Appendix

Contents

A Dataset Details **2**
A.1 Pretrain Data Demographics . . . 2
A.2 ICD10-Code Mapping . . . 2
A.3 Sample Size and Label for Each Downstream Dataset . . . 4

B Algorithm and Evaluation Details **8**
B.1 Multiscale Masking . . . 8
B.2 Multiscale Masking Visualization . . . 9
B.3 Feature Map Visualization . . . 9
B.4 Modified Auxiliary Loss Free Load Balancing . . . 13
B.5 Multi-Modal Learning Methods . . . 13
B.6 Model Pretraining Details . . . 15
B.7 Model Evaluation Details . . . 15
B.8 Evaluation Scope . . . 15

C Unimodal Learning Experiments **18**
C.1 Average Performance Across Tasks . . . 18
C.2 AUPRC for Public Datasets and Best Achievable Unimodal Performance Across Datasets . . . 18
C.3 AUROC and AUPRC for Health System Datasets with Per Modality Performance . . . 18
C.4 Kaplan Meier Curve and C-index for Time-to-Event Tasks . . . 18
C.5 Age Prediction Performance as Regression . . . 18

D Multi-Modal Learning Experiments **20**
D.1 Average Performance Across Tasks . . . 20
D.2 Multi-Modal Learning Result Across Fusion Methods - Public Datasets . . . 20
D.3 Multi-Modal Gain Result - Public Datasets . . . 20
D.4 Multi-Modal Learning Result Across Fusion Methods - BIND-MGH . . . 20
D.5 Multi-Modal Gain Result - BIND-MGH . . . 20

E Ablation Studies Experiments **36**
E.1 Ablation Study on Number of Experts . . . 36
E.2 Ablation Study on Methods . . . 36
E.3 Ablation Study on NeuroVFM . . . 36
E.4 Ablation Study on Different Percent of Pretrain Data . . . 36
E.5 Ablation Study on Model Size . . . 36
E.6 Ablation Study on Patch Size . . . 36
E.7 Ablation Study on Pretraining with Uncurated vs. Curated Data . . . 36
E.8 Ablation Study on MAE vs. JEPA Pre-training Performance . . . 36

F Generalization Under Cohort and Modality Shifts **42**
F.1 Cross-Cohort Transfer Across Matched Clinical Endpoints . . . 42
F.2 Cross-Modality Generalization to DWI . . . 42

G Pre-Training Optimization Dynamics **49**

H Additional Validation Analyses **50**
H.1 Comparison with Simple CNN Baseline . . . 50
H.2 Few-shot Analysis . . . 51
AUROC and AUPRC on Few-shot Performance Averaged Across Modalities • Per-modality AUROC and AUPRC Few-short Performance
H.3 Fairness Analysis . . . 60

H.4 TSNE Visualization on Age Groups . . . 60
I Additional Result for MoE Routing Analysis and Visualization **62**
I.1 MoE Foreground vs. Background Routing . . . 62
I.2 MoE Different Modalities Routing . . . 62
I.3 MoE Heatmap . . . 62
I.4 MoE Visualization . . . 62
I.5 Attention Map Visualization . . . 62
References **74**

A Dataset Details

A.1 Pretrain Data Demographics

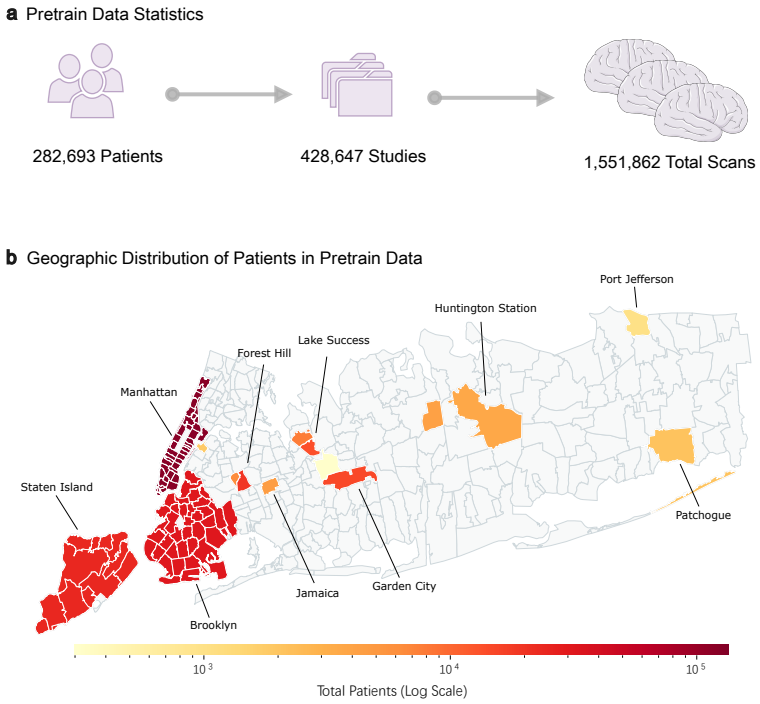
We present patient demographics information on age, ethnicity and gender for our pretrained data in Supplementary Table 1 from available EHR data. Due to the asynchronous update of the institutional EHR and PACS systems, demographic metadata were successfully integrated for $155,064/282,693$ of the cohort (before $01/01/2023$). The remaining represents the most recent imaging acquisitions $(2023-2025)$. The entire cohort was utilized for pretraining, while the matched subset was used for demographic-stratified examination and analysis. In total, our pretrain data contains 282,693 Patients, 428,647 Studies and 1,551,862 Total Scans across multiple regions in New York Area as shown in Supplementary Figure 1.

Supplementary Table 1. Pre-training Data Demographics

Category	Count	%
Age		
Child (0–12)	1,891	1.2
Teenager (13–17)	2,694	1.7
Young Adult (18–24)	5,160	3.3
Adult (25–34)	15,085	9.6
Adult (35–44)	18,985	12.1
Middle-Aged (45–54)	20,436	13.0
Older Adult (55–64)	24,737	15.7
Senior (65+)	56,842	36.1
Unknown	9,234	5.9
Ethnicity		
White	81,182	51.6
African / Black	14,486	9.2
Asian / Pacific Islander	6,971	4.4
Excluded / Other / Unknown / Refused	52,425	33.3
Gender		
Female	91,812	59.5
Male	54,723	35.5
Unknown	8,529	5.5
Total (unique)	**157,064**	

A.2 ICD10-Code Mapping

We present ICD10 Code Mapping details in Table 2, where the mapped labels are used for NYU-Langone and Longisland downstream evaluation.



Supplementary Figure 1. Pretrain Data Patients Statistics - **a,** Number of patients, studies and total scans used in pretrain data. **b,** Geographic distribution of patients in pretrain data

Supplementary Table 2. The definition of diseases in EHR by diagnosis codes and medications.

Disease	Definition in EHR
IPH	I61.0, I61.1, I61.2, I61.3, I61.4, I61.8, I61.9
IVH	I61.5, P52.1, P52.2, P52.3
ICH	IPH + IVH + I61.6, I62.9, P10.9, P52.4, P52.9
SDH	S06.5, I62.0
SAH	I60.*, S06.6, P52.5, P10.3
Cancer	C71.*, C79.3, D33.0, D33.1, D33.2, D33.3, D33.7, D33.9
Hydrocephalus	G91.*
Edema	G93.1, G93.5, G93.6, G93.82, S06.1
Dementia	G23.1, G30.*, G31.01, G31.09, G31.83, G31.85, G31.9, F01.*, F02.*, F03.*, G31.84, G31.1, **Medication:** DONEPEZIL, RIVASTIGMINE, GALANTAMINE, MEMANTINE, TACRINE
Major Depression Disorder	F32.*, F33.*

A.3 Sample Size and Label for Each Downstream Dataset

We report the number of patients and samples for each downstream evaluation dataset in Supplementary Table 3. For public datasets, the unimodal and multimodal evaluations include different sample sizes because multimodal analyses were restricted to subjects with all required modalities available, enabling balanced and controlled comparisons across modalities. The exact dataset splits used for evaluation are provided in our public code repository.

Supplementary Table 3. Benchmark dataset statistics. *Patients*: number of unique individuals. *Samples*: total imaging sessions aggregated across training, validation, and test splits. For ICSPR, MCSA, and time-to-event tasks, counts are task-specific after excluding sessions with missing labels or missing event/censoring information.* *Positive labels*: for binary classification and time-to-event tasks, positive-class or event sample count and prevalence (%); for 3-class tasks (CNP Psychiatric diagnosis, ICSPR Lesion type, MCSA Cognitive impairment, PPMI Parkinson diagnosis), class-0 / class-1 / class-2 sample counts (Healthy Control/Bipolar Disorder or Schizophrenia/ADHD for CNP Psychiatric diagnosis; Ischemic/Hemorrhagic/Absent for Lesion type; Cognitive Normal/Mild Cognitive Impairment/Alzheimer for MCSA Cognitive impairment; Healthy Control/Prodromal/Parkinson for PPMI diagnosis); for BIND-MGH multi-label findings see Table 4.

Dataset	Task	Modality	Patients	Samples	Positive labels
UNIMODAL					
ABIDE	ASD diagnosis	T1w	1,099	1,099	528 (48.1%)
ADHD-200	ADHD diagnosis	T1w	776	776	285 (36.7%)
ADNI	AD diagnosis	T1w	455	1,632	784 (48.0%)
	Amyloid-PET status	FLAIR	317	318	145 (45.6%)
		T1w	167	167	78 (46.7%)
		T2w	205	206	97 (47.1%)
	MCI to AD conversion	T1w	209	209	94 (45.0%)
BIND-MGH	Multi-label† (15 conditions)	FLAIR	11,802	22,142	See Table 4
		T1w	11,802	22,142	See Table 4
		T2w	11,802	22,142	See Table 4
CNP	Psychiatric diagnosis	T1w	265	265	125 / 99 / 41
ICSPR-Stroke	90-day mRS*	DWI	1,139	1,139	439 (38.5%)
		FLAIR	1,099	1,099	421 (38.3%)
		T1w	1,011	1,011	386 (38.2%)
		T2w	1,019	1,019	376 (36.9%)
	Lesion type*	DWI	2,643	2,643	1,878 / 295 / 470
		FLAIR	2,508	2,508	1,770 / 283 / 455
		T1w	2,281	2,281	1,612 / 267 / 402
		T2w	2,342	2,342	1,645 / 269 / 428
	LOS $>$ 8 days*	DWI	1,892	1,892	398 (21.0%)
		FLAIR	1,823	1,823	372 (20.4%)
		T1w	1,655	1,655	339 (20.5%)
		T2w	1,675	1,675	327 (19.5%)
MCSA	Cognitive impairment*	FLAIR	1,713	2,866	2,486 / 337 / 43
		T1w	1,713	2,866	2,486 / 337 / 43
	Stroke*	FLAIR	1,712	2,847	98 (3.4%)
		T1w	1,712	2,847	98 (3.4%)
	Hypertension*	FLAIR	1,712	2,847	1,849 (64.9%)
		T1w	1,712	2,847	1,849 (64.9%)

Continued on next page

	Dyslipidaemia*	FLAIR	1,712	2,847	2,319 (81.5%)
		T1w	1,712	2,847	2,319 (81.5%)
NACC	AD diagnosis	FLAIR	3,024	3,755	710 (18.9%)
		T1w	3,841	4,994	1,053 (21.1%)
		T2w	2,538	3,062	535 (17.5%)
	Amyloid-PET status	FLAIR	155	159	70 (44.0%)
		T1w	176	182	83 (45.6%)
		T2w	93	97	44 (45.4%)
OASIS-3	AD diagnosis	FLAIR	727	1,028	107 (10.4%)
		T1w	1,126	1,924	250 (13.0%)
		T2w	1,004	1,665	210 (12.6%)
OpenBHB	Age regression	T1w	3,984	3,984	—
PPMI	PD diagnosis	FLAIR	1,577	2,365	655 / 796 / 914
		T1w	1,000	1,763	570 / 335 / 858
		T2w	346	1,296	278 / 120 / 898
	Prodromal to PD conversion	FLAIR	882	882	86 (9.8%)
		T1w	397	397	30 (7.6%)
SOOP	mRS (binary)	FLAIR	647	647	302 (46.7%)
		T1w	647	647	302 (46.7%)
UCSF-PDGM	IDH mutation	DWI	495	495	103 (20.8%)
		FLAIR	495	495	103 (20.8%)
		T1w	495	495	103 (20.8%)
		T2w	495	495	103 (20.8%)
	Overall survival	FLAIR	495	495	248 (50.1%)
		T1w	495	495	248 (50.1%)
		T2w	495	495	248 (50.1%)
MULTIMODAL					
BIND-MGH	Multi-label† (15 conditions)	T1w + T2w/FLAIR	11,802	22,142	See Table 4
ICSPR-Stroke	90-day mRS*	T1w + FLAIR	993	993	377 (38.0%)
	Lesion type*		2,212	2,212	1,561 / 259 / 392
	LoS > 8 days*		1,620	1,620	326 (20.1%)
MCSA	Cognitive impairment*	T1w + FLAIR	1,713	2,866	2,486 / 337 / 43
	Stroke*		1,712	2,847	98 (3.4%)
	Hypertension*		1,712	2,847	1,849 (64.9%)
	Dyslipidaemia*		1,712	2,847	2,319 (81.5%)
NACC	AD diagnosis	T1w + T2w	3,132	3,772	525 (13.9%)
OASIS-3	AD diagnosis	T1w + T2w	883	1,384	163 (11.8%)
PPMI	PD diagnosis	T1w + T2w	302	1,339	286 / 121 / 932
SOOP	mRS (binary)	T1w + FLAIR	647	647	302 (46.7%)
UCSF-PDGM	IDH mutation	T1w + FLAIR	495	495	103 (20.8%)
INTERNAL COHORTS					
NYU-Langone	Multi-label‡ (11 conditions)	T1w	10,004	19,325	See Table 5
		T2w	19,132	29,237	See Table 5

Continued on next page

	FLAIR	15,937	23,302	See Table 5
NYU-LongIslandMulti-label[‡] (11 conditions)	T1w	3,306	8,024	See Table 6
	T2w	2,828	5,050	See Table 6
	FLAIR	2,596	3,376	See Table 6

[†] BIND-MGH comprises 15 binary radiology-finding labels: Astrocytoma, Atrophy, Cyst, Edema, Enhancement, Hematoma, Infarct, Ischemic, Mass effect, Midline shift, Multiple sclerosis, Schwannoma, Cancer, Glioblastoma multiforme, and Gliosis. Per-finding prevalence reported in Table 4.

[‡] NYU-Langone and NYU-Longisland each comprise 10 binary clinical-finding labels: Cancer, Hydrocephalus, Edema, Dementia, Intraparenchymal haemorrhage (IPH), Intraventricular haemorrhage (IVH), Subdural haematoma (SDH), Subarachnoid haemorrhage (SAH), Intracranial haemorrhage (ICH), and Major depressive disorder. There is also 1 combined finding Haemorrhage from all haemorrhage subtypes. Per-finding prevalence reported in Tables 5–6.

[*] ICSPR (90-day mRS, Lesion type, LoS $>$ 8 days) and MCSA (Cognitive impairment, Stroke, Hypertension, Dyslipidaemia): sessions with missing labels for a given task are excluded; patient and sample totals are task-specific.

Abbreviations. AD, Alzheimer's disease; MCI, mild cognitive impairment; ADHD, attention-deficit/hyperactivity disorder; ASD, autism spectrum disorder; DWI, diffusion-weighted imaging; FLAIR, fluid-attenuated inversion recovery; IDH, isocitrate dehydrogenase; LOS, length of hospital stay; mRS, modified Rankin Scale; PD, Parkinson's disease.

Supplementary Table 4. BIND-MGH radiology finding prevalence. Positive counts and prevalence are identical across T1w, T2w, and FLAIR unimodal splits, and across the multimodal split, because all modalities correspond to the same patient sessions with shared labels. Total: 22,142 sessions from 11,802 unique patients.

Finding	**Positive (*n*)**	**Prevalence (%)**
Astrocytoma	1,152	5.2
Atrophy	2,375	10.7
Cyst	1,700	7.7
Edema	2,119	9.6
Enhancement	5,169	23.3
Hematoma	1,733	7.8
Infarct	4,932	22.3
Ischemic	2,849	12.9
Mass effect	743	3.4
Midline shift	877	4.0
Multiple sclerosis	1,382	6.2
Schwannoma	282	1.3
Cancer	5,713	25.8
Glioblastoma multiforme	545	2.5
Gliosis	722	3.3

Prevalence = positive sessions / total sessions. Findings listed in source-dataset column order.

Supplementary Table 5. NYU-Langone radiology and clinical finding prevalence by modality. Each cell reports the number of positive sessions and prevalence (positive sessions / total sessions for that modality). T1w: 9,692 patients, 19,325 sessions. T2w: 17,485 patients, 29,237 sessions. FLAIR: 14,555 patients, 23,302 sessions.

Finding	T1w	T2w	FLAIR
Cancer	8,596 (44.5%)	9,360 (32.0%)	9,715 (41.7%)
Hydrocephalus	1,305 (6.8%)	1,753 (6.0%)	1,817 (7.8%)
Edema	2,707 (14.0%)	3,441 (11.8%)	3,099 (13.3%)
Dementia	2,193 (11.3%)	6,771 (23.2%)	3,694 (15.9%)
IPH	1,146 (5.9%)	1,424 (4.9%)	1,148 (4.9%)
IVH	709 (3.7%)	861 (2.9%)	718 (3.1%)
SDH	742 (3.8%)	1,067 (3.6%)	985 (4.2%)
SAH	622 (3.2%)	969 (3.3%)	880 (3.8%)
ICH	1,318 (6.8%)	1,621 (5.5%)	1,311 (5.6%)
Major depressive disorder	2,597 (13.4%)	5,099 (17.4%)	3,641 (15.6%)
Haemorrhage	2,006 (10.4%)	2,678 (9.2%)	2,267 (9.7%)

Prevalence = positive sessions / total sessions per modality. IPH, intraparenchymal haemorrhage; IVH, intraventricular haemorrhage; SDH, subdural haematoma; SAH, subarachnoid haemorrhage; ICH, intracranial haemorrhage.

Supplementary Table 6. NYU-Long Island radiology and clinical finding prevalence by modality. Each cell reports the number of positive sessions and prevalence (positive sessions / total sessions for that modality). T1w: 3,306 patients, 8,024 sessions. T2w: 2,828 patients, 5,050 sessions. FLAIR: 2,596 patients, 3,376 sessions.

Finding	T1w	T2w	FLAIR
Cancer	1,762 (22.0%)	587 (11.6%)	452 (13.4%)
Hydrocephalus	345 (4.3%)	199 (3.9%)	104 (3.1%)
Edema	1,377 (17.2%)	662 (13.1%)	416 (12.3%)
Dementia	1,314 (16.4%)	1,192 (23.6%)	601 (17.8%)
IPH	719 (9.0%)	363 (7.2%)	195 (5.8%)
IVH	419 (5.2%)	205 (4.1%)	111 (3.3%)
SDH	323 (4.0%)	202 (4.0%)	103 (3.1%)
SAH	260 (3.2%)	162 (3.2%)	80 (2.4%)
ICH	825 (10.3%)	413 (8.2%)	226 (6.7%)
Major depressive disorder	1,589 (19.8%)	1,035 (20.5%)	590 (17.5%)
Haemorrhage	1,030 (12.8%)	583 (11.5%)	311 (9.2%)

Prevalence = positive sessions / total sessions per modality. IPH, intraparenchymal haemorrhage; IVH, intraventricular haemorrhage; SDH, subdural haematoma; SAH, subarachnoid haemorrhage; ICH, intracranial haemorrhage.

B Algorithm and Evaluation Details

B.1 Multiscale Masking

We empirically observed that masking strategies can largely influence JEPA pre-training generalization for neuroimaging. Supplementary Algorithm 1 presents one of masking implementations we attempted that presents stable training and improved performance over original V-JEPA 2 masking implementation under our experimental setups. While it does not represent the optimal masking strategy for neuroimaging, it indicates the importance of careful masking design for special data type such as neuroimaging, where model can easily learn shortcut in latent space such as interpolating nearby pixels (anatomy of the brain is highly structured).

The algorithm operates on a 3D patch grid derived from a volume. Phase 1 places K multiscale blocks with log-uniform aspect ratios. Phase 2 reaches the exact target context size by eroding/dilating along block boundaries, preserving spatial coherence. Fixed output lengths is applied for direct batch stacking without padding. Phase 3 shuffles indices to ensure spatial information is conveyed only through position embeddings, not through sequence order.

