Multi-Task Collaborative Pre-Training and Adaptive Token Selection: A Unified Framework for Brain Representation Learning

Ning Jiang, Gongshu Wang, Chuyang Ye, Tiantian Liu, and Tianyi Yan, *Member, IEEE*

***Abstract*—Structural magnetic resonance imaging (sMRI) reveals the structural organization of the brain. Learning general brain representations from sMRI is an enduring topic in neuroscience. Previous deep learning models neglect that the brain, as the core of cognition, is distinct from other organs whose primary attribute is anatomy. Capturing the high-level representation associated with inter-individual cognitive variability is key to appropriately represent the brain. Given that this cognition-related information is subtle, mixed, and distributed in the brain structure, sMRI-based models need to both capture fine-grained details and understand how they relate to the overall global structure. Additionally, it is also necessary to explicitly express the cognitive information that implicitly embedded in local-global image features. Therefore, we propose MCPATS, a brain representation learning framework that combines Multi-task Collaborative Pre-training (MCP) and Adaptive Token Selection (ATS). First, we develop MCP, including mask-reconstruction to understand global context, distort-restoration to capture fine-grained local details, adversarial learning to integrate features at different granularities, and age-prediction, using age as a surrogate for cognition to explicitly encode cognition-related information from local-global image features. This co-training allows progressive learning of implicit and explicit cognition-related representations. Then, we develop ATS based on mutual attention for downstream use of the learned representation. During fine-tuning, the ATS highlights discriminative features and reduces the impact of irrelevant information. MCPATS was validated on three different public datasets for brain disease diagnosis, outperforming competing methods and achieving accurate diagnosis. Further, we performed detailed analysis to confirm that the MCPATS-learned representation captures cognition-related information.**

***Index Terms*—Brain representation learning, multi-task collaborative pre-training, adaptive token selection, brain structure-behavior relationships, MRI.**

I. Introduction

LEARNING general and robust brain representations from structural magnetic resonance imaging (sMRI) is an enduring problem that underlies many neuroscientific analyses (e.g., brain disease diagnosis, structural segmentation, brain structure-behavior association establishment). Previous studies have focused on manually extracting and selecting morphological estimates of brain structure, such as grey matter (GM) volume, cortical thickness, and surface area [1], [2], [3]. To precisely depict the complex geometric patterns of brain structure, deep brain representation learning has recently gained traction in brain imaging analysis [4], [5], [6], [7]. These approaches tend to follow a common research paradigm: combining a popular deep representation learning model (generative adversarial networks (GANs) [8], variational autoencoder (VAE) [9], etc.) with a biological prior, e.g., topological invariance of the anatomical structure [6], which is inserted into deep learning (DL) frameworks as a customized regularization or module. Afterward, they verify the learned representations via general approaches, such as downstream tasks (e.g., Alzheimer's disease (AD) diagnosis, tumor segmentation), location of important brain regions, and inner-group comparison.

However, the neuroanatomical basis of the brain structure supports extensive dynamic activities and ultimately generates a wide range of cognitive functions, such as thinking, reasoning, decision-making [10]. Neurological diseases, psychiatric diseases, brain injuries, strokes, cerebral haemorrhages, and other brain abnormalities all exhibit cognitive deficits as their outward indicators. Therefore, capturing representations that can reflect the brain's built-in patterns and encode rich cognitive information is key to brain representation learning. During model evaluation, verification of the learned brain representations is also need to be grounded in cognition.

Recently, several convincing articles highlighted that the geometry of brain structure shape brain function and provide a significantly greater amount of advanced bioinformatics than one would anticipate [11], [12], [13]. This enables the extraction of cognition-related brain representations from sMRI. Considering that the brain is made up of countless interacting neurons

Manuscript received 18 November 2023; revised 8 May 2024; accepted 8 June 2024. Date of publication 18 June 2024; date of current version 6 September 2024. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2023B0303030002, in part by STI 2030-Major Projects under Grant 2022ZD0208500, in part by the National Natural Science Foundation of China under Grant U20A20191 and Grant 62336002, and in part by China Postdoctoral Science Foundation under Grant 2023TQ0027. *(Ning Jiang and Gongshu Wang are co-first authors.) (Corresponding author: Tianyi Yan.)*

Ning Jiang, Gongshu Wang, Tiantian Liu, and Tianyi Yan are with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: ningjiang@bit.edu.cn; gongshu@bit.edu.cn; tiantian2bit@bit.edu.cn; yantianyi@bit.edu.cn).

Chuyang Ye is with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: chuyang.ye@bit.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2024.3416038, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2024.3416038

(approximately 10^{15} paired connected neurons), the portion of the structural variations that dominate inter-individual cognitive variations are subtle, mixed, and distributed in brain sMRI [14], [15], [16], [17]. To comprehensively capture informative features, it's essential for DL models to identify fine-grained features and understand how they relate to the overall global structure. In addition, DL models need to expose the cognitive information that is implicitly embedded within these structural features. In other words, a deep brain representation model should take into account local detail recognition, global context understanding, and cognitive information anchoring simultaneously.

Therefore, the existing studies have three limitations:

1) A single training task domain with a highly biased optimization objective is not conducive to capturing both local detail and global context, much less explicitly encoding cognitive information.
2) The researchers focused on training an efficient feature extractor, but overlooked the selecting of individual task-specific features, which is crucial given the heterogeneity of diverse downstream tasks and the inter-individual differences in the human brain.
3) The explained variance of brain representations for cognition is neglected. Their representation analysis, such as localization of brain regions or inter-group comparison, are not cognitively grounded and thus not sufficient to directly justify the plausibility of the learned representations.

A series of studies have confirmed the power of the Vison Transformer (ViT) [18] to capture both global and local features. This ability makes ViT innately suitable for brain representation learning. The long-range "receptive field" of ViT enables it to locate subtle changes and establish their contextual relationships to address the abovementioned challenges. We propose a ViT-based MCPATS, which leverages Multi-task Collaborative Pre-training (MCP) that enables the encoder to optimize for multiple objectives and acquire diverse and complementary representations from various task domains. Having already extracted the robust representations through the pre-trained encoder, Adaptive Token Selection (ATS) realized by a mutual attention mechanism is utilized to filter task-specific features for different downstream tasks.

Specifically, we first reshape an input sMRI into several patches, and a random mask divides the patches into visible patches and masked patches. The pre-training stage has four proxy tasks: (1) Infer the mask target from the visible context, forcing the model to understand long-range contextual semantic information. (2) Restore the distorted visible patches back to the original voxels, encouraging the model to focus on the structural information hidden in local image details. (3) Adversarial learning to integrate the global and local features from the first two tasks, driving the model to establish a latent space that matches the distribution of realistic brain data. (4) Predict the age of the input sample from the visible patches, naturally exposes information generated during brain development and aging, which is explicitly and significantly shared with cognitive function [19], [20]. Through the synergy and co-training of the four proxy tasks, the model is incorporated into an effective constraint and is optimized for a more complex and difficult objective, which facilitates complementary learning of general brain representation. A shared ViT encoder, three independent lightweight decoders, and a convolutional neural network (CNN) discriminator are used for pre-training. Then, we transfer the pre-trained encoder to downstream tasks. For fine-tuning, we propose a mutual-attention-based method to adaptively and individually select important tokens from each transformer layer. Finally, the selected tokens are fed into the classifier as identified discriminative features. Three challenging downstream tasks are leveraged to validate the capacity of our proposed MCPATS: (1) Attention deficit hyperactivity disorder (ADHD) diagnosis. (2) Schizophrenia (SZ) diagnosis. (3) Preclinical AD identification. Unlike neurological diseases, e.g., AD, with obvious morphological changes caused by brain atrophy, psychiatric diseases, such as ADHD and SZ, show subtle structural alterations and thus are difficult to distinguish. For subjects we defined as "preclinical AD", all the parameters, physical assessments, and clinical data state that the patient is healthy and showing no symptoms, but brain neuronal structure has started to deteriorate [21]. The MCPATS outperforms several compared methods in challenging downstream tasks, showing superiority in representation learning from sMRI and fine-grained categorization. Furthermore, we verify the interpretability of the proposed model, using not only common representation analysis but also validation of the associations between brain representations and cognitive functions. Our main contributions are summarized as follows:

1) We propose MCPATS, a unified framework that combines multi-task collaborative pre-training to mutually reinforce the encoder in extracting general and robust representations from brain sMRI, and adaptive token selection to filter task-specific features. The MCPATS exhibits superior performance in challenging downstream tasks.
2) Given that the purpose of representing the brain is to accurately delineate its properties and, ultimately, encode cognitive information, we studied 12 behavioral tasks to explore the direct causal correlation between learned brain representation and cognitive functions. Our results demonstrate that the proposed model represents the high-level information relevant to cognitive functions involving comprehension, reasoning, memory, emotion, and learning.
3) We explore how different types of pre-training components affect the representation learning capacity of ViT, revealing the promising potential of ViT for both competitive performance and interpretability. Such endeavors, we envision, will inspire future studies of tailor-made deep brain representation learning frameworks using ViT as the backbone instead of CNNs.

