ATLAS-GUIDED REGIONAL REPRESENTATION ADAPTATION FOR BRAIN FOUNDATION MODELS CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain foundation models provide a powerful starting point for transferring knowledge across neuroimaging tasks, yet their learned representations are often organized around generic image tokens rather than anatomically defined brain regions. This makes it challenging to directly exploit the regional structure of the brain when adapting pretrained representations to downstream applications. We introduce AnatoQ, an efficient anatomy-aware adaptation framework that converts generic foundation-model features into region-specific representations using anatomical priors. At its center is an Atlas-Guided Q-Former (AGQ), which uses atlas-defined regions as queries to selectively aggregate informative features from pretrained image tokens. Rather than altering the architecture or retraining the foundation model, AnatoQ operates as a lightweight adaptation module, allowing anatomical structure to be incorporated while retaining the knowledge encoded by the original backbone. The resulting regional representations provide a common interface for downstream brain imaging tasks and can be combined with different pretrained backbones. Experiments on a broad range of downstream tasks and multi-session neuroimaging datasets show general gains across foundation models and aggregation methods. Additional analyses demonstrate that the learned representations exhibit stable correspondence with anatomically meaningful regions and support task-dependent interpretation. These results suggest that atlas-guided regional adaptation offers an effective and general approach for injecting anatomical structure into pretrained brain representations.

1 INTRODUCTION

The human brain is one of the most anatomically and functionally complex organs, making brain image analysis a long-standing challenge in medical computer vision Ghosh et al. (2026). With the rapid development of medical imaging foundation models, learning transferable and anatomically meaningful brain representations has become increasingly important for downstream applications Deng et al. (2025)Liu et al. (2025)Barbano et al. (2026)Wu et al. (2026). Structural magnetic resonance imaging (sMRI) is widely used for computer-aided diagnosis, disease staging, and biomarker discovery because of its non-invasive acquisition, high spatial resolution, and broad clinical availability Tae et al. (2025). In representative neurodegenerative disorders such as Alzheimer's disease (AD), disease progression is often accompanied by region-specific structural alterations, including hippocampal atrophy, gray matter loss, and ventricular enlargement Filippi et al. (2020). However, the high dimensionality and complex anatomical organization of brain MRI make it challenging to learn representations that are both discriminative and anatomically interpretable.

Recent advances in large-scale pretraining have led to the emergence of brain foundation models, such as MedicalNet Chen et al. (2019), SAM-Med3D Wang et al. (2025), SwinUNETR Hatamizadeh et al. (2022), BrainIAC Tak et al. (2026), and BrainSegFounder Cox et al. (2024). Despite differences in network architectures and pretraining objectives, these models generally represent brain volumes as sequences of regularly partitioned patch tokens followed by global aggregation for downstream tasks. Although patch-based representations have shown strong transferability across diverse neuroimage tasks, neurological disorders are often characterized by anatomically localized structural alterations, making explicit modeling of anatomy important for MRI analysis.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, regular patches are computational units rather than anatomical units. A single patch may span multiple anatomical regions, while one brain region may be fragmented across multiple patches. Consequently, disease-related structural changes can become distributed across multiple patch representations. Subsequent global aggregation further ignores explicit anatomical boundaries and regional identities, making it difficult to establish stable correspondence between learned representations and brain anatomy. For neurological dementia, whose brain patterns exhibit strong regional specificity, this mismatch may limit both downstream transferability and anatomical interpretability. These limitations motivate the need for anatomically structured representations that better bridge pretrained visual features and brain anatomy.

Existing approaches attempt to alleviate this limitation from several perspectives. Fine-tuning improves task adaptation but typically requires updating many parameters and is prone to overfitting in small-scale medical imaging datasets Chen et al. (2022). Token aggregation methods, such as CLS tokens, mean pooling, and learnable token aggregation, produce compact global representations but still lack explicit anatomical constraints Dosovitskiy et al. (2020)Jaegle et al. (2021). Traditional atlas-based neuroimaging analysis provides anatomically meaningful regional priors, yet generally relies on predefined regional statistics and remains largely independent of the high-dimensional representations learned by foundation models Yang et al. (2026)Barbano et al. (2026). Therefore, there remains a need for a parameter-efficient and plug-and-play framework that incorporates anatomical priors into pretrained representations while preserving the generalization capability of foundation models. Overall, grid-based representations preserve pretrained knowledge but lack explicit anatomical correspondence, whereas atlas-based pooling provides regional identities but may lose fine-grained information and restrict adaptive interactions across regions. An effective brain representation should therefore simultaneously achieve anatomical correspondence, representation preservation, and contextual flexibility.

To address this challenge, we propose AnatoQ, a lightweight framework that bridges pretrained patch representations and anatomical priors through a core module, AGQ. Without modifying the architecture of the pretrained backbone, AGQ reorganizes patch-level representations into anatomically meaningful region-level representations under the guidance of a standard brain atlas. By integrating AGQ into different pretrained encoders, AnatoQ framework can be seamlessly applied to diverse foundation models and downstream neuroimaging tasks, enabling more anatomically informed and transferable brain representations. The main contributions are summarized as follows:

- We identify the structural mismatch between patch-based representations and anatomical brain organization as a fundamental limitation of current brain MRI foundation models for downstream neuroimaging analysis.
- We propose AnatoQ, a lightweight framework with AGQ as the core module, which injects anatomical atlas priors into pretrained representations to construct anatomically meaningful region-level representations without modifying the pretrained backbone.
- AnatoQ framework is compatible with diverse foundation models and downstream prediction heads, generally improving performance on AD diagnosis, brain age prediction, and multiple pretrained encoders while introducing only minimal trainable parameters.

2 Related Work

Pre-trained and Foundation Models for Brain MRI. Large-scale pretraining has become an effective strategy for learning transferable representations from medical images, reducing the dependence on extensive annotations in downstream tasks. Early studies, such as Med3D Chen et al. (2019) and Models Genesis Zhou et al. (2021), demonstrated the potential of supervised and self-supervised pretraining for volumetric medical imaging. Recent 3D foundation models further extend this paradigm through large-scale pretraining and Transformer-based architectures. For brain imaging, models such as BrainSegFounder and BrainIAC have been developed to learn generalizable representations across multiple sequences, diseases, and downstream applications. These models significantly improve the capability of brain MRI representation learning; however, their learned features are still primarily organized as generic patch-level representations. Therefore, beyond developing stronger encoders, an important yet underexplored problem is how to adapt pretrained representations into anatomically meaningful representations for brain analysis.

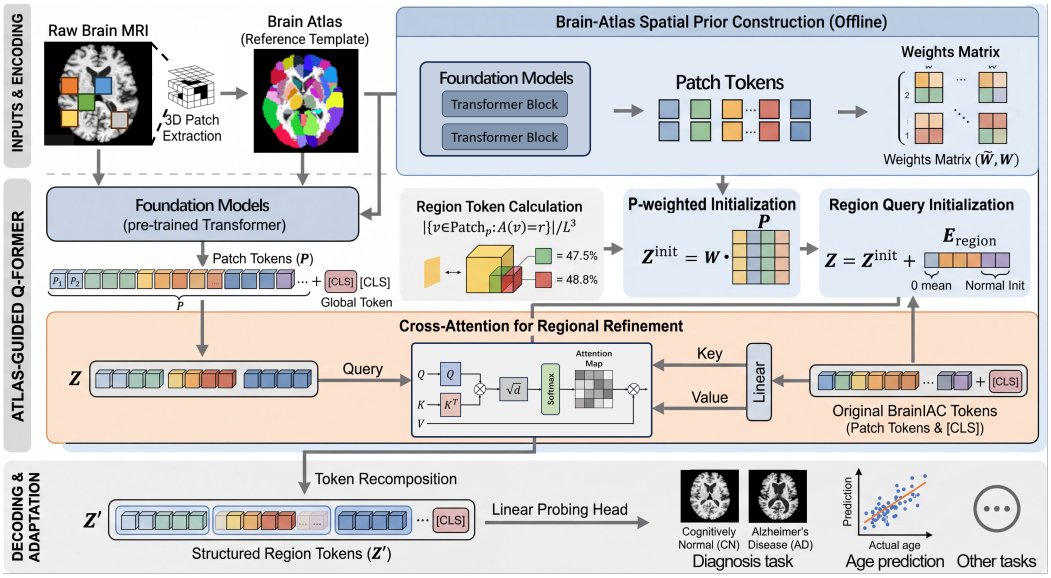

Figure 1: Overview of the proposed AnatoQ framework. (a) Input & Encoding: A pretrained volumetric encoder extracts patch tokens from 3D brain MRI, while a brain atlas provides region-to-patch spatial priors. (b) Atlas-guided Q-Former: AGQ reorganizes patch tokens into anatomically grounded region tokens through atlas-guided initialization and cross-attention. (c) Decoding & adaptation: The resulting region tokens are fed into linear prediction heads for downstream tasks.

