compression while preserving substantially more fine-grained spatial information than conventional parcellation; (3) we systematically study the tradeoff between fully learned and prior-guided hierarchical organization, including a hybrid approach combining adaptive tokenization with structured neurobiological priors; and (4) we demonstrate that the resulting representations support reconstruction, autoregressive prediction of brain dynamics, and demographic prediction tasks, while showing that structured atlas priors provide a strong inductive bias for hierarchical representation learning in fMRI.

2 Related Work

2.1 Tokenization and Hierarchical Tokenization

Modern tokenization emerged from variational autoencoders [15] and VQ-VAE [22], which introduced discrete latent representations compatible with autoregressive modeling. VQ-VAE-2 [25]

established the now-standard multiscale formulation of coarse and fine latents, later adopted by VQGAN [10], DALL-E [24], ImageGPT [7], and AudioLM [5]. Together, these works demonstrated that strong tokenizers can compress high-dimensional signals into compact representations suitable for large-scale generative modeling.

Most hierarchical tokenizers define hierarchy through fixed spatial mechanisms such as multiresolution grids, residual quantization, or decoder pyramids. HR-VQVAE [2], RQ-VAE [19], and VAR [29] progressively refine representations across scales and show that multiscale tokenization improves efficiency and inductive bias. However, these hierarchies remain fundamentally tied to Euclidean image structure rather than structured scientific domains whose meaningful units may be spatially distributed. A second line of work instead learns aggregation directly. Perceiver [14] compresses large inputs into latent bottlenecks via cross-attention, while DiffPool [36] learns soft hierarchical graph coarsening. These works suggest that tokenization need not rely on fixed partitions, but can instead emerge through learned many-to-few aggregation.

2.2 Tokenization and Autoregressive Modeling for fMRI

Recent fMRI foundation models primarily focus on pretraining objectives rather than on learning the tokenizer itself. BrainLM [6] applies large-scale masked-prediction training to rs-fMRI and demonstrates transfer to clinical and demographic tasks, while Brain-JEPA [9] introduces latent-target prediction with spatiotemporal masking. SLIM-Brain [32] moves toward atlas-free voxel-level modeling using selective hierarchical encoding. These works show that both atlas-based and voxel-level pretraining can transfer effectively, but generally treat representational units as fixed.

Several recent works move closer to explicit tokenization and compression. BrainCodec [21] imports neural codec ideas into fMRI for denoising and reconstruction, while HST [33] combines hierarchical state-space modeling with clustered VQ-VAE quantization for brain-state analysis. Earlier voxelwise VAE work [13] similarly demonstrated that learned latent rs-fMRI representations preserve information useful for downstream prediction. Complementary autoregressive approaches include FARM [12], continuous-space autoregression [23], voxel-level transformer prediction [27], and multimodal next-token modeling [35]. Together, these frameworks establish that compression and autoregressive modeling of fMRI signals are both feasible and useful.

2.3 Recent Tokenization and Quantization Techniques

Recent tokenizer research increasingly studies the geometry and scalability of the quantization bottleneck itself. LFQ/MAGVIT-v2 [37], FSQ [20], BSQ [39], and spherical Leech quantization [40] demonstrate that bottleneck geometry strongly influences tokenizer stability, compression efficiency, and scalability. Other approaches focus on compactness and codebook utilization. TiTok [38] compresses images into extremely short token sequences, VQGAN-LC [41] scales to very large vocabularies, IBQ [26] introduces differentiable code indices, and FQGAN [3] factorizes vocabularies into semantically enriched sub-codebooks. Recent work also increasingly co-designs tokenizers and autoregressive models. VAR [29], for example, replaces raster next-token prediction with next-scale prediction over multiscale latent maps generated by a hierarchical VQ-VAE. However, these approaches remain largely tied to predefined image pyramids and grid-structured data, whereas rs-fMRI exhibits distributed low-SNR structure spanning voxel, parcel, and network scales.