II. Related Work

A. Multi-Task Learning of Brain Representations

Multi-task learning (MKL) and collaborative learning have been widely employed in visual semantic learning [22], [23],

and there have been attempts to deploy MKL for medical image analysis, especially for medical segmentation [24], [25]. In recent years, MKL has also been involved in brain representation learning, such as brain stroke lesion segmentation [26], brain tumor segmentation [27], [28], and brain disease diagnosis [29]. Specifically, [26] adopted stroke lesion mask-prediction as the main task and stroke lesion edge-prediction as the auxiliary task for segmentation of stroke lesions. [27] integrated three tasks (i.e., coarse segmentation of the complete tumor, refined segmentation for the complete tumor and its intra-tumoral classes and precise segmentation for the enhancing tumor) into one framework and achieved state-of-the-art performance on two datasets. [28] utilized a shared encoder to extract features and two task-specific decoders, a segmentation decoder and an auxiliary fusion decoder for better segmentation performance. [29] added two tasks (prediction of the Mini-Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog)) to their basic AD diagnosis task for higher classification accuracy.

The abovementioned multi-task brain representation learning approaches, in fact, trained their models via multiple tasks but were limited in a single domain and optimized by biased objectives. For example, all the tasks in [26] and [27] are mask-prediction tasks whose essence is pixel-classification, and all the tasks in [29] are label-prediction tasks. Given that tasks of different domains (e.g., restorative learning, adversarial learning) have different implications for representation learning, this limitation may hinder their models from learning effective and diverse brain representations.

B. Attention-Guided Feature Selection

Inspired by the successes of attention mechanisms in CNNs, several studies have employed attention guidance to identify anatomically meaningful regions in the brain [30], [31], [32]. The H-FCN [30] generated location proposals using anatomical biomarkers as prior knowledge and then selected discriminative regions from a whole brain MRI via an attention mechanism for accurate AD diagnosis. Similar to [30], HybNet [31] also selected discriminative regions for AD diagnosis, but its feature selection was guided by a class activation map (CAM), which was extracted from pre-trained models. In addition, the selected multi-scale features were further fused for the classification task. The DA-MIDL [32] model consists of patch-net with a spatial attention block, attention multi-instance learning pooling and an attention-aware global classifier to construct an effective classification model for AD diagnosis.

However, existing attention-guided methods were designed for CNNs and introduced extra parameters thus are inapplicable to ViT based models. At the core of ViT, the self-attention mechanism may be inherently capable to generate guidance for feature selection, without introducing any extra parameters.

C. Linking Brain Structure to Behavior

Understanding the brain structural correlations of inter-individual differences in behavior (especially cognition) is an important objective in neuroscience. Recently, searching for direct associations between local measurements of neuroanatomy and behavioral variables has gained traction in brain structure-behavior studies [33], [34], [35]. [33] attempted to link the Big Five personality traits of openness-to-experience with variability in brain structural features (cortical thickness, surface area, subcortical volume and white matter microstructural integrity) but producing inconsistent results. [34] studied how children's family income gaps in cognition relate to the volume of the anterior and posterior hippocampus. They observed that the anterior, but not posterior hippocampus mediates income-related differences in cognitive scores. [35] used a fully probabilistic approach to explore the variation in regional GM volume depending on sex and social traits and provide a population-level window into the brain associations with social behavior. These methods usually use manual features, whereas the high-level features extracted by DL models are coarse, abstract, and hard to understand at the level of human knowledge; hence, current research on the link between deep brain representation and behavior is scarce.

III. Method

A. MCPATS Framework

1) Multi-Task Collaborative Pre-Training: Restorative Learning Our restorative learning framework consists of two branches: parallel distillation masked image modeling (named branch A) and visible portion restoration (named branch B). Parallel distillation, using semantic diversity regularization inspired by the brain's prior knowledge to guide masked image modeling, was shown to be effective to learn the brain's built-in pattern in our previous work [36]. Therefore, we introduce branch A to encourage the encoder to distill global semantic information for comprehensive understanding of the brain. Conversely, branch B is exploited to enhance fine-grained representation learning, especially for subtle, local variations. As shown in Fig. 1(a), we first reshape an input sMRI x into 3D patches x_{all} without overlapping. Then, random masking divides x_{all} into two portions: x_{vis} and x_{mask}.

For branch A, x_{vis} is processed by the shared encoder E to generate the latent feature $y_{vis} = E(x_{vis})$. The projection heads h_{A1} and h_{A2} map y_{vis}into latent embeddings $z_{vis_{A1}}$ and $z_{vis_{A2}}$. We expect the two latent embeddings to represent opposing semantic information, and a semantic diversity loss is designed to disentangle the two in the latent space. Specifically, we compute the Gram matrices $G_{vis_{A1}} = z_{vis_{A1}} \bullet z^T_{vis_{A1}}$ and $G_{vis_{A2}} = z_{vis_{A2}} \bullet z^T_{vis_{A2}}$, and the semantic diversity loss is:

$$L_{sd} = -log\left(\sigma\left(G_{vis_{A1}}\right)\right) - log\left(1 - \sigma\left(G_{vis_{A2}}\right)\right) \quad (1)$$

This loss results in the mean of $G_{vis_{A1}}$ infinitely large and the mean of $G_{vis_{A2}}$ infinitely small. Consequently, $z_{vis_{A1}}$ focuses on regions that are interconnected and extract semantically similar components from them. $z_{vis_{A2}}$ focuses on regions that are relatively independent and extract semantically dissimilar components from them. This semantic diversity loss enables the model to achieve a more comprehensive and in-depth understanding of the brain.

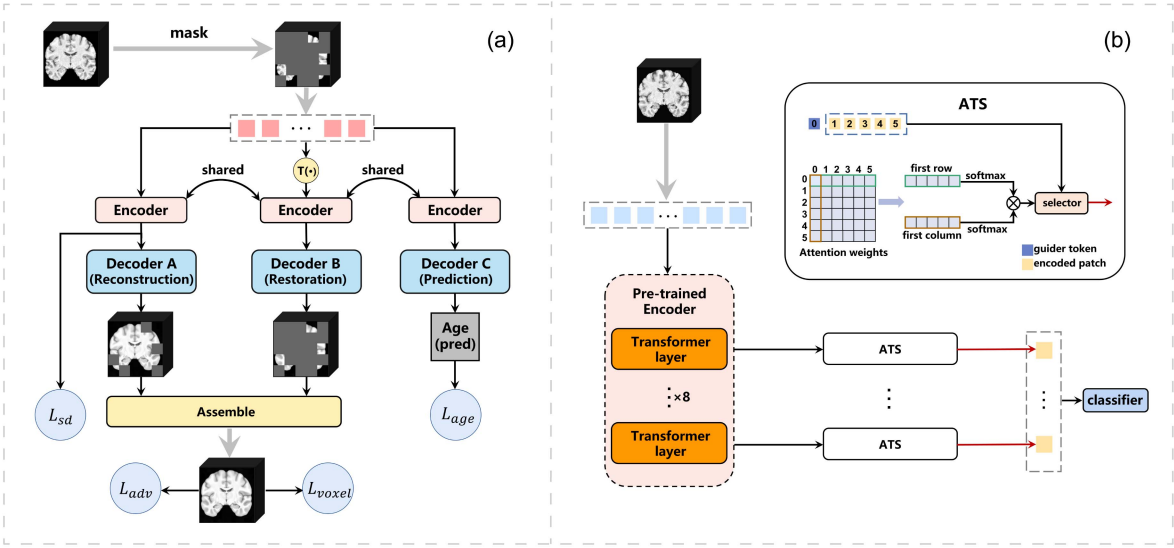

Fig. 1. The overall architecture of MCPATS. (a) Multi-task collaborative pre-training (MCP), where $T(\cdot)$ is a transformation function to distort the visible patches and "Assemble" denotes merging the reconstructed masked patches and the restored visible patches into their original positions. L_{sd} denotes the semantic diversity loss, L_{voxel} denotes the voxel-prediction loss, L_{adv} denotes the adversarial loss, and L_{age} denotes the age-prediction loss. (b) Mutual attention mechanism based adaptive token selection (ATS) for fine-tuning. $\otimes$ denotes Hadamard product.