Patch-Based Representation Learning for Volumetric Brain Imaging. Most modern volumetric foundation models adopt patch-based representations inherited from vision Transformers, where 3D images are divided into regular patches and encoded as token sequences. Existing approaches improve volumetric representation learning through enhanced attention mechanisms, hierarchical architectures, and multi-scale feature modeling. However, patch tokens are defined by computational grids rather than anatomical structures. For brain MRI, this design may result in a mismatch between representation units and anatomical regions: a single patch can involve multiple brain regions, while one anatomical region can be distributed across multiple patches. Although positional encoding preserves spatial information, it does not provide explicit anatomical identity. Therefore, existing patch-based representations lack direct correspondence between learned tokens and brain regions.

Brain Atlases and Anatomy-Aware Representation Learning. Brain atlases provide explicit anatomical correspondence by defining consistent regional identities across individuals. They have been widely used in neuroimaging for regional analysis, connectivity modeling, and brain network construction. Previous anatomy-aware approaches incorporate atlas information through ROI-based modeling, graph representation learning, or anatomical constraints, demonstrating the importance of structural priors in brain analysis. However, atlas-based approaches are typically built upon predefined regional measurements or extracted features and are largely independent of modern foundation model representations. In contrast, foundation models learn rich high-dimensional features but generally lack explicit regional alignment. How to integrate anatomical priors with pretrained representations remains an important challenge.

Query-Based Token Aggregation and Adaptation. Query-based aggregation provides an efficient mechanism for compressing high-dimensional representations and adapting pretrained models. By introducing learnable queries, existing approaches can extract compact latent representations from large visual encoders while avoiding extensive modification of the backbone. Nevertheless, current query-based methods usually rely on unconstrained learnable queries, whose semantic meanings are determined by data-driven optimization. Such generic queries do not have predefined correspondence with meaningful structures, limiting their interpretability and consistency in brain imaging. Incorporating anatomical guidance into query-based aggregation provides a potential direction for constructing more structured representations from pretrained brain models.

3 METHODOLOGY

3.1 PROBLEM SETUP

Typically, a pretrained volumetric foundation model is employed as a frozen backbone to extract transferable spatial representations from brain MRI. Given an input brain volume X and brain atlas template with K regions, the encoder produces a sequence of spatial tokens:

$$\mathbf{P} = T(f_\theta(X)) = [\mathbf{p}_1, \cdots, \mathbf{p}_N] \in \mathbb{R}^{N \times d}, \tag{1}$$

where f_θ denotes the frozen encoder, $T(\cdot)$ represents the encoder-specific tokenization operation, N is the number of spatial tokens, and d is the feature dimension.

Although these representations encode rich semantic information learned by large-scale pretraining, they are typically defined over regular computational units, such as image patches or feature grids, rather than anatomically defined brain regions. Consequently, the extracted spatial representations lack explicit anatomical correspondence, limiting their ability to capture anatomy-specific brain disease patterns.

To bridge this gap, we propose **Anatomy-Grounded Query Aggregation (AnatoQ)**, a lightweight post-encoder adaptation framework that aligns pretrained spatial representations with brain anatomical structures without modifying or retraining the foundation model. Its core component, **Atlas-guided Q-Former (AGQ)**, leverages atlas priors to construct anatomy-grounded region queries and refines them through cross-attention with global spatial representations. The overall framework is illustrated in Figure 1, which consists of three stages: (1) atlas-to-patch soft assignment, (2) anatomy-grounded query initialization, and (3) anatomy-anchored contextual refinement.

3.2 ATLAS-TO-PATCH SOFT ASSIGNMENT

To establish anatomical correspondence between atlas regions and spatial tokens of a pretrained encoder, we construct an atlas-guided soft assignment matrix. The brain atlas and MRI volume are aligned in the same reference space. Let $A(v) \in \{0, 1, \cdots, K\}$ denote the atlas label of voxel v, where K is the number of atlas regions. For each spatial token p, let $\mathcal{V}_p$ denote its corresponding voxel set. The overlap between anatomical region r and token p is defined as:

$$\widetilde{W}_{r,p} = \frac{1}{|\mathcal{V}_p|} \sum_{v \in \mathcal{V}_p} \mathbb{I}(A(v) = r). \tag{2}$$

where $\mathbb{I}[\cdot]$ denotes indicator function. $\widetilde{W}_{r,p} \in [0, 1]$ quantifies the proportion of the voxel support of token p assigned to region r. To eliminate the dependence of the subsequent aggregation scale on regional volume, the overlap coefficients are normalized over the token dimension for each region:

$$W_{r,p} = \frac{\widetilde{W}_{r,p}}{\sum_{j=1}^{N} \widetilde{W}_{r,j} + \epsilon}, \tag{3}$$

where ϵ is a small constant. This produces a soft anatomical correspondence matrix $\mathbf{W} \in \mathbb{R}^{K \times N}$, in which each r-th row characterizes the relative association between region r and all spatial tokens.

By construction, $\mathbf{W}$ is a deterministic and feature-independent anatomical operator derived solely from the spatial geometry of the atlas and pre-defined encoder token grid. It retains the many-to-many correspondence induced by anatomical boundaries and resolution differences: a token can contribute to multiple regions, and a region can receive contributions from multiple tokens. Within the AnatoQ framework, $\mathbf{W}$ parameterizes the anatomical aggregation of subject-specific encoder features that links grid-based encoder representations to atlas-defined regions.

3.3 ANATOMY-GROUNDED QUERY INITIALIZATION

Given the anatomical correspondence matrix $\mathbf{W}$, we aggregate the pretrained spatial representations into atlas-defined regional features:

$$\mathbf{Z} = \mathbf{W}\mathbf{P}, \qquad \mathbf{z}_r = \sum_{p=1}^{N} W_{r,p}\mathbf{p}_p, \tag{4}$$

From the formula, each $\mathbf{z}_r$ is a normalized, overlap-weighted aggregation of the encoder tokens associated with region r. It therefore preserves subject-specific information from the pretrained representation while establishing an explicit correspondence between each regional feature and an atlas-defined anatomical structure.

The regional feature $\mathbf{z}$ is utilized for query initialization. The initialization differs from conventional query-based aggregation, in which learnable queries have no predefined spatial or anatomical correspondence. Rather than learning query semantics solely from downstream supervision, AGQ uses atlas geometry to determine the initial support of each query. Consequently, the same query index is associated with the same anatomical region across subjects, providing a structured initialization for subsequent feature refinement.

However, the atlas-defined regional features alone may not sufficiently distinguish regions with similar aggregated representations. We therefore introduce a learnable region identity embedding $\mathbf{E}_{\text{region}} \in \mathbb{R}^{K\times d}$:

$$\mathbf{Q} = \mathbf{Z} + \mathbf{E}_{region}. \tag{5}$$

Here, $\mathbf{Z}$ encodes subject-specific regional evidence, while $\mathbf{E}_{\text{region}}$ represents the subject-independent identity of each atlas region. Their combination produces region queries that are both image-conditioned and anatomically identifiable. These queries provide the anatomical anchors for the contextual refinement described in the following subsection.

3.4 ANATOMY-ANCHORED CONTEXTUAL REFINEMENT

Although atlas-guided initialization establishes explicit anatomical correspondence, it remains a fixed aggregation determined by spatial overlap and cannot incorporate dependencies among spatially distributed regions. We therefore refine the anatomy-grounded queries via global multi-head cross-attention over the complete encoder token sequence:

$$\mathbf{R} = \mathrm{LN}\left(\mathrm{MHA}(\mathbf{Q}, \mathbf{P}, \mathbf{P})\right), \tag{6}$$

where $\mathbf{Q}$ provides the queries and $\mathbf{P}$ serves as both keys and values. The resulting representation $\mathbf{R} \in \mathbb{R}^{K\times d}$ contains one refined token for each region. Unlike the fixed coefficients $\mathbf{W}$, the cross-attention weights are conditioned on both the anatomy-grounded queries and the subject-specific encoder features, enabling content-adaptive feature aggregation.