3 HBAR: Hierarchical Tokenization for fMRI

3.1 Problem Formulation

We consider data defined over a structured domain represented as a graph $\mathcal{G} = (V, E)$, where V denotes a set of nodes and E encodes relationships between nodes. Each observation is a signal:

$$X \in \mathbb{R}^{|V| \times d_{\text{in}}}$$

where nodes $i \in V$ are associated with feature vectors X_i . In rs-fMRI, nodes correspond to voxels, and edges reflect spatial adjacency or implicit functional relationships. Our goal is to construct a *hierarchical tokenization* of X , producing a sequence of progressively coarser representations:

$$Z^{(0)} \rightarrow Z^{(1)} \rightarrow \dots \rightarrow Z^{(L)}$$

where $Z^{(0)} = X$, and $Z^{(l)} \in \mathbb{R}^{N_l \times d}$ denotes a set of N_l tokens at level l , with $N_0 > N_1 > \dots > N_L$.

3.2 Hierarchical Tokenization as Aggregation

We define a general tokenization operator at level l as $Z^{(l+1)} = \mathcal{T}^{(l)}(Z^{(l)}, \mathcal{G})$ which maps fine-scale representations to coarser tokens. We consider two complementary regimes:

Fixed (prior-based) tokenization. Tokenization may be defined by a partition of nodes into groups $\{S_k \subset V\}_{k=1}^{N_{l+1}}$, yielding:

$$Z_k^{(l+1)} = \text{Pool}(\{Z_i^{(l)} : i \in S_k\})$$

where Pool is an aggregation operator (e.g., mean or attention pooling). In structured domains such as the brain, these partitions may be derived from large-scale datasets (e.g., functional atlases), providing a strong prior over meaningful organization. A large body of resting-state fMRI literature focuses on deriving large scale groupings of brain regions at various scales: as small as seven networks [28] and as large as 1024 [8] or even 3000 parcels [4]. Similarly, we instantiate our framework for resting-state fMRI using a three-level hierarchy spanning voxel, parcel, and network scales. For each subject, we assume a preprocessed 4D resting-state fMRI volume

$$X \in \mathbb{R}^{X \times Y \times Z \times T}$$

in standard space (e.g., 2 mm MNI), along with a gray-matter mask $M \in \{0, 1\}^{X \times Y \times Z}$. Only voxels with $M = 1$ are retained, yielding V voxels per subject. In our experiments, a typical scan contains approximately $V \approx 8.1 \times 10^5$ voxels and $T \in [971, 1200]$ timepoints.

Learned tokenization via cross-attention. Alternatively, tokenization can be defined as a learned aggregation:

$$Z^{(l+1)} = \text{Attn}(Q^{(l)}, Z^{(l)}),$$

where $Q^{(l)} \in \mathbb{R}^{N_{l+1} \times d}$ are learned query embeddings, and Attn denotes multi-head cross-attention. Each token aggregates information from all tokens at level l with learned, input-dependent weights, allowing the model to learn distributed and non-local structure directly from data.

In our architecture, at the bottom level, voxel-level sub-parcel tokens are first encoded using a temporal encoder operating on local voxel groups over short temporal windows. The resulting token representations are then aggregated hierarchically to produce parcel- and network-level representations. Finally, we also consider a soft hierarchy setting, in which the same learned cross-attention architecture is used, but representations are regularized toward atlas-derived groupings.

Overall, these two regimes represent complementary design choices: fixed tokenization can leverage strong domain priors, while learned tokenization offers flexibility and the ability to capture data-driven structure. In domains with well-established large-scale organization, prior-based tokenization can provide a particularly strong inductive bias. Our formulation unifies and extends several existing paradigms. Convolutional neural networks aggregate over fixed local neighborhoods on a grid, while graph neural networks aggregate over predefined neighborhoods in $\mathcal{G}$. In contrast, hierarchical tokenization defines aggregation operators that may be either fixed or learned, and organizes them across multiple scales. This flexibility is especially important in domains such as brain data, where meaningful structure is often distributed and not strictly localized in space.

3.3 Multiscale Representation

Each level of the hierarchy induces an implicit coarsening of the underlying graph. Tokens at level $l + 1$ can be interpreted as nodes in a coarsened graph $\mathcal{G}^{(l+1)}$, where each coarse node represents a subset or distribution over nodes at level l . Unlike classical graph coarsening methods, which rely on fixed clustering heuristics, our formulation allows tokenization to be defined either through explicit priors or learned aggregation, enabling flexible multiscale representations.