All of the tokens are fed into the reconstruction decoder D_A, including $z_{vis_{A1}}$, $z_{vis_{A2}}$, and the masked tokens z_{mask}, which are learnable tokens put in the positions of x_{mask}, and the reconstructed masked patches are output:

$$x_{mask\_recon} = D_A\ (z_{vis_{A1}} \oplus z_{mask}) - D_A\,(z_{vis_{A2}} \oplus z_{mask}) \tag{2}$$

where $\oplus$ denotes the concatenation operation of tensors.

For branch B, we first distort the visible patches x_{vis} by a function T. The goal of the encoder E, the projection head h_B and the restoration decoder D_B is to map the distorted patches back to the original ones. The output can be described as:

$$x_{vis\_restore} = D_B\ (h_B\,(E\,(T\,(x_{vis})))) \tag{3}$$

Note that T is a nonlinear transformation function realized by Bézier curves [37].

Finally, an entire image is assembled with the reconstructed masked patches and the restored visible patches:

$$x_{full\_restore} = ASS[x_{mask\_recon}\ ;\ x_{vis\_restore}] \tag{4}$$

where ASS denotes the patches are assembled in their original positions to form an entire image.

The voxel-prediction loss can be described as:

$$L_{voxel} = \ MSE\,(x_{full\_restore}, x) \tag{5}$$

Adversarial Learning: Considering the voxel-prediction loss is calculated at each isolated voxel, this hinders the information interaction and knowledge sharing between the two restorative learning branches in the output stage. We employ adversarial learning to integrate features learned by different branches, using a discriminator D_{adv} to distinguish the restored images (assembled with the outputs of the two branches) from the original ones. The adversarial loss [8] is:

$$L_{adv} = \ log\,(D_{adv}\,(x))\ + log\,(1 - D_{adv}\,(x_{full\_restore})) \tag{6}$$

Age Prediction Learning: Age information is easily available, and cognitively meaningful phenotypes emerge naturally through age prediction [20]. Information associated with aging and development is also related to cognition and pathology [19]. Given that obtaining estimates of subjects' various cognitive abilities is costly and difficult, we use age as a surrogate for cognitive ability to help the model explicitly encode cognition-related information. The goal of E and the age-prediction decoder D_{age} is to predict the age of the input sample using the visible patches x_{vis}, which can be described as:

$$x_{age\_pred} = D_{age}\ (E\,(x_{vis})) \tag{7}$$

$$L_{age} = \ MSE\,(x_{age\_pred}, x_{age\_target}) \tag{8}$$

Multi-task Collaborative Pre-training: Finally, the overall objective of pre-training becomes:

$$\begin{aligned} L = \lambda_{sd}\ * L_{sd} + \lambda_{voxel} * L_{voxel} \\ +\lambda_{adv} * L_{adv} + \lambda_{age} * L_{age} \end{aligned} \tag{9}$$

where λ_{sd}, λ_{voxel}, λ_{adv} and λ_{age} are hyperparameters that determine the importance of losses. Through collaborative pre-training, E can extract comprehensive representations containing not only local-global image features but also cognitive information that is a fundamental property of the brain. In particular, λ_{sd} helps the model more deeply learn diverse semantic representations. λ_{voxel} simultaneously forces the model to infer the masked target by aggregating visible context, which facilitates an understanding of the global context, and to restore the distorted voxels in visible patches, which facilitates the

capture of fine-grained local features. λ_{adv} integrates features at different granularities in different restorative learning branches to capture more significant features. Finally, λ_{age} enhances the cognitive plausibility of the learned representations.

2) Adaptive Token Selection: For fine-tuning, a new classifier with initialized weights is added to the pre-trained encoder, which can extract subtle yet discriminative features but needs to further highlight and strengthen task-specific features in fine-tuning. We propose a mutual-attention-based adaptive token selection method and insert it into the pre-trained ViT backbone, where we adaptively and individually select the informative tokens from each transformer layer (see Fig. 1(b) for more details).

Our proposed token selection method introduces an additional independent guider token G to identify important tokens. Similar to the class token in [18], the initialized G does not contain any information about the input but interacts with all patch tokens layer by layer. Unlike ViT, G is only used to guide token-selection and is not fed into the classifier. The input of the classifier is the set of selected tokens. We are motivated by this assumption: the guider token G aggregates the global context from all patch tokens, thus containing information that can describe the holistic semantics and style of the input sMRI. Meanwhile, G stays out of the classifier, so it has no preference for tokens containing more low-level structural information but becomes an impartial guider that is rich in global information. Therefore, a set of tokens, selected from each transformer layer, which is the most relative to G, can be seen as the most discriminative representation of the input. A mutual attention mechanism is used to compute the importance score of tokens:

$$A = \begin{bmatrix} a_{00} & \cdots & a_{0j} & \cdots & a_{0N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i0} & \dots & a_{ij} & \cdots & a_{iN} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{N0} & \cdots & a_{Nj} & \cdots & a_{NN} \end{bmatrix} \tag{10}$$

$$r_i = [a_{i0},\ a_{i1},\ a_{i2}, \dots,\ a_{ij}, \dots,\ a_{iN}] \tag{11}$$

$$c_j = [a_{0j},\ a_{1j},\ a_{2j}, \dots,\ a_{i,j}, \dots,\ a_{Nj}]^T \tag{12}$$

where $A \in \mathbb{R}^{(N+1)*(N+1)}$ denotes an attention weight matrix for one attention head, and N is the number of tokens. r_i denotes row i of A, c_j denotes column j of A, and a_{ij} is the attention weight between token i and token j in the context of token i. We define the guider token G as token 0, in which case a_{0j} denotes the attention weight between token G and token j in the context of token G, and a_{i0} denotes the attention weight between token G and token i in the context of token i. Therefore, r_0 is the **attention weight vector** between token G and other tokens in the context of **token** G compared with c_0 in the context of **other tokens**. We perform Softmax operation on r_0 and c_0 to compute two normalized attention score vectors:

$$S_{G-to-n} = \frac{e^{a_{0n}}}{\sum_{j=0}^{N} e^{a_{0j}}} \tag{13}$$

$$S_{n-to-G} = \frac{e^{a_{n0}}}{\sum_{i=0}^{N} e^{a_{i0}}} \tag{14}$$

and the mutual attention score S_n between token G and token n is:

$$S_n = S_{G-to-n} \ * S_{n-to-G} \tag{15}$$

A higher mutual attention score S_n indicates that token n is more similar to the guider token G that is computed interactively with all tokens at each layer. This implies that token n contains richer global information. We average the attention weights of all the heads to compute the mutual attention scores. Then, the tokens are sorted by their scores, and we select the top tokens from each layer according to their indices. Finally, all selected tokens from a sample x are fed into the classifier, but not the guider token G. It is essential to clarify that the input to each transformer layer is the full set of tokens from the previous layer, rather than the tokens that were selected in the previous layer.

B. Experiments

1) Dataset Description and Data Partition: The brain structural images (T1w) were provided by the Cam-CAN (http://www.cam-can.com/.) [38], ADHD-200 (http://fcon_1000.projects.nitrc.org/indi/adhd200/) [39], MCIC (http://www.schizconnect.org/) [40], OASIS-3 (http://www.oasis-brains.org/) [41] and CNP (https://openneuro.org/datasets/ds000030/) [42]. We performed image preprocessing using sMRIprep. The T1w image was corrected for intensity nonuniformity with N4 bias correction and was skull-stripped using the antsBrainExtraction workflow. Volume-based spatial normalization to standard space (MNI152NLin2009cAsym, $181 \times 217 \times 181$ voxels) was performed through nonlinear registration with antsRegistration. Then, we cut off part of the black background at the edges, and a main area ($150 \times 180 \times 150$ voxels) was reserved.

In this study, 636 cognitively healthy adults from Cam-CAN were used for collaborative pre-training. For ADHD-200, we used the same official training/testing sets as in previous studies: 768 training subjects (488 normal control subjects (NCs), 280 ADHD patients) and 171 test subjects (94 NCs, 77 ADHD patients) collected from eight independent sites. For the MCIC and OASIS datasets, we performed ten-fold cross-validation on 190 subjects (99 NCs, 91 SZ patients) and 280 subjects (190 NCs, 90 preclinical ADs) and reported the averaged accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic (ROC) curve (AUC).