Require: Patch-grid dimensions (H, W, D); total mask ratio $\rho \in (0,1)$
Require: Number of blocks K; spatial-scale range $[s_{\min}, s_{\max}]$, depth-scale range $[\delta_{\min}, \delta_{\max}]$, aspect-ratio range $[\alpha_{\min}, \alpha_{\max}]$, $\mathcal{N}_6(v) = \{u \in [H]\times[W]\times[D] : \|u-v\|_1 = 1\}$, the 6-connected neighbourhood of patch v.
Require: (Optional) Foreground map $\mathcal{F} \in \{0,1\}^{H\times W\times D}$
Ensure: Shuffled context indices $\mathbf{I}_{\text{enc}}$ and target indices $\mathbf{I}_{\text{pred}}$ of fixed lengths N_{enc} and N_{pred}

$N_{\text{pred}} \leftarrow \lfloor \rho HWD \rfloor$; $\quad N_{\text{enc}} \leftarrow HWD - N_{\text{pred}}$
$M \leftarrow \mathbf{1}^{H\times W\times D}$ ▷ M_v=1: context (kept); M_v=0: target (masked)

Phase 1. Multiscale block placement
for $k = 1, \ldots, K$ **do**
  $s \sim \mathcal{U}(s_{\min}, s_{\max})$, $\quad \delta \sim \mathcal{U}(\delta_{\min}, \delta_{\max})$
  $A \leftarrow \max(1, \lfloor sHW \rfloor)$, $\quad d \leftarrow \max\big(1, \min(\lfloor \delta D \rfloor, D)\big)$
  $\log \alpha \sim \mathcal{U}(\log \alpha_{\min}, \log \alpha_{\max})$ ▷ log-uniform aspect ratio
  $h \leftarrow \text{clip}\big(\lfloor \sqrt{A\,\alpha} \rceil, 1, H\big)$, $\quad w \leftarrow \text{clip}\big(\lfloor \sqrt{A/\alpha} \rceil, 1, W\big)$
  Sample origin (y, x, z) uniformly from valid placements
  $M[y{:}y{+}h,\ x{:}x{+}w,\ z{:}z{+}d] \leftarrow 0$
end for

Phase 2. Boundary-coherent count adjustment
$n \leftarrow \sum_v M_v$
while $n > N_{\text{enc}}$ **do** ▷ Erode: expand masked region inward
  $\mathcal{B} \leftarrow \{\, v : M_v{=}1 \text{ and } \exists u \in \mathcal{N}_6(v),\ M_u{=}0 \,\}$ ▷ outer boundary
  if $\mathcal{B} = \emptyset$ **then break**
  end if
  if $\mathcal{F}$ provided **then**
    $\mathcal{B}_0 \leftarrow \{b \in \mathcal{B} : \mathcal{F}_b{=}0\}$, $\quad \mathcal{B}_1 \leftarrow \{b \in \mathcal{B} : \mathcal{F}_b{=}1\}$
    $\mathcal{B} \leftarrow \text{shuffle}(\mathcal{B}_0) \,\|\, \text{shuffle}(\mathcal{B}_1)$ ▷ background removed first
  else
    $\mathcal{B} \leftarrow \text{shuffle}(\mathcal{B})$
  end if
  $\Delta \leftarrow \min(n - N_{\text{enc}},\ |\mathcal{B}|)$
  $M[\mathcal{B}[1{:}\Delta]] \leftarrow 0$, $\quad n \leftarrow n - \Delta$
end while
while $n < N_{\text{enc}}$ **do** ▷ Dilate: shrink masked region
  $\mathcal{I} \leftarrow \{\, v : M_v{=}0 \text{ and } \exists u \in \mathcal{N}_6(v),\ M_u{=}1 \,\}$ ▷ inner boundary
  if $\mathcal{I} = \emptyset$ **then break**
  end if
  if $\mathcal{F}$ provided **then**
    $\mathcal{I}_1 \leftarrow \{i \in \mathcal{I} : \mathcal{F}_i{=}1\}$, $\quad \mathcal{I}_0 \leftarrow \{i \in \mathcal{I} : \mathcal{F}_i{=}0\}$
    $\mathcal{I} \leftarrow \text{shuffle}(\mathcal{I}_1) \,\|\, \text{shuffle}(\mathcal{I}_0)$ ▷ foreground restored first
  else
    $\mathcal{I} \leftarrow \text{shuffle}(\mathcal{I})$
  end if
  $\Delta \leftarrow \min(N_{\text{enc}} - n,\ |\mathcal{I}|)$
  $M[\mathcal{I}[1{:}\Delta]] \leftarrow 1$, $\quad n \leftarrow n + \Delta$
end while

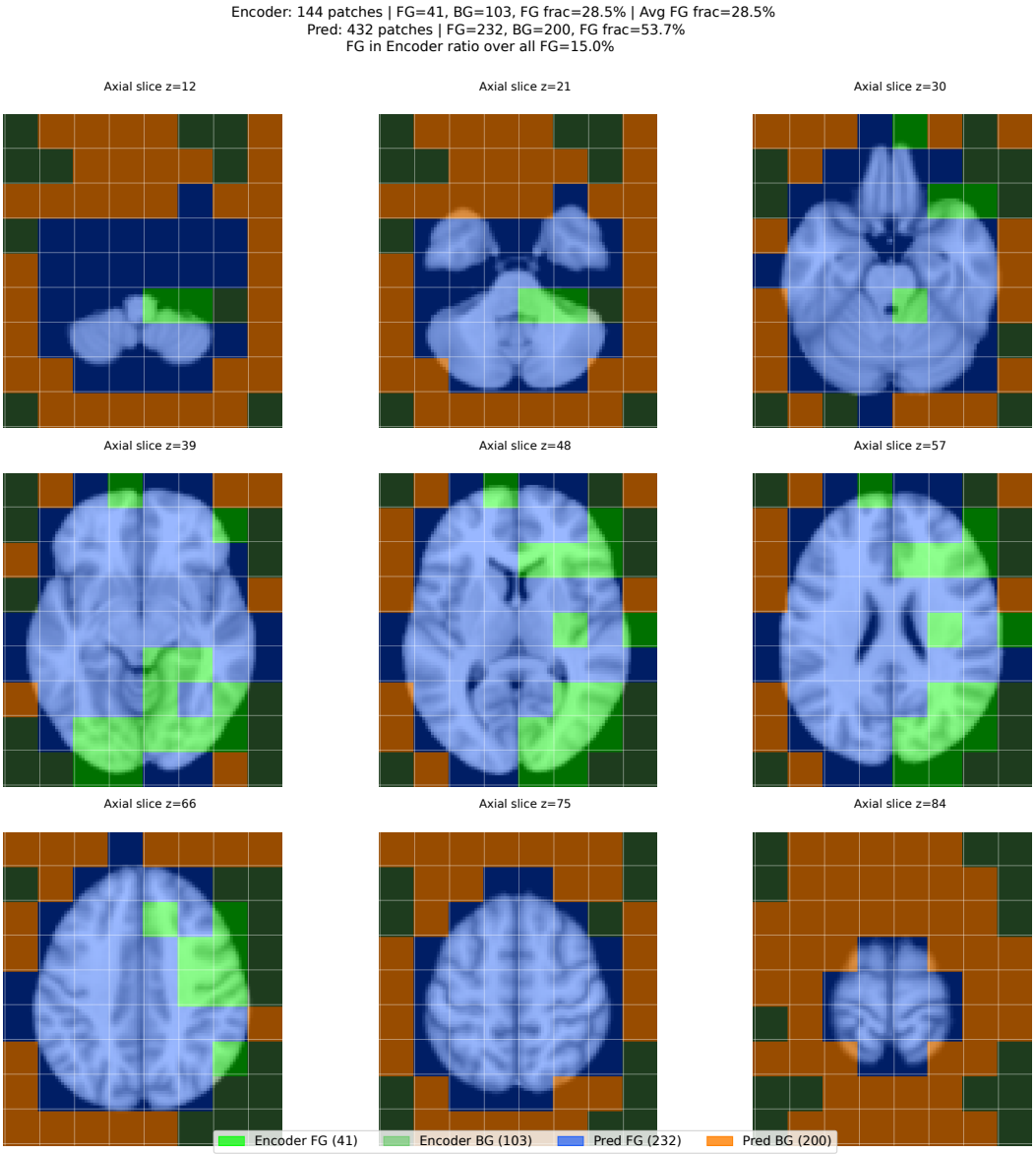
Phase 2b. Random fallback (guarantees exact target counts)
if $n > N_{\text{enc}}$ **then**
  Randomly select $n - N_{\text{enc}}$ patches from $\{v : M_v{=}1\}$ and set to 0
else if $n < N_{\text{enc}}$ **then**
  Randomly select $N_{\text{enc}} - n$ patches from $\{v : M_v{=}0\}$ and set to 1
end if

Phase 3. Index extraction
$\mathbf{I}_{\text{enc}} \leftarrow \text{shuffle}\big(\{v : M_v{=}1\}\big)$
$\mathbf{I}_{\text{pred}} \leftarrow \text{shuffle}\big(\{v : M_v{=}0\}\big)$
return $\mathbf{I}_{\text{enc}}$, $\mathbf{I}_{\text{pred}}$

Algorithm 1. Multiscale Volumetric Block Masking.

B.2 Multiscale Masking Visualization

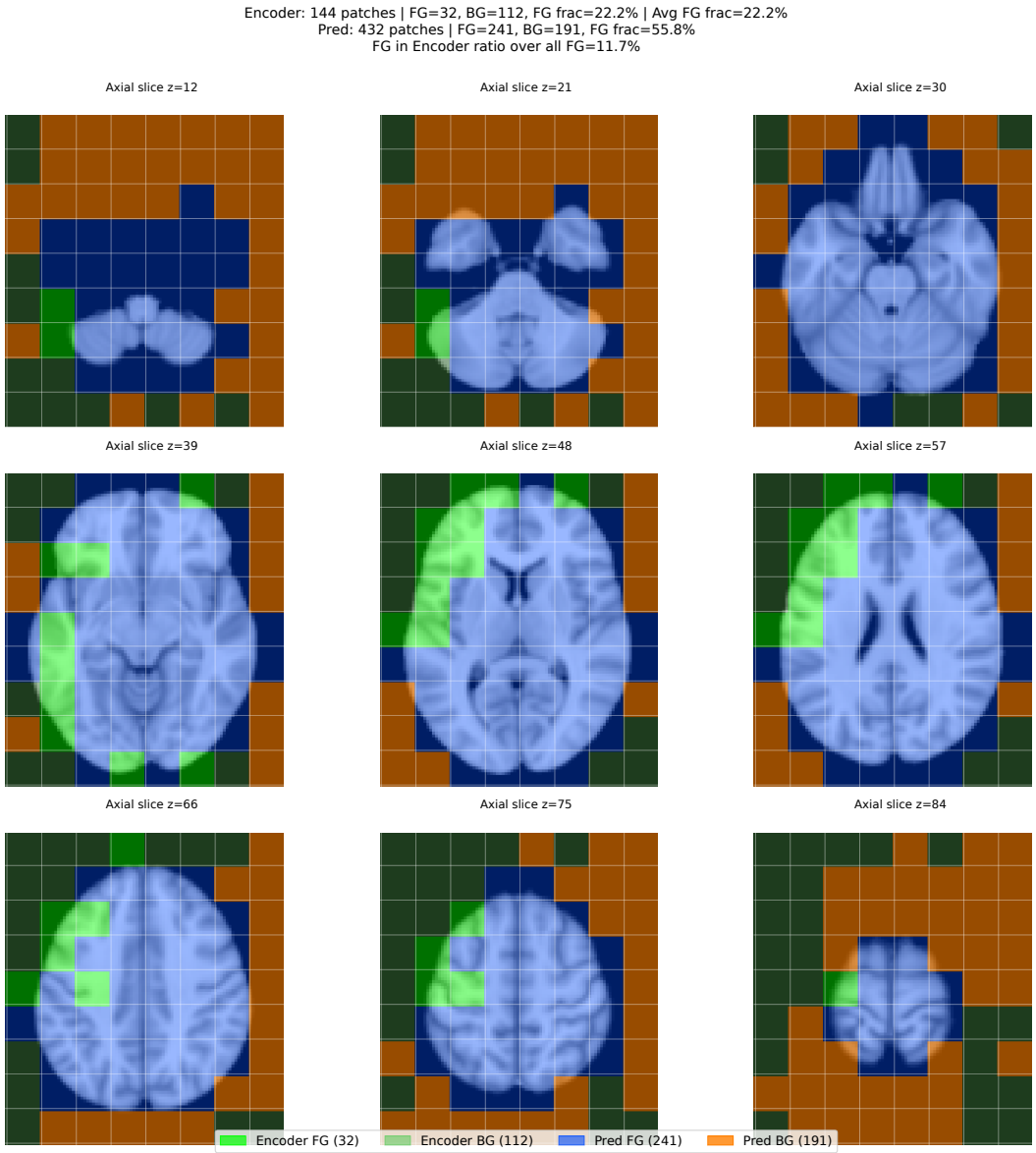
Supplementary Figures 2 to 4 show sample visualization for multiscale masking, with foreground and background in encoder/predictor shown as different colors. The ratio of foreground in encoder and predictor is also calculated to monitor the given context.



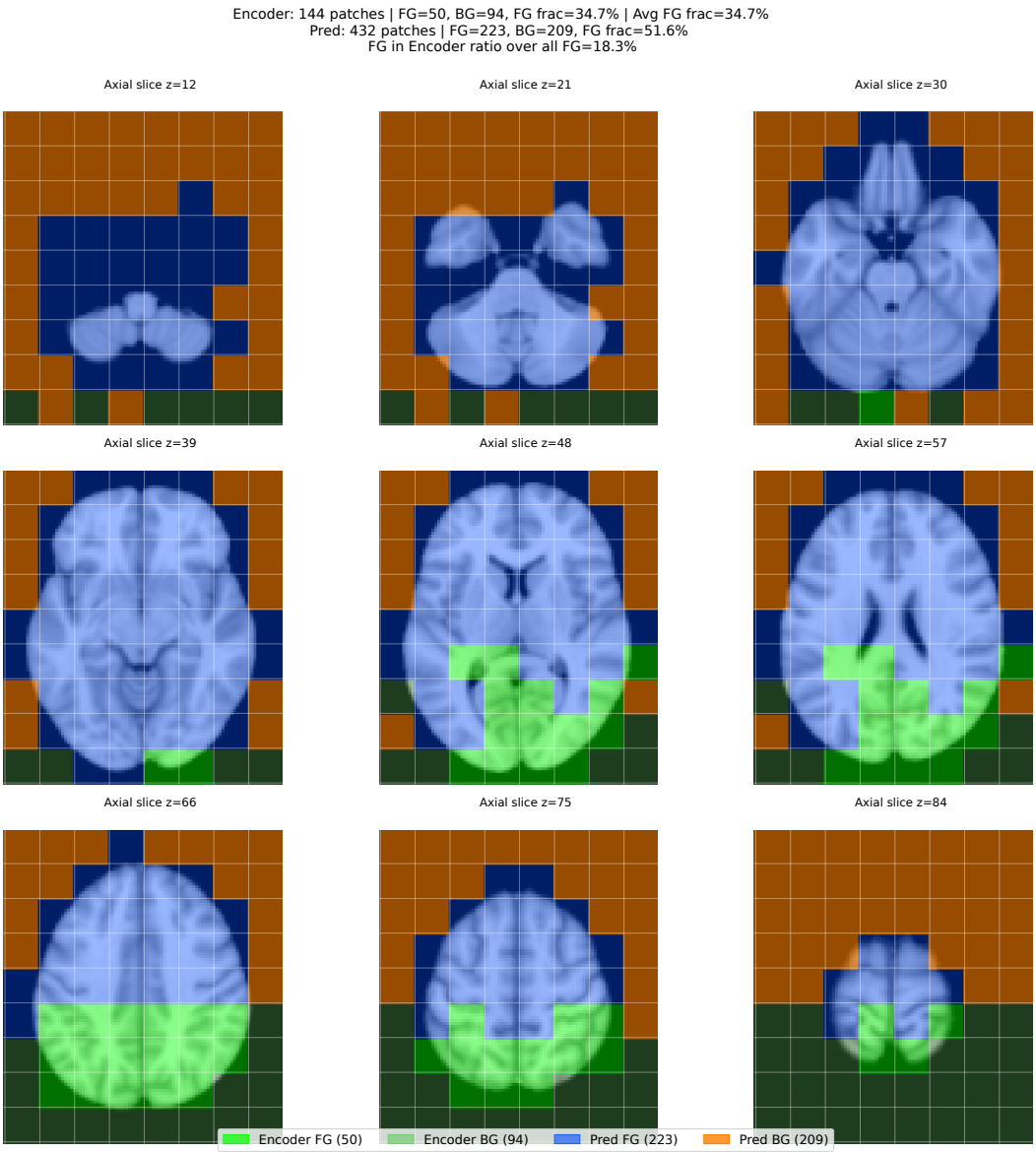
Supplementary Figure 2. Multiscale Masking Sample. Foreground is denoted as FG, background is denoted as BG and the foreground ratio for encoder/predictor are separately shown.

B.3 Feature Map Visualization

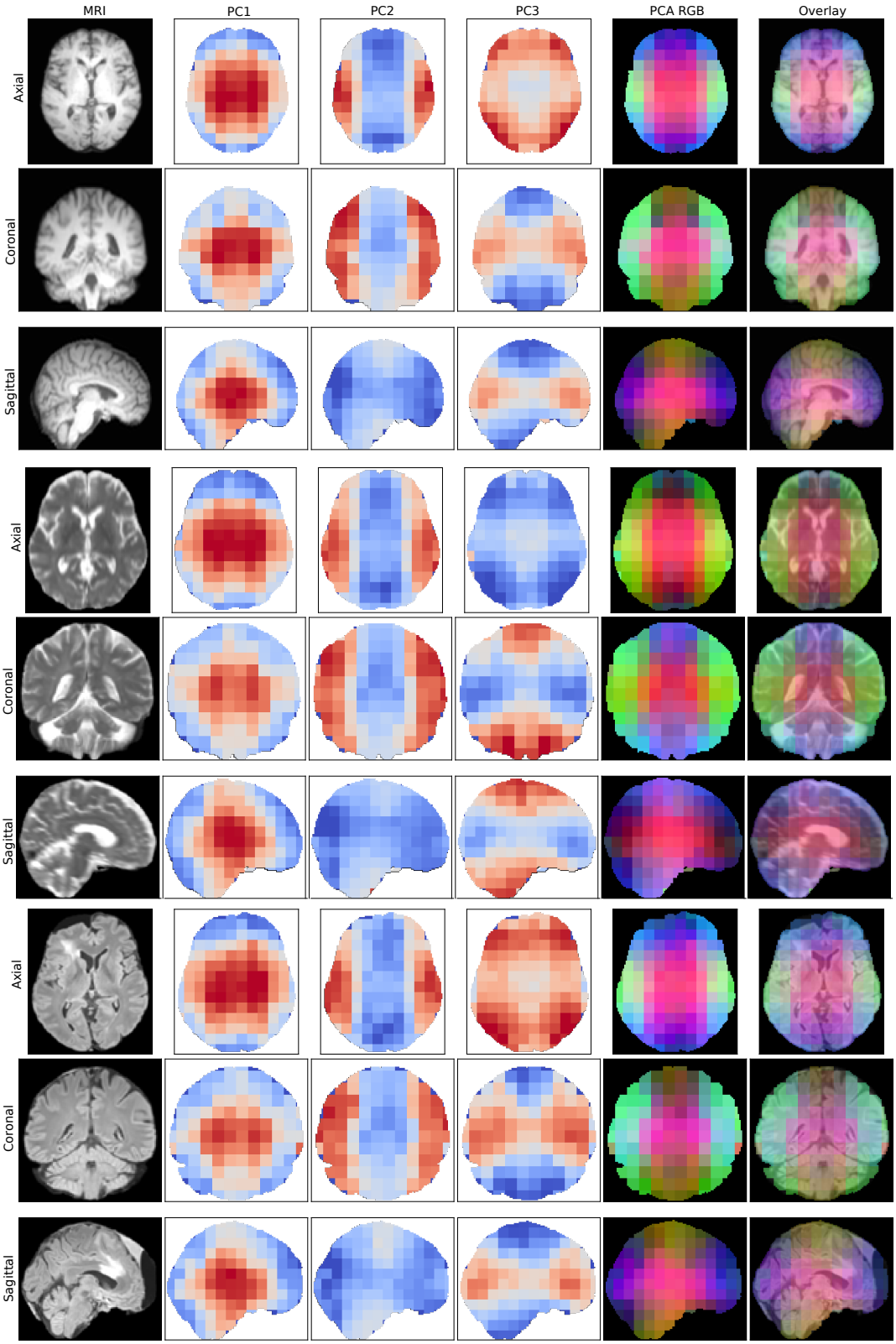
Supplementary Figure 5 displays feature map visualizations obtained via Principal Component Analysis (PCA) on output token features, following the methodology of DINO [1, 2]. For each scan (T1w, T2w, and FLAIR), features were extracted using a ViT-Large backbone with Mixture-of-Experts (MoE) and a patch size of 8 to capture fine-grained spatial details. The resulting visualizations demonstrate that the learned representations sharply delineate distinct anatomical structures within the brain.



Supplementary Figure 3. Multiscale Masking Sample. Foreground is denoted as FG, background is denoted as BG and the foreground ratio for encoder/predictor are separately shown.



Supplementary Figure 4. Multiscale Masking Sample. Foreground is denoted as FG, background is denoted as BG and the foreground ratio for encoder/predictor are separately shown.



Supplementary Figure 5. **Feature Map Visualization with PCA - ViT-L-MoE with Path Size 8**

B.4 Modified Auxiliary Loss Free Load Balancing

We observe original auxiliary loss free load balancing does not preserve ideal load balancing during pre-training. We therefore modified the algorithm with new bias update rule based on error correction, zero-mean projection and bias clipping to preserve more balanced bias update as demonstrated in Supplementary Algorithm 2.

Algorithm 2 Modified MoE Load Balancing Bias Update

Require: Tokens routed to expert i: C_i, Number of experts: N, Learning rate: η, Smoothing factor: δ, Clipping threshold: $b_{\max}$

$\bar{C} \leftarrow \frac{1}{N}\sum_{i=1}^{N} C_i$ ▷ Calculate average tokens per expert

for each expert $i \in \{1,\ldots,N\}$ **do**

  $\varepsilon_i \leftarrow \bar{C} - C_i$ ▷ Error Calculation

  Original Aux Loss Free Bias Update: ~~$\Delta b_i \leftarrow \eta \cdot \mathrm{sgn}(\varepsilon_i)$~~

  New Bias Update Step: $\Delta b_i \leftarrow \eta \cdot \frac{\varepsilon_i}{\bar{C}+\delta}$

  $b_i \leftarrow b_i + \Delta b_i$

end for

// Zero-Mean Projection to prevent infinite drift

for each expert $i \in \{1,\ldots,N\}$ **do**

  $b_i \leftarrow b_i - \frac{1}{N}\sum_{j=1}^{N} b_j$

end for

// Bias Clipping to avoid exploding bias terms

for each expert $i \in \{1,\ldots,N\}$ **do**

  $b_i \leftarrow \max(-b_{\max}, \min(b_i, b_{\max}))$

end for

B.5 Multi-Modal Learning Methods

There are five multimodal learning methods we attempted on investigating optimal achievable performance for each task. The formulation for each of them is detailed below:

Logits-Averaging Ensemble This is the simplest baseline where independent classification followed by logit-level averaging. Each modality m has its own linear head $f_m : \mathbb{R}^D \to R^C$ for hidden dimension D and number of classes C. For each modality m, given weights $W_m \in \mathbb{R}^{C\times D}$, latent representation $z^{(m)} \in \mathbb{R}^D$, the linear head map latent representation to logits by

$$\ell^{(m)} = W_m \mathbf{z}^{(m)}$$

The, the final prediction is the average of per-modality logits as

$$\hat{y} = \frac{1}{m}\sum_{m=1}^{M} \ell^{(m)}$$

This method serves as a strong baseline because it makes no assumptions about inter-modal interactions and ensemble averaging reduces variance in the logit space.