Additionally, the global cross-attention allows each region query to retrieve relevant information from the entire spatial token sequence. The atlas therefore acts as an initialization prior rather than a hard spatial constraint. With each query assigned with a predefined anatomical identity, global attention permits the corresponding representation to incorporate complementary evidence from other regions. This design preserves the atlas-defined organization of the output tokens while supporting the cross-regional dependency modeling.

Collectively, the aforementioned three components constitute the proposed AGQ, and its complete transformation can be expressed as:

$$\mathbf{R} = \Phi_{\mathrm{AGQ}}\left(T(f_\theta(X)), A\right), \tag{7}$$

where Φ_{AGQ} denotes atlas-guided query initialization followed by contextual refinement. By applying this transformation to the representations of a frozen encoder, AGQ incorporates anatomical

Table 1: AD diagnosis performance on ADNI and OASIS using different foundation model encoders. Results are reported as mean (std) across data splits. MP denotes mean pooling. For AGQ, bold values indicate improvement relative to MP.

Encoder	Method	ADNI				OASIS			
		ACC (%)	AUC (%)	SEN (%)	SPE (%)	ACC (%)	AUC (%)	SEN (%)	SPE (%)
SAM-Med3D	MP	90.9 (1.5)	96.5 (0.5)	91.7 (4.5)	90.5 (4.2)	90.2 (1.1)	96.2 (0.1)	90.0 (4.7)	90.3 (5.9)
	AGQ	**91.9 (5.7)**	**97.4 (2.2)**	**94.0 (3.5)**	89.8 (8.9)	89.7 (1.1)	**96.3 (1.4)**	**93.4 (4.7)**	89.7 (2.0)
BrainIAC	MP	87.4 (4.6)	95.3 (1.7)	91.7 (4.5)	84.1 (6.1)	84.9 (2.2)	94.6 (1.3)	90.0 (4.7)	80.6 (0.0)
	AGQ	**88.9 (3.5)**	**96.1 (1.0)**	86.8 (6.6)	**95.0 (2.0)**	**87.9 (4.3)**	**95.9 (1.3)**	83.3 (7.1)	**95.8 (2.0)**
MedicalNet	MP	86.4 (5.3)	93.2 (3.0)	85.9 (6.5)	85.7 (7.8)	84.9 (6.4)	91.7 (2.1)	88.3 (7.1)	82.0 (5.9)
	AGQ	**94.9 (3.2)**	**98.4 (1.7)**	**95.6 (5.1)**	**94.6 (3.1)**	**93.2 (1.1)**	**97.6 (1.3)**	**93.4 (4.7)**	**93.1 (1.9)**
SwinUNETR	MP	88.9 (6.3)	96.6 (2.2)	78.6 (13.2)	96.5 (1.8)	86.4 (6.4)	95.7 (1.9)	75.0 (16.5)	95.8 (2.0)
	AGQ	**90.4 (4.6)**	96.6 (3.0)	**88.9 (9.6)**	92.2 (2.9)	**87.9 (2.1)**	94.9 (0.6)	**83.3 (0.0)**	91.7 (3.9)

Table 2: Regression performance on continuous demographic and cognitive outcomes in ADNI. Results are reported as mean (std) across data splits. MAE is lower-is-better and r is higher-is-better. MP denotes mean pooling. Bold values indicate improvement of AGQ relative to MP.

Encoder	Method	Age		MMSE		ADAS11		ADASQ4		RAVLT-imm.	
		MAE ↓	r ↑	MAE ↓	r ↑	MAE ↓	r ↑	MAE ↓	r ↑	MAE ↓	r ↑
SAM-Med3D	MP	4.43 (0.22)	0.44 (0.12)	2.14 (0.02)	0.69 (0.01)	4.81 (0.46)	0.72 (0.07)	1.96 (0.12)	0.73 (0.02)	8.14 (0.39)	0.68 (0.05)
	AGQ	**4.38 (0.09)**	0.44 (0.12)	**1.93 (0.15)**	**0.73 (0.04)**	5.24 (0.54)	**0.74 (0.03)**	**1.93 (0.17)**	0.69 (0.07)	**8.11 (0.57)**	**0.69 (0.01)**
BrainIAC	MP	4.78 (0.22)	0.37 (0.09)	2.17 (0.13)	0.65 (0.07)	5.53 (0.67)	0.66 (0.09)	1.96 (0.16)	0.70 (0.06)	8.50 (0.44)	0.64 (0.04)
	AGQ	**4.57 (0.27)**	0.29 (0.22)	2.17 (0.11)	**0.66 (0.04)**	**5.38 (0.61)**	**0.68 (0.05)**	**1.93 (0.27)**	0.67 (0.10)	**8.20 (0.22)**	**0.69 (0.05)**
MedicalNet	MP	4.79 (0.34)	0.22 (0.14)	2.35 (0.35)	0.55 (0.22)	6.14 (1.27)	0.65 (0.09)	2.24 (0.08)	0.64 (0.02)	10.56 (0.84)	0.47 (0.02)
	AGQ	**4.72 (0.39)**	**0.27 (0.15)**	**2.16 (0.46)**	**0.74 (0.07)**	**5.24 (0.82)**	**0.74 (0.03)**	**2.00 (0.18)**	**0.73 (0.03)**	**8.85 (0.99)**	**0.67 (0.03)**
SwinUNETR	MP	4.10 (0.59)	0.52 (0.16)	1.77 (0.28)	0.79 (0.04)	4.60 (0.24)	0.80 (0.02)	1.89 (0.05)	0.74 (0.02)	7.82 (0.30)	0.71 (0.05)
	AGQ	**3.99 (0.35)**	**0.55 (0.12)**	**1.60 (0.18)**	**0.82 (0.02)**	**4.57 (0.70)**	0.80 (0.02)	1.94 (0.24)	**0.76 (0.01)**	8.27 (0.75)	**0.73 (0.03)**

structure without full-backbone fine-tuning and introduces only a modest number of additional trainable parameters (see Appendix C).

4 EXPERIMENTS

We evaluate AGQ on structural brain MRI (sMRI) from the ADNI Weiner et al. (2010) and OASIS Marcus et al. (2007) cohorts. To obtain a sufficiently large and fairly class-balanced dataset, we pool subjects across the available releases of each cohort. All volumes are affine-registered to the MNI152 template and skull-stripped; detailed cohort compositions and data preprocessing procedures are provided in Appendix A. Four publicly available pretrained volumetric encoders are used as frozen backbones: SAM-Med3D Wang et al. (2025), BrainIAC Tak et al. (2026), MedicalNet Chen et al. (2019), and SwinUNETR Hatamizadeh et al. (2022). For anatomical guidance, we adopt the second version of the Automated Anatomical Labeling atlas (AAL2) Rolls et al. (2015), which parcellates the brain into $K = 120$ brain regions. AAL2 serves as the default parcellation and is held fixed throughout all experiments, i.e., the same atlas, and hence the same K region tokens, is used for every backbone, dataset, and downstream task.

AGQ is first pretrained for 30 epochs on a region-volume regression objective. For each downstream task, both the backbone and the pretrained AGQ are kept frozen, and only a linear probe is trained for an additional 30 epochs. Optimization uses AdamW with linear warmup followed by cosine learning-rate decay; complete hyperparameters and training settings are reported in Appendix B.

We consider two families of downstream tasks: (i) AD diagnosis, distinguishing Alzheimer's disease (AD) patients from cognitively normal (CN) controls; and (ii) regression of multiple continuous cognitive and demographic phenotypes, namely chronological age, MMSE, ADAS11, ADASQ4, and RAVLT-immediate (RAVLT-imm.). The diagnosis task is assessed by accuracy (ACC), area under the ROC curve (AUC), sensitivity (SEN), and specificity (SPE); the regression tasks by mean absolute error (MAE) and Pearson's r.

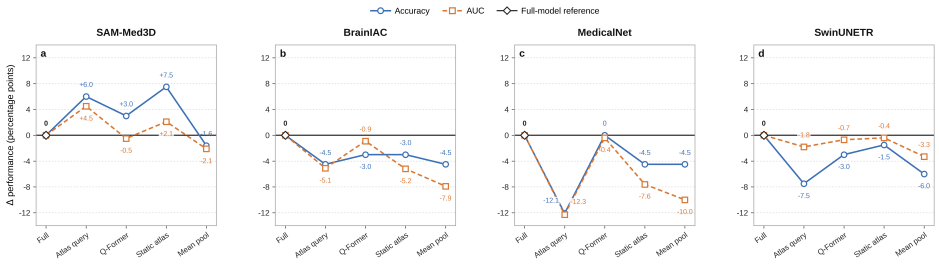

Figure 2: Ablation results of AGQ across pretrained encoders on ADNI dataset. The full model is compared with Atlas Query Only (without contextual refinement), Q-Former (with generic learnable queries instead of anatomy-grounded query initialization), Static Atlas (without region identity embedding or contextual refinement), and Mean Pooling (without anatomy-grounded query initialization or contextual refinement). The ablations are tested across three random seeds. Average changes in Accuracy and AUC are reported in percentage points relative to the full model, with negative and positive values indicating performance degradation and improvement, respectively.