2) Experimental Settings: Training Protocol In pre-training, the model is optimized by AdamW with a batch size of 16 and a weight decay parameter of 0.05 on four NVIDIA Tesla P100 GPUs using PyTorch. The model is pre-trained for 3000 epochs (40 warm-up epochs) with a cosine linear-rate scheduler, and the initial learning rate is set as $lr \times BatchSize/4$, where $lr = 1.5e^{-4}$. We empirically set the mask ratio to 0.75, the patch size to $30 \times 30 \times 30$, the nonlinear rate for the function T to 0.9, and λ_{sd}, λ_{voxel}, λ_{adv}, λ_{age} to 0.005, 0.8, 0.1, and 0.1, respectively. The backbone of the shared encoder E is a ViT composed of 8 transformer blocks using 12 heads for multi-head self-attention (MSA) and 512 as the hidden size. The decoders D_A, D_B and

TABLE I
EXPERIMENTAL RESULTS (%) ON ADHD-200

Method	Features	ACC	AUC	SEN/SPE
Quotes results from the literature (sMRI-based)				
3D-CNN[50]	*GM density*	*65.86*	-	-
Tensor Boosting[51]	*sMRI*	*69.01*	-	-
3D-CNN[2]	*FDCM of GM density*	*69.01*	-	-
Quotes results from the literature (fMRI-based)				
3D-CNN[50]	*fALFF*	*66.04*	-	-
FC-HAT[52]	*FBN*	*69.2*	-	*83.0/46.8*
EM-MI[53]	*fMRI*	*70.4*	-	-
4D-CNN[54]	*fMRI*	*71.3*	-	-
SASNI[55]	*fMRI*	*72.5*	-	-
Quotes results from the literature (Multimodality)				
MKL[56]	*ReHo,GM and cortical thickness*	*61.54*	-	-
3D-CNN[50]	*fALFF and GM density*	*69.15*	-	-
MGF[57]	*sMRI, fMRI*	*71.9*	-	-
Results obtained by this paper				
ResNet50	sMRI	69.01	65.06	57.14/78.72
ViT	sMRI	64.91	64.49	40.26/85.11
DA-MIDL	sMRI	61.99	62.59	62.34/61.70
Incep_Resnet	sMRI	65.50	69.23	38.96/87.23
Top-K	sMRI	70.76	69.61	53.25/85.11
TransFG	sMRI	71.35	72.42	62.34/78.72
$MCPATS_2$	sMRI	72.51	74.14	55.84/86.17
$MCPATS_3$	sMRI	**74.27**	74.79	61.04/85.11
$MCPATS_4$	sMRI	73.68	**75.79**	57.14/87.23

* The rows in italics have their numbers copied from the studies using the same training/testing sets as ours. Note that $MCPATS_N$ denotes that the model selects N tokens per layer (totally $8N$ tokens per sample) in fine-tuning. Values in bold represent the best performance for each metric.

TABLE II
HYPERPARAMETER ANALYSIS OF MCPATS

Dataset	λ_{sd}	λ_{voxel}	λ_{age}	λ_{adv}	T_N	ACC	AUC	SEN/SPE
ADHD 200	0.05	0.70	0.15	0.15	2	71.35	74.52	64.94/76.60
					3	**73.68**	**75.48**	61.04/84.04
					4	70.76	74.61	55.84/82.98
	0.05	0.80	0.10	0.10	2	72.51	74.14	55.84/86.17
					3	**74.27**	74.79	61.04/85.11
					4	73.68	**75.79**	57.14/87.23
	0.05	0.90	0.05	0.05	2	**73.68**	**75.97**	54.55/89.36
					3	72.51	74.05	53.25/88.30
					4	73.10	75.09	54.55/88.30

* T_N denotes that the model selects N tokens per layer in fine-tuning.
Values in bold represent the best performance for each metric.

TABLE III
EXPERIMENTAL RESULTS (%) ON MCIC

Method	Features	ACC	AUC	SEN/SPE
ResNet50	sMRI	77.37±7.87	69.22±16.14	78.00±22.50/76.67±19.21
ViT	sMRI	70.00±9.62	72.89±9.27	81.00±11.01/57.78±20.15
DA-MIDL	sMRI	64.74±6.10	66.25±10.29	64.00±22.71/65.56±17.72
Incep_Resnet	sMRI	72.63±6.93	69.44±10.05	73.00±18.89/72.22±13.10
Top-K	sMRI	76.32±6.20	75.78±10.52	75.00±13.54/77.78±12.82
TransFG	sMRI	78.95±7.45	76.55±10.05	78.00±11.60/78.89±11.05
$MCPATS_3$	sMRI	**80.00±6.47**	**78.67±11.28**	79.00±13.70/81.11±12.88

Values in bold represent the best performance for each metric.

TABLE IV
EXPERIMENTAL RESULTS (%) ON OASIS

Method	Features	ACC	AUC	SEN/SPE
ResNet50	sMRI	78.93±6.62	74.21±17.75	62.22±23.54/76.67±10.89
ViT	sMRI	74.64±2.63	63.71±12.34	45.68±16.94/87.72±8.52
DA-MIDL	sMRI	75.00±5.32	61.08±15.57	50.00±24.71/86.84±7.94
Incep_Resnet	sMRI	78.21±8.82	77.95±9.62	55.56±19.60/86.84±7.55
Top-K	sMRI	82.14±2.38	78.19±8.54	70.37±12.88/87.72±7.46
TransFG	sMRI	81.35±2.82	77.93±9.34	67.90±17.73/87.72±7.87
$MCPATS_3$	sMRI	**83.57±1.84**	**79.98±8.05**	71.61±19.03/89.47±7.85

Values in bold represent the best performance for each metric.

D_{age} are composed of a single transformer block using 12 heads for MSA and 512 as the hidden size. The discriminator D_{adv} is ResNet-10, which is strictly the same as in [43].

In fine-tuning, we transfer the pre-trained encoder E to downstream tasks by fine-tuning all of the parameters. Unlike previous methods [44], [45], [46], each layer is equally important in MCPATS, so layer-decay is not used in fine-tuning.

Baselines. Our proposed MCPATS is compared with four methods, including two general baselines (ResNet50 [43] and ViT [18]) and two state-of-the-art models designed for a specific brain disease (DA-MIDL [32] and Inception-Resnet [47]). Additionally, the proposed ATS is compared with two attention matrix-based feature selection method, Top-K [48] and TransFG [49]. We used Top-K and TransFG for feature selection during model fine-tuning and adopted the same pre-trained model as our proposed ATS (the encoder pre-trained by MCP) for a fair comparison.

Implementation of baselines and our proposed MCPATS is available at: https://github.com/NingJiang-git/MCPATS

IV. RESULTS AND DISCUSSION

A. Classification Performances

The classification results of ADHD-200 are shown in Table I. The MCPATS achieves the best accuracy (74.27%) compared to baselines that adopted the official partition of training/testing sets using sMRI, with up to 5.26% improvement in accuracy. Our proposed MCPATS even surpasses most of the approaches from the literature that use functional features of the brain (usually functional connectivity estimates) or multimodality features, such as 4D-CNN, SASNI and MGF.

Given that weights of various losses (the values of λ_{sd}, λ_{voxel}, λ_{adv} and λ_{age}) and the number of selected tokens may simultaneously influence the model's performance, we have provided our search space and sensitivity analysis results for these hyperparameters in Table II. For ADHD classification, our method achieves the best performance when setting λ_{sd}, λ_{voxel}, λ_{adv}, λ_{age} to 0.005, 0.8, 0.1, and 0.1 and selecting three tokens per layer, which is the default setting in subsequent experiments.

Note that MCPATS always reports a competitive accuracy (over 70.76%) and AUC (over 74.05%) with different pre-training or fine-tuning settings, indicating that the model has established robust representations for brain sMRI images.