Gated Cross-Attention Late Fusion Given two modalities producing patch-level feature sequences $\mathbf{Z}^{(1)} \in \mathbb{R}^{B\times N_1\times D}$ and $\mathbf{Z}^{(2)} \in \mathbb{R}^{B\times N_2\times D}$ from a shared backbone, each is first mapped through a modality-specific projection head with residual connection:

$$\mathbf{H}^{(m)} = \mathrm{Proj}_m(\mathbf{Z}^{(m)}) = \sigma\big(W_1^{(m)}\mathbf{Z}^{(m)}\big)W_2^{(m)} + W_1^{(m)}\mathbf{Z}^{(m)}, \quad m \in \{1,2\}$$

where σ is GELU activation. Bidirectional cross-attention then enables each modality to attend to the other. For modality 1 attending to modality 2:

$$\mathrm{CrossAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}$$

where $\mathbf{Q} = \mathbf{H}^{(1)}W_Q$, $\mathbf{K} = \mathbf{H}^{(2)}W_K$, $\mathbf{V} = \mathbf{H}^{(2)}W_V$, and $d_h = D/H$ is the per-head dimension across H heads. A symmetric operation computes modality 2's attention over modality 1. Residual connections and layer normalization (LN) are applied:

$$\hat{\mathbf{H}}^{(1)} = \mathrm{LN}\Big(\mathbf{H}^{(1)} + \mathrm{LN}\Big(\mathrm{CrossAttn}_{1\to 2}(\mathbf{H}^{(1)},\mathbf{H}^{(2)})\Big)\Big)$$

$$\hat{\mathbf{H}}^{(2)} = \mathrm{LN}\Big(\mathbf{H}^{(2)} + \mathrm{LN}\Big(\mathrm{CrossAttn}_{2\to 1}(\mathbf{H}^{(2)}, \mathbf{H}^{(1)})\Big)\Big)$$

After mean-pooling over the token dimension to obtain $\bar{\mathbf{h}}^{(m)} = \frac{1}{N_m}\sum_{i=1}^{N_m} \hat{\mathbf{H}}_i^{(m)} \in \mathbb{R}^D$, a learned gating performs soft feature selection by

$$\mathbf{g} = \tanh\Big(W_g^{(2)}\,\mathrm{ReLU}\Big(W_g^{(1)}[\bar{\mathbf{h}}^{(1)}; \bar{\mathbf{h}}^{(2)}]\Big)\Big) \in \mathbb{R}^D$$
$$\mathbf{f} = \mathbf{g} \odot \bar{\mathbf{h}}^{(1)} + (1-\mathbf{g}) \odot \bar{\mathbf{h}}^{(2)}$$

where $[\cdot;\cdot]$ denotes concatenation, $\odot$ is the Hadamard product, and $\mathbf{g}$ acts as a dimension-wise interpolation coefficient between the two modality representations. The final prediction is $\hat{y} = W_c\mathbf{f} + b_c$.

This design allows fine-grained, element-wise arbitration of how much each modality contributes to each feature dimension.

Classify-Then-Aggregate Multiple Instance Learning [3] This method treats the multimodal problem as a Multiple Instance Learning (MIL) task. Patch tokens from all modalities are concatenated into a single bag by

$$\mathbf{X} = [\mathbf{Z}_1^{(1)}, \ldots, \mathbf{Z}_{N_1}^{(1)}, \mathbf{Z}_1^{(2)}, \ldots, \mathbf{Z}_{N_2}^{(2)}] \in \mathbb{R}^{N\times D}$$

where $N = N_1 + N_2$. Each patch (instance) $\mathbf{x}_i$ undergoes two parallel computations

1) *Gated attention scoring.* An attention score per class c is computed via a gated attention mechanism:

$$a_{i,c} = \mathbf{w}_c^\top\,[\tanh(V\mathbf{x}_i) \odot \sigma(U\mathbf{x}_i)]$$

where $V, U \in \mathbb{R}^{D'\times D}$ are shared attention and gating projections, σ is the sigmoid function providing a soft gate, and $\mathbf{w}_c \in \mathbb{R}^{D'}$ is a class-specific attention vector. Per-class attention weights are obtained via segment softmax over each bag $\mathscr{B}_b$:

$$\alpha_{i,c} = \frac{\exp(a_{i,c})}{\sum_{j\in\mathscr{B}_b}\exp(a_{j,c})}$$

2) *Instance-level classification.* A shared MLP produces patch-level logits:

$$\hat{y}_{i,c} = \mathrm{MLP}(\mathbf{x}_i)_c$$

Bag-level aggregation. The final bag prediction for class c is the attention-weighted sum of instance predictions.

$$\hat{Y}_{b,c} = \gamma_c\left(\sum_{i\in\mathscr{B}_b}\alpha_{i,c}\cdot\hat{y}_{i,c}\right) + \beta_c$$

where γ_c and β_c are learnable scale and bias parameters. This "classify-then-aggregate" formulation enables interpretable per-patch importance attribution across modalities.

Product-of-Experts Fusion This method frames multimodal fusion as a Product of Experts (PoE) [4], where each modality acts as an independent expert that contributes a probabilistic opinion over the label space. The joint posterior is obtained by multiplying per-modality softmax distributions and renormalizing — equivalent to summation in log-probability space.

Given M modalities, a shared backbone f_θ extracts features $\mathbf{z}^{(m)} = f_\theta(\mathbf{x}^{(m)})$, and each modality m has a dedicated classification head g_m that produces logits $\ell^{(m)} = g_m(\mathbf{z}^{(m)}) \in \mathbb{R}^C$. Each expert's belief is a categorical distribution:

$$p_m(y \mid \mathbf{x}^{(m)}) = \mathrm{softmax}(\ell^{(m)})$$

The PoE joint distribution is defined as the normalized product of all expert distributions:

$$p_{\mathrm{PoE}}(y \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) = \frac{\prod_{m=1}^M p_m(y \mid \mathbf{x}^{(m)})}{\sum_{c=1}^C \prod_{m=1}^M p_m(y=c \mid \mathbf{x}^{(m)})}$$

For numerical stability, all computations are carried out in log-space. Defining $\lambda^{(m)} = \log\mathrm{softmax}(\ell^{(m)})$:

$$\log p_{\mathrm{PoE}}(y=c) = \sum_{m=1}^M \lambda_c^{(m)} - \log\sum_{c'=1}^C \exp\left(\sum_{m=1}^M \lambda_{c'}^{(m)}\right)$$

The model is trained with negative log-likelihood loss (NLLLoss) on the fused log-probabilities:

$$\mathscr{L} = -\log p_{\mathrm{PoE}}(y = y^* \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$$

For ViT backbones, each per-modality head is Attentive Classifier (cross-attention pooler + linear layer); for CNN/Swin-UNETR/BrainIAC backbones, each head is a simple linear layer applied to mean-pooled features.

Product-of-Experts with Joint Head [5] This variant extends the standard PoE by introducing an additional joint expert g_{joint} that operates on the concatenated features of all modalities, capturing cross-modal interactions that unimodal experts cannot model.

For ViT backbones, the joint head receives the token-level concatenation along the sequence dimension $\mathbf{Z}_{\text{cat}} = [\mathbf{Z}^{(1)}; \mathbf{Z}^{(2)}; \ldots; \mathbf{Z}^{(M)}] \in \mathbb{R}^{B\times(N\cdot M)\times D}$, processed by an Attentive Classifier. For non-ViT backbones, pool-level features are concatenated along the feature dimension $\mathbf{z}_{\text{cat}} = [\mathbf{z}^{(1)}; \ldots; \mathbf{z}^{(M)}] \in \mathbb{R}^{D\cdot M}$, processed by a linear head.

The joint expert produces logits $\ell^{(\text{joint})} = g_{\text{joint}}(\mathbf{Z}_{\text{cat}})$ and participates in the PoE as an $(M{+}1)$-th expert that is always considered valid:

$$p_{\text{PoE+Joint}}(y=c) \propto \left(\prod_{m=1}^{M} p_m(y=c \mid \mathbf{x}^{(m)})\right) \cdot p_{\text{joint}}(y=c \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$$

In log-space, with same definition on λ as Product-of-Experts,

$$\log p_{\text{PoE+Joint}}(y=c) = \left(\sum_{m=1}^{M} \tilde{\lambda}_c^{(m)} + \lambda_c^{(\text{joint})}\right) - \log \sum_{c'=1}^{C} \exp\left(\sum_{m=1}^{M} \tilde{\lambda}_{c'}^{(m)} + \lambda_{c'}^{(\text{joint})}\right)$$

The joint expert captures multimodal synergies that arise only when multiple modalities are observed together, while the unimodal experts ensure that informative predictions can still be made when synergies are missing.

B.6 Model Pretraining Details

We present our pretraining hyperparameter configuration in Supplementary Tables 7 and 8 for our base model. The learning rate linearly increase from 1.0×10^{-4} to 5.25×10^{-4} in first 40 epochs, stay on 5.25×10^{-4} for 160 (ViT-Base) or 200 (ViT-Large) epochs and then cooldown with cosine decay to 1.0×10^{-6} with 40 (ViT-Base) or 80 (ViT-Large) epochs. Three different block size sampling is performed on multiscale masking with masking size demonstrated in Masking Spatial Scale and Masking Depth Scale.

B.7 Model Evaluation Details

To ensure a fair comparison across models and downstream tasks, we performed model-specific hyperparameter sweeps and reported the test-set performance obtained from the configuration selected by validation-set performance. This procedure accounts for the fact that different foundation models may require different optimization settings to achieve their best downstream transfer performance. For both unimodal and multimodal evaluations, we swept learning rate, weight decay and batch size for each model. For BrainIAC, we evaluated learning rates of $\{3\times10^{-4}, 1\times10^{-4}, 1\times10^{-3}\}$, weight decay values of $\{1\times10^{-5}, 1\times10^{-2}, 1\times10^{-1}\}$, epochs number of $\{15, 30\}$ and batch sizes of $\{16, 32, 64\}$. For VoCo, we evaluated learning rates of $\{1.5\times10^{-4}, 1.5\times10^{-3}\}$, weight decay values of $\{1\times10^{-2}, 1\times10^{-1}\}$ and a batch size of 4. For NeuroVFM, we evaluated learning rates of $\{5\times10^{-4}, 1\times10^{-4}, 1\times10^{-3}\}$, weight decay values of $\{5\times10^{-2}, 1\times10^{-2}, 1\times10^{-1}\}$, epochs number of $\{30, 50\}$ and a batch size of 8. For Neuro-JEPA, we evaluated learning rates of $\{1.5\times10^{-5}, 1.5\times10^{-4}, 3\times10^{-5}\}$, weight decay values of $\{1\times10^{-5}, 1\times10^{-2}, 1\times10^{-1}\}$, epochs number of $\{15, 30\}$ and batch sizes of $\{16, 32, 64\}$. For CNN, we we evaluated learning rates of $\{3\times10^{-4}, 1\times10^{-4}, 1\times10^{-3}\}$, weight decay values of $\{1\times10^{-5}, 1\times10^{-2}, 1\times10^{-1}\}$, epochs number of $\{30, 50\}$ and batch sizes of $\{16, 32, 64\}$. The search space for VoCo was necessarily smaller because of its substantially higher computational cost. For VoCo and NeuroVFM, the batch sizes were constrained by the maximum feasible per-GPU memory capacity. For NeuroVFM, we consistently used multiple instance learning (MIL) pooling across all evaluations with frozen backbone following the optimal setup from original manuscript and the suggestions by consulting the authors. While full fine-tuning is also experimented for NeuroVFM, we observe diminishing or not improved performance across the tasks.

B.8 Evaluation Scope

The objective of Neuro-JEPA is to learn image-level representations for clinically relevant tasks. We therefore benchmarked the model on diagnosis, prognosis, time-to-event and age prediction, and deliberately restricted claims to this intended use. We did not include segmentation or open-ended vision–language evaluation because these task families require distinct supervision, inference interfaces, baselines and validity criteria. Medical image segmentation is a dense-prediction problem with strong task-specific standards; recent benchmarking [6, 7] shows that fair segmentation claims require carefully configured baselines, adequate dataset diversity and resource-matched comparisons, since many apparent architectural gains disappear under rigorous validation. Moreover, dedicated segmentation foundation models such as MedSAM [8], BiomedParse [9], and VISTA3D [10] are trained and evaluated with mask-level objectives and segmentation-specific metrics, making segmentation a separate methodological question rather than an auxiliary image-level diagnostic task. We also did not report generic medical VQA metrics because recent studies [11, 12, 13] show that high Vision-Language Model (VLM) accuracy can persist when images are absent, blank or mismatched, and therefore may not establish causal use of visual information without explicit grounding controls. Thus, adding superficial segmentation or

VQA experiments would expand the apparent scope of the study without providing a valid test of the proposed methods. We therefore view segmentation transfer and grounded vision–language reasoning as important future directions requiring dedicated protocols and claims.

Supplementary Table 7. Neuro-JEPA-Base pretraining configuration.

Parameter	**Value**
Architecture	ViT-Base-MoE (128M total params w. 86M activated params)
Number of total experts	16
Number of shared experts	2
Number of activated experts	6
Auxiliary loss free bias update	$1e$-4
MoE score function	softmax
Decoder depth	4
Exponential Moving Average Ratio	$(0.99925, 0.99925)$
Per-GPU batch	48
Number of GPUs	NVIDIA L40s $\times$ 12
Gradient accumulation	$1 \rightarrow$ global batch $= 576$
Iteration per epoch	$\lceil 921,600/576 \rceil = 1600$
Warmup epochs	$\lceil 64,000/1600 \rceil = 40$
Scheduling epochs	$\lceil 256,000/1600 \rceil = 160$
Cooldown epochs	$\lceil 64,000/1600 \rceil = 40$
Start learning rate	1.0×10^{-4}
Max learning rate	5.25×10^{-4}
Min learning rate	1.0×10^{-6}
Weight decay	0.04
Optimizer	AdamW, $\beta = (0.9, 0.999)$
Gradient Clip	3.0
Background Weight	0.1
Precision	bfloat16
Masking Aspect Ratio	$(0.75, 1.5)$
Masking Spatial Scale	$(0.0, 0.2), (0.2, 0.5), (0.5, 0.7)$
Masking Depth Scale	$(0.0, 1.0), (0.0, 1.0), (0.0, 1.0)$

Supplementary Table 8. Neuro-JEPA-Large pretraining configuration.

Parameter	Value
Architecture	ViT-Large-MoE (429M total params w. 303M activated params)
Number of total experts	16
Number of shared experts	2
Number of activated experts	6
Auxiliary loss free bias update	$1e$-4
MoE score function	softmax
Decoder depth	4
Exponential Moving Average Ratio	$(0.99925, 0.99925)$
Per-GPU batch	30
Number of GPUs	NVIDIA L40s $\times$ 12
Gradient accumulation	$1 \rightarrow$ global batch $= 360$
Iteration per epoch	$\lceil 288,000/360 \rceil = 800$
Warmup epochs	$\lceil 32,000/800 \rceil = 40$
Scheduling epochs	$\lceil 160,000/800 \rceil = 200$
Cooldown epochs	$\lceil 64,000/800 \rceil = 80$
Start learning rate	1.0×10^{-4}
Max learning rate	5.25×10^{-4}
Min learning rate	1.0×10^{-6}
Weight decay	0.04
Optimizer	Muon, $\beta_{muon} = 0.95$, $\beta_{adamw} = (0.9, 0.999)$
Gradient Clip	0.0
Background Weight	0.1
Precision	bfloat16
Masking Aspect Ratio	$(0.75, 1.5)$
Masking Spatial Scale	$(0.0, 0.2), (0.2, 0.5), (0.5, 0.7)$
Masking Depth Scale	$(0.0, 1.0), (0.0, 1.0), (0.0, 1.0)$

Supplementary Table 9. Average unimodal performance across cohorts and task types. Values are mean [95% CI]. **Bold**: best per metric. ↑ higher is better; ↓ lower is better. "combs" here refer to the number of combinations on dataset-task-modality. "n" for brain age represents number of samples in the test set. underlining indicates the second-best model.

	Public (41 combs)		**NYU** (30 combs)		**LI** (30 combs)		**MGH** (45 combs)	
Model	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑
VoCo	0.721 [0.689, 0.754]	0.573 [0.513, 0.635]	0.762 [0.726, 0.793]	0.338 [0.265, 0.412]	0.735 [0.702, 0.764]	0.295 [0.240, 0.354]	0.717 [0.693, 0.741]	0.233 [0.196, 0.269]
BrainIAC	0.728 [0.705, 0.751]	0.555 [0.494, 0.618]	0.730 [0.697, 0.761]	0.296 [0.226, 0.370]	0.674 [0.643, 0.704]	0.212 [0.162, 0.266]	0.655 [0.636, 0.673]	0.179 [0.148, 0.211]
NeuroVFM	0.741 [0.706, 0.774]	0.585 [0.526, 0.648]	0.772 [0.743, 0.801]	0.343 [0.271, 0.422]	0.739 [0.713, 0.762]	0.217 [0.175, 0.260]	0.725 [0.705, 0.745]	0.237 [0.199, 0.277]
Neuro-JEPA-B	0.785 [0.760, 0.811]	0.649 [0.588, 0.705]	0.825 [0.796, 0.850]	0.457 [0.382, 0.535]	0.806 [0.780, 0.828]	0.383 [0.322, 0.442]	0.741 [0.720, 0.761]	0.253 [0.216, 0.292]
Neuro-JEPA-L	**0.801** [0.775, 0.828]	**0.687** [0.631, 0.740]	**0.838** [0.812, 0.862]	**0.498** [0.430, 0.567]	**0.810** [0.779, 0.838]	**0.407** [0.344, 0.465]	**0.748** [0.726, 0.770]	**0.277** [0.239, 0.317]

	Time-to-Event (6 combs)	**Brain Age** ($n = 757$)		
Model	C-index↑	R^2 ↑	MAE↓ (yr)	RMSE↓ (yr)
VoCo	0.663 [0.616, 0.699]	0.111 [0.065, 0.165]	6.22 [5.46, 6.94]	12.03 [10.54, 13.36]
BrainIAC	0.650 [0.615, 0.686]	0.522 [0.447, 0.591]	5.42 [4.93, 5.95]	8.82 [7.77, 9.78]
NeuroVFM	0.629 [0.584, 0.673]	0.673 [0.611, 0.735]	4.36 [3.91, 4.77]	7.29 [6.35, 8.12]
Neuro-JEPA-B	0.695 [0.664, 0.718]	**0.894** [0.860, 0.917]	2.78 [2.56, 3.02]	**4.15** [3.69, 4.70]
Neuro-JEPA-L	**0.705** [0.6501, 0.754]	0.893 [0.862, 0.915]	**2.74** [2.50, 2.95]	4.18 [3.73, 4.61]

C Unimodal Learning Experiments

C.1 Average Performance Across Tasks

Supplementary Table 9 reports mean unimodal performance averaged across all dataset–task–modality combinations within each cohort group: Public datasets (41 tasks), NYU Langone (30 tasks), NYU Long Island (30 tasks), BIND-MGH (45 tasks), time-to-event tasks (6 tasks), and brain-age prediction (OpenBHB; $n = 757$). Diagnosis and prognosis tasks are evaluated by AUROC and AUPRC; time-to-event tasks by C-index; and brain-age prediction by R^2, MAE, and RMSE. Across all cohorts and metrics, Neuro-JEPA consistently achieves the highest average performance relative to all baselines.

C.2 AUPRC for Public Datasets and Best Achievable Unimodal Performance Across Datasets

Extended Data Fig.1 presents AUPRC for (a) model performance across different modalities on public datasets for different evaluated models (b-c) all evaluated models on best achievable unimodal performance across all evaluated tasks and datasets (Public datasets, NYU-Langone, NYU-Longisland and BIND-MGH).

C.3 AUROC and AUPRC for Health System Datasets with Per Modality Performance

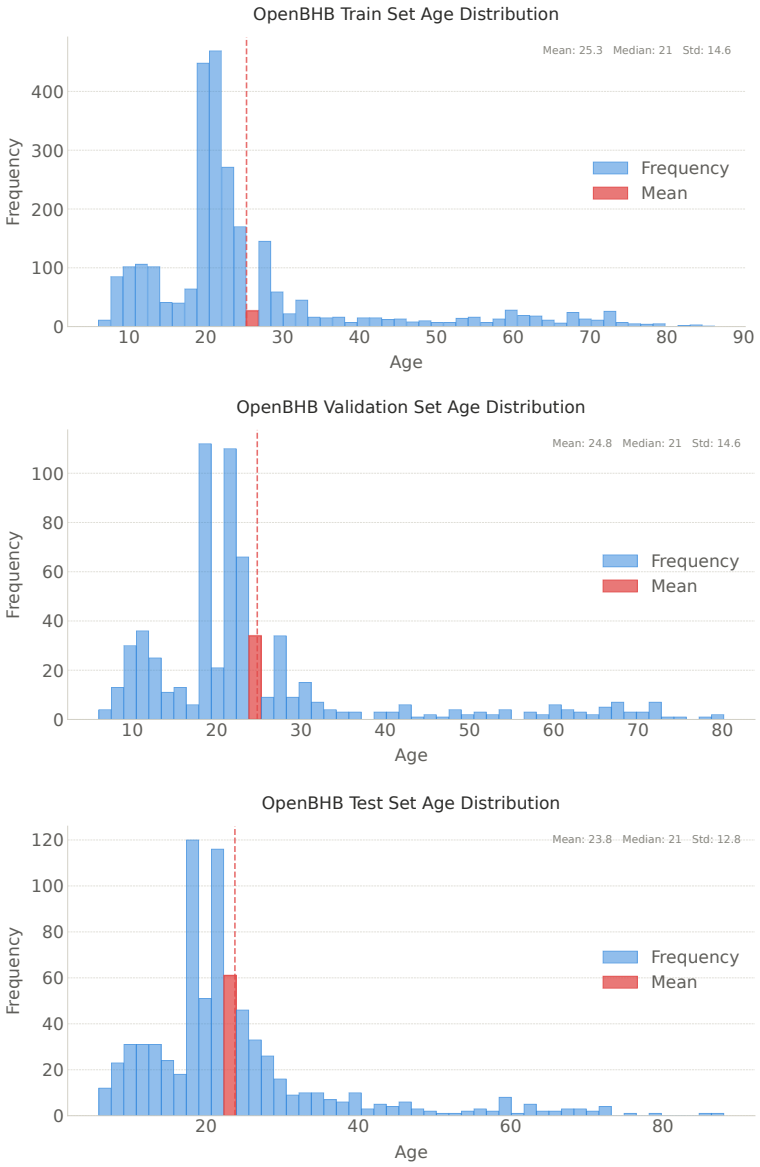
Extended Data Fig.2 presents detailed unimodal performance on all modalities (T1w, T2w, FLAIR) across clinical cohorts datasets (NYU-Langone, NYU-Longisland, BIND-MGH). The result is present with both AUROC and AUPRC. The result shows that Neuro-JEPA consistently perform among the best across tasks and modalities.

C.4 Kaplan Meier Curve and C-index for Time-to-Event Tasks

Extended Data Fig.3 presents Kaplan Meier Curve for evaluated time-to-event tasks on all modalities.

C.5 Age Prediction Performance as Regression

Extended Data Fig.4 reports age regression on the healthy OpenBHB cohort for all evaluated models, using quasi-raw scans. Neuro-JEPA generalizes best across R^2, MAE, and RMSE. Notably, models pre-trained without clinical neuroimaging data (BrainIAC, VoCo) fail to generalize beyond a certain age, corresponding to the long tail of the OpenBHB age distribution (Supplementary Figure 6), whereas Neuro-JEPA and NeuroVFM do not—underscoring the value of large-scale clinical pre-training. (NeuroVFM originally reports 2.8 years MAE, which we could not reproduce under our protocol of quasi-raw scans and predefined splits; since its preprocessing and splits are unspecified, we report our reproduced numbers to keep the evaluation setting consistent across all models.)