Voxel-level At the lowest level, we construct voxel-level tokens by partitioning each parcel into fixed-size voxel groups of size $K = 16$, preserving fine-grained spatial information that is typically lost in standard parcellations that average much larger groups of voxels. For each parcel, voxel coordinates are normalized and ordered deterministically, after which they are divided into contiguous groups, padding the final group if necessary. This results in approximately $N_0 \approx 5.1 \times 10^4$ bottom-level tokens per subject for our data.

Parcel-level At the intermediate level, we use DiFuMo-1024 parcels [8].

Network-level At the top level, we use these large-scale functional networks (specifically the Yeo-17 atlas [28]). Importantly, this hierarchy is anchored in well-established functional atlases, which provide a strong prior over brain organization.

Discretizations at multiple levels Vector [22], binary spherical [39], or other quantization [40] can be applied at intermediate levels to obtain discrete tokens. In practice, BSQ discretizes representations by projecting features onto a binary hypersphere using a straight-through sign operator, producing compact discrete tokens while maintaining stable optimization and efficient codebook utilization. Compared to conventional approaches, BSQ enables substantially higher compression without codebook collapse, making it especially attractive for extremely high-dimensional data such as fMRI.

3.4 Optimization and Regularization

Decoding proceeds hierarchically across scales: network-level predictions are first reconstructed from top-level tokens, parcel-level predictions are modeled as residuals relative to broadcast network predictions, and voxel-level reconstructions are modeled as residuals relative to parcel-level signals. This residual formulation encourages coarse representations to capture low-frequency global structure, while finer scales model higher-frequency local deviations. The full objective is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\delta} \mathcal{L}_{\delta} + \lambda_{\text{vq}} \mathcal{L}_{\text{vq}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \quad (1)$$

Reconstruction terms Mean-squared-error (MSE) supervision individually at voxel, parcel, and network scales encourages the hierarchy to preserve both fine-grained and large-scale functional organization. The temporal consistency term $\mathcal{L}_{\delta}$ encourages preserving local temporal dynamics in noisy voxel signals.

$$\mathcal{L}_{\text{recon}} = \sum_{\ell \in \{\text{voxel}, \text{parcel}, \text{network}\}} \text{MSE}(\hat{\mathbf{X}}_{(\ell)}, \mathbf{X}_{(\ell)}), \quad \mathcal{L}_{\delta} = \text{MSE}(\Delta_t \hat{\mathbf{X}}_{\text{voxel}}, \Delta_t \mathbf{X}_{\text{voxel}}). \quad (2)$$

Soft cross-scale alignment In the soft hierarchy regime, atlas-derived representations are treated as fixed targets and do not receive gradients. The alignment objective therefore acts as a soft structural prior, encouraging learned groupings to remain close to known functional organization while still permitting flexible reassignment through attention-based pooling. In contrast, the fully learned hierarchy omits this term and learns cross-scale organization purely from reconstruction objectives.

$$\mathcal{L}_{\text{align}} = \|\mathbf{Z}_{\text{parcel}}^{\text{learned}} - \mathbf{Z}_{\text{parcel}}^{\text{atlas}}\|_2^2 + \|\mathbf{Z}_{\text{network}}^{\text{learned}} - \mathbf{Z}_{\text{network}}^{\text{atlas}}\|_2^2. \quad (3)$$

BSQ objective The quantization term is summed across all levels:

$$\mathcal{L}_{\text{vq}} = \sum_{\ell \in \{\text{voxel}, \text{parcel}, \text{network}\}} \mathcal{L}_{\text{BSQ}}^{(\ell)}. \quad (4)$$