4.1 Main Results

First, we systematically compare AGQ with mean pooling across different pretrained encoders on the AD diagnosis task. As shown in Table 1, AGQ outperforms mean pooling in 23 of the 32 encoder, dataset, and metric combinations, including 7 of 8 for ACC and 6 of 8 for AUC, indicating that anatomically structured aggregation can capture disease-related information that is not effectively preserved by mean pooling. Since these improvements are observed across encoders with different architectures and pretraining objectives, AGQ appears to exploit information already encoded in their spatial representations rather than relying on a specific encoder design. The varying magnitude of improvement across backbones can be attributed to differences in spatial resolution, pretraining supervision, and baseline transferability. In contrast to ACC and AUC, SEN and SPE fluctuate more noticeably across encoders and splits, and account for most of the remaining cases in which AGQ does not surpass mean pooling. This variability is largely a consequence of the limited test set size. With only a few dozen AD and CN subjects in each test split, a single misclassification alters SEN or SPE by several percentage points.

Table 2 further evaluates whether the same representation benefits continuous demographic and cognitive outcome prediction. AGQ yields the most prominent improvements on cognition-related targets, where regional structural patterns are likely more informative than uniformly pooled whole-brain representations. The general gains across multiple cognitive assessments suggest that the anatomical adapter is not specialized for AD diagnosis but can support downstream tasks with diverse prediction objectives.

Across both task families, the improvements are most pronounced on MedicalNet, which also exhibits the weakest mean-pooled baseline. We attribute this primarily to differences in token resolution. As listed in Table A3, MedicalNet produces a 12^3 token grid, whereas SAM-Med3D produces an 8^3 grid and BrainIAC and SwinUNETR produce 6^3 grids. A finer grid allows each atlas region to be supported by a more distinct subset of tokens, so that the atlas-guided queries encode more region-specific information. Conversely, on coarser grids, neighboring AAL2 regions are largely supported by the same tokens, which constrains the regional specificity that AGQ can exploit and consequently yields more moderate gains.

In summary, the diagnosis and regression results demonstrate that AGQ serves as a general representation adapter, whose effectiveness arises from the interaction among pretrained feature spaces, anatomical aggregation, and downstream tasks.

4.2 Ablation Analysis

To examine the contribution of each component of AGQ, we construct a series of variants that progressively introduce anatomical structure and adaptive aggregation, as summarized in Table A7. Mean pooling averages all spatial tokens and uses neither atlas information nor learnable aggregation. Static atlas pooling aggregates tokens through the fixed assignment matrix $\mathbf{W}$, which intro-

Table 3: Attribution characteristics of AnatoQ across pretrained encoders. Top-20 attribution mass and attribution stability (ρ) measure attribution concentration and stability, respectively. Mean $\pm$ std across subject-level splits.

Metric	SAM-Med3D	BrainIAC	MedicalNet	SwinUNETR	Avg.
Top-20 Mass (%)	40.2 ± 0.2	37.6 ± 0.2	27.6 ± 0.1	25.4 ± 0.2	$\mathbf{32.7 \pm 7.3}$
Attribution stability ρ	0.995 ± 0.001	0.981 ± 0.002	0.997 ± 0.002	0.998 ± 0.002	$\mathbf{0.993 \pm 0.008}$

duces regional organization without any trainable parameters. Atlas Query Only adds the learnable region identity embedding to the atlas-pooled features but omits cross-attention refinement. The generic Q-Former refines learnable queries through cross-attention, which enables adaptive aggregation but provides no explicit anatomical correspondence. The full AGQ combines anatomy-grounded query initialization with cross-attention refinement. Since mean pooling and the Q-Former are widely used aggregation strategies for pretrained encoders, this comparison serves both as an ablation of AGQ and as a comparison with existing aggregation methods. All variants share the same frozen backbone features, data splits, prediction head, and optimization settings, and all region-based variants produce K tokens, so that performance differences arise solely from the aggregation module. Mean pooling and static atlas pooling contain no trainable aggregation parameters and are evaluated directly, whereas the remaining variants undergo the same anatomical pretraining as the full model. All results are averaged over three random seeds on ADNI.

As shown in Figure 2, performance generally increases as anatomical structure and adaptive aggregation are introduced. Static atlas pooling performs better than or comparably to mean pooling across encoders, indicating that preserving regional organization is beneficial even without learnable parameters. Building on this regional organization, AGQ further incorporates cross-attention refinement and achieves the best or comparable performance on BrainIAC, MedicalNet, and SwinUNETR, outperforming the generic Q-Former that lacks anatomical initialization. On SAM-Med3D, however, the simpler atlas-based variants surpass the full model, suggesting that the benefit of contextual refinement depends on the backbone. Overall, these results indicate that anatomical initialization and contextual refinement contribute progressively to the performance of AGQ, although the magnitude of each contribution varies across encoders.

4.3 Anatomical Link and Task-Specific Interpretation

Since each output token of AnatoQ corresponds to a predefined atlas region, the learned representations can be interpreted at the regional level. For each subject, we compute a gradient-based attribution score for every region token, where a larger score indicates a stronger region-specific influence on the prediction; the formal definition is given in Appendix E.1.

Table 3 summarizes two statistics derived from these scores, which are computed on the held-out test subjects of each downstream task and reported as the mean and standard deviation across the six tasks. Top-20 attribution mass measures concentration and is defined as the fraction of each subject's total attribution captured by its 20 highest-scoring regions out of the $K = 120$ atlas regions, averaged across subjects. Attribution stability measures reproducibility. For each task, subjects are randomly divided into two disjoint halves, the region-wise mean attribution is computed within each half, and the Spearman correlation between the two mean vectors is averaged over 20 random partitions. Across all encoders, the Top-20 attribution mass clearly exceeds the uniform baseline of 16.7%, indicating that predictive evidence is concentrated on a limited subset of regions. Attribution stability remains high, with ρ between 0.981 and 0.998, indicating that the regional importance ranking estimated from one group of subjects generalizes to an independent group.

As shown in Figure 3, we further visualize the anatomical distributions of the Top-20 brain regions with the highest mean attribution mass for each downstream task. The saliency maps reveal both shared and task-specific anatomical patterns. For AD diagnosis, the most salient regions primarily involve temporal association cortices and subcortical nuclei, including the superior temporal cortex, temporal pole, caudate, thalamus, orbitofrontal cortex, and inferior parietal regions, which have been widely recognized as neuroimaging biomarkers associated with Alzheimer's progression Kim et al. (2023)Zhou et al. (2016). In contrast, age prediction places greater emphasis on subcortical and cerebellar structures, reflecting age-related structural variation across distributed brain regions.

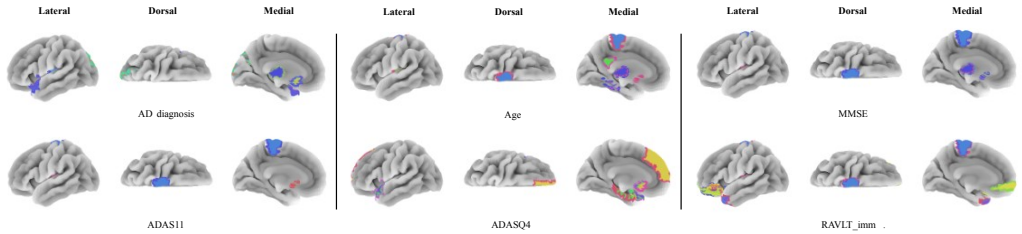

Figure 3: Task-specific Top-20 salient brain regions identified by AGQ on various tasks. Saliency scores are obtained using MedicalNet as the pretrained encoder and averaged across subjects in the test set for each task. The 20 regions with the highest average saliency are mapped to their corresponding regions and displayed from lateral, dorsal, and medial views. Each color uniquely identifies one selected brain region.

Cognitive outcome prediction tasks, including MMSE, ADAS11, ADASQ4, and RAVLT-immediate, consistently highlight cerebellar, orbitofrontal, temporal, and pallidal regions, which are closely associated with memory, executive function, and global cognitive performance Raz & Daugherty (2021)Bartrés-Faz & Arenaza-Urquijo (2011)Sintini & Whitwell (2021).