Supplementary Figure 6. Age Distribution on OpenBHB Dataset. We show age distribution on train, validation and test set on OpenBHB dataset. As it demonstrates, the dataset presents heavy long-tailed distribution on elder age, where the evaluated models not training on large scale clinical dataset fail to generalize.

Supplementary Table 10. Average multimodal performance across cohorts. Values are mean [95% CI]. **Bold**: best per metric. ↑ higher is better. The number of tasks here refer to the number of combination on dataset-task-modality. underlining indicates the second-best model.

	Public (12 combs)		**MGH** (30 combs)	
Model	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑
VoCo	0.743 [0.698, 0.789]	0.562 [0.443, 0.673]	0.729 [0.701, 0.757]	0.241 [0.194, 0.288]
BrainIAC	0.730 [0.693, 0.766]	0.552 [0.428, 0.673]	0.684 [0.662, 0.707]	0.203 [0.160, 0.245]
NeuroVFM	0.748 [0.684, 0.804]	0.574 [0.449, 0.685]	0.742 [0.721, 0.765]	0.255 [0.205, 0.305]
Neuro-JEPA-B	**0.805** [0.759, 0.849]	**0.637** [0.505, 0.749]	**0.763** [0.739, 0.789]	**0.295** [0.248, 0.343]

D Multi-Modal Learning Experiments

D.1 Average Performance Across Tasks

Supplementary Table 10 reports mean performance for dataset-task combinations (combs) in which two modalities were jointly available, specifically T1w+T2w and T1w+FLAIR. The result is averaged across all dataset–task combinations within each cohort (Public datasets, 12 tasks; BIND-MGH, 30 tasks). Performance is evaluated by AUROC and AUPRC. Neuro-JEPA achieves the highest average AUROC and AUPRC in both cohorts, demonstrating that its representational advantage extends to multimodal settings.

D.2 Multi-Modal Learning Result Across Fusion Methods - Public Datasets

Supplementary Figures 7 to 9 present multi-modal learning performance on different models across different fusion methods on different public datasets tasks and multi-modal combinations (T1w+T2w or T1w+FLAIR) reported in AUROC and AUPRC. Each row in the plot shows present modalities for multimodal fusion and the corresponding performance on different fusion methods. The method with best performance is selected for the main evaluation. The result shows that Neuro-JEPA present higher multi-modal gain over other models in majority of the tasks.

D.3 Multi-Modal Gain Result - Public Datasets

Extended Data Fig.5 present (a) performance on best multimodal method across tasks for each model reported in AUPRC. (b,c) multi-modal learning gain over uni-modal on public datasets for NeuroVFM and Neuro-JEPA reported reported in AUPRC.

Extended Data Fig.6 represent multi-modal learning gain over uni-modal on public datasets for BrainIAC and VoCo reported reported in both AUROC and AUPRC.

D.4 Multi-Modal Learning Result Across Fusion Methods - BIND-MGH

Supplementary Figures 10 to 17 show multimodal learning performance (T1w+T2w or T1w+FLAIR) on BIND-MGH dataset across different fusion methods for different models reported in both AUROC and AUPRC. Each row in the plot shows present modalities for multimodal fusion and the corresponding performance on different fusion methods.

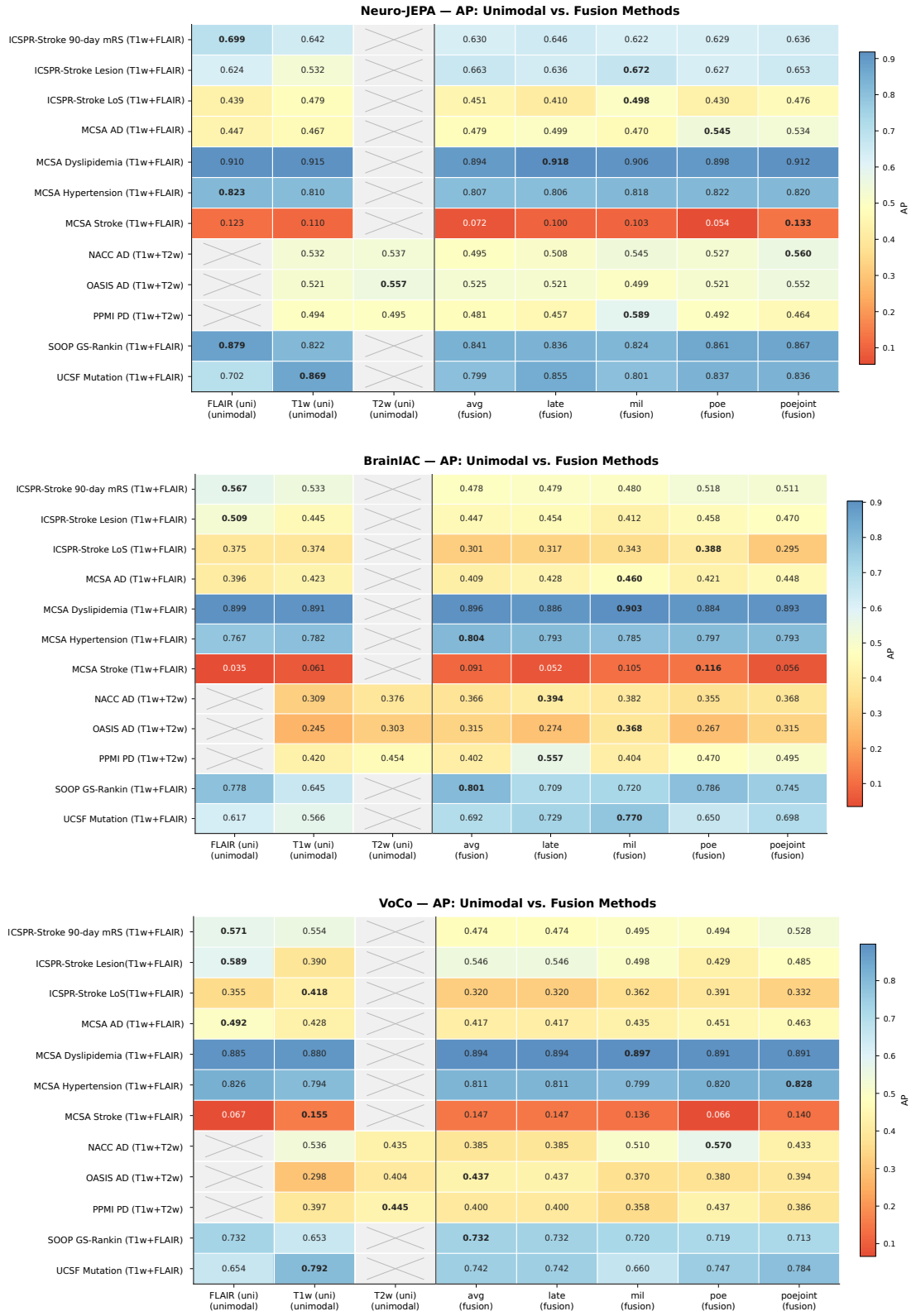
D.5 Multi-Modal Gain Result - BIND-MGH

Extended Data Fig.7 shows best multimodal learning performance across all 15 tasks on BIND-MGH datasets for all evaluated models. The result is reported in both AUROC and AUPRC. The result shows that Neuro-JEPA present larger multi-modal gain over other models in majority of the tasks.

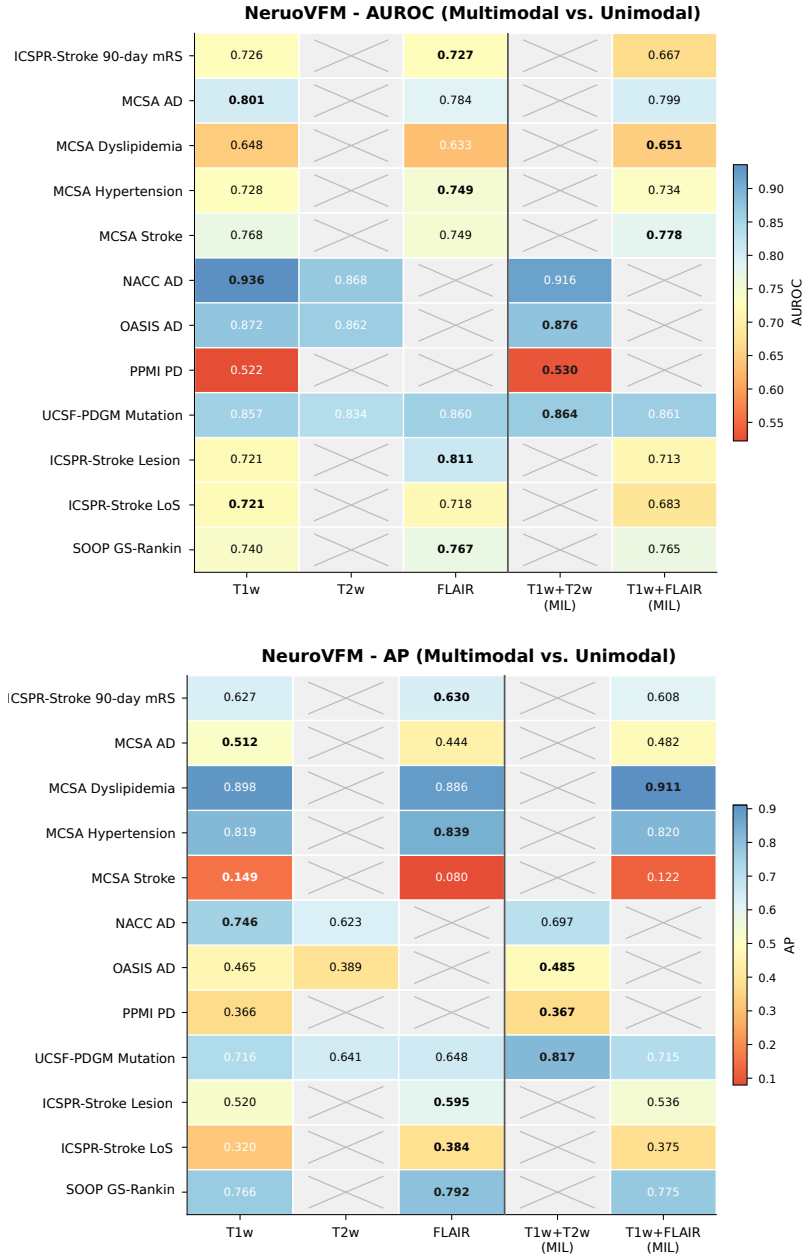
Supplementary Figures 18 to 21 present multimodal gain over unimodal for all fifteen tasks on BIND-MGH dataset for all evaluated models. The result is reported in both AUROC and AUPRC. The result shows that Neuro-JEPA present most positive transfer while maintaining highest overall performance across the tasks.



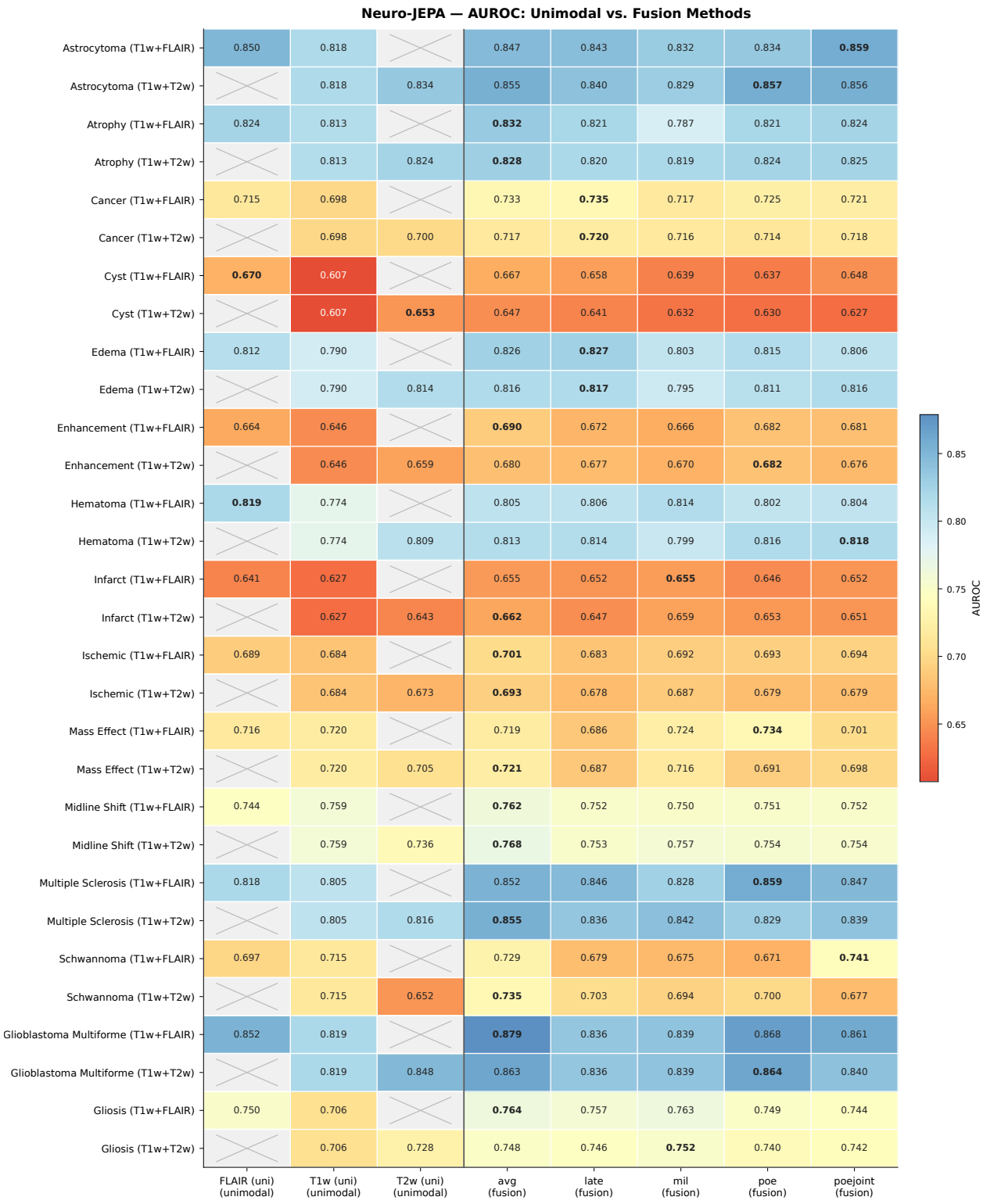
Supplementary Figure 7. **Multimodal Learning Performance Across Fusion Methods for BrainIAC, VoCo and Neuro-JEPA on Public Datasets - AUROC.**



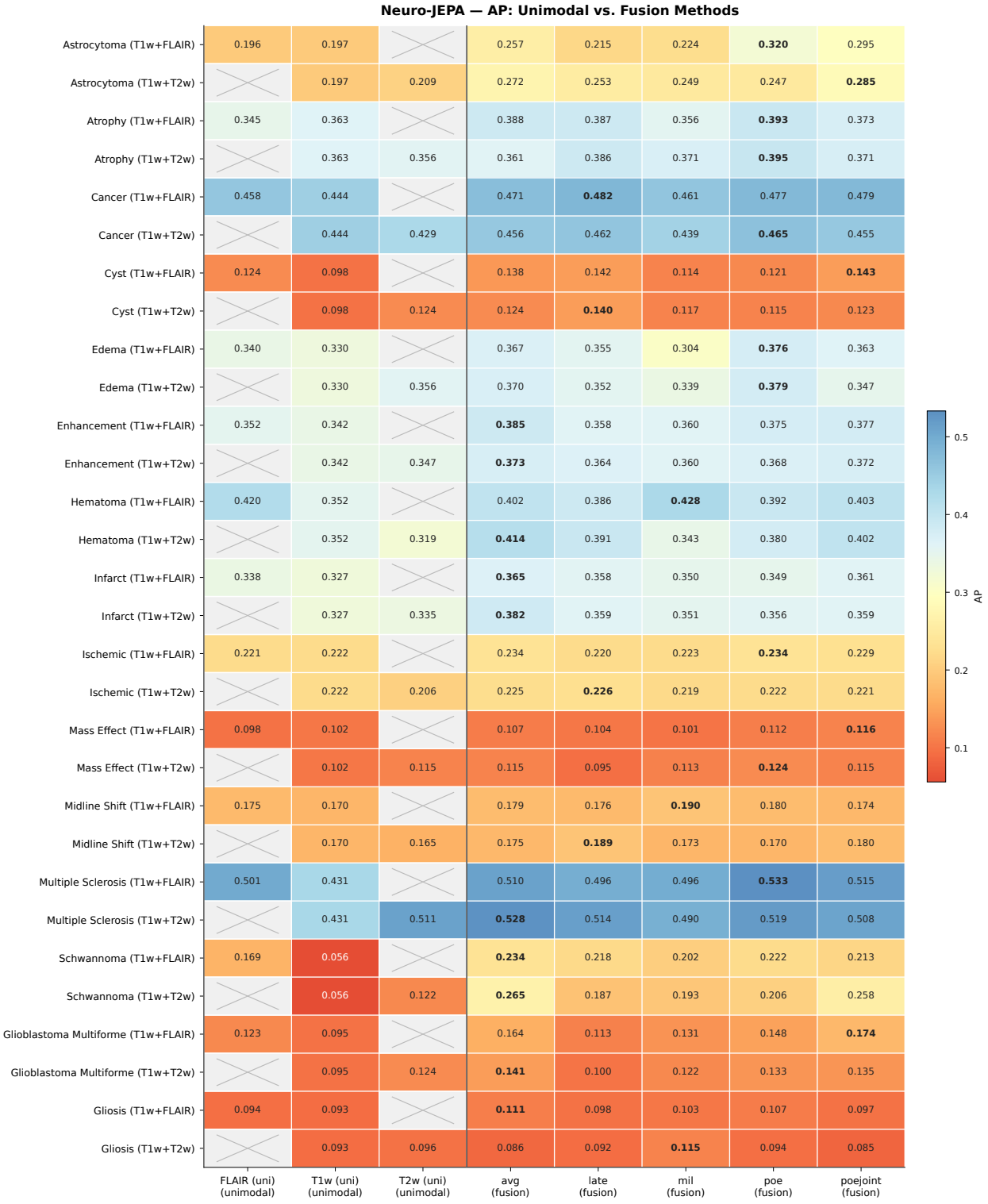
Supplementary Figure 8. Multimodal Learning Performance Across Fusion Methods for BrainIAC, VoCo and Neuro-JEPA on Public Datasets - AP.



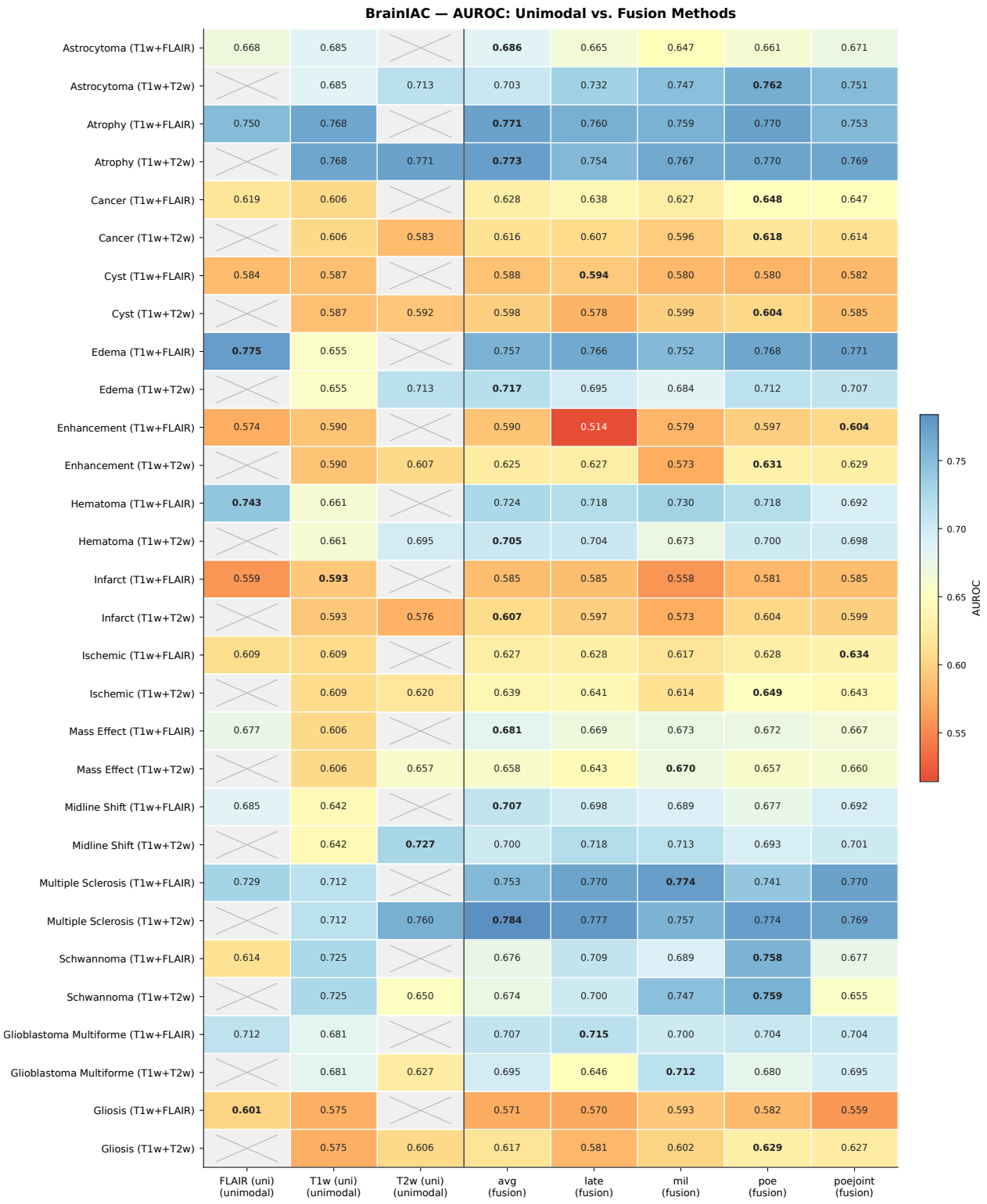
Supplementary Figure 9. Multimodal Learning Performance Across Fusion Methods for NeuroVFM on Public Datasets - AUROC and AP. AUROC and AUPRC for unimodal and multimodal performance for NeuroVFM. The best suggested method (MIL) from original paper for multimodal fusion is applied in this evaluation.