For each level ℓ , with $\hat{\mathbf{z}}^{(\ell)} = \text{sign}(\mathbf{z}^{(\ell)})$, namely $\mathcal{L}_{\text{BSQ}}^{(\ell)}$ is a weighted average of the commitment loss and the entropy penalty, $\mathcal{L}_{\text{BSQ}}^{(\ell)} = \mathcal{L}_{\text{commit}}^{(\ell)} + \frac{\zeta^{(\ell)}}{\tau_{\text{inv}}^{(\ell)}} \mathcal{L}_{\text{entropy}}^{(\ell)}$:

$$\mathcal{L}_{\text{commit}}^{(\ell)} = \beta^{(\ell)} \left\| \text{sg}(\hat{\mathbf{z}}_{\text{scaled}}^{(\ell)}) - \mathbf{z}^{(\ell)} \right\|_2^2, \quad \mathcal{L}_{\text{entropy}}^{(\ell)} = \gamma_0^{(\ell)} H_{\text{persample}}^{(\ell)} - \gamma^{(\ell)} H_{\text{codebook}}^{(\ell)}, \quad (5)$$

$$(6)$$

where $\text{sg}(\cdot)$ is stop-gradient and β, γ_0, γ and ζ are hyperparameters whose value we follow suggestions without further tuning. BSQ quantization is applied independently at the voxel-, parcel-, and network-token levels prior to hierarchical decoding (Figure 1).

4 Experiments

4.1 Data and Preprocessing

We use $N = 1056$ resting-state fMRI BOLD recordings for pre-training HBAR from the Human Connectome Project Young-Adults (HCP-YA) [30]. We reserve 85% of the data for training and 15%

for testing, and train for 10 epochs with a $2e-4$ learning rate. The datasets are partitioned at the subject level to prevent any repeated or longitudinal recordings from the same individual from appearing in both training and testing sets. All datasets are preprocessed using standardized pipelines [1, 11], and all models and baselines are trained on identical data splits using approximately compute-matched model sizes and compression ratios to ensure fair comparison across tokenizer families.

4.2 Baselines

We compare against several tokenizer baselines spanning both conventional parcellation and hierarchical generative modeling approaches.

Raw-Voxel VQ-VAE / BSQ. Single-scale tokenizers operating directly on voxel-level chunks without explicit hierarchical organization.

MS-VQVAE. A hierarchical VQ-VAE baseline inspired by multiscale autoregressive image models such as VAR [29]. This baseline progressively compresses representations across scales using learned hierarchical latent maps, but does not incorporate structured neurobiological priors or explicit atlas-based organization. To provide a compute-matched multiscale VQ baseline, we implement a two-level hierarchical VQ-VAE following the progressive latent compression strategy used in VAR [29]. The model operates on chunk-scale middle tokens and parcel-scale top tokens, using 12-bit and 13-bit EMA-VQ codebooks respectively. This yields an effective compression ratio of approximately $564\times$, representing a stable compression, close to that of the hierarchical-BSQ models, achievable without codebook collapse in standard EMA-VQ training. Unlike our hierarchical BSQ framework, which uses three explicit scales (voxel, parcel, and network), the MS-VQVAE baseline uses only two discrete latent scales and does not incorporate structured priors or cross-scale alignment objectives.

DiFuMo Baselines. Conventional atlas-based representations derived directly from DiFuMo parcel time series, optionally compressed using PCA or BSQ.

4.3 Temporal Windowing

Training is performed on overlapping temporal windows of length $W = 32$ with stride $s = 16$. For a sequence of length T , this yields $N_{\text{win}} = \lfloor \frac{T-W}{s} \rfloor + 1$ training windows per scan, corresponding to approximately 59–74 windows in our setting. Each training sample consists of voxel-level windows, sub-parcel token groups, and corresponding parcel- and network-level summaries. Within the encoder, temporal windows are partitioned into smaller segments that are summarized into a fixed number of latent temporal tokens at each scale.

5 Results

5.1 Reconstruction

Hierarchical tokenization substantially improves reconstruction quality across voxel, parcel, and network scales compared to both conventional parcellation-based representations and prior multiscale generative approaches (Table 1). In particular, all hierarchical variants outperform raw-voxel VQ-VAE, BSQ, and MS-VQVAE baselines across nearly all metrics, highlighting the importance of explicitly modeling multiscale brain organization. For raw-voxel baselines that do not explicitly model parcel- or network-level representations, parcel and network predictions are computed by averaging reconstructed voxel signals according to the same DiFuMo-1024 and Yeo-17 atlas mappings used throughout the hierarchical models. This allows reconstruction quality to be evaluated consistently across scales for all tokenizer families. Among hierarchical approaches, the fixed atlas-based hierarchy consistently achieves the strongest overall performance, particularly at parcel and network scales. While the soft hierarchy provides a marginal improvement in voxel-level reconstruction, this gain comes at the cost of degraded reconstruction quality at coarser scales.