These shared and task-specific attribution patterns demonstrate the effectiveness of AGQ in learning clinically meaningful and task-adaptive regional representations. The alignment between highly attributed regions and established neuroimaging biomarkers indicates that atlas-guided queries effectively localize predictive evidence to anatomically relevant brain structures. Meanwhile, the distinct regional patterns observed across different objectives suggest that AnatoQ adapts its anatomical focus according to task-specific clinical requirements rather than relying on fixed regional preferences. Together with the stable attribution patterns across diverse pretrained encoders, these results demonstrate that AGQ enhances brain foundation models with robust, interpretable, and task-specific anatomical representations.

4.4 Discussion

Our results suggest that the gains of AGQ arise from progressively combining anatomy-grounded query initialization with content-adaptive contextual refinement. Anatomical initialization assigns each query a stable regional identity, and contextual refinement further integrates subject-specific evidence from the full spatial representation. The benefit is largest on MedicalNet, which has the finest token grid among the evaluated encoders, whereas coarser grids limit the regional specificity that the atlas prior can provide. This resolution limit is a property of the backbone rather than of the adapter, suggesting that AGQ may benefit further from higher-resolution foundation models.

Beyond performance, AnatoQ provides anatomically grounded evidence for interpretable medical AI. Since each output token corresponds to a named atlas region, the attribution score s_r directly identifies the structures that drive a prediction, and these attributions are both concentrated and reproducible across subjects. Such traceable regional evidence offers a potential interface for future evidence-grounded clinical reasoning systems, in which model decisions can be verified against established anatomical knowledge.

5 Conclusion

We addressed the mismatch between patch-based representations and anatomical organization in brain MRI foundation models by proposing AnatoQ, a lightweight framework centered on AGQ. Through atlas-to-patch soft assignment, anatomy-grounded query initialization, and anatomy-anchored context refinement, AGQ reorganizes frozen encoder features into region tokens with explicit anatomical correspondence, while global cross-attention preserves regional identity and models cross-regional dependencies. Across various foundation models, AD diagnosis, and continuous cognitive regression tasks, AnatoQ generally improves performance with a small number of trainable parameters, while its region-level saliency maps provide anatomically grounded evidence consistent with established AD pathology. Overall, AnatoQ bridges foundation model representations and brain anatomy, enabling more interpretable and clinically meaningful neuroimaging analysis.

AI USE STATEMENT

No generative AI tools were used in the preparation of this work.

ETHICS STATEMENT

We did not identify any specific ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we will release the source code on GitHub upon publication.

ACKNOWLEDGMENTS

This work received no external funding.

REFERENCES

Carlo Alberto Barbano, Matteo Brunello, Benoit Dufumier, and Marco Grangetto. Anatomical foundation models for brain MRIs. *Pattern Recognition Letters*, 199:178–184, 2026. doi: 10.1016/j.patrec.2025.11.028. URL https://www.sciencedirect.com/science/article/pii/S0167865525003848.

David Bartrés-Faz and Eider M Arenaza-Urquijo. Structural and functional imaging correlates of cognitive and brain reserve hypotheses in healthy and pathological aging. *Brain topography*, 24 (3):340–357, 2011.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Joseph Cox, Peng Liu, Skylar E Stolte, Yunchao Yang, Kang Liu, Kyle B See, Huiwen Ju, and Ruogu Fang. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301, 2024.

Zhongying Deng, Haoyu Wang, Ziyan Huang, Lipei Zhang, Angelica I Aviles-Rivero, Chaoyu Liu, Junjun He, Zoe Kourtzi, and Carola-Bibiane Schönlieb. Brain foundation models with hypergraph dynamic adapter for brain disease analysis. *Pattern Recognition*, pp. 112595, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Massimo Filippi, Silvia Basaia, Elisa Canu, Francesca Imperiale, Giuseppe Magnani, Monica Falautano, Giancarlo Comi, Andrea Falini, and Federica Agosta. Changes in functional and structural brain connectome along the alzheimer's disease continuum. *Molecular psychiatry*, 25(1):230–239, 2020.

Tapotosh Ghosh, Farnaz Sheikhi, Junlin Guo, Yashbir Singh, Khaled Younis, Shiba Kuanar, Shahriar Faghani, Eduardo Moreno Judice de Mattos Farina, Yuankai Huo, and Farhad Maleki. Foundation models for volumetric medical imaging: Opportunities, challenges, and future directions. *Electronics*, 15(6):1245, 2026.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Soyun Kim, Jenna N Adams, Lea Stith, Lisa M Taylor, Alyssa L Harris, Liv McMillan, and Michael A Yassa. Cerebellar-cortical functional connectivity is associated with alzheimer's disease pathology and cerebellar atrophy. *Alzheimer's & Dementia*, 19:e081860, 2023.

Peirong Liu, Oula Puonti, Xiaoling Hu, Karthik Gopinath, Annabel Sorby-Adams, and Juan Eugenio Iglesias. A modality-agnostic multi-task foundation model for human brain imaging, 2025. URL https://arxiv.org/abs/2509.00549.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

Naftali Raz and Ana M. Daugherty. Aging of the brain and cognitive functions: Structural and functional perspectives. *Annual Review of Psychology*, 72:393–421, 2021.

Edmund T Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5, 2015.

Irene Sintini and Jennifer L. Whitwell. Update on neuroimaging in alzheimer's disease. *Current Opinion in Neurology*, 34(4):525–531, 2021.

Woo-Suk Tae, Byung-Joo Ham, Sung-Bom Pyun, and Byung-Jo Kim. Current clinical applications of structural mri in neurological disorders. *Journal of Clinical Neurology (Seoul, Korea)*, 21(4): 277, 2025.

Divyanshu Tak, Biniam Garomsa, Anna Zapaishchykova, Tafadzwa Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, Maryam Mahootiha, Mitchell Parker, Luke Pike, Ceilidh Smith, Ariana Familiar, Kevin Liu, Sanjay Prabhu, Omar Arnaout, Pratiti Bandopadhayay, Ali Nabavizadeh, Sabine Mueller, Hugo Aerts, Raymond Huang, Tina Poussaint, and Benjamin H. Kann. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*, 2026. doi: 10.1038/s41593-026-02202-6. URL https://www.nature.com/articles/s41593-026-02202-6.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, and Yiqing Shen. Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 36:17599–17612, 2025.

Michael W Weiner, Paul S Aisen, Clifford R Jack Jr, William J Jagust, John Q Trojanowski, Leslie Shaw, Andrew J Saykin, John C Morris, Nigel Cairns, Laurel A Beckett, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, 6(3): 202–211, 2010.

Yizhou Wu, Shansong Wang, Yuheng Li, Mojtaba Safari, Mingzhe Hu, Chih-Wei Chang, Harini Veeraraghavan, and Xiaofeng Yang. Braindino: A brain mri foundation model for generalizable clinical representation learning. *arXiv preprint arXiv:2604.27277*, 2026.

Guangqian Yang, Tong Ding, Wenlong Hou, Yue Xun, Ye Du, Qian Niu, and Shujun Wang. Brainanytime: Anatomy-aware cross-modal pretraining for brain image analysis with arbitrary modality availability. *arXiv preprint arXiv:2605.13059*, 2026.

Yanhong Zhou, Chuangchuang Tan, Dong Wen, Hongmin Sun, Wei Han, and Yuchen Xu. The biomarkers for identifying preclinical alzheimer's disease via structural and functional magnetic resonance imaging, 2016.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

Cohort	Group	N	Sex (M/F)	Age	Education	MMSE	ADAS11	ADASQ4	RAVLT-imm.
ADNI	CN	360	174 / 186	74.6 (5.8)	16.2 (2.8)	29.1 (1.1)	6.0 (3.0)	2.9 (1.7)	44.5 (10.0)
	AD	307	169 / 138	75.1 (7.7)	15.2 (3.0)	23.2 (2.1)	19.7 (6.7)	8.6 (1.5)	22.7 (7.5)
	sMCI	316	191 / 125	72.6 (7.5)	16.1 (2.8)	28.0 (1.7)	8.6 (3.6)	4.6 (2.2)	37.8 (10.5)
	pMCI	220	130 / 90	74.0 (7.1)	15.8 (2.9)	26.8 (1.7)	13.1 (4.5)	7.0 (2.1)	28.0 (6.8)
	MCI	213	114 / 99	72.5 (8.2)	15.9 (3.0)	27.9 (1.8)	8.9 (4.2)	4.7 (2.5)	37.2 (11.8)
	All	**1,416**	**778 / 638**	**73.9 (7.3)**	**15.8 (2.9)**	**27.0 (2.7)**	**11.1 (6.8)**	**5.4 (2.9)**	**34.6 (12.5)**

Table A1: Demographic and cognitive characteristics of the ADNI cohort. Continuous variables are reported as mean (standard deviation) across subjects; age and education are measured in years. M and F denote male and female, respectively.