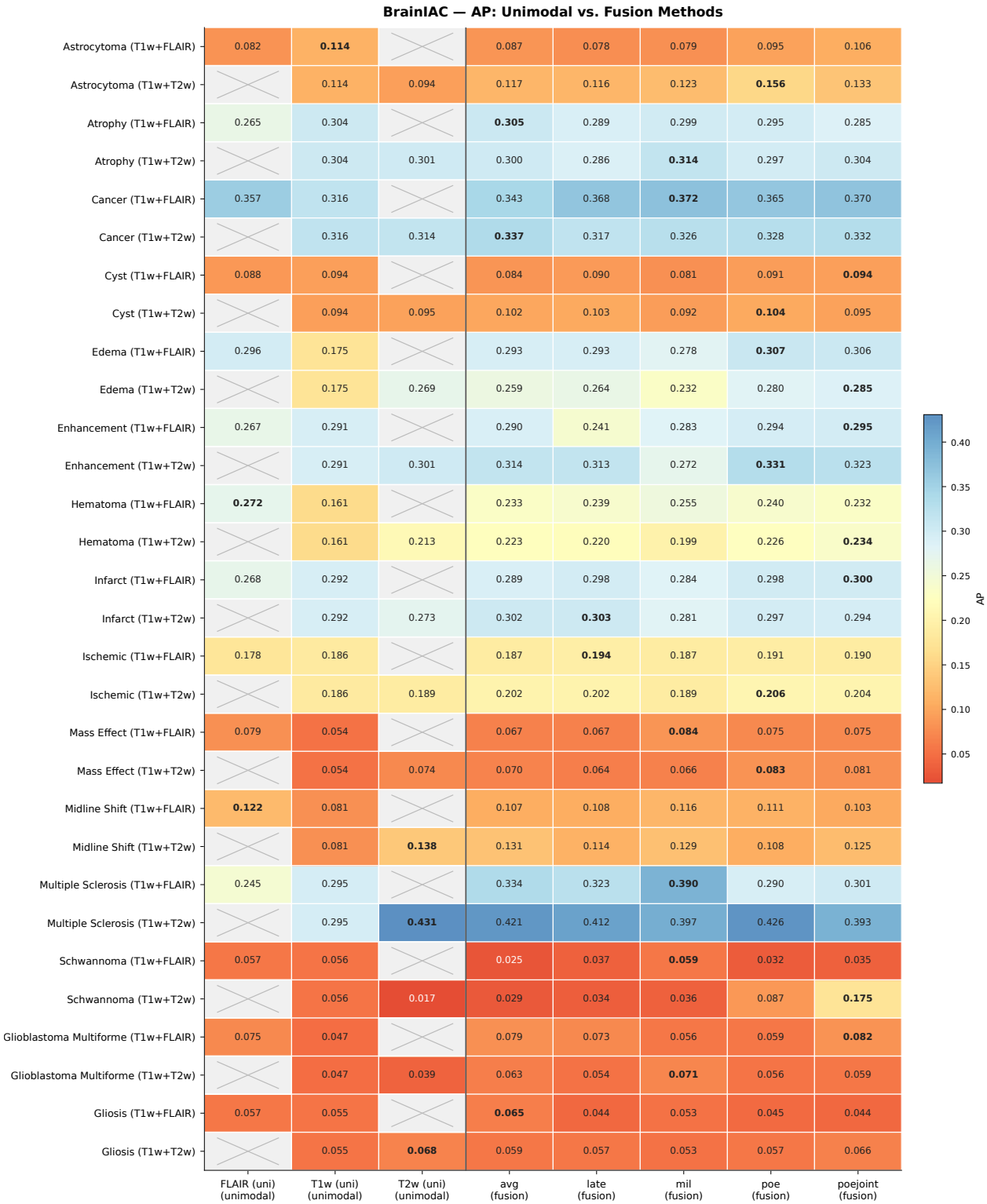
Supplementary Figure 10. Multimodal Learning Performance Across Fusion Methods for Neuro-JEPA on BIND-MGH - AUROC.



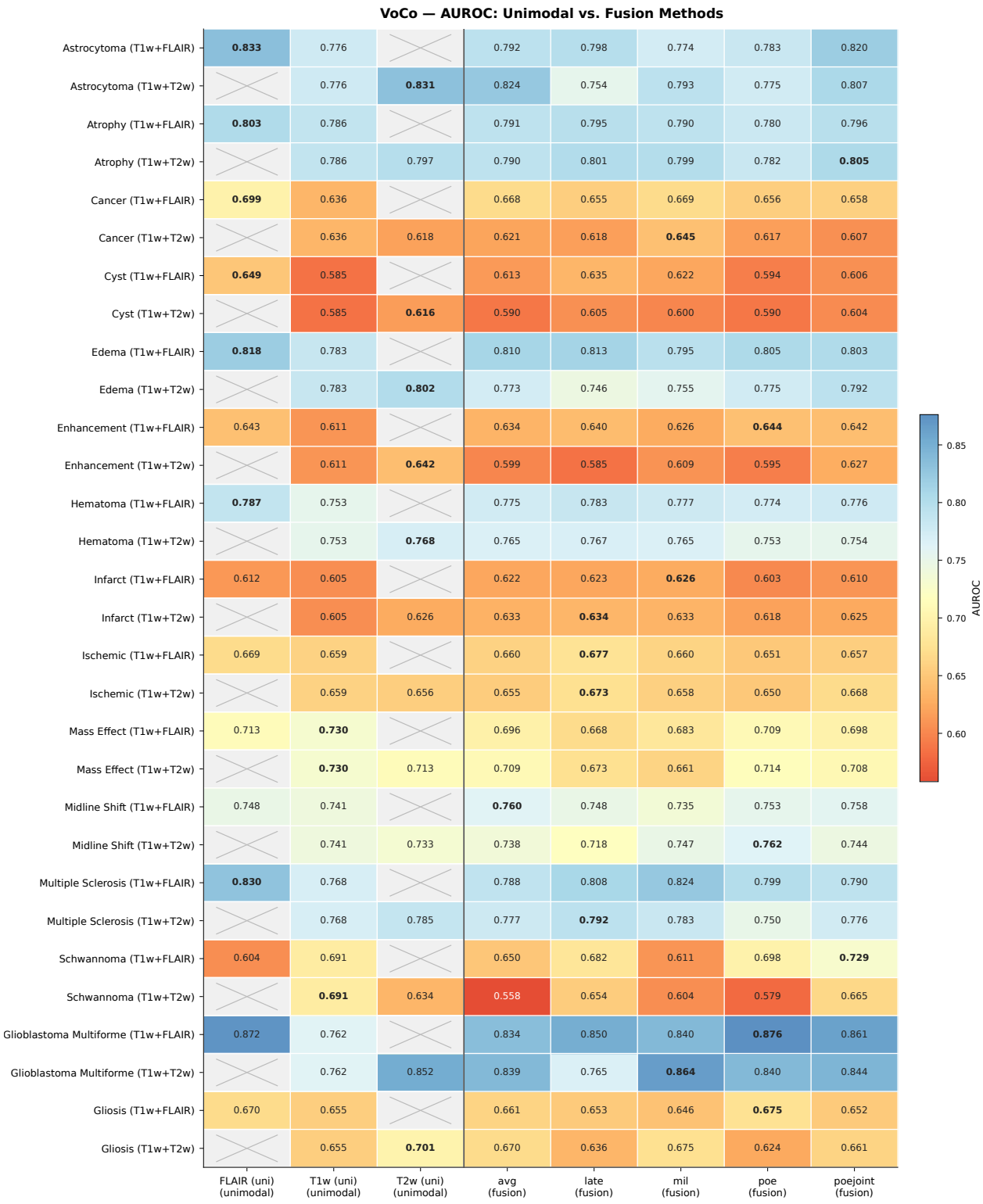
Supplementary Figure 11. Multimodal Learning Performance Across Fusion Methods for Neuro-JEPA on BIND-MGH - AP.



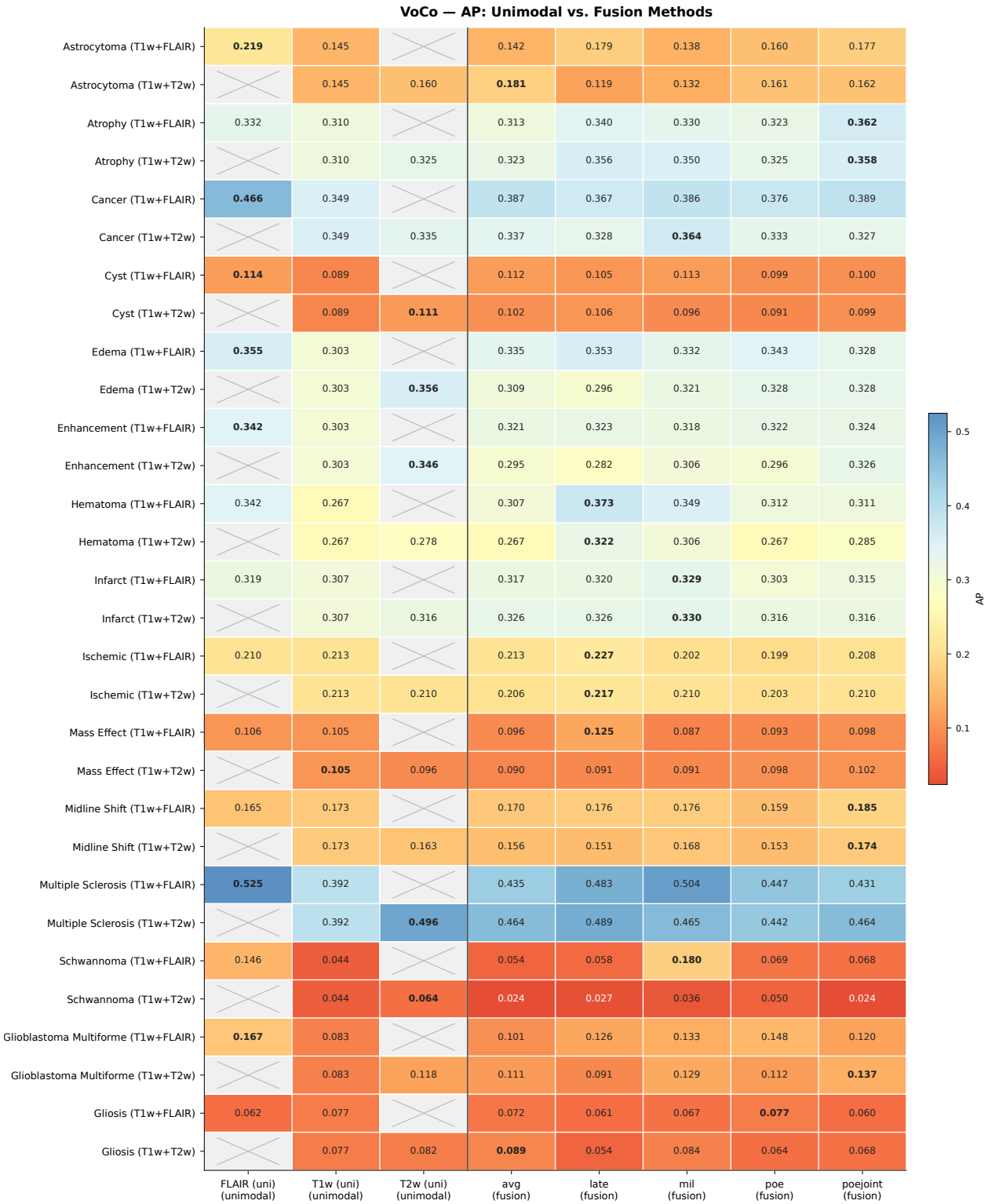
Supplementary Figure 12. Multimodal Learning Performance Across Fusion Methods for BrainIAC on BIND-MGH - AUROC.



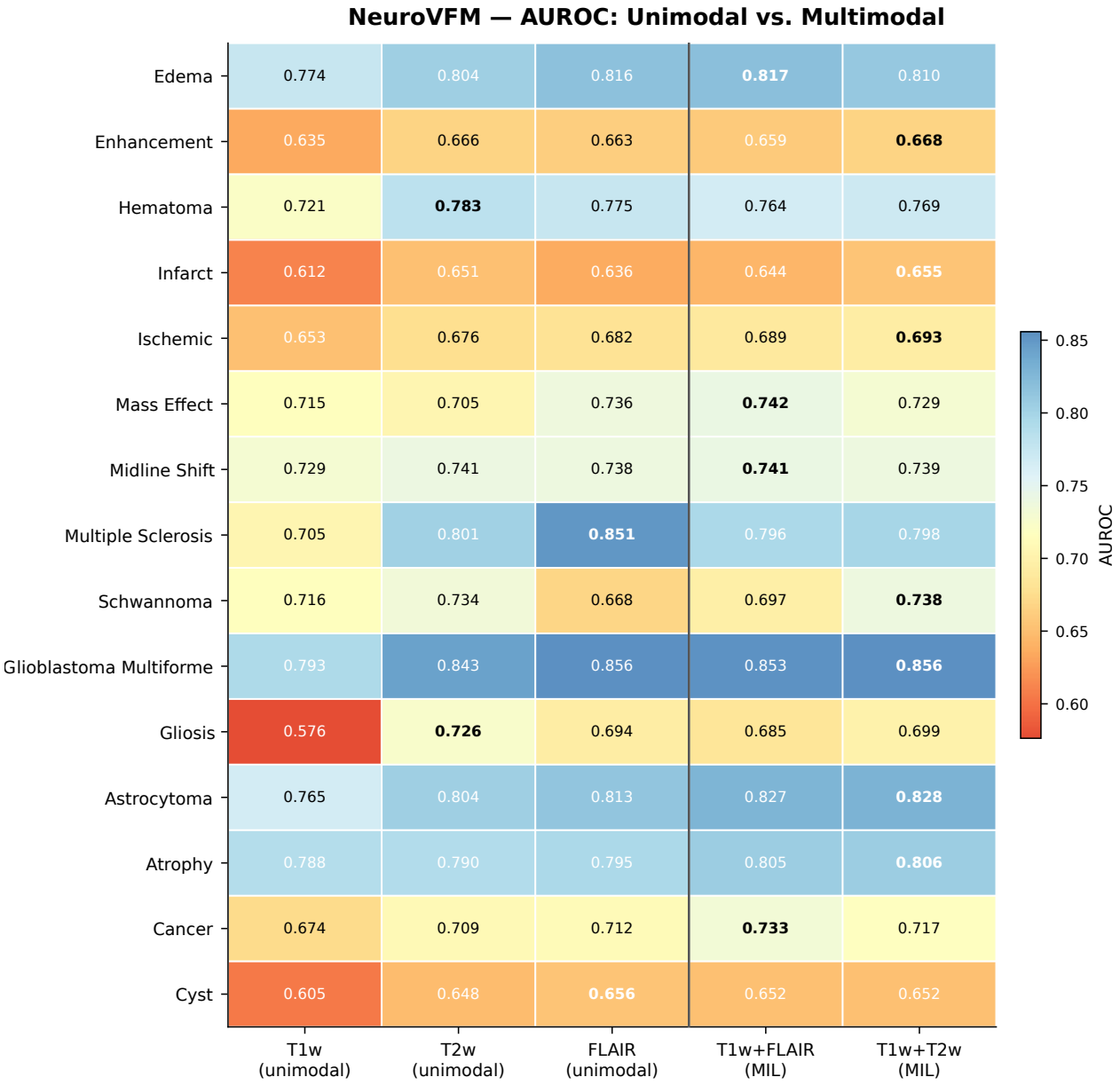
Supplementary Figure 13. **Multimodal Learning Performance Across Fusion Methods for BrainIAC on BIND-MGH - AP.**



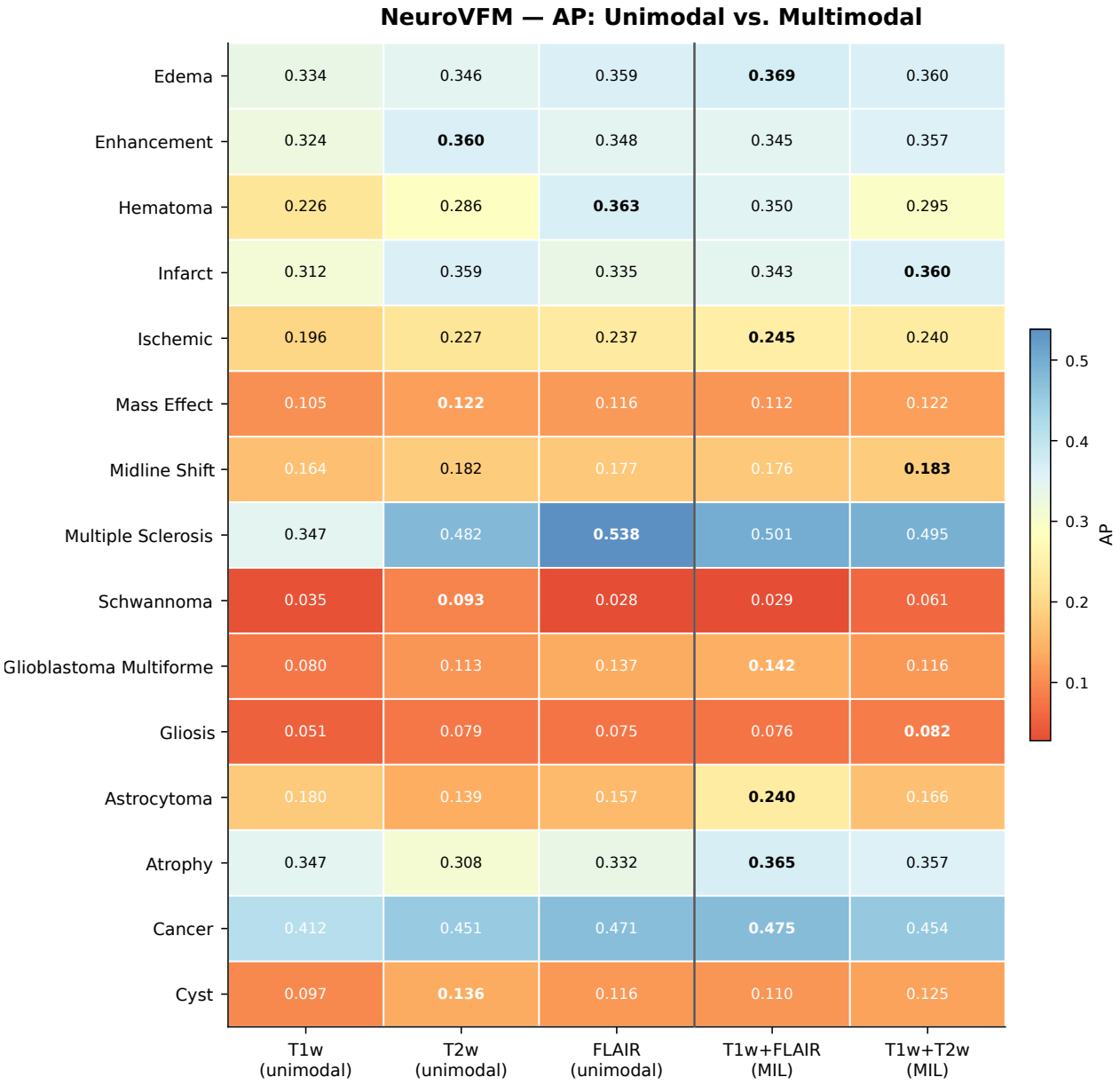
Supplementary Figure 14. Multimodal Learning Performance Across Fusion Methods for VoCo on BIND-MGH - AUROC.



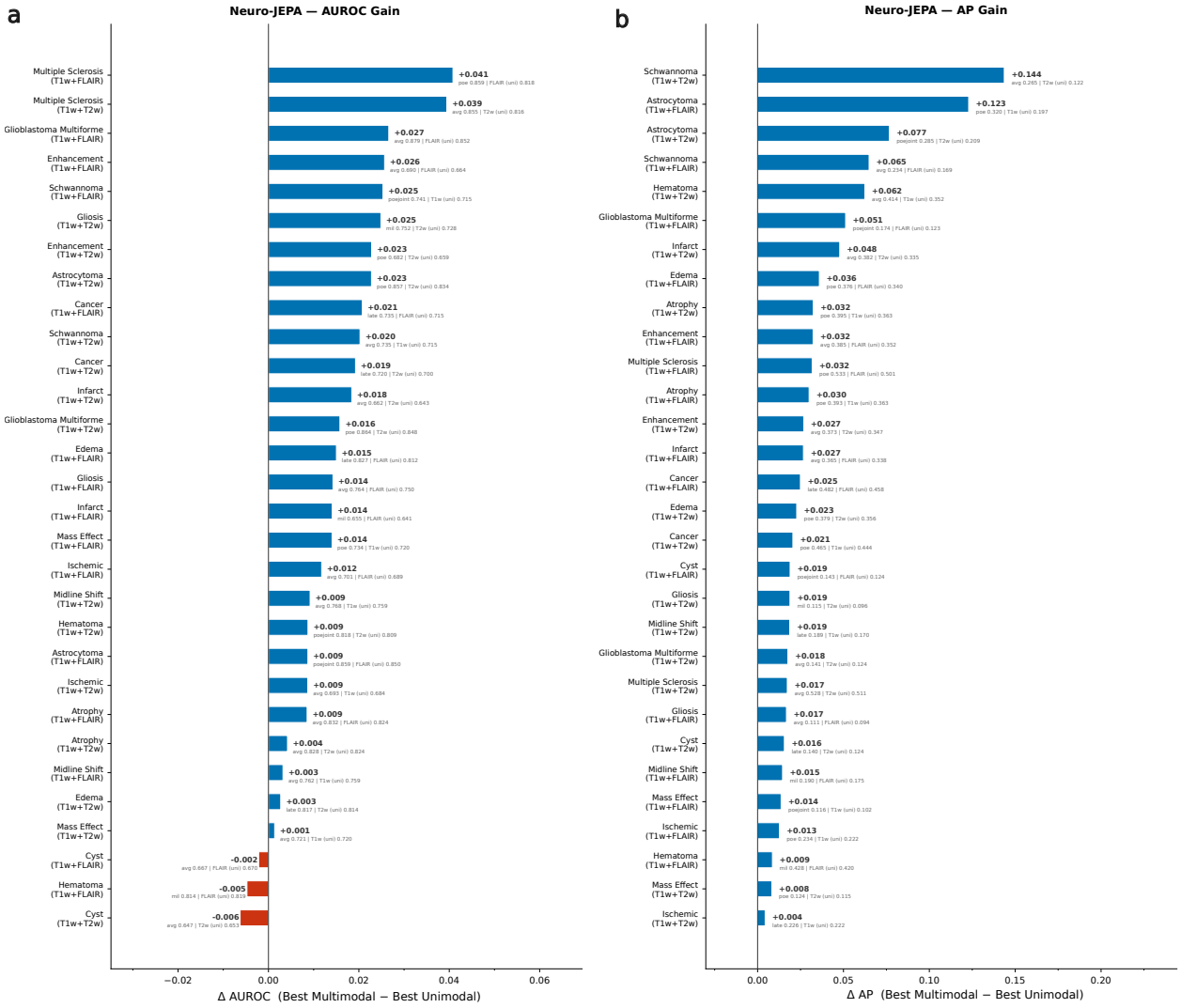
Supplementary Figure 15. **Multimodal Learning Performance Across Fusion Methods for VoCo on BIND-MGH - AP.**