5.2 Downstream Performance

We next evaluate whether the learned representations support downstream modeling beyond reconstruction. We consider two evaluation settings: (i) linear probing using mean-pooled token embeddings, and (ii) autoregressive modeling using a causal sequence model trained directly on the

Model	Voxel Corr $\uparrow$	Voxel MSE $\downarrow$	Parcel Corr $\uparrow$	Parcel MSE $\downarrow$	Net Corr $\uparrow$	Net MSE $\downarrow$
Mean Baseline	0.2882	0.7992	0.6223	0.0320	0.6336	0.0084
Random Baseline	0.0000	1.9856	0.0000	0.0620	-0.0010	0.0312
DiFuMo BSQ	—	—	0.6723	0.0323	—	—
Raw-Voxel VQVAE	0.3751	0.8426	0.6464	0.0331	0.7491	0.0072
Raw-Voxel BSQ	0.4199	0.8093	0.5929	0.0355	0.6339	0.0090
MS-VQVAE	0.3576	0.7933	0.6568	0.0299	0.7170	0.0067
Soft-Hier BSQ	0.5377	0.7001	0.8613	0.0165	0.9543	0.0022
Learned-Hier BSQ	0.5278	0.7116	0.8543	0.0176	0.9433	0.0024
Fixed-Hier BSQ	<u>0.5367</u>	<u>0.7006</u>	0.8882	0.0134	0.9688	0.0019

Table 1: Reconstruction metrics across six tokenizer families (higher correlation is better; lower MSE is better) and mean/random baselines. The best model for each metric is bolded; the second-best is underlined. Paired significance testing for *soft* vs. *fixed* hierarchical models (across 3200 windows; bootstrap CI + sign-flip permutation) shows a small but statistically significant ($p < 0.05$) voxel-level gain for soft with degradation at coarser scales.

Model	Linear Probing		Autoregressive Modeling	
	Age F1 (%) $\uparrow$	Sex Acc (%) $\uparrow$	Age F1 (%) $\uparrow$	Sex Acc (%) $\uparrow$
Yeo-Networks	38.9%	59.2%	—	—
DiFuMo	29.1%	59.8%	—	—
DiFuMo $\rightarrow$ PCA	31.0%	56.7%	—	—
DiFuMo BSQ	46.2%	66.5%	35.9%	58.6%
Raw-Voxel VQVAE	<u>42.9%</u>	64.6%	35.2%	52.1%
Raw-Voxel BSQ	32.4%	63.4%	43.2%	68.6%
MS-VQVAE	33.3%	66.6%	33.2%	57.1%
Soft-Hier BSQ	31.4%	68.9%	<u>43.4%</u>	61.4%
Learned-Hier BSQ	37.2%	<u>70.7%</u>	42.3%	<u>70.0%</u>
Fixed-Hier BSQ	39.3%	71.3%	44.6%	73.0%

Table 2: Downstream performance across baselines (using pooled Yeo-Network, DiFuMo, and DiFuMo $\rightarrow$ PCA features), and several tokenizer families under two evaluation paradigms. For age regression, higher F1 score is better; for sex classification, higher accuracy is better. Linear Probing uses mean-pooled token features with a linear head, while Autoregressive Modeling reports readout performance from a causal token model trained on the tokenized representations from each model.

tokenized representations. As shown in Table 2, hierarchical tokenization substantially improves downstream performance over conventional atlas-based baselines. In particular, the hierarchical BSQ models consistently achieve the strongest performance on sex classification, with the fixed hierarchical tokenizer achieving the best overall accuracy (**73.0%**) under autoregressive modeling. These results suggest that multiscale tokenization preserves demographic information more effectively than both standard parcellation and prior multiscale generative baselines.

Interestingly, the relative ordering of models differs between linear probing and autoregressive evaluation. While DiFuMo BSQ performs strongly under linear probing, hierarchical tokenizers consistently outperform competing methods when evaluated through autoregressive sequence modeling. This suggests that the hierarchical representations are particularly effective for modeling temporal brain dynamics, rather than simply producing linearly separable static embeddings.