Cohort	Group	N	Sex (M/F)	Age	MMSE
OASIS	CN	451	259 / 192	71.8 (7.4)	28.5 (1.3)
	AD	362	207 / 155	76.9 (7.8)	21.6 (3.9)
	All	**813**	**466 / 347**	**74.1 (7.9)**	**25.8 (4.2)**

Table A2: Demographic and cognitive characteristics of the OASIS cohort. Age and MMSE are reported as mean (standard deviation) across subjects, with age measured in years. M and F denote male and female, respectively.

A Datasets and Preprocessing

A.1 Study Cohorts

AGQ is evaluated on two cohorts, ADNI (adni.loni.usc.edu) and OASIS (sites.wustl.edu/oasisbrains/). Their group-wise demographic and cognitive statistics are given in Tables A1 and A2, respectively.

ADNI. The cohort consists of T1-weighted sMRI scans from the ADNI 1, ADNI GO and ADNI 2 baseline and comprises 1,416 subjects. Only the month-0 visit of each subject enters the dataset, so no individual contributes more than one scan. This cohort supplies region-level pretraining, AD diagnosis and all regression targets.

OASIS. The cohort comprises 1,013 subjects drawn from OASIS. Cognitive measurements other than MMSE are incomplete, so it is used for AD diagnosis only.

A.2 Image Preprocessing

This work preprocesses brain imaging data via conventional procedures. Both cohorts are processed through the Clinica platform (www.clinica.run). Each scan first undergoes anterior commissure–posterior commissure correction and is then affine-registered to a normalised template in MNI152 space. Skull stripping is performed with SynthStrip (w3id.org/synthstrip). The resulting volumes are resampled to 1.5 mm isotropic resolution, producing a spatial extent of $112 \times 128 \times 112$ voxels.

For each brain imaging scan, intensity is standardised within the brain region: voxels with non-zero intensity are mapped to zero mean and unit variance per subject, and background voxels are set to zero, so that padding introduced later cannot shift the distribution. The AAL2 label map Rolls et al. (2015) undergoes the same padding and resampling as the imaging volumes, with nearest-neighbour interpolation.

A.3 Region-Level Supervision

Anatomical pretraining requires one supervisory value per atlas region and subject. We overlay the AAL2 parcellation on each subject's gray matter (GM) segmentation and record the volumes of the $K = 120$ regions, yielding a volume target vector $\mathbf{y} \in \mathbb{R}^K$. The targets are scaled by 100 so that their range falls above the Huber transition point and the loss operates in its quadratic regime for

Encoder	Backbone	Input	Tokens N	d	Token extent
SAM-Med3D	3D ViT-B	128^3	8^3	384	$24\times24\times24$
BrainIAC	3D ViT-B	96^3	6^3	768	$28\times32\times28$
MedicalNet	3D ResNet-18	96^3	12^3	512	$14\times16\times14$
SwinUNETR	SwinViT	96^3	6^3	384	$28\times32\times28$

Table A3: Token geometry of the four frozen backbones. The native $112 \times 128 \times 112$ volume at 1.5 mm is centre-padded to 128^3 for SAM-Med3D and resampled to 96^3 for the other three. Token extent is the physical size of one spatial token in millimetres, obtained by mapping the token grid back to the original field of view. MedicalNet exposes a residual feature map at stride 8 and SwinUNETR the stage-3 output of its Swin encoder; both are flattened channel-last, in the same order used to build $\mathbf{W}$.

Algorithm A1 Atlas-to-Patch Soft Assignment

Require: Atlas label map A; token voxel supports $\{\mathcal{V}_p\}_{p=1}^{N}$; region count K; stability constant ϵ
Ensure: Soft assignment matrix $\mathbf{W} \in \mathbb{R}^{K\times N}$

```
Apply to A the spatial transform used for the MRI volumes, with nearest-neighbour interpolation
for p = 1, ..., N do
    for r = 1, ..., K do
        W̃_{r,p} ← (1/|V_p|) Σ_{v∈V_p} 𝕀(A(v) = r)
    end for
end for
for r = 1, ..., K do
    W_{r,p} ← W̃_{r,p} / ( Σ_{j=1}^{N} W̃_{r,j} + ε )
end for
return W
```

typical residuals. Because this signal is derived from the imaging data alone, region-level pretraining consumes no diagnostic label and thus can use all subjects.

B Implementation Details and Hyperparameters

B.1 Backbones and Token Geometry

We selected four open-source models as the backbone of AnatoQ. The four pretrained encoders are frozen in every experiment. In Table A3, the details of each encoder are given. Since encoders discretise a brain volume very differently, AnatoQ derives the assignment matrix $\mathbf{W}$ according to the backbone's token grid.

Token extent also bounds how finely anatomy can be resolved. On the 6^3 grids of BrainIAC and SwinUNETR a single token occupies roughly 25 cm^3 and overlaps many AAL2 regions at once, so the rows of $\mathbf{W}$ belonging to small structures draw on a handful of tokens that they share with several neighbours. MedicalNet's 12^3 grid reduces token volume by about eightfold and separates regional support correspondingly. This is the resolution limit referred to in the Discussion, and it is a property of the backbone rather than of the adapter.

B.2 Atlas Assignment and Region Tokenisation

The soft assignment matrix depends only on atlas geometry and the token grid, so it is computed once per backbone and reused for every subject and every task. Algorithm A1 states that computation and Algorithm A2 the AGQ forward pass, both in the notation of the Methodology section.

Algorithm A2 Atlas-Guided Q-Former (AGQ) Forward Pass

Require: Spatial tokens $\mathbf{P} \in \mathbb{R}^{N\times d}$; assignment $\mathbf{W} \in \mathbb{R}^{K\times N}$; region identity embedding $\mathbf{E}_{\text{region}} \in \mathbb{R}^{K\times d}$
Ensure: Region tokens $\mathbf{R} \in \mathbb{R}^{K\times d}$
$\mathbf{Z} \leftarrow \mathbf{W}\mathbf{P}$ {anatomy-grounded init., Eq. 4}
$\mathbf{Q} \leftarrow \mathbf{Z} + \mathbf{E}_{\text{region}}$ {region identity, Eq. 5}
$\mathbf{R} \leftarrow \mathrm{LN}\big(\mathrm{MHA}(\mathbf{Q}, \mathbf{P}, \mathbf{P})\big)$ {refinement, Eq. 6}
return $\mathbf{R}$

Algorithm A3 Two-Stage Training of AnatoQ

Require: Spatial tokens $\{\mathbf{P}^{(i)}\}$; region GM targets $\{\mathbf{y}^{(i)}\}$; downstream labels $\{t^{(i)}\}$; masking ratio ρ
Stage 1: anatomical pretraining
for $e = 1, \ldots, E_1$ **do**
  for each minibatch **do**
    $\mathbf{P} \leftarrow \mathbf{P} \odot \mathbf{m},\ \mathbf{m}_p \sim \mathrm{Bernoulli}(1-\rho)$
    $\mathbf{R} \leftarrow \Phi_{\mathrm{AGQ}}(\mathbf{P}, \mathbf{W});\ \hat{\mathbf{y}} \leftarrow g(\mathbf{R})$
    $\mathcal{L} \leftarrow \sum_{r=1}^{K} \mathrm{Huber}_{\delta}(\hat{y}_r,\, y_r)$
    Update Φ_{AGQ} and g
  end for
end for
Keep the last checkpoint; discard g
Stage 2: downstream probing
Freeze f_θ and Φ_{AGQ}
for $e = 1, \ldots, E_2$ **do**
  for each minibatch **do**
    $\mathbf{R} \leftarrow \Phi_{\mathrm{AGQ}}(T(f_\theta(X)), \mathbf{W})$
    $\mathbf{h} \leftarrow \frac{1}{K}\sum_{r=1}^{K} \mathbf{R}_r;\ \hat{t} \leftarrow \mathrm{head}(\mathbf{h})$
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{task}}(\hat{t}, t)$; update the head only
  end for
end for
return downstream head

B.3 Training Procedure

The two-stage training procedure is summarised in Algorithm A3, and the complete configurations for both stages are reported in Table A4. Two design choices specific to AnatoQ warrant further clarification.