The classification results of the MCIC and OASIS datasets are shown in Tables III and IV. The impressive performance demonstrates the generalization of our proposed MCPATS and suggests that the model establishes a strong latent space that semantically interprets the brain's built-in pattern to encode those implicit features into an explicit form. As mentioned in Section I, the capacity of MCPATS to discriminate preclinical

TABLE V
ABLATION STUDY RESULTS (%) OF MCPATS

Ablation Mode	Pre-training.				Fine-tuning.	ACC		
	Res_A	Age	Res_B	Adv	IAT	ADHD-200	MCIC	OASIS
mode0	×	×	×	×	✓	65.50	74.21±4.61	74.64±2.63
mode1	✓	×	×	×	✓	66.08↑	77.37±7.47↑	77.86±3.69↑
mode2	✓	✓	×	×	✓	66.67↑	78.42±8.02↑	79.64±2.94↑
mode3	✓	✓	✓	×	✓	70.18↑	76.32±7.55↓	79.28±4.70↓
mode4 (w/o ATS)	✓	✓	✓	✓	×	69.01	78.42±8.02	78.21±4.59
mode4 (proposed)	✓	✓	✓	✓	✓	**74.27↑**	**80.00±6.47↑**	**83.57±1.84↑**

* Res_A and Res_B denote restorative learning branches A and B. ATS denotes adaptive token selection in fine-tuning.
Values in bold represent the best performance for each metric.

AD from healthy people demonstrates that MCPATS has captured the deterioration of brain neuronal structure at its nascent stage, which has important clinical significance.

B. Ablation Study

We perform a thorough ablation study to show how each component contributes to MCPATS. For pre-training, we start with a ViT backbone and incrementally add restorative learning branch A, age prediction learning, restorative learning branch B, and adversarial learning. For fine-tuning, we compare MCPATS with ViT without ATS when all four pre-training components are applied.

The ablation results are shown in Table V. We make the following observations: (1) In pre-training, each component significantly improves the performance of ADHD diagnosis, whereas in SZ diagnosis and preclinical AD identification, there is slight performance degradation when adding restorative learning branch B. Note that this gap is later compensated after adding adversarial learning, which indicates that collaborative pre-training, especially the unification of restorative learning and adversarial learning, achieves a complementary harmony and thus has superior performance in downstream tasks. This conclusion could be verified by ablation on ADHD-200. After adding adversarial learning, there is an obvious improvement in accuracy (+4.09%). (2) The adaptive token selection strategy enhances all of the pre-trained models in downstream tasks. The proposed MCPATS is an end-to-end model with T1w images as its input and automatically and adaptively selects features across individuals. The ablation results demonstrate the effectiveness of our proposed ATS method.

The comparison results between our proposed ATS with other feature selection methods (Top-K and TransFG) are shown in Tables I, III and IV. The results demonstrate that, despite utilizing the same pre-trained model, the integration of ATS during model fine-tuning can yield superior classification performance compared to other feature selection methods. This highlights the advantages of our proposed ATS.

To provide theoretical guidance for subsequent research on how to tailor a brain representation learning framework for ViT to specific task/data domains, we further explore the following three questions: (1) Are the representations extracted by the pre-trained models associated with cognitive functions? How can the cognitive interpretability of the model be verified? (2) For each of the components in pre-training, what representations are promoted, and what representations are damaged? (3) For medical image analysis, is feature selection and feature fusion for different transformer layers (e.g., ATS) a more appropriate approach than using features from the last layer (e.g., vanilla ViT) for downstream tasks?

C. Representation Analysis for the Pre-trained Models

1) Association Between Representation and Cognition: Partial least square regression (PLSR) is widely used to establish brain-behavior relationships [19]. In this study, we employed PLSR to build prediction models for individual brain representations and behavior metric scores. We study 12 behavioral tasks (Emotion Memory, Fluid Intelligence (FI), Picture Priming, Motor Learning, Sentence Comprehension, Face Recognition: Familiar Faces, Face Recognition: Unfamiliar Faces, Emotional Regulation, Visual Short-Term Memory (VSTM), RT Simple, Force Matching and Tip-of-the-Tongue Task (TOT)), which are described in detail by [38] (see the study protocol in [58]), and follow the definition in [59] that RT Simple, Force Matching and TOT are basic active tasks and the other nine are cognitive tasks.

Specifically, for all subjects in the Cam-CAN dataset, we first freeze the parameters of our pre-trained encoder and three pre-trained ablation models and extract the encoded features from each transformer layer during a forward propagation process. Then, the encoded features are used to predict each metric score of the 12 behavioral tasks. We use ten-fold cross-validation in the prediction analysis. Concretely, a random division of data folds is performed, and 90% of the data are used for training and 10% are used for testing. The prediction model built on the training set is directly tested on the testing set, and the Pearson correlation coefficient between the actual and predicted scores is used to estimate the interpretability of learned representations for cognition. We repeat the ten-fold cross-validation procedure 20 times and report the averaged performance in Fig. 2. We make the following observations:

All correlations between cognitive functions (involving comprehension, reasoning, memory, emotion, and learning) and representations extracted by MCP are statistically significant at false discovery rate (FDR)-corrected $p < 0.05$ (except for the VSTM task at the first layer; see details in Fig. 2), whereas the correlations of some basic active tasks (RT Simple, Force Matching, and TOT) are not significant at FDR-corrected $p > 0.05$. Specifically, Emotion Memory shows the highest predictability, with mean correlations between actual versus predicted scores reaching $r >= 0.559 \pm 0.012$ for features in every transformer layer, averaging across all cross-validation repetitions. For other metrics scores of cognitive functions, the correlations vary from $r(VSTM) >= 0.159 \pm 0.021$ to $r(FI) >= 0.484 \pm 0.013$, and most of the correlations tend to increase as the transformer layers become deeper. The significant interpretability demonstrates that MCP encodes cognition-relevant features and represents intrinsic properties of the brain.

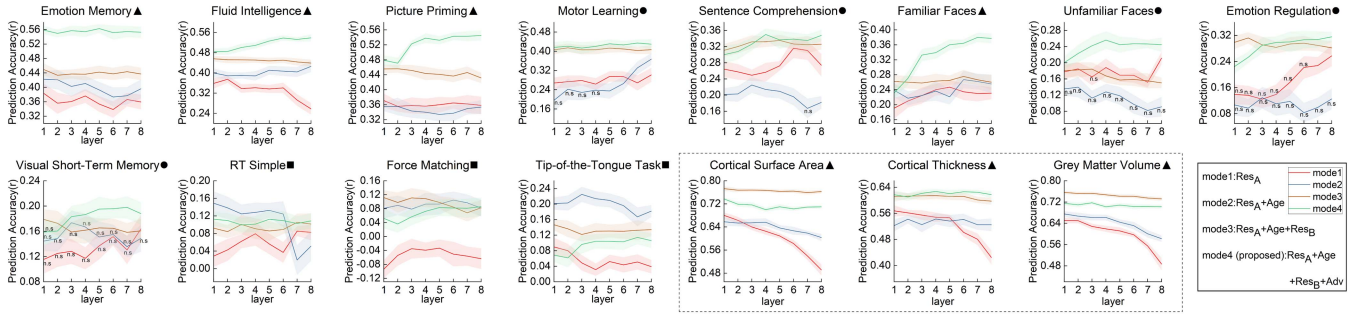

Fig. 2. Prediction results for representations and behavior metric scores and morphological estimates (inside the box). We first extract the encoded representations from each transformer layer of the pre-trained encoder, then use PLSR to build prediction models for individual brain representations and behavior metric scores/morphological estimates, and finally report the Pearson correlation coefficient between the actual and predicted scores (ten-fold cross-validation, repeated for 20 times). Note that ▲ denotes there are significant correlations (FDR-corrected $p < 0.05$) for every mode in every layer, ■ denotes that all the correlations are not significant, and ● denotes that some of the correlations are not significant (labeled n.s).

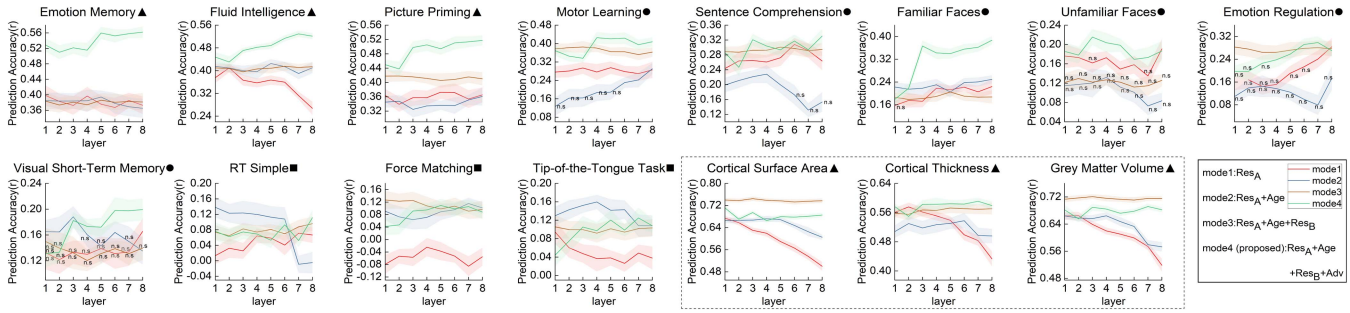

Fig. 3. Prediction results for representations and behavior metric scores and morphological estimates (inside the box). when controlling for the effect of age. We excluded age-association features (whose correlation with age survived the Bonferroni correction) in prediction model building stage, and 20 repetitions of ten-fold cross-validation were performed same as Fig. 2. Note that ▲, ■, and ● have the same meaning as in Fig. 2.