Performance on age prediction is comparatively modest across all methods. However, it is important to note that HCP-YA contains a relatively narrow age distribution and provides age only in discretized bins rather than exact continuous values. We therefore formulate age prediction as a multi-class classification problem and report macro F1 scores. Despite the challenging setting, the hierarchical tokenizers remain among the strongest-performing approaches, with the fixed hierarchy consistently outperforming fully learned alternatives.

5.3 Learned vs. Fixed Hierarchies

To better understand the role of structured priors, we compare fully learned, fixed atlas-based, and softly aligned hierarchical tokenization strategies (Fig. 2). The soft hierarchy introduces a tunable alignment parameter σ controlling the tradeoff between learned flexibility and adherence to atlas structure. As shown in Fig. 2 (left), decreasing alignment strength slightly improves voxel-level reconstruction but consistently degrades parcel- and network-level reconstruction. The best-performing configurations at coarser scales use predefined atlas organization, reinforcing the importance of structured priors in modeling large-scale brain activity.

Because learned network indices are permutation-invariant, direct comparison between learned and atlas-derived hierarchies is not meaningful. Instead, we analyze global assignment structure and align learned network slots to atlas networks using Hungarian matching [17]. Interestingly, the learned hierarchy exhibits assignment distributions broadly similar to the fixed atlas hierarchy, with several dominant network slots containing disproportionately large mass (Fig. 2, top right). This suggests that the model partially recovers large-scale organizational structure even without explicit supervision. However, the aligned soft assignment matrix reveals that learned parcel-to-network mappings remain substantially more diffuse than atlas-derived organization (Fig. 2, bottom right). Rather than recovering sharply localized atlas-like partitions, learned hierarchies distribute mass across multiple networks, indicating that reconstruction objectives alone are insufficient to recover canonical functional organization.

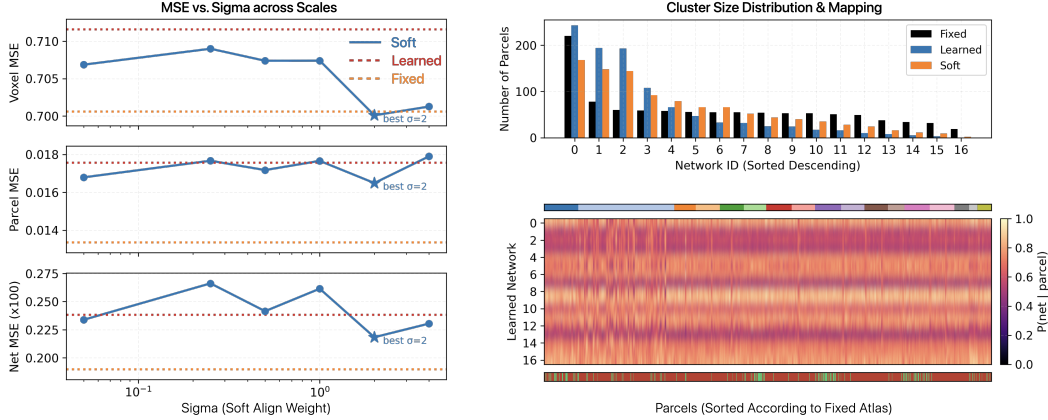


Figure 2: **Comparison of fixed, learned, and softly aligned hierarchical tokenization strategies.** (Left) Reconstruction error across voxel, parcel, and network scales as a function of the soft-alignment weight σ . Smaller σ values encourage stronger alignment to atlas priors, while larger values permit more flexible learned structure. The soft hierarchy achieves a marginal improvement in voxel-level reconstruction, but atlas-based fixed hierarchies consistently provide the best performance at coarser parcel and network scales. (Top right) Distribution of parcel assignments across learned network slots for fixed, learned, and soft hierarchies. Because learned network indices are permutation-invariant, we compare distributions rather than direct one-to-one correspondence. Both learned and soft hierarchies exhibit skewed assignment patterns similar to the atlas-based hierarchy, suggesting convergence toward structured large-scale organization. (Bottom right) Soft parcel-to-network assignment matrix after Hungarian alignment between learned and atlas-derived network indices. Parcels are sorted according to fixed atlas labels, and colors indicate the probability of assignment to each aligned learned network. The diffuse structure and sparse hard matches indicate that fully learned hierarchies only partially recover atlas-like organization, despite exhibiting similar global assignment statistics.