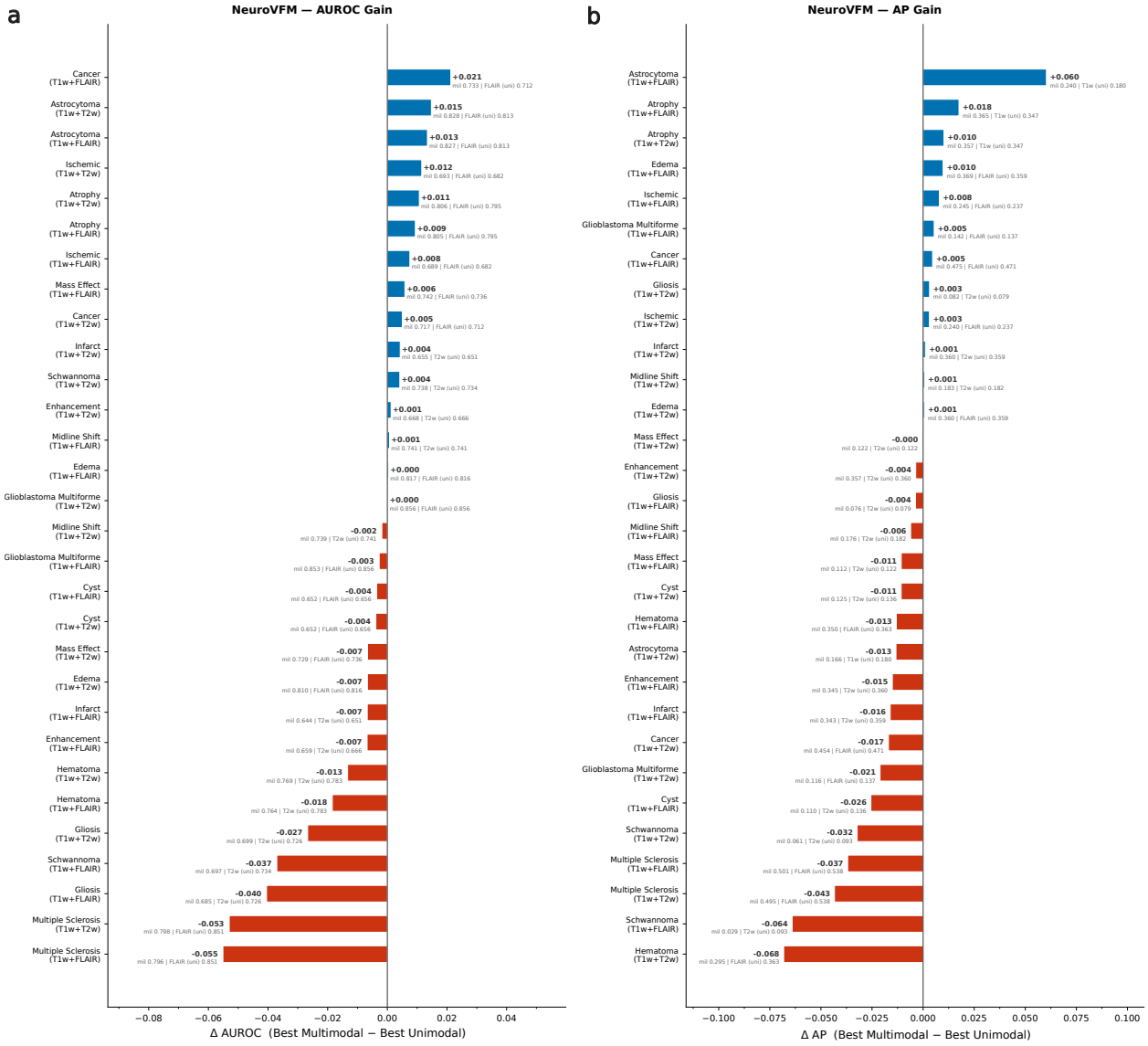
Supplementary Figure 16. Multimodal Learning Performance Across Fusion Methods for NeuroVFM on BIND-MGH - AUROC.



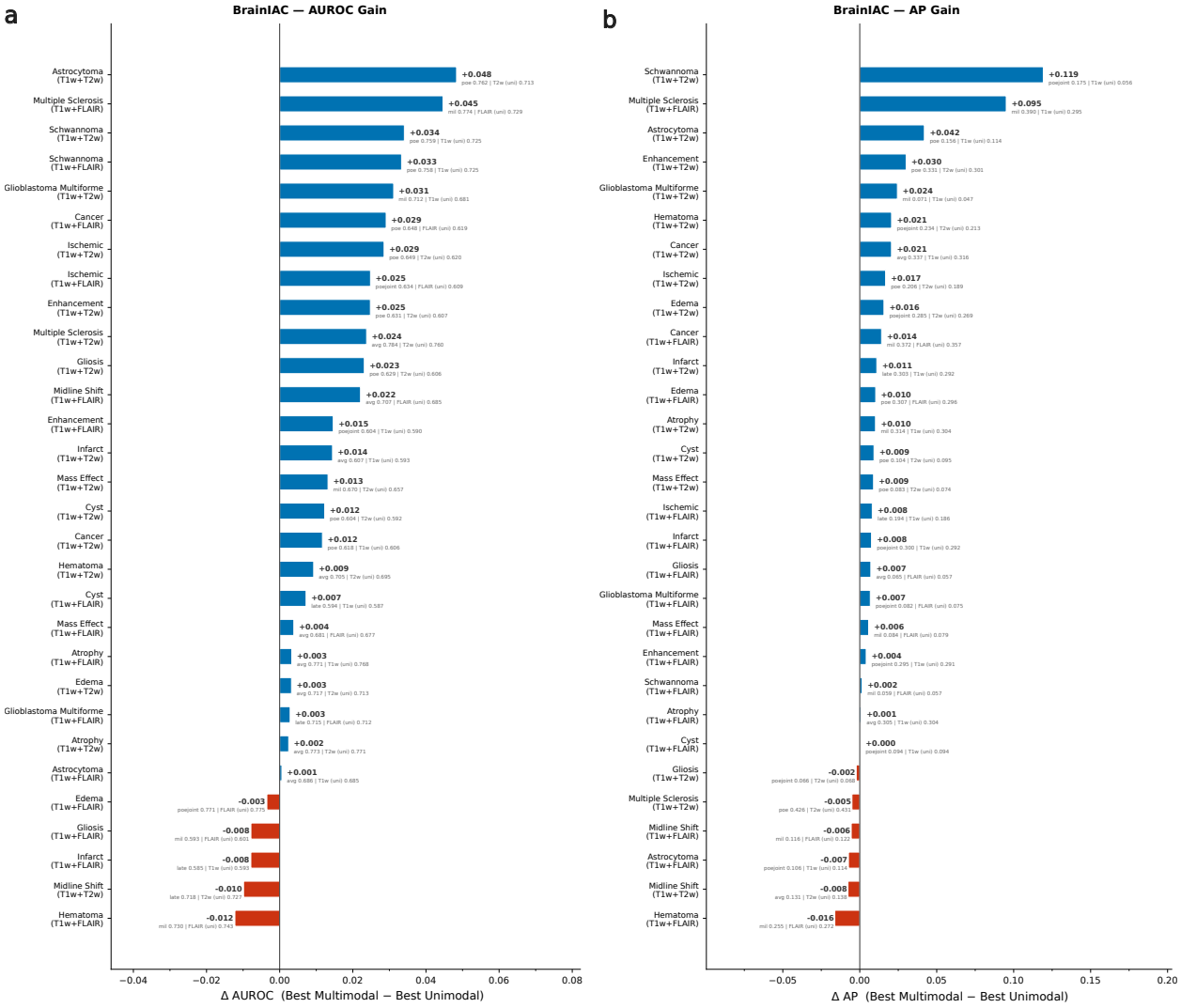
Supplementary Figure 17. **Multimodal Learning Performance Across Fusion Methods for NeuroVFM on BIND-MGH - AP.**



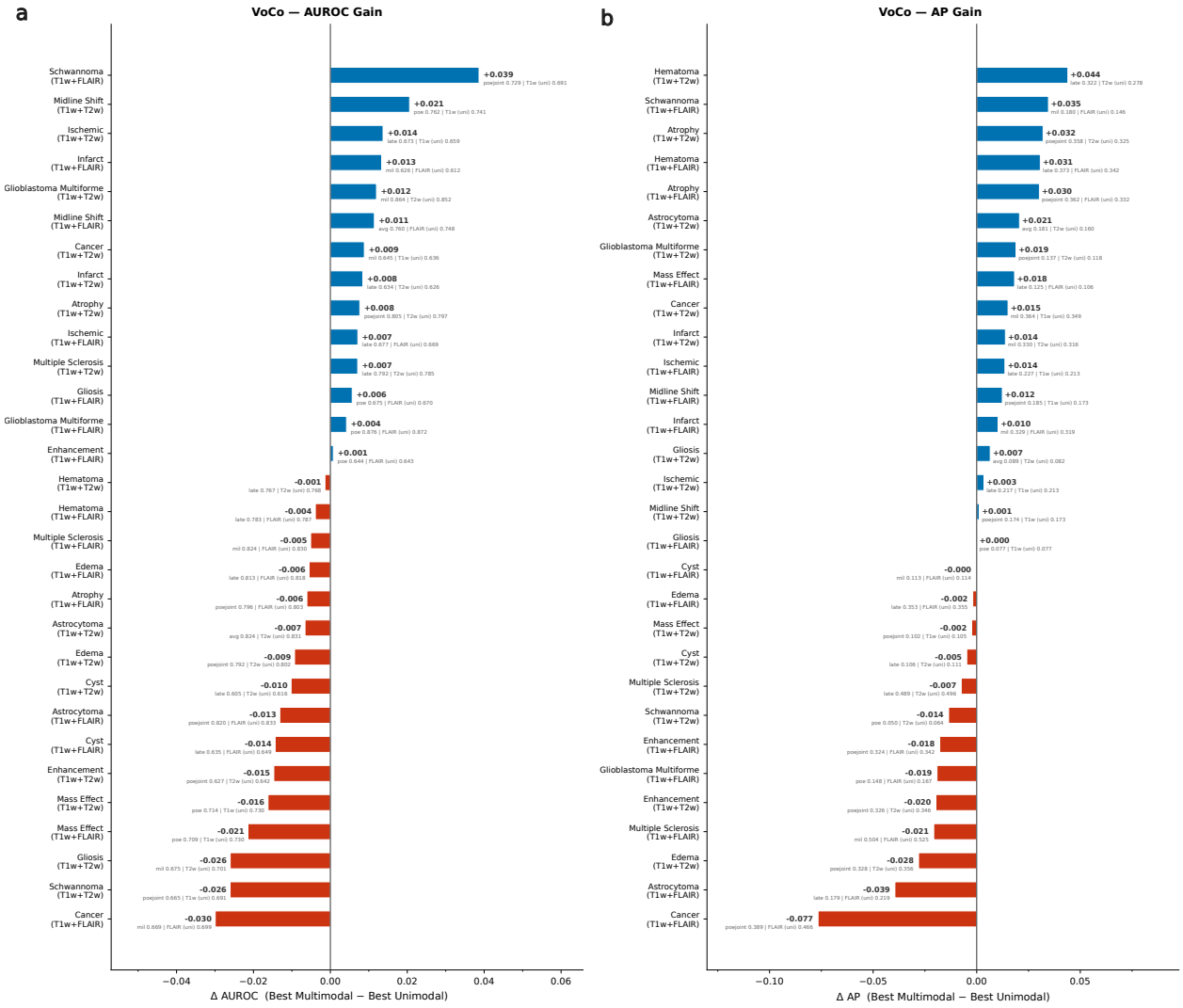
Supplementary Figure 18. Multimodal Gain Over Unimodal for Neuro-JEPA on BIND-MGH. The difference between best multimodal fusion vs. best unimodal performance on Neuro-JEPA reported with AUROC and AUPRC. **a,** AUROC and AUPRC multimodal performance gain on the difference. **b,** AUPRC multimodal performance gain on the difference.



Supplementary Figure 19. Multimodal Gain Over Unimodal for NeuroVFM on BIND-MGH. The difference between best multimodal fusion vs. best unimodal performance on NeuroVFM reported with AUROC and AUPRC. **a,** AUROC and AUPRC multimodal performance gain on the difference. **b,** AUPRC multimodal performance gain on the difference.



Supplementary Figure 20. Multimodal Gain Over Unimodal for BrainIAC on BIND-MGH. The difference between best multimodal fusion vs. best unimodal performance on BrainIAC reported with AUROC and AUPRC. **a,** AUROC and AUPRC multimodal performance gain on the difference. **b,** AUPRC multimodal performance gain on the difference.



Supplementary Figure 21. Multimodal Gain Over Unimodal for VoCo on BIND-MGH. The difference between best multimodal fusion vs. best unimodal performance on NeuroVFM reported with AUROC and AUPRC. **a,** AUROC and AUPRC multimodal performance gain on the difference. **b,** AUPRC multimodal performance gain on the difference.

E Ablation Studies Experiments

E.1 Ablation Study on Number of Experts

Supplementary Figure 22 presents detailed result for mixture of experts and dense model across evaluated clinical cohorts datasets with attentive probing. As the result shown, changing from dense model to 16 total experts give most performance improvement, while increasing number of total experts presents diminishing returns.

E.2 Ablation Study on Methods

Supplementary Figure 23 presents result for the impact of modifying the algorithm details starting from original V-JEPA 2 across the evaluated clinical cohorts datasets. The result is present as incrementally adding Multiscale Masking, Mixture of Experts and Foreground-aware L1 Loss. As the result shown, every modification presents meaningful improvement on the model overall performance.

E.3 Ablation Study on NeuroVFM

Supplementary Figure 24 presents per-dataset model performance comparison with clinical cohorts datasets for Neuro-JEPA and NeuroVFM with attentive probing at per modality level (T1w, T2w, FLAIR). As the result demonstrates, Neuro-JEPA show improved performance over NeuroVFM on a majority of cases, highlighting the important of algorithmic improvement over data scaling.

E.4 Ablation Study on Different Percent of Pretrain Data

Supplementary Figure 25 presents model per dataset and modality performance on model pre-trained on different percentage of pre-training data evaluated on clinical cohorts datasets with attentive probing. The result shows that our method is scalable with increasing data size across the tasks with diminishing return on data scaling similar to V-JEPA 2 [14].

E.5 Ablation Study on Model Size

We present ablation study on model performance for different parameter size on ViT-B-MoE (121 million total parameters) and ViT-L-MoE (429 million total parameters) both pretrained with same Neuro-JEPA receipt on Supplementary Figures 26 and 27. The result presents that ViT-L-MoE outperform ViT-B-MoE in majorty of cases, highlighting the effectiveness of our pretraining receipt unnder model parameters size scaling.

E.6 Ablation Study on Patch Size

We present ablation study on model patch size for ViT-Large with 12 patches versus 8 patches in Figure 28. We observe the significant performance difference across tasks with the more fine-grained patch size of 8.

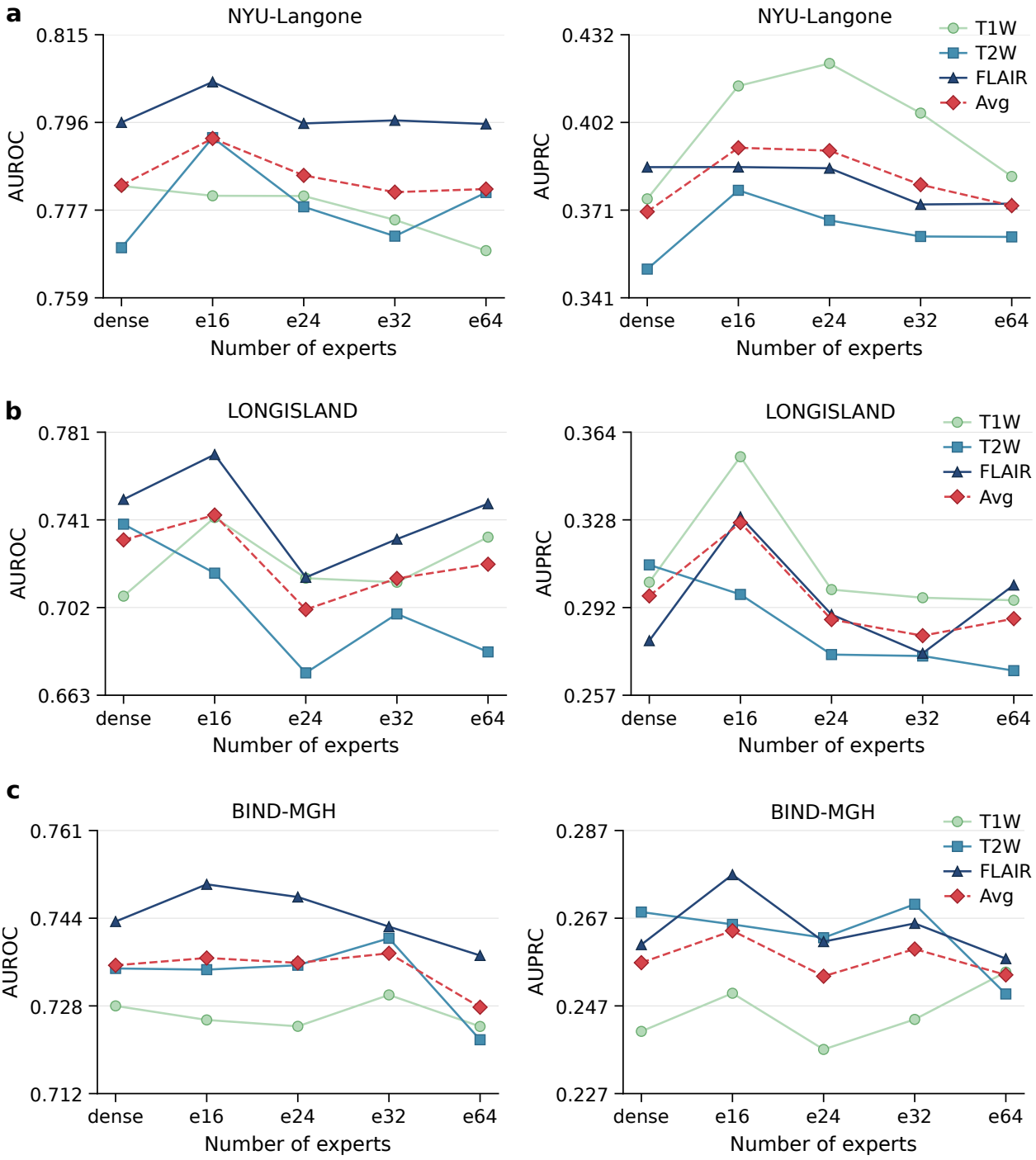
E.7 Ablation Study on Pretraining with Uncurated vs. Curated Data

Supplementary Figure 29 presents the performance difference on pretraining model with curated vs. uncurated data with 30% pretraining data. Uncurated data present 5% more noisy samples that are filtered out from curated data. The result demonstrates that including noisy samples for neuroimaging pre-training can present negative impact on overall model performance.

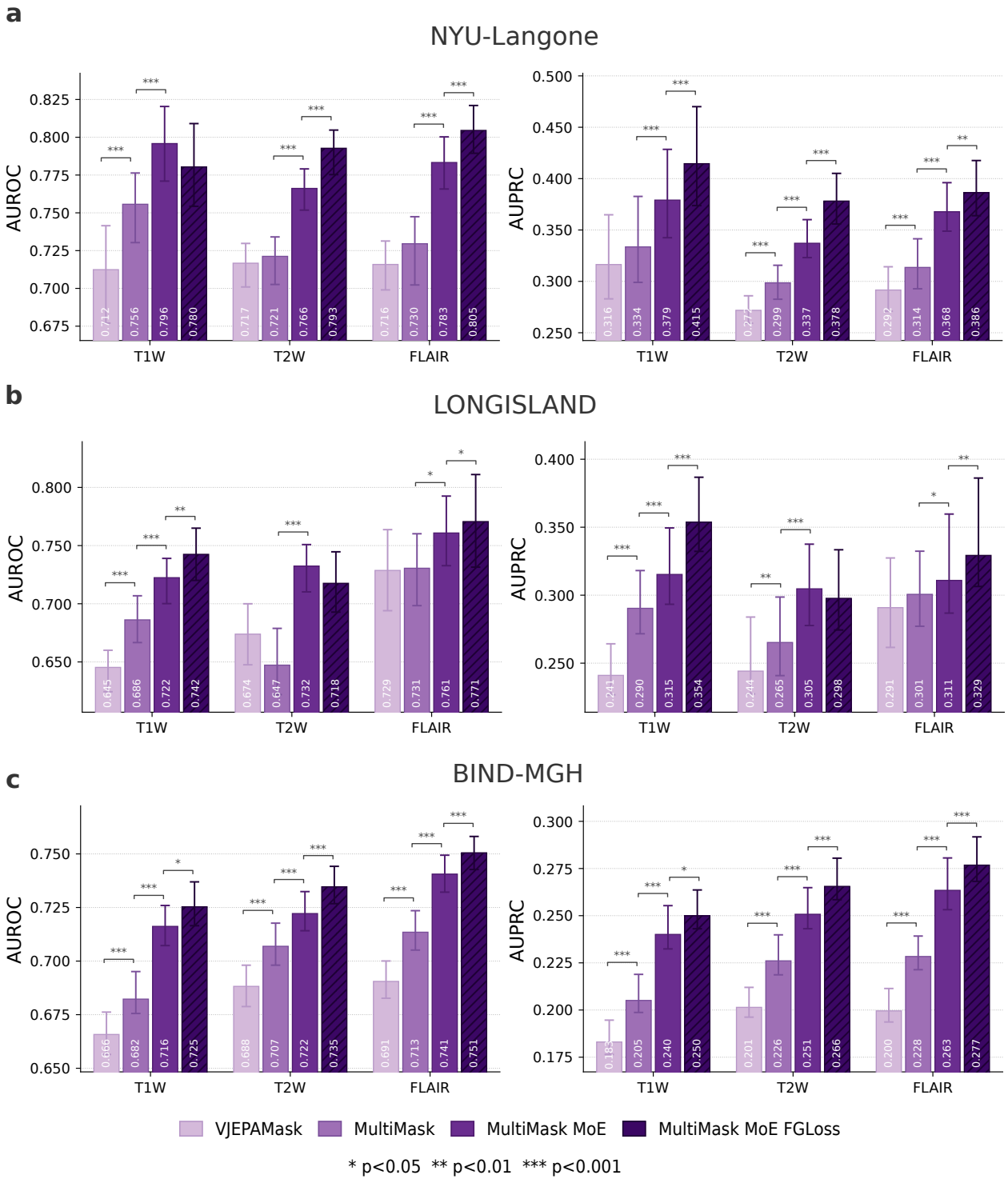
Supplementary Figure 30 shows examples on some filtered out samples. Most samples present misaligned or missing anatomical brain structure that can potentially introduce pure noisy signals on model pre-training and make JEPA training trajectory to be unstable due to lack of visible context information on predicting the masked regions.