2) Association Between Representation and Cognition When Controlling the Effect of Age: To verify that the proposed MCP learned not only age-shared but also age-independent cognitive information, we excluded age-association features in model building. Concretely, we first computed the Pearson correlation between individual encoded features and ages, and the features whose correlation survived the Bonferroni correction were removed ($p < 0.05$). Then, the individually remaining features were used to build prediction models, and repetitions of ten-fold cross-validation were performed, as described in Section IV-C.

After controlling for the effect of age, correlations of almost all cognitive function scores were still significant at FDR-corrected $p < 0.05$ (except for the Emotion Regulation task at the first three layers and the VSTM task at the first two layers; see details in Fig. 3), and the correlations of basic active tasks were still not significant (FDR-corrected $p > 0.05$). This result suggests that our proposed MCP encodes cognition-relevant representations independent of age. Notably, for the first three layers of the model, the prediction performance slightly decreases after controlling for the effect of age, whereas there is little or no effect on the deeper layers, which indicates that the ViT encoder may tend to extract low-level features that is directly related to age (e.g., density of brain tissues) in the first few layers but encode high-level cognition-relevant semantic information in the deeper layers.

3) Association Between Representation and Morphological Estimates: We study the association between the brain representations and three common morphological estimates, including cortical surface area, cortical thickness and GM volume, using the abovementioned repetitive ten-fold cross-validation framework. The results with/without controlling the effect of age are illustrated in Figs. 2 and 3. There are significant correlations between the encoded features and morphological estimates (FDR-corrected $p < 2.0 \times 10^{-21}$), and the prediction performance trends oscillate downward as the transformer layers become deeper. In this context, the superior performance of MCPATS in downstream tasks can be theoretically interpreted from the following two perspectives: (1) The pre-trained encoder represents both morphological features and high-level semantic information of the brain, thus making a holistic comprehension of structural characteristics and cognitive functions. (2) Each transformer block has different preferences for the representing of the brain, so selecting and fusing the important features of each layer in fine-tuning, which fosters the combination of multi-level, multi-stage representations, can facilitate the classification performance of the model.

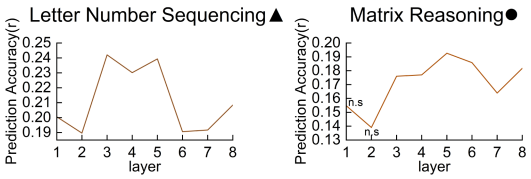

Fig. 4. Transfer results on the CNP dataset. We build a FI prediction model on the cam-CAN dataset (using PLSR, same as Figs. 2 and 3) and test the model on the CNP dataset to predict FI-related scores (letter number sequencing and matrix reasoning). This figure presents the Pearson correlation coefficient between the actual and predicted scores. Note that▲, ● and n.s are the same as in Fig. 2.

4) External Validation on an Independent Dataset: Considering the broad replication crisis in neuroscience, especially in brain structure-behavior studies [60], external validation on an independent cohort is required to reliably identify the association between brain representations and cognition. We transfer the FI prediction model built on the Cam-CAN dataset to the CNP dataset to predict the scores of two WAIS-IV subtests: letter number sequencing and matrix reasoning, which are the most relevant to fluid intelligence. (The standard form of the Cattell Culture Fair is used to measure the "fluid intelligence" for subjects from the Cam-CAN dataset, but the subjects from the CNP dataset are tested by WAIS-IV.) We observe a significant correlation between the actual and predicted scores, as shown in Fig. 4. The transfer-validation result demonstrates that MCP has learned cognition-relevant representations and enables conceptual replications and extrapolations for subsequent research.

5) Ablation Study for Representation Analysis: We conduct a representation ablation study to explore how each component affects the representation learning preferences of MCP. The results are illustrated in Figs. 2 and 3. We can observe that: (1) Significant correlations with cognition were found for the encoded features extracted from the model pre-trained by a single proxy task, parallel distillation masked image modeling (ablation mode 1). This suggests that after incorporating a biologically meaningful prior, a simple mask-reconstruction task could enable the model to capture high-level semantic information from the brain structure. (2) The age prediction task improves the correlations of representations with morphological estimates, suggesting that age prediction enhances the model's ability to understand the geometry of brain structure. (3) Visible portion restoration (restorative learning branch B) improves both the correlations of learned representations with cognitive functions and morphological estimates, suggesting that it can serve as a basic self-supervised pre-training component in representation learning frameworks. Previous studies have indeed taken this transformation-restoration approach [37], [61], [62], although they have not made considerations based on comprehensive representation analysis. (4) The addition of adversarial learning significantly improves the correlations between representations and cognitive functions, suggesting that adversarial learning, especially the synergy between adversarial and restorative learning, facilitates the model to capture high-level semantics and learn general brain representations. Surprisingly, previous computer vision literature has often assumed that adversarial learning can enhance the ability of models to capture spatial features such as shape and texture [63], [64], but we reach a different conclusion. The fact that adversarial learning reduces the correlations between representations and morphological estimates of the brain exemplifies the heterogeneity of natural and medical images (specifically brain sMRI) and highlights the importance of re-exploring the impact of various pre-training methods on model capabilities in brain representation learning.

TABLE VI
LINEAR-PROB ACCURACY (%) ON ADHD-200

Layer	1	2	3	4	5	6	7	8
mode1	60.23	**60.82**	57.31	57.89	59.65	58.48	56.14	59.06
mode2	61.40	59.65	**64.33**	62.57	56.14	62.57	63.74	62.57
mode3	60.82	**63.16**	62.57	56.14	**63.16**	57.31	59.06	60.82
mode4	62.57	61.40	60.23	57.31	63.16	**66.67**	57.89	59.65

Values in bold represent the best performance for each metric.

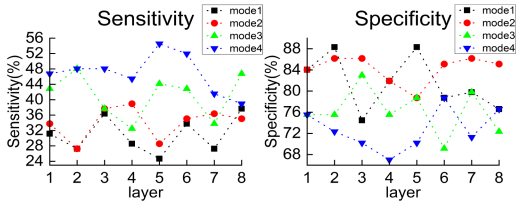

Fig. 5. Linear-prob sensitivity and specificity of ADHD-200.

6) Linear-Prob Classification on ADHD-200: The linear-prob classification results of ADHD-200 are shown in Table VI. In particular, we freeze the pre-trained models as the feature extractors and feed them with all the data in the ADHD-200 training and testing sets. Then, the encoded features of the training set are used to train a linear SVM as a classifier, which is tested on the encoded features of the testing set in each transformer layer. Our proposed MCP achieves a 66.67% linear-prob accuracy, surpassing most of the compared methods, indicating that the pre-trained encoder has learned how to represent biological patterns of the brain organization in healthy people and thus can discriminate intra-class variation and inter-class variation. Notably, we visualize the tendencies of sensitivity and specificity in Fig. 5, finding that the model pre-trained by a simple proxy task (ablation mode 1) tends to identify most of the test samples as healthy people (SPE $>$ 74%, SEN $<$ 38%), echoing our perspective that a single biased optimization objective does not allow the model to extract sufficiently discriminative features. In fact, the sensitivity of MCP increases as the pre-training components are added one by one, suggesting that each component makes the model more sensitive to disease-related alterations, and the important role of multi-task collaborative pre-training was again emphasized.

D. Representation Analysis for the Fine-Tuned Models

In Section IV-C, we have observed that: (1) Correlations between MCP-extracted representations and cognitive functions

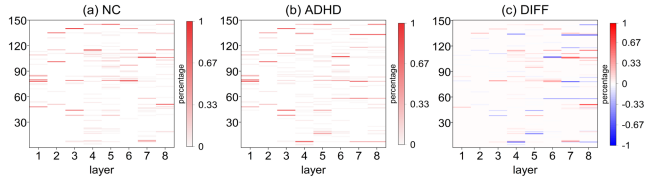

Fig. 6. Selected frequency of each token in each transformer layer. We choose the model that performed best in ADHD classification task and compute the selected frequency of each token in each layer (what percentage of subjects select it). The difference of the selected frequency between NC and ADHD is shown in Fig. 6(c) (c = a –b).