5.4 Model Efficiency

HBAR contains 5.6M parameters and processes a temporal window in 276 ms on a single H100 GPU in our implementation. This corresponds to roughly 18 seconds to tokenize an entire resting-state scan, making the approach practical for patient-level analysis and large-scale neuroimaging datasets alike while operating directly on voxel-level signals.

6 Discussion

We introduced a hierarchical tokenization framework for structured data and instantiated it for resting-state fMRI through a voxel-to-parcel-to-network hierarchy. Across reconstruction and downstream evaluation, hierarchical BSQ tokenizers substantially outperform conventional atlas-only representations and single-scale voxel tokenizers, demonstrating that compact discrete representations can retain information from voxel-level fMRI while remaining practical for large-scale modeling. Notably, hierarchical voxel-preserving tokenization consistently outperforms direct DiFuMo-based representations, suggesting that conventional parcel averaging discards fine-grained spatial variation that remains highly informative for both reconstruction and downstream modeling.

A second key finding is that structured neurobiological priors remain highly valuable. Although learned and soft hierarchies achieve competitive voxel-level reconstruction, they do not consistently recover atlas-like organization and generally underperform fixed atlas-based hierarchies at parcel/network scales and downstream tasks. This suggests that learned hierarchies may be sufficient for compression, but structured priors appear important when the goal is to obtain representations that are both biologically aligned and useful for downstream modeling. Our results therefore suggest that large-scale functional atlases encode organizational structure that remains difficult to recover through self-supervised reconstruction objectives, even in large-scale hierarchical representation learning.

More broadly, these findings highlight a useful design principle for representation learning in scientific domains. Unlike images or videos, where multiscale structure is often defined by spatial resolution, fMRI has meaningful organization across distributed functional systems. Our results indicate that hierarchical tokenization should not simply import grid-based multiscale designs, but should instead exploit domain structure. In this sense, atlas-based hierarchies are not merely preprocessing choices; they provide powerful inductive biases that can guide interpretable and effective token formation.

Limitations This study focuses on resting-state fMRI from HCP-YA, which has a relatively narrow young-adult age range and discretized age labels. As a result, age prediction is a challenging and limited downstream task. Our experiments also use a fixed atlas hierarchy based on DiFuMo-1024 and Yeo-17, and although these are well-established priors, other parcellations or subject-specific functional networks may yield different tradeoffs. Finally, while we evaluate autoregressive modeling and demographic prediction, we do not yet test broader clinical, cognitive, or cross-dataset transfer settings where the value of voxel-preserving tokenization may be even larger.

Future work Future work should evaluate hierarchical tokenization across more diverse neuroimaging cohorts, including lifespan, aging, and disease datasets. A particularly important direction is subject-adaptive tokenization, where atlas priors provide initialization or regularization while allowing individualized deviations in functional organization. Another promising direction is to couple the tokenizer more tightly to downstream autoregressive objectives, so that the hierarchy directly shapes temporal brain dynamics modeling. Finally, while this paper focuses on fMRI, the framework is naturally applicable to other structured scientific signals, including electrophysiology, spatial transcriptomics, and sensor networks, where meaningful units are distributed and multiscale.

Conclusion Hierarchical tokenization provides a compact and expressive way to represent high-dimensional voxel-level fMRI. By preserving sub-parcel information while organizing representations across parcel and network scales, our approach achieves strong compression, improves over conventional parcellation-based baselines, and supports downstream autoregressive and demographic prediction tasks. At the same time, our comparison of fixed, learned, and soft hierarchies shows that strong domain priors remain critical: established functional atlases provide structure that is difficult to recover from reconstruction alone. These results argue for a hybrid view of scientific representation learning, where learned tokenizers are most effective when paired with principled domain structure.

References

- [1] Mohammad Hassan Abbasi and Ehsan Adeli. fMRI Processing Pipeline: A lightweight, end-to-end framework for BIDS-compatible fMRI preprocessing and quality control, 2025.