During anatomical pretraining, a fraction $\rho = 0.3$ of the spatial tokens in each sample is randomly masked. This strategy encourages the region queries to infer regional information from distributed contextual evidence rather than relying exclusively on individual local tokens. The region decoder g is a bottlenecked two-layer MLP applied independently to each region token. It is used solely to instantiate the auxiliary pretraining objective and is discarded after pretraining.

During downstream probing, both the backbone and the pretrained AGQ remain frozen; task-specific prediction heads are attached to the aggregated region representation, and only their parameters are optimized.

B.4 Evaluation

Both task families consume the same representation: the K region tokens are averaged into one d-dimensional vector, which is exactly what a head would receive under mean pooling of the raw patch tokens. Head capacity and input dimensionality are therefore constant across every row of Tables 1 and 2, and any difference between AGQ and mean pooling must come from the representation.

	Stage 1: pretraining	Stage 2: cls.	Stage 2: reg.
Trainable	AGQ + region dec.	downstream head	downstream head
Frozen	backbone	backbone, AGQ	backbone, AGQ
Objective	Huber (δ=5)	CE	MSE
Optimizer	AdamW	AdamW	AdamW
LR	5×10^{-5} (AGQ), 1×10^{-4} (dec.)	1×10^{-3}	1×10^{-3}
WD	0.01 (AGQ), 0 (dec.)	0.01	0.01
LR schedule	cosine	cosine	cosine
Warmup ratio	0.05	0.1	0.1
Epochs	30	30	30
Batch size	32	64	64
Grad. clip	1.0	—	—
Token masking ρ	0.3	—	—
Region tokens K	120	120	120
Attention heads	8	—	—
Precision	FP32	FP32	FP32
Seed	42	42	42

Table A4: Training configurations for both stages. AGQ denotes Atlas-Guided Q-Former; cls., classification; reg., regression; MLP, multilayer perceptron; CE, cross-entropy; MSE, mean squared error; LR, learning rate; WD, weight decay; dec., decoder; and FP32, 32-bit floating-point precision. Entries marked "—" represent not applicable to the corresponding stage.

For the metrics, AD is the positive class throughout. Accuracy is the proportion of correctly classified test subjects, sensitivity and specificity are the recalls of the AD and CN classes, and AUC is computed from the predicted probability of the AD class. Regression is scored by MAE and by Pearson's r between predictions and ground truth.

B.5 Computing Environment

All experiments were run on a single NVIDIA A100 GPU with 40 GB of memory. Since the backbone is frozen, each subject is passed through the encoder once and both training stages then operate on the resulting fixed token sequence; peak memory is therefore governed by the token count N of Table A3 rather than by the depth of the volumetric encoder, and every configuration in Table A4 fits on a single device at the batch size listed there. Table A5 specifies the software environment. Distributed launches are delegated to Accelerate, so the same script runs unchanged on one or several devices, and the learning-rate schedule of Table A4 is defined against the number of optimiser steps rather than the number of devices, leaving results invariant to the launch configuration.

C Parameter Efficiency

AGQ introduces a region identity embedding ($K\times d$ parameters), one multi-head cross-attention layer ($4d^2 + 4d$) and an output layer normalisation ($2d$). The assignment matrix $\mathbf{W}$ adds none, being fully determined by the atlas and the token grid. The module therefore holds between 0.64M and 2.46M parameters depending on the width of the backbone it adapts, one to two orders of magnitude below the volumetric encoders themselves.

Table A6 reports the parameter counts of AGQ, the auxiliary pretraining decoder, and the downstream prediction heads. Adapting AnatoQ to a new backbone requires only recomputing $\mathbf{W}$ for the

Component	Version	Role
Python	3.9	—
CUDA	12.1	GPU runtime
PyTorch	≥ 2.1	models, optimisation
Accelerate	≥ 0.30	distributed launch
Transformers	≥ 4.45	cosine schedule with warmup
MONAI	1.3.2	BrainIAC, SwinUNETR backbones
NiBabel	≥ 5.0	NIfTI volume and atlas I/O
SciPy	≥ 1.10	resampling, Spearman ρ
scikit-learn	≥ 1.3	ROC-AUC
NumPy, pandas	—	array and tabular handling
Matplotlib, Nilearn	—	region rendering (Fig. 3)

Table A5: Software environment. Versions are the lower bounds the implementation is written against, except MONAI, which is pinned because the two backbones it provides are version-sensitive.

Encoder	d	AGQ	Stage-1 dec.	Cls. head	Reg. head
SAM-Med3D	384	0.64 M	167 k	1.5 k	75 k
BrainIAC	768	2.46 M	481 k	3.1 k	298 k
MedicalNet	512	1.11 M	255 k	2.1 k	133 k
SwinUNETR	384	0.64 M	167 k	1.5 k	75 k

Table A6: Trainable parameter counts introduced by AnatoQ with $K = 120$ region tokens and eight attention heads. Here, d denotes the encoder feature dimension; AGQ, Atlas-Guided Q-Former; dec., decoder; Cls., classification; Reg., regression; and k and M, thousand and million parameters, respectively. The Stage-1 region decoder is used only for anatomical pretraining and discarded thereafter; downstream inference uses AGQ with the corresponding task-specific head.

corresponding token geometry and pretraining a lightweight AGQ instance. Once AGQ has been pretrained, adaptation to a new downstream task requires optimizing only a task-specific prediction head, while both the backbone and AGQ remain frozen.

D Ablation Protocol

AGQ composes three operations — (A) atlas-guided pooling, (B) the learnable region identity embedding, and (C) cross-attention refinement — and the ablation of Figure 2 removes them one at a time. Table A7 records the query each variant delivers to the head.

Three controls keep the comparison interpretable. All four tokenizer variants emit $K = 120$ region tokens, and mean pooling emits one token that the head consumes at the same dimensionality, so no row gains from token count or head capacity. Every row draws on the same backbone features, splits, head and optimiser settings, leaving the tokenizer module as the only difference. The output layer normalisation is kept in all four tokenizer variants, being a normalisation rather than a region-discriminative operation.

Static Atlas and Mean Pooling have no subject-dependent trainable parameters, so pretraining cannot change them and they are evaluated directly as frozen probes. The remaining variants are given a trained cross-attention before they are compared. All rows in Figure 2 use the same pretraining stage as the main results.

E Regional Attribution and Task-Specific Anatomy

E.1 Regional Attribution

Every output token of AGQ carries a fixed atlas identity, allowing predictions to be attributed to named anatomical regions rather than individual voxels. For a trained model and a given subject, let $\mathbf{R}_r$ denote the refined token of region r, and let o denote the scalar output under analysis. For

Variant	(A)	(B)	(C)	Query fed to the head
AGQ (full)	✓	✓	✓	$\mathrm{MHA}(\mathbf{WP}+\mathbf{E},\mathbf{P})$
Atlas Query Only	✓	✓	×	$\mathbf{WP}+\mathbf{E}$
Q-Former	×	✓	✓	$\mathrm{MHA}(\mathbf{E},\mathbf{P})$
Static Atlas	✓	×	×	$\mathbf{WP}$
Mean Pooling	×	×	×	$\frac{1}{N}\sum_p \mathbf{p}_p$

Table A7: Ablation variants. (A) atlas-guided pooling, (B) region identity embedding $\mathbf{E} \equiv \mathbf{E}_{\mathrm{region}}$, (C) cross-attention refinement. Removing (A) turns the anatomy-grounded queries into generic learnable queries, recovering a conventional Q-Former; removing (C) reduces the module to a fixed, geometry-determined aggregation.

diagnosis, o is the logit of the ground-truth class; for regression, o is the model prediction, which is standardised with training-set statistics since targets are Z-normalised during training. Regional attribution is defined as

$$s_r = \left\| \bar{\mathbf{R}}_r \odot \nabla_{\mathbf{R}_r} o \right\|_1 = \sum_{c=1}^{d} \left| \left(R_{r,c} - \bar{R}_c\right) \frac{\partial o}{\partial R_{r,c}} \right|, \qquad \bar{\mathbf{R}}_r = \mathbf{R}_r - \tfrac{1}{K}\textstyle\sum_{r'} \mathbf{R}_{r'}, \tag{8}$$

where $\odot$ denotes element-wise multiplication, $\bar{R}_c$ is the mean of channel c over the K region tokens of the subject, and the ℓ_1 norm is taken over feature channels. The across-region mean is subtracted to isolate region-specific evidence from the component shared by all region tokens. Attribution is computed on the held-out test subjects of each downstream task, and all test subjects are included regardless of whether they are correctly predicted.