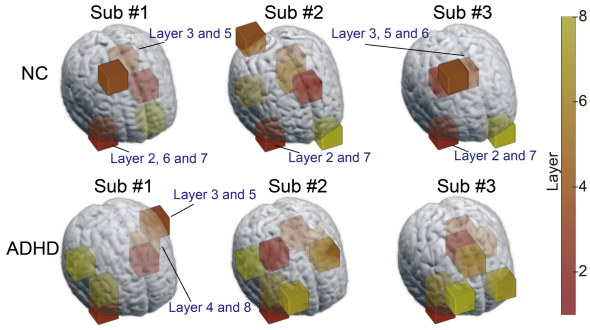

Fig. 7. The locations of the first-selected tokens in each transformer layer.

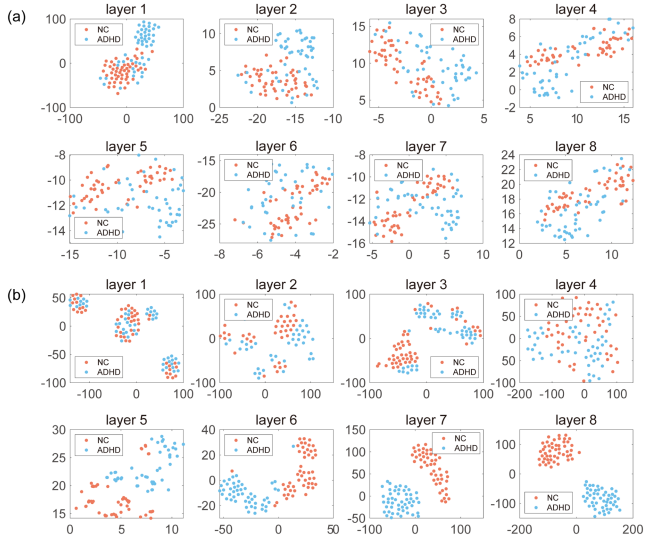

Fig. 8. t-SNE visualization. We choose the models that performed best in ADHD classification task for ablation mode 4 (w/o ATS) and ablation mode 4 (proposed) and visualize the distribution of NC and ADHD in the latent space. (a) Latent space of ablation mode 4 (w/o ATS) (all 150 tokens of each subject). (b) Latent space of our proposed model (24 selected tokens of each subject; 8 transformer layers in the encoder, 3 tokens are selected in each layer).

tend to strengthen with increasing depth of the transformer layers, whereas for morphological estimates, the correlation strength tends to oscillate downward or is relatively stable, whether or not the effects of age are controlled for. (2) After controlling for the effect of age, correlations between representations and cognitive functions slightly decrease in the first three layers, but there is little or no effect in deeper layers. Therefore, we theoretically answer the question that features from different transformer layers represent different brain representations, so feature selection and fusion is more suitable for downstream tasks than using features from the last layer. Here, we further validate this observation in the fine-tuning stage and demonstrate the validity of the proposed ATS method from the perspective of representation analysis.

1) Selected Frequency and Location of Tokens: We computed the selected frequency of each token in each transformer layer (what percentage of subjects selected it) using the model that performed best in ADHD classification. Fig. 6(a) and (b) show that the distribution of selected tokens is wide and discrete, and dominant tokens are very scarce, indicating that there is no strong consistency among individuals. Moreover, Fig. 6(c) shows that the selected tokens are different between NC and ADHD, and the differences increase layer by layer. This demonstrates that MCPATS can locate informative regions, even if they are displaced by disease or cognitive variation, and later separate patients from healthy people using the selected and fused discriminative features.

In Fig. 7, we choose 6 subjects (3 ADHDs, 3 NCs) and visualize the locations of the tokens that are the first to be selected for each subject in each layer. It suggests that there are both inter-class variation and intra-class variation, indicating that the token selection strategy is highly individualized.

2) Visualization in the Latent Space: We visualize all 150 tokens and three ATS-selected tokens for each transformer layer in the latent space to demonstrate the validity of the ATS method. Specifically, we choose the models that performed best in ADHD classification for ablation mode 4 (w/o ATS) and ablation mode 4 (proposed) and visualized the distribution of ADHD and NC in the latent space via t-Distributed Stochastic Neighbor Embedding (t-SNE) for features in each layer. The results are illustrated in Fig. 8.

We can observe that after adding ATS, ADHDs and NCs are better separated in the latent space, especially for the last four layers. Notably, for ablation mode 4 (w/o ATS), the classifier is fed with all 150 tokens in the last layer, but ADHDs and NCs are not roughly separated even in the last layer, which explains the large gap in classification accuracy (69.01% vs. 74.27%) between ablation mode 4 (w/o ATS) and ablation mode 4 (proposed) to some extent. Therefore, we demonstrate that ATS not only fuses the features from different layers (different brain representations) but also reduces redundant features to prevent the model from being confused by disease-independent features, making it more suitable for downstream tasks.

V. Conclusion

We propose MCPATS, a unified end-to-end framework for brain representation learning, combining multi-task collaborative pre-training and adaptive token selection. The key insight is that encode cognition-related information is crucial for brain representation learning. We have demonstrated the efficacy of our proposed MCPATS across three types of brain disease with

six compared methods. The proposed MCPATS shows remarkable performance compared to previous studies. Comprehensive ablation studies and hyperparameter analysis demonstrate the plausibility and robustness of the proposed MCPATS.

We investigate the correlation between metric scores of 12 behavioral tasks with MCP-extracted representations, demonstrating that our MCP produced high-quality cognitive interpretations. Our proposed MCP method can accurately capture information from sMRI related to various cognitive functions, and the prediction accuracy is comparable to that of previous studies which predict cognition using functional networks [19], [65]. This suggests that MCP method can extract the information interaction pattern between neurons, which can be regarded as the high-level semantics in the brain (see details in the Supplementary Material).

Analysis of the selected frequency, location, and t-SNE visualization of tokens indicate that the model flexibly captures a set of discriminative regions, no matter how individually their location changes, reflecting the importance of a suitable strategy (e.g., ATS) for feature selection and fusion. The comparison results between ATS with other feature selection methods highlight the advantages of our proposed ATS.

There are several inspirations to be considered in the future. First, we employ only one fixed patch size in this study, and it is reasonable to explore the influence of patch size to determine the optimal one. Second, we verify the effectiveness of the MCPATS on multiple classification tasks, while a wide range of downstream tasks (e.g., brain tumor segmentation, multiple MRI modality synthesis) may make the performance of the framework more convincing. Third, given the interpretability of the learned representations, it is a natural idea to generate explainable guides by the pre-trained model to reinforce the fine-tuning procedure. Moreover, an attempt was made to increase the scale of the pre-training data in order to enhance the model's performance. However, the desired results were not achieved (see details in the Supplementary Material). It is postulated that data heterogeneity and anisotropy from multi-source domain data may negatively impact the performance of pre-trained models. However, this effect can often be mitigated by increasing the data scale to a minimum of 5000 [20], [66]. In future research, we will collect additional data to extend our study into a foundational model for the brain that can be efficiently applied to a range of tasks.

We envisage that MCPATS will inspire future work on collaborative training, feature selection and feature fusion to improve the learning of universal representations for medical imaging.

REFERENCES

[1] A. D. Nostro, V. I. Muller, A. T. Reid, and S. B. Eickhoff, "Correlations between personality and brain structure: A crucial role of gender," *Cereb. Cortex*, vol. 27, no. 7, pp. 3698–3712, Jul. 2017.

[2] T. Wang and S.-I. Kamata, "Classification of structural MRI images in Adhd using 3D fractal dimension complexity map," in *Proc. IEEE Int. Conf. Inf. Process.*, 2019, pp. 215–219.

[3] J. L. Winterburn et al., "Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study," *Schizophrenia Res.*, vol. 214, pp. 3–10, 2019.

[4] H. Cai, Y. Gao, and M. Liu, "Graph transformer geometric learning of brain networks using multimodal MR images for brain age estimation," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 456–466, Feb. 2023.

[5] P. Huang et al., "Common feature learning for brain tumor MRI synthesis by context-aware generative adversarial network," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102472.

[6] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2965–2975, Sep. 2020.