- [2] Mohammad Adiban, Kalin Stefanov, Sabato Marco Siniscalchi, and Giampiero Salvi. Hierarchical Residual Learning Based Vector Quantized Variational Autoencoder for Image Reconstruction and Generation. 8 2022.
- [3] Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. Factorized Visual Tokenization and Generation. 11 2024.
- [4] Thomas Blumensath, Saad Jbabdi, Matthew F. Glasser, David C. Van Essen, Kamil Ugurbil, Timothy E.J. Behrens, and Stephen M. Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *NeuroImage*, 76:313–324, 8 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.03.024.
- [5] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: a Language Modeling Approach to Audio Generation. 7 2023.
- [6] 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, Insu Han, Amin Karbasi, Chadi G. Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings, 9 2023.
- [7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining From Pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 5 2020. URL <https://proceedings.mlr.press/v119/chen20s.html>.
- [8] Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J. Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, 221:117126, 11 2020. ISSN 10538119. doi: 10.1016/j.neuroimage.2020.117126.
- [9] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking. 9 2024.
- [10] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. 6 2021.
- [11] Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jesse Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 1 2019. ISSN 1548-7091. doi: 10.1038/s41592-018-0235-4.
- [12] Rahul Garg, Guillermo A. Cecchi, and A. Ravishankar Rao. Full-brain auto-regressive modeling (FARM) using fMRI. *NeuroImage*, 58(2):416–441, 9 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2011.02.074.
- [13] Eloy Geenjaer, Tonya White, and Vince Calhoun. Variational voxelwise rs-fMRI representation learning: Evaluation of sex, age, and neuropsychiatric signatures. 8 2021.
- [14] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention. 6 2021.
- [15] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. 12 2022.
- [16] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 2 2017.
- [17] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 3 1955. ISSN 0028-1441. doi: 10.1002/nav.3800020109.

- [18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [19] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive Image Generation using Residual Quantization. 3 2022.
- [20] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite Scalar Quantization: VQ-VAE Made Simple. 10 2023.
- [21] Yuto Nishimura, Masataka Sawayama, Ayumu Yamashita, Hideki Nakayama, and Kaoru Amano. BrainCodec: Neural fMRI codec for the decoding of cognitive brain states. 10 2024.
- [22] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. 5 2018.
- [23] François Paugam, Basile Pinsard, Guillaume Lajoie, and Pierre Bellec. A benchmark of individual auto-regressive models in a massive fMRI dataset. *Imaging Neuroscience*, 2, 7 2024. ISSN 2837-6056. doi: 10.1162/imag{_}a{_}00228.
- [24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 5 2021. URL <https://proceedings.mlr.press/v139/ramesh21a.html>.
- [25] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. 6 2019.
- [26] Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable Image Tokenization with Index Backpropagation Quantization. 3 2025.
- [27] 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. 6 2025.
- [28] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3): 1125–1165, 9 2011. ISSN 0022-3077. doi: 10.1152/jn.00338.2011.
- [29] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. 6 2024.
- [30] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, 10 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.041.
- [31] Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Xue-Jun Kong, Quanzheng Li, Daniel S. Barron, Anqi Qiu, Randy Hirschtick, Byung-Hoon Kim, Hongbin Han, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for fMRI analysis. 3 2026.
- [32] 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. 1 2026.
- [33] Yanwu Yang and Thomas Wolfers. Hierarchical Characterization of Brain Dynamics via State Space-based Vector Quantization. 6 2025.
- [34] 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*, 43(11):4004–4016, 11 2024. ISSN 0278-0062. doi: 10.1109/TMI.2024.3414476.

- [35] Xuanhua Yin, Runkai Zhao, Lina Yao, and Weidong Cai. BrainVista: Modeling Naturalistic Brain Dynamics as Multimodal Next-Token Prediction. 2 2026.
- [36] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical Graph Representation Learning with Differentiable Pooling. 2 2019.
- [37] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. 3 2024.
- [38] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An Image is Worth 32 Tokens for Reconstruction and Generation. 6 2024.
- [39] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and Video Tokenization with Binary Spherical Quantization. 6 2024.
- [40] Yue Zhao, Hanwen Jiang, Zhenlin Xu, Chutong Yang, Ehsan Adeli, and Philipp Krähenbühl. Spherical Leech Quantization for Visual Tokenization and Generation. 12 2025.
- [41] Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99%. 6 2024.