E.2 Concentration and Stability

Table 3 summarises attribution with two statistics, both computed separately for each of the six downstream tasks and reported as the mean and standard deviation across tasks for each encoder.

Top-k attribution mass measures concentration. The per-subject attribution vector is normalised to sum to one, $\tilde{s}_r = s_r / \sum_{r'} s_{r'}$, and the mass of its k largest entries is accumulated,

$$M_k = \sum_{r \in \mathcal{T}_k} \tilde{s}_r, \tag{9}$$

where $\mathcal{T}_k$ indexes the k regions with the largest $\tilde{s}_r$, with $k = 20$ of $K = 120$. M_k is then averaged across subjects. A model spreading evidence uniformly would score $k/K \approx 16.7\%$, so the observed 25.4% to 40.2% reflects genuine concentration.

Attribution stability asks whether the anatomical ranking transfers to other subjects. For each task, the test subjects are randomly partitioned into two *disjoint* halves, the subject-averaged attribution vector is formed within each half from the raw scores s_r without per-subject normalisation, and the Spearman rank correlation between the two vectors is computed. This procedure is repeated over 20 random partitions and the resulting correlations are averaged. Disjointness is what makes the number meaningful: correlating two overlapping resamples drives ρ towards 1 merely because the groups share members, whereas a split-half design with no shared subjects tests whether a ranking estimated from one set of subjects predicts the ranking obtained from a different set. Since raw scores are used, ρ reflects the consistency of the ranking of average attribution strength across the two halves.

E.3 Task-Specific Salient Regions

Tables A8 and A9 report the twenty regions rendered in Figure 3 together with their scores, for all six downstream tasks. The displayed quantity $\bar{s}_r$ is a variant of s_r adapted for display: channel contributions are summed before the absolute value is taken, and the per-subject vector is min–max normalised to $[0, 1]$ before averaging over subjects. Values are therefore comparable within a task but not across tasks, and a high $\bar{s}_r$ indicates not only strong influence but consistent influence, since a region reaches a high average only by ranking near the top in most subjects.

	AD diagnosis		Age		MMSE	
Rank	Region	$\bar{s}_r$	Region	$\bar{s}_r$	Region	$\bar{s}_r$
1	Caudate L	0.599	Pallidum L	0.745	Cerebellum 10 L	0.927
2	Thalamus R	0.596	Cerebellum 10 L	0.740	Heschl L	0.918
3	OFC lat R	0.579	Cerebellum 10 R	0.720	Cerebellum 10 R	0.877
4	Parietal Inf R	0.573	Vermis 1/2	0.712	Pallidum L	0.842
5	Temporal Pole Sup R	0.571	Thalamus L	0.696	OFC lat R	0.823
6	Heschl L	0.570	Vermis 3	0.694	Paracentral Lobule L	0.814
7	Heschl R	0.568	Thalamus R	0.693	Parietal Sup R	0.808
8	Temporal Sup R	0.565	Cerebellum 3 L	0.692	Caudate L	0.770
9	Frontal Inf Oper R	0.564	Caudate L	0.675	Paracentral Lobule R	0.763
10	Pallidum L	0.564	Cerebellum 3 R	0.672	Temporal Pole Sup R	0.758
11	Thalamus L	0.563	Vermis 4/5	0.666	Cerebellum 3 L	0.757
12	Precentral R	0.561	Cingulate Post L	0.653	Frontal Inf Tri R	0.754
13	Caudate R	0.561	Vermis 6	0.638	Vermis 1/2	0.753
14	Occipital Sup L	0.561	Vermis 7	0.631	Precentral R	0.732
15	Paracentral Lobule R	0.559	Cerebellum 4/5 L	0.631	Parietal Inf R	0.732
16	Vermis 1/2	0.558	Parietal Sup R	0.630	Postcentral R	0.713
17	Frontal Inf Orb 2 R	0.558	Calcarine R	0.629	Frontal Inf Orb 2 R	0.705
18	Cerebellum 10 R	0.558	Heschl L	0.625	Vermis 3	0.700
19	Temporal Pole Sup L	0.557	Cerebellum 4/5 R	0.617	Heschl R	0.693
20	Olfactory L	0.557	Paracentral Lobule L	0.613	Thalamus L	0.686

Table A8: Top-20 salient regions for AD diagnosis, age, and MMSE, obtained with MedicalNet as the frozen encoder. $\bar{s}_r$ denotes the subject-averaged display saliency defined in Appendix E.3 and averaged over held-out test subjects ($n = 32$ for AD diagnosis and $n = 40$ for each regression task). Region names follow the AAL2 parcellation, with L and R denoting the left and right hemispheres, respectively. These regions are rendered in Figure 3.

The tables expose structure that the renderings cannot. Two regions, the right cerebellar lobule X and the left pallidum, appear in the top twenty of all six tasks, and a further five — left caudate, left cerebellar lobule X, left cerebellar lobule III, left Heschl's gyrus and vermis lobules I–II — appear in five of six, forming a subcortical and cerebellar core that the model consults regardless of objective. Against that shared background the tasks separate sharply. AD diagnosis is the only one whose leading ranks are not cerebellar: its first three regions are the left caudate, right thalamus and right lateral orbitofrontal cortex, and its first cerebellar entry appears only at rank sixteen, whereas the five demographic and cognitive targets all place cerebellar structures within their top three.

The two memory-specific scales separate further still, and in anatomically coherent directions. ADASQ4, a delayed word-recall item, is the only task to recruit medial temporal memory structures: the bilateral parahippocampal gyri and amygdalae enter its top twenty and appear in no other task. RAVLT-immediate instead recruits a contiguous orbitofrontal and ventromedial prefrontal cluster spanning ten of its twenty regions — the lateral, medial, anterior and posterior orbitofrontal subdivisions, the medial orbital and inferior orbital gyri, and the gyrus rectus — of which nine are unique to that task. These dissociations were not supervised in any way; they follow from applying the same frozen adapter to different prediction targets.

	ADAS11		ADASQ4		RAVLT-immediate	
Rank	Region	$\bar{s}_r$	Region	$\bar{s}_r$	Region	$\bar{s}_r$
1	Cerebellum 10 L	0.838	Cerebellum 10 L	0.839	Cerebellum 10 L	0.698
2	Heschl L	0.805	Vermis 1/2	0.713	Heschl L	0.657
3	Cerebellum 10 R	0.784	Cerebellum 10 R	0.705	Cerebellum 10 R	0.649
4	Pallidum L	0.740	Cerebellum 3 L	0.680	OFC lat R	0.621
5	Vermis 9	0.701	Cerebellum 3 R	0.644	Vermis 7	0.610
6	Vermis 7	0.662	Vermis 7	0.637	Pallidum L	0.609
7	Paracentral Lobule L	0.658	Pallidum L	0.583	Parietal Sup R	0.608
8	Parietal Sup R	0.657	Olfactory L	0.581	OFC lat L	0.607
9	Cerebellum 3 L	0.628	Frontal Sup Medial L	0.581	OFC med R	0.600
10	OFC lat R	0.617	ParaHippocampal L	0.575	Paracentral Lobule L	0.599
11	Frontal Inf Tri R	0.605	ParaHippocampal R	0.567	Frontal Med Orb R	0.592
12	Temporal Pole Sup R	0.602	Olfactory R	0.561	Frontal Med Orb L	0.591
13	Heschl R	0.592	Vermis 3	0.558	Rectus R	0.590
14	Vermis 1/2	0.578	Caudate L	0.557	Vermis 9	0.586
15	Precentral R	0.575	Temporal Pole Sup L	0.548	Temporal Pole Mid L	0.581
16	Caudate L	0.573	Vermis 9	0.548	Cerebellum 3 L	0.575
17	Vermis 8	0.571	Amygdala L	0.542	OFC post L	0.571
18	Vermis 3	0.564	Vermis 6	0.541	Frontal Inf Orb 2 L	0.565
19	Cerebellum 9 R	0.560	Amygdala R	0.539	OFC med L	0.564
20	Postcentral R	0.558	Thalamus R	0.535	OFC ant L	0.562

Table A9: Top-20 salient regions for ADAS11, ADASQ4, and RAVLT-immediate, obtained with MedicalNet as the frozen encoder. $\bar{s}_r$ denotes the subject-averaged display saliency defined in Appendix E.3 and averaged over held-out test subjects ($n = 40$ for each regression task). Region names follow the AAL2 parcellation, with L and R denoting the left and right hemispheres, respectively. These regions are rendered in Figure 3.