[7] B. M. Cobbinah et al., "Reducing variations in multi-center Alzheimer's disease classification with convolutional adversarial autoencoder," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102585.

[8] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. *arXiv:1312.6114*.

[10] A. E. Brown and B. De Bivort, "Ethology as a physical science," *Nature Phys.*, vol. 14, no. 7, pp. 653–657, 2018.

[11] J. C. Pang et al., "Geometric constraints on human brain function," *Nature*, vol. 618, no. 7965, pp. 566–574, 2023.

[12] J. C. Pang, J. K. Rilling, J. A. Roberts, M. P. Van Den Heuvel, and L. Cocchi, "Evolutionary shaping of human brain dynamics," *Elife*, vol. 11, 2022, Art. no. e80627.

[13] E. Schwartz et al., "Evolution of cortical geometry and its link to function, behaviour and ecology," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 2252.

[14] M. Thiebaut de Schotten, C. Foulon, and P. Nachev, "Brain disconnections link structural connectivity with function and behaviour," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 5094.

[15] J. M. Shine et al., "Human cognition involves the dynamic integration of neural activity and neuromodulatory systems," *Nature Neurosci.*, vol. 22, no. 2, pp. 289–296, 2019.

[16] D. Wu et al., "Hierarchy of connectivity–function relationship of the human cortex revealed through predicting activity across functional domains," *Cereb. Cortex*, vol. 30, no. 8, pp. 4607–4616, 2020.

[17] V. J. Sydnor et al., "Intrinsic activity development unfolds along a sensorimotor–association cortical axis in youth," *Nature Neurosci.*, vol. 26, no. 4, pp. 638–649, 2023.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[19] R. Jiang et al., "A neuroimaging signature of cognitive aging from whole-brain functional connectivity," *Adv. Sci.*, vol. 9, no. 24, Aug. 2022, Art. no. 2201621.

[20] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 213–224, Jan. 2022.

[21] F. Altay, G. R. Sanchez, Y. James, S. V. Faraone, S. Velipasalar, and A. Salekin, "Preclinical stage Alzheimer's disease detection using magnetic resonance image scans," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 15088–15097.

[22] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.

[23] T. Isobe et al., "Multi-target domain adaptation with collaborative consistency learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8183–8192.

[24] T. He, J. Hu, Y. Song, J. Guo, and Z. Yi, "Multi-task learning for the segmentation of organs at risk with label dependence," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101666.

[25] C. Playout, R. Duval, and F. Cheriet, "A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 101–108.

[26] L. Liu, C. Huang, C. Cai, X. Zhang, and Q. Hu, "Multi-task learning improves the brain stoke lesion segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 2385–2389, 2022.

[27] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, "One-pass multi-task networks with cross-task guided attention for brain tumor segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 4516–4529, 2020.

[28] Y. Liu, F. Mu, Y. Shi, and X. Chen, "SF-net: A multi-task model for brain tumor segmentation in multimodal MRI via image fusion," *IEEE Signal Process. Lett.*, vol. 29, pp. 1799–1803, 2022.

[29] N. Zeng, H. Li, and Y. Peng, "A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease," *Neural Comput. Appl.*, vol. 35, no. 16, pp. 11599–11610, Jun. 2023.

[30] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2020.
[31] C. Lian, M. Liu, Y. Pan, and D. Shen, "Attention-guided hybrid network for dementia diagnosis with structural MR images," *IEEE Trans. Cybern.*, vol. 52, no. 4, pp. 1992–2003, Apr. 2022.
[32] W. Zhu, L. Sun, J. Huang, L. Han, and D. Zhang, "Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI," *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2354–2366, Sep. 2021.
[33] R. Avinun, S. Israel, A. R. Knodt, and A. R. Hariri, "Little evidence for associations between the big five personality traits and variability in brain gray or white matter," *NeuroImage*, vol. 220, 2020, Art. no. 117092.
[34] A. L. Decker, K. Duncan, A. S. Finn, and D. J. Mabbott, "Children's family income is associated with cognitive function and volume of anterior not posterior hippocampus," *Nature Commun.*, vol. 11, no. 1, Aug. 2020, Art. no. 4040.
[35] H. Kiesow et al., "10,000 social brains: Sex differentiation in human brain anatomy," *Sci. Adv.*, vol. 6, no. 12, Mar. 2020, Art. no. eaaz1170.
[36] G. Wang et al., "Connectional-style-guided contextual representation learning for brain disease diagnosis," *Neural Netw.*, vol. 175, 2024, Art. no. 106296.
[37] Z. Zhou et al., "Models genesis: Generic autodidactic models for 3D medical image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 384–393.
[38] J. R. Taylor et al., "The cambridge centre for ageing and neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample," *NeuroImage*, vol. 144, pp. 262–269, Jan. 2017.
[39] F. X. Castellanos et al., "The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience," *Front. Syst. Neurosci.*, vol. 6, 2012, Art. no. 62.
[40] R. L. Gollub et al., "The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, pp. 367–388, 2013.
[41] P. J. LaMontagne et al., "OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease," *MedRxiv*, 2019.
[42] R. A. Poldrack et al., "A phenome-wide examination of neural and cognitive function," *Sci. Data*, vol. 3, no. 1, pp. 1–12, 2016.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[44] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[45] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," 2021, *arXiv:2106.08254*.
[46] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.
[47] M. Hu, K. Sim, J. H. Zhou, X. Jiang, and C. Guan, "Brain MRI-based 3D convolutional neural networks for classification of schizophrenia and controls," in *Proc. IEEE 42nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2020, pp. 1742–1745.
[48] M. Fayyaz et al., "Adaptive token sampling for efficient vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 396–414.
[49] J. He et al., "Transfg: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 852–860.
[50] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.
[51] B. Zhang, H. Zhou, L. Wang, and C. Sung, "Classification based on neuroimaging data by tensor boosting," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2017, pp. 1174–1179.
[52] J. Ji, Y. Ren, and M. Lei, "FC-HAT: Hypergraph attention network for functional brain network classification," *Inf. Sci.*, vol. 608, pp. 1301–1316, Aug. 2022.
[53] C. Dou, S. Zhang, H. Wang, L. Sun, Y. Huang, and W. Yue, "ADHD fMRI short-time analysis method for edge computing based on multi-instance learning," *J. Syst. Architecture*, vol. 111, Dec. 2020, Art. no. 101834.
[54] Z. Mao et al., "Spatio-temporal deep learning method for ADHD fMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, Oct. 2019.
[55] J. Zhang, L. Zhou, and L. Wang, "Subject-adaptive integration of multiple sice brain networks with different sparsity," *Pattern Recognit.*, vol. 63, pp. 642–652, 2017.
[56] D. Dai, J. Wang, J. Hua, and H. He, "Classification of ADHD children through multimodal magnetic resonance imaging," *Front. Syst. Neurosci.*, vol. 6, pp. 63–63, 2012.
[57] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7627–7641, Jun. 2024.
[58] M. A. Shafto et al., "The cambridge centre for ageing and neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing," *BMC Neurol.*, vol. 14, 2014, Art. no. 204.
[59] G. S. Borgeest, R. N. Henson, M. Shafto, D. Samu, and R. A. Kievit, "Greater lifestyle engagement is associated with better age-adjusted cognitive abilities," *PLoS One*, vol. 15, no. 5, May 2020, Art. no. e0230077.
[60] S. Genon, S. B. Eickhoff, and S. Kharabian, "Linking interindividual variability in brain structure to behaviour," *Nature Rev. Neurosci.*, vol. 23, no. 5, pp. 307–318, May 2022.
[61] F. Haghighi, M. R. H. Taher, Z. Zhou, M. B. Gotway, and J. Liang, "Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2857–2868, Oct. 2021.
[62] Y. He et al., "Geometric visual similarity learning in 3D medical image self-supervised pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9538–9547.
[63] H. Huang et al., "AGTGAN: Unpaired image translation for photographic ancient character generation," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 5456–5467.
[64] S. Yang et al., "Controllable artistic text style transfer via shape-matching GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4441–4450.
[65] K. A. Tsvetanov et al., "Extrinsic and intrinsic brain network connectivity maintains cognition across the lifespan despite accelerated decay of regional brain activation," *J. Neurosci.*, vol. 36, no. 11, pp. 3115–3126, 2016.
[66] S. He, Y. Feng, P. E. Grant, and Y. Ou, "Deep relation learning for regression and its application to brain age estimation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2304–2317, Sep. 2022.