E.8 Ablation Study on MAE vs. JEPA Pre-training Performance

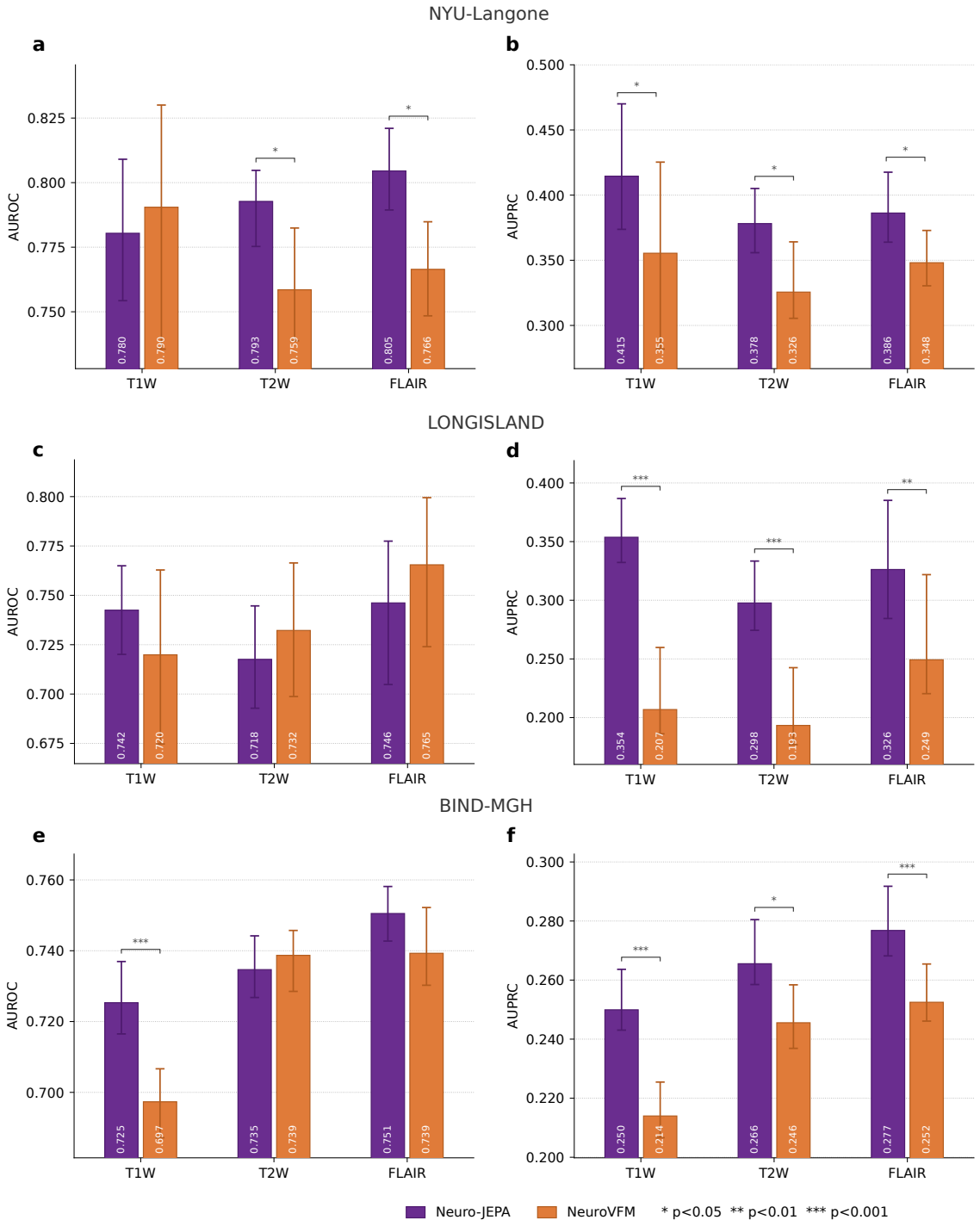
Supplementary Figure 31 presents comparison on MAE vs. JEPA pre-training on 30% data with same architecture. The result is evaluated on clinical cohorts datasets with attentive probing. The result demonstrates that our improved JEPA pre-training can effectively improve model overall performance over MAE across the datasets and modalities.



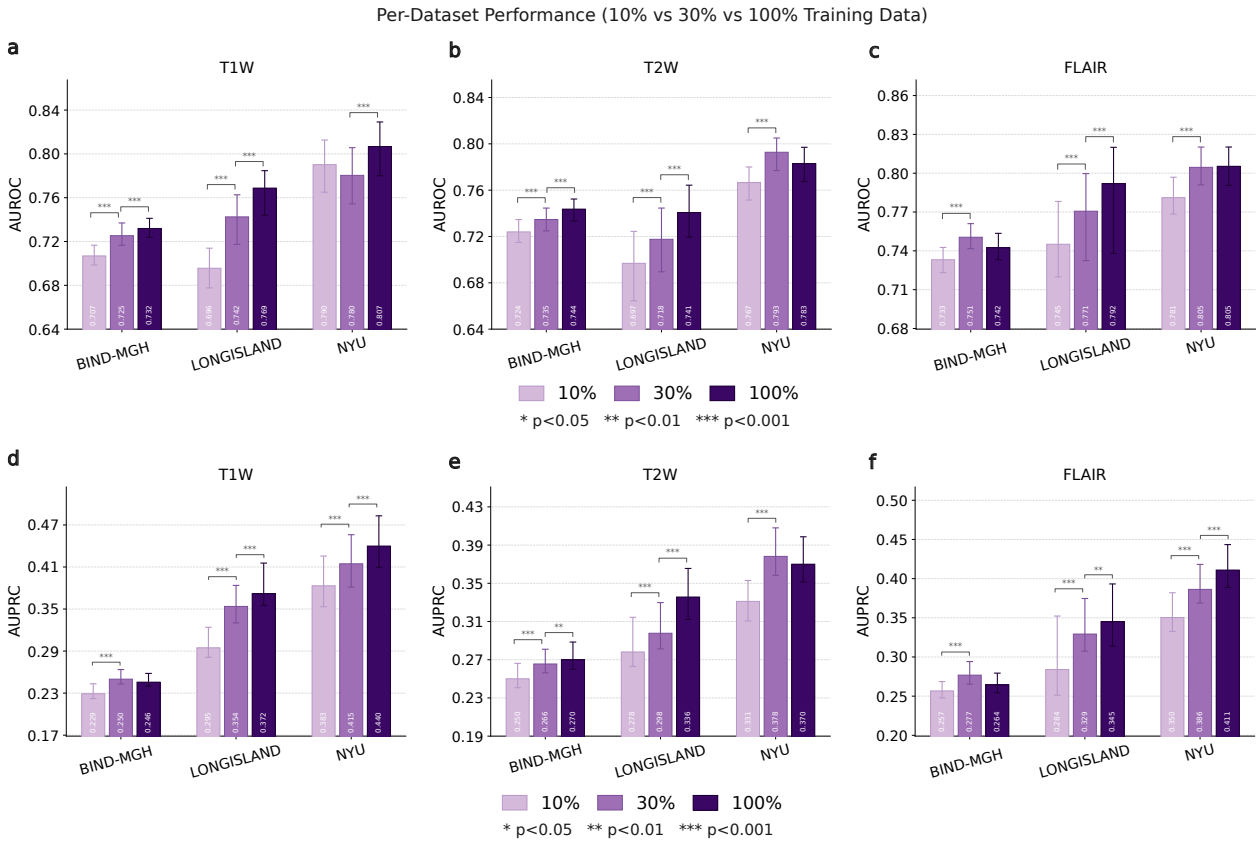
Supplementary Figure 22. Per Dataset Result for Number of Expert Ablation Study. Per dataset result for number of experts with attentive probing across three evaluated datasets (NYU Langone, Longisland and BIND-MGH). Consistent with average performance show in main manuscript [ref], we observed improved performance on both AUROC and AUPRC when comparing model performance on dense model vs. model with same setting on 16 total experts. However, the performance improvement is diminishing when number of experts is increase from 24 to 64.



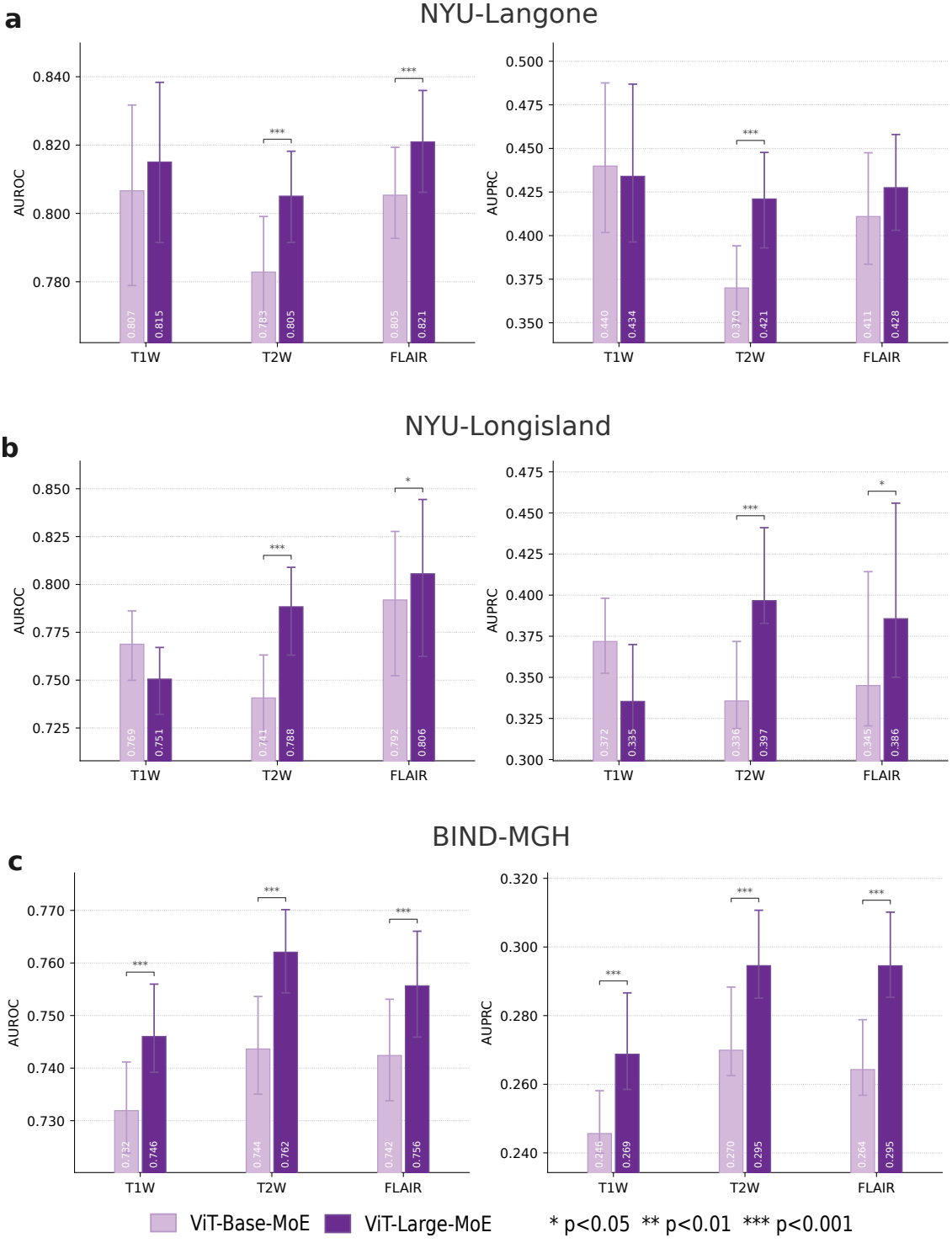
Supplementary Figure 23. Per Dataset Result for Design Choices Ablation Study. Per dataset result for design choices with attentive probing across three evaluated datasets (NYU Langone, Longisland and BIND-MGH). The color indicates different combination of methods used during pretraining. Consistent with average performance show in main manuscript [ref], incrementally adding multiscale masking, mixture of experts and foreground aware L1 loss improve overall probing performance across the datasets.



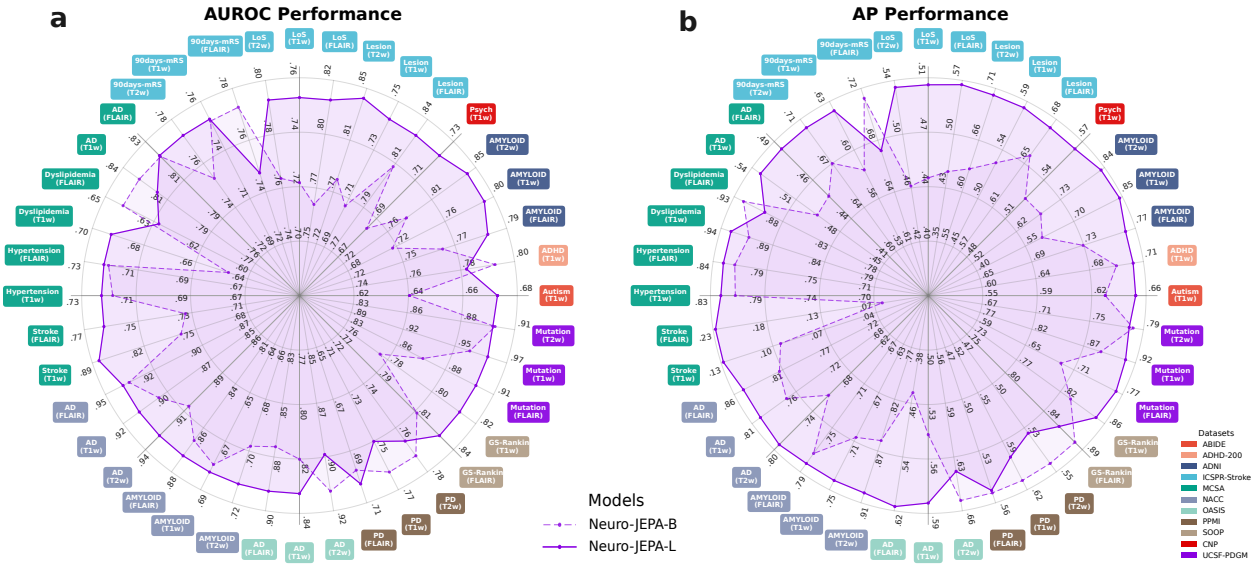
Supplementary Figure 24. Per Dataset Comparison with NeuroVFM. Per dataset performance comparison with NeuroVFM. The performance is reported as AUROC and AUPRC average across all tasks on each dataset. The result shows that although both models have similar performance on AUROC, our model shows large performance improvement on AUPRC. This is critical on indicating the superior of our model as AUPRC better reflect the correctness on classification for the diagnosis where the positive cases are usually rare.



Supplementary Figure 25. Model Performance on Different Percentage of Pretrain Data. AUROC and AUPRC averaged across tasks for each dataset and modality. **a-c,** AUROC for different datasets on different percentage for T1w, T2w and FLAIR. **e-f,** AUPRC for different datasets on different percentage for T1w, T2w and FLAIR.



Supplementary Figure 26. Model Size Ablation - Health System Datasets. This ablation experiment show per health-system dataset and modality performance (AUROC and AUPRC for Neuro-JEPA with Vit-Base and ViT-Large with **a,** NYU-Langone **b,** NYU-Longisland **c,** MGH



Supplementary Figure 27. Model Size Ablation - Public Datasets. This ablation experiment show per public dataset and modality performance for Neuro-JEPA with ViT-Base and ViT-Large with **a,** AUROC **b,** AUPRC

F Generalization Under Cohort and Modality Shifts

F.1 Cross-Cohort Transfer Across Matched Clinical Endpoints

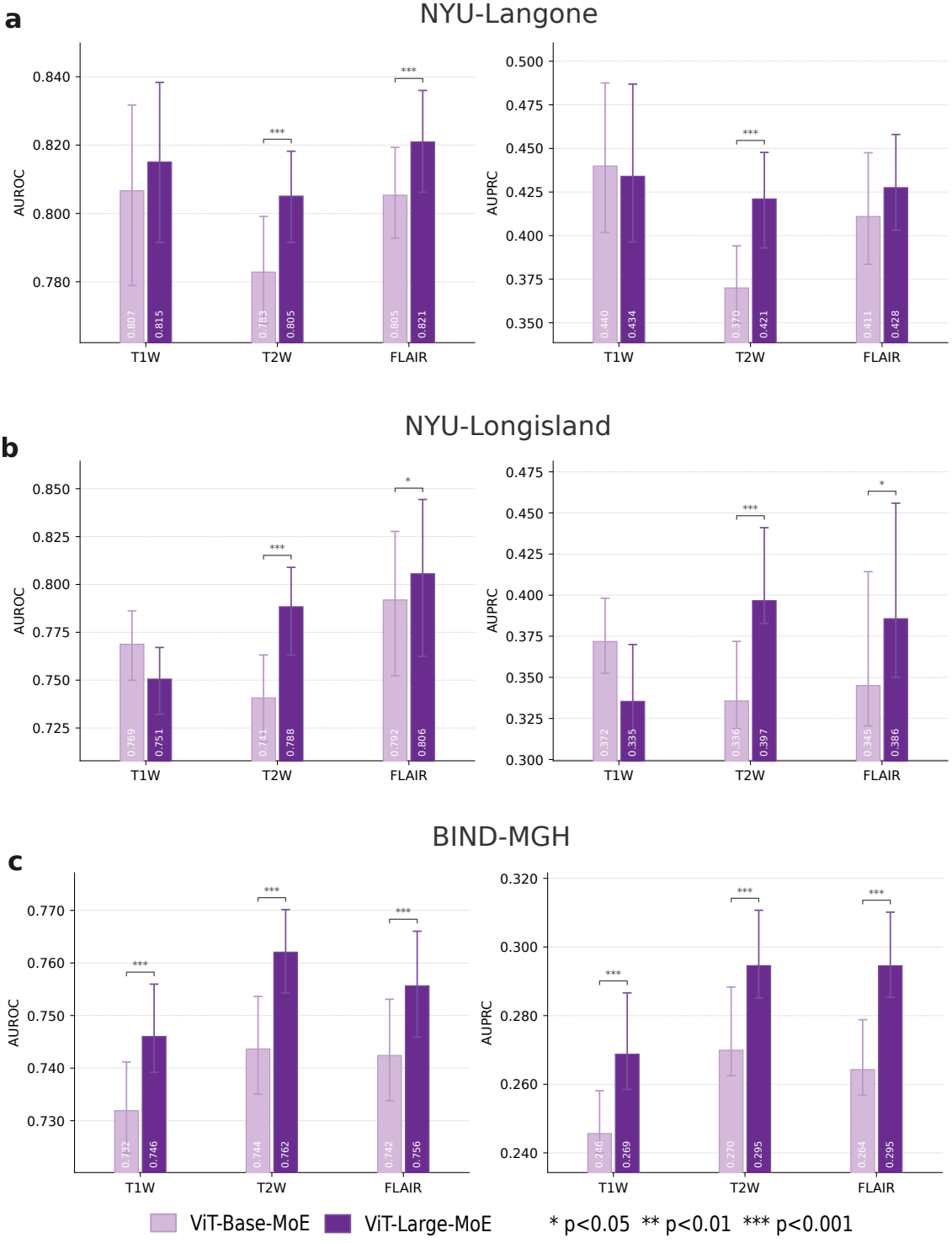
Supplementary Figure 32 evaluates the out-of-domain transfer performance of Neuro-JEPA across independent clinical cohorts. In this setting, the model is fine-tuned on a source cohort and evaluated directly on an external target cohort with the same label definition, providing a clinically relevant assessment of robustness under cohort and institutional distribution shifts. Across the evaluated tasks, Neuro-JEPA maintains strong transfer performance relative to in-domain evaluation, demonstrating effective generalization beyond the source-domain data.

F.2 Cross-Modality Generalization to DWI

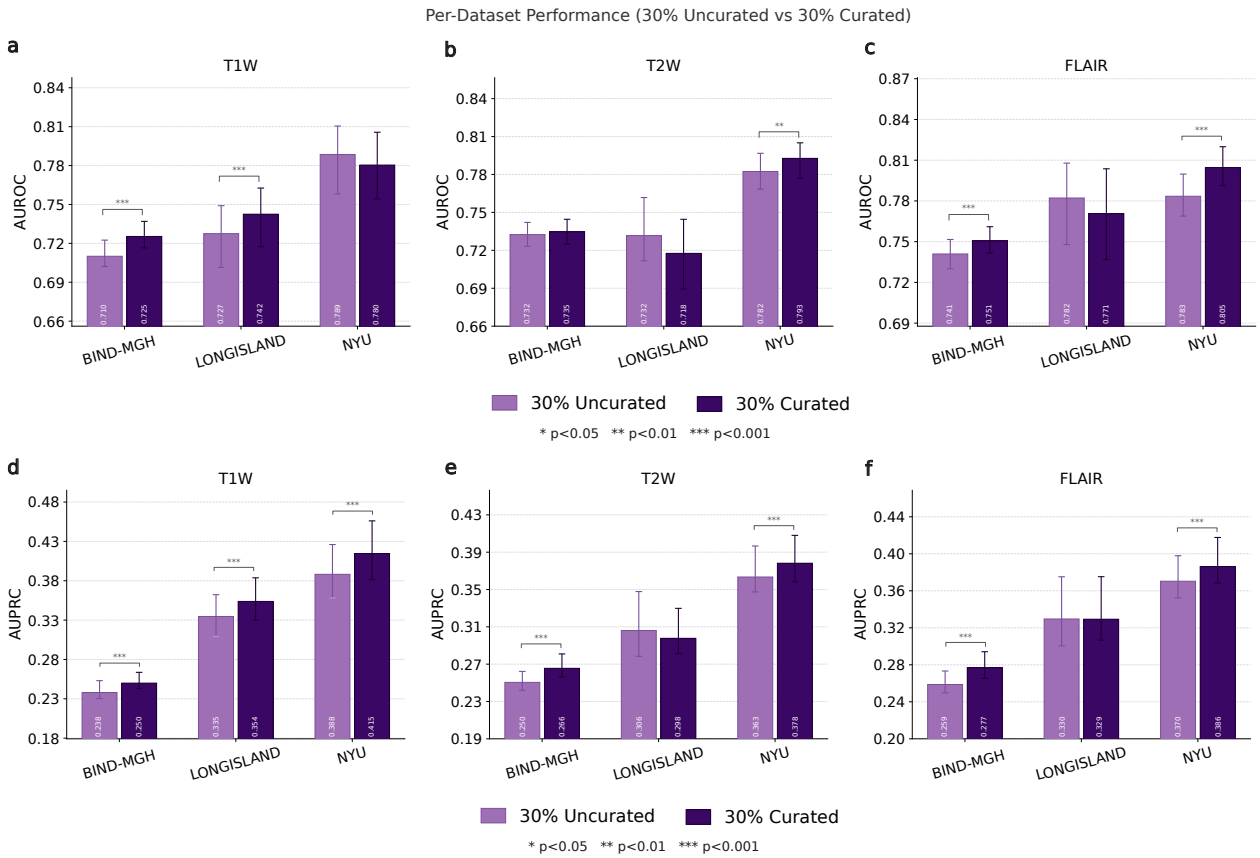
Supplementary Figure 33 and table 11 evaluates the ability of Neuro-JEPA to generalize to diffusion-weighted MRI (DWI), a modality that was not included during pretraining. Models are fine-tuned on DWI scans and evaluated on downstream tasks from the ICSPR-Stroke and UCSF-PDGM datasets. Despite the absence of DWI during pretraining, Neuro-JEPA achieves competitive performance across tasks and obtains the best macro-averaged AUROC and AUPRC among the evaluated models. These results indicate that Neuro-JEPA learns representations that transfer beyond the imaging modalities observed during pretraining.

Supplementary Table 11. Performance comparison on DWI downstream tasks. Metrics are macro-averaged AUROC and AUPRC. For the three-class lesion type task, both metrics are computed using a one-versus-rest macro average. **Bold** indicates the best-performing model for each metric, and underlining indicates the second-best model.

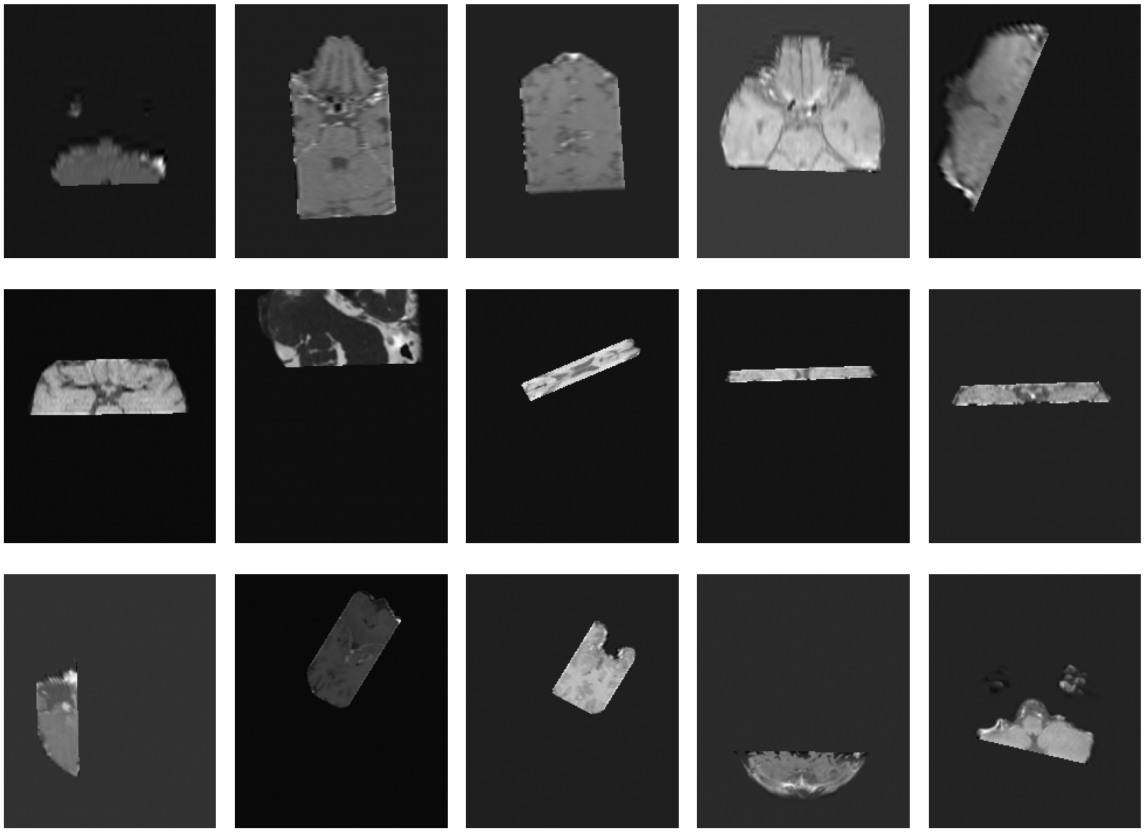
	Neuro-JEPA-B		VoCo		BrainIAC		NeuroVFM	
Task	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
ICSPR-Stroke								
90-Day mRS	0.710	**0.643**	0.687	0.608	0.694	0.627	**0.714**	0.628
Lesion Type	0.884	0.762	0.885	0.753	0.717	0.521	**0.902**	**0.805**
Length of Stay	**0.754**	**0.471**	0.710	0.411	0.718	0.394	0.687	0.407
UCSF-PDGM								
IDH Mutation	**0.832**	**0.650**	0.792	0.573	0.772	0.521	0.811	0.613
Average	**0.795**	**0.631**	0.769	0.586	0.725	0.516	0.778	0.613



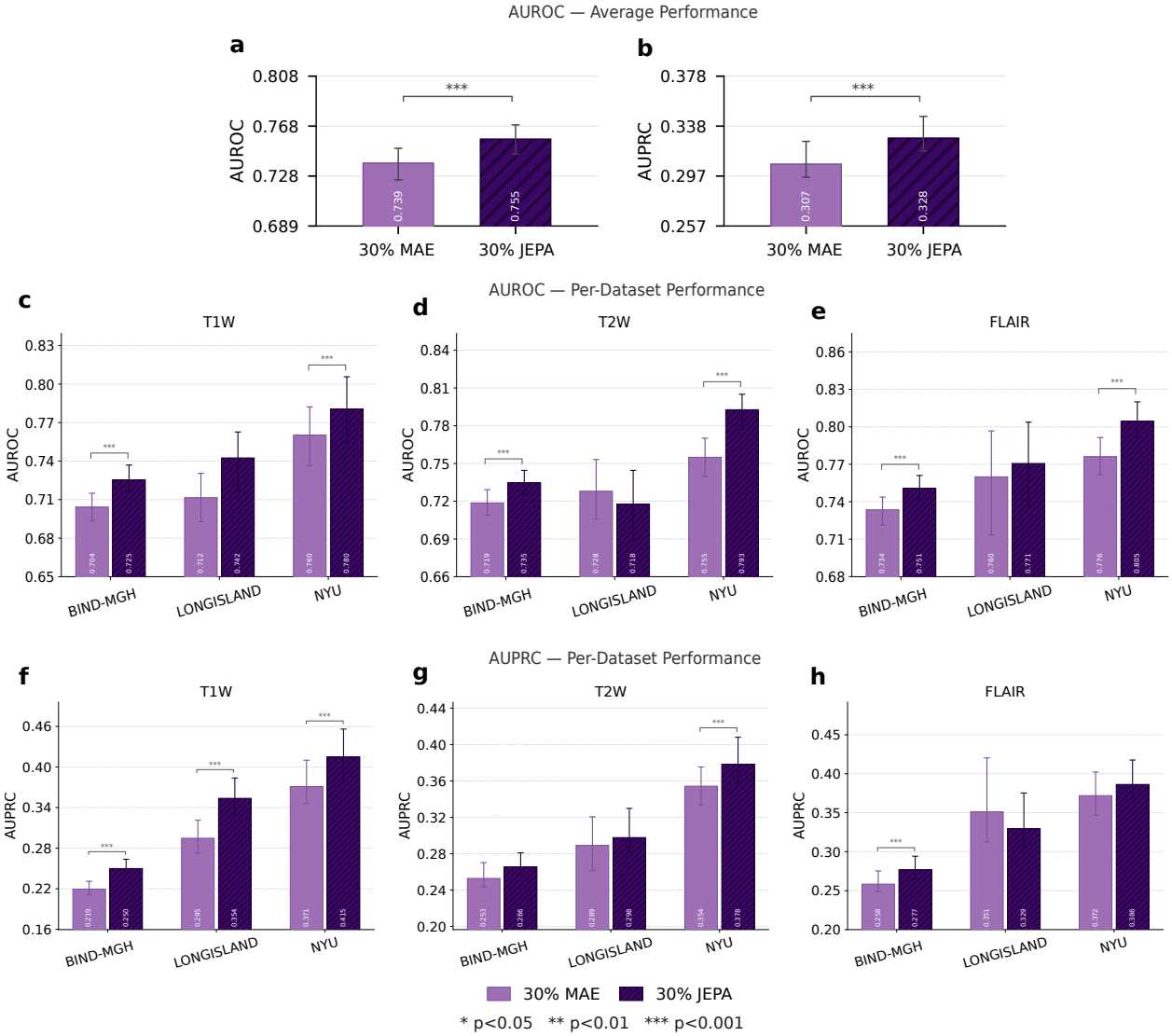
Supplementary Figure 28. Patch Size Ablation - Health System Datasets. This ablation experiment show per health-system dataset and modality performance (AUROC and AUPRC for Neuro-JEPA with Vit-Base and ViT-Large with **a,** NYU-Langone **b,** NYU-Longisland **c,** MGH



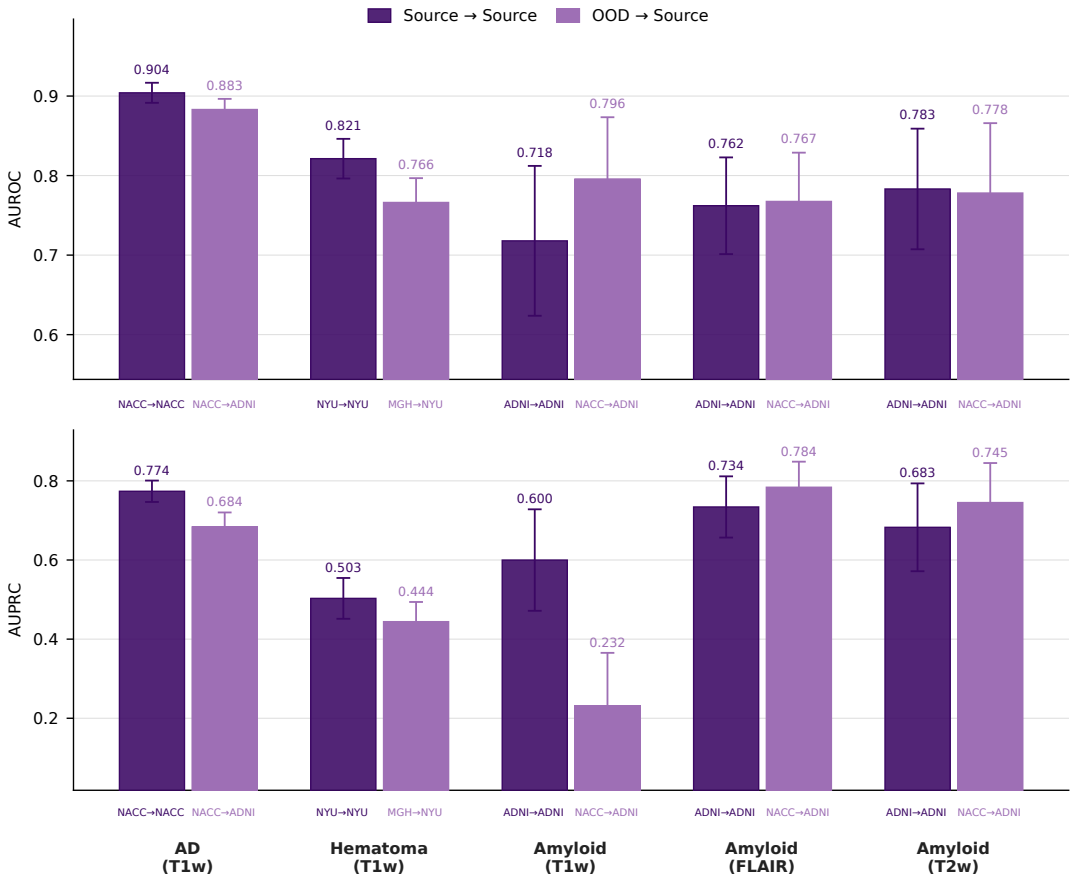
Supplementary Figure 29. Model Performance on Pretrain with Uncurated vs. Curated Data. This ablation experiments are run on 30% of pretrain data with Multi-Scale Masking, MoE and Foreground-Aware Masking all enabled. AUROC and AUPRC are averaged across tasks for each dataset and modality. **a-c,** AUROC on different datasets for model pretrained with uncurated vs.curated data for T1w, T2w and FLAIR. **e-f,** AUPRC on different datasets for model pretrained with uncurated vs.curated data for T1w, T2w and FLAIR.



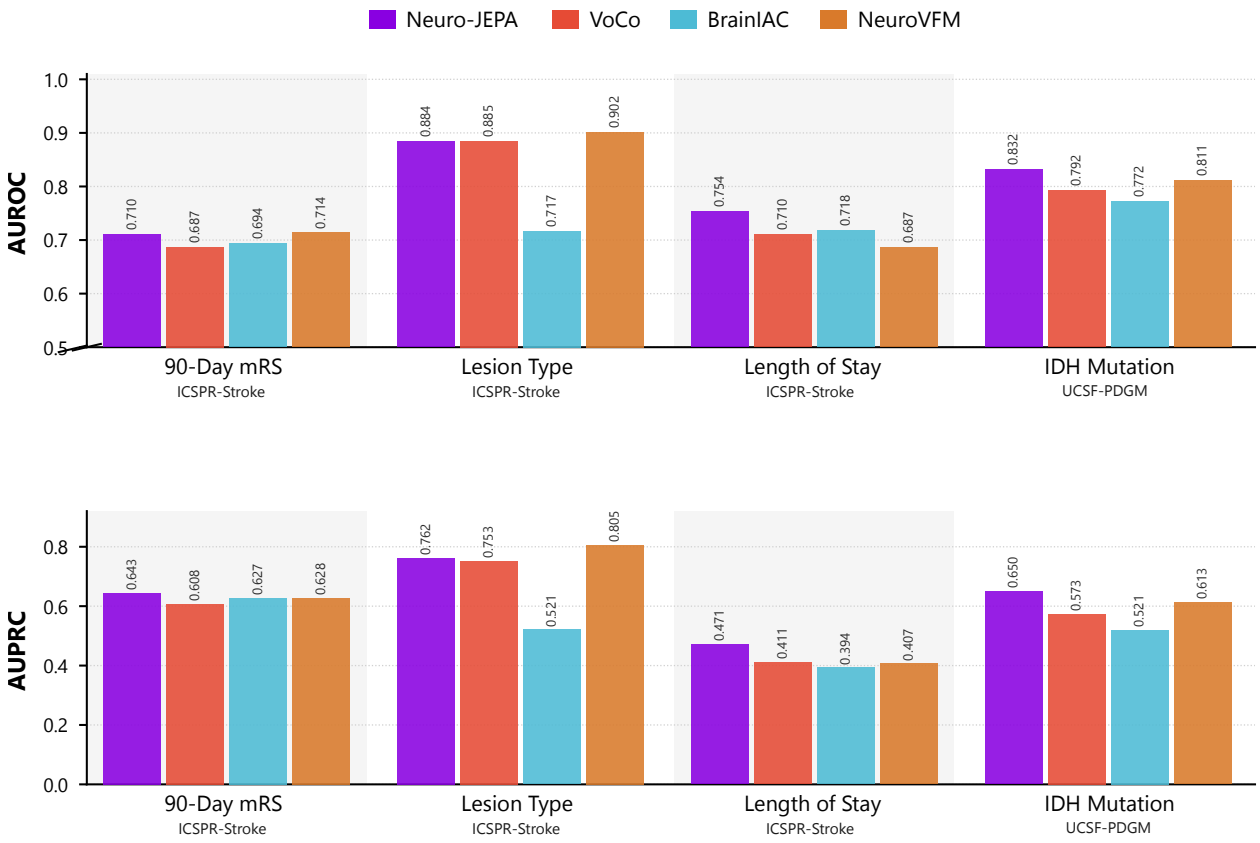
Supplementary Figure 30. Samples of Filtered Out Noisy Scans. We show representative slices from scans excluded during scan-level quality control. Many of these scans contained limited usable anatomical information after registration, primarily due to restricted or incorrect fields of view, severe motion artifacts, acquisition or reconstruction failures, and other technical issues. Empirically, we found that pretraining with such scans can lead to unstable training dynamics and degraded downstream model performance. These observations underscore the importance of systematic data curation when developing neuroimaging foundation models from large-scale clinical cohorts collected directly from routine hospital practice.



Supplementary Figure 31. MAE vs. JEPA Performance Across Datasets. We compare the performance trained on 30% of pretrain data, where we show JEPA on our full configurations with Multi-Scale Masking, MoE and Foreground Aware L1 Loss consistently outperform MAE. **a,b,** AUROC and AUPRC performance comparison averaged across all datasets and modalities **c-e,** AUROC performance comparison on each dataset and modality. **f-g,** AUPRC performance comparison on each dataset and modality.



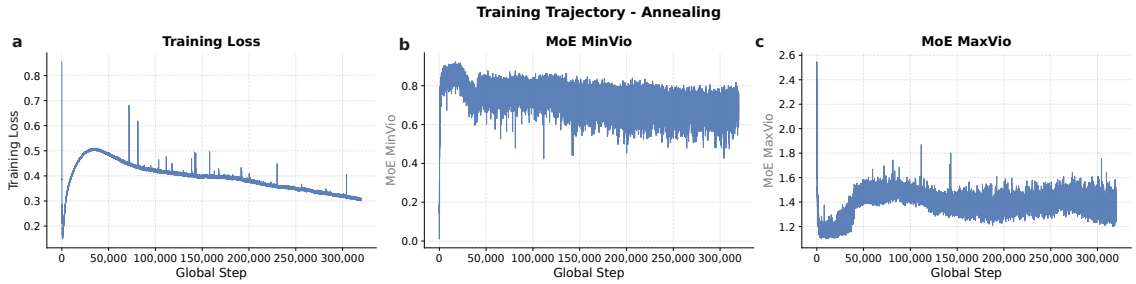
Supplementary Figure 32. Cross-cohort out-of-domain transfer performance. AUROC and AUPRC are reported for models fine-tuned on one cohort and evaluated on an external cohort with matched task definitions. The evaluated transfer settings include NACC to ADNI for Alzheimer's disease and amyloid prediction, and MGH to NYU for hematoma prediction. Transfer performance is compared with in-domain performance, where models are trained and evaluated within the same cohort. Neuro-JEPA preserves strong predictive performance under external cohort shift in a majority of tasks, supporting its robustness in out-of-domain transfer settings.



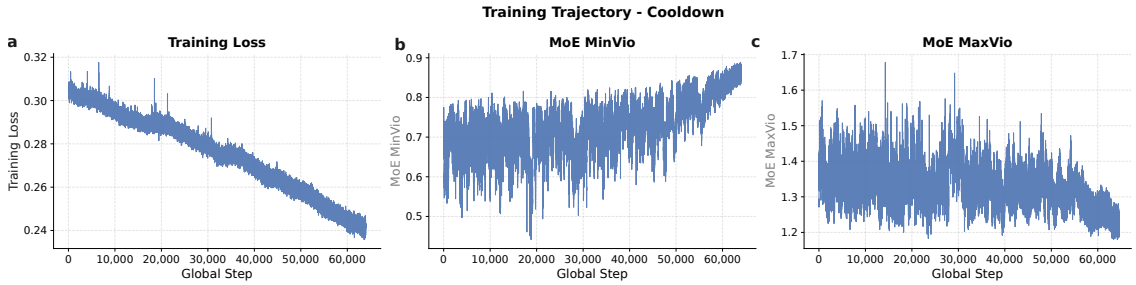
Supplementary Figure 33. Performance comparison on DWI downstream tasks. Metrics are macro-averaged AUROC and AUPRC. For the three-class lesion type task, both metrics are computed using a one-versus-rest macro average. **Bold** indicates the best-performing model for each metric, and underlining indicates the second-best model.

G Pre-Training Optimization Dynamics

Supplementary Figures 34 and 35 presents training dynamics on training loss and Mixture of Experts load balancing (Minimum Violation and Maximum Violation) with 100% pretraining data on our base model for 200 epochs annealing and 40 epochs cooldown. The result demonstrates that our model can be stably pre-trained under proper hyper-parameters setup. More details on pretraining dynamics can be found in the public wandb report https://api.wandb.ai/links/notody/7t7d4dks.



Supplementary Figure 34. Training Dynamics on Full-data Annealing Pretraining. We show the optimization trajectory of Neuro-JEPA trained for 200 epochs annealing on the full pretraining dataset with our base model, using multiscale masking, Mixture-of-Experts (MoE) routing, and foreground-aware L1 latent predictive loss. **a,** Latent predictive L1 loss over training steps. **b,** Minimum MoE load-balancing violation, used to monitor under-utilization of experts. **c,** Maximum MoE load-balancing violation, used to monitor over-utilization of experts.



Supplementary Figure 35. Training Dynamics on Full-data Cooldown Pretraining. We show the optimization trajectory of Neuro-JEPA trained for 40 epochs cooldown on the full pretraining dataset with our base model, using multiscale masking, Mixture-of-Experts (MoE) routing, and foreground-aware L1 latent predictive loss. **a,** Latent predictive L1 loss over training steps. **b,** Minimum MoE load-balancing violation, used to monitor under-utilization of experts. **c,** Maximum MoE load-balancing violation, used to monitor over-utilization of experts.

Supplementary Table 12. Average unimodal performance across cohorts and task types (Public datasets). Values are mean [95% CI]. **Bold**: best per metric. ↑ higher is better; ↓ lower is better. "combs" refer to the number of dataset-task-modality combinations. "n" for brain age represents number of samples in the test set. underlining indicates the second-best model.

(a) Classification / Diagnosis / Prognosis (41 combs)

Model	AUROC↑	AUPRC↑
VoCo-B	0.721 [0.689, 0.754]	0.573 [0.513, 0.635]
BrainIAC	0.728 [0.705, 0.751]	0.555 [0.494, 0.618]
NeuroVFM	0.741 [0.706, 0.774]	0.585 [0.526, 0.648]
CNN	0.748 [0.718, 0.779]	0.604 [0.539, 0.668]
Neuro-JEPA	**0.785** [0.760, 0.811]	**0.649** [0.588, 0.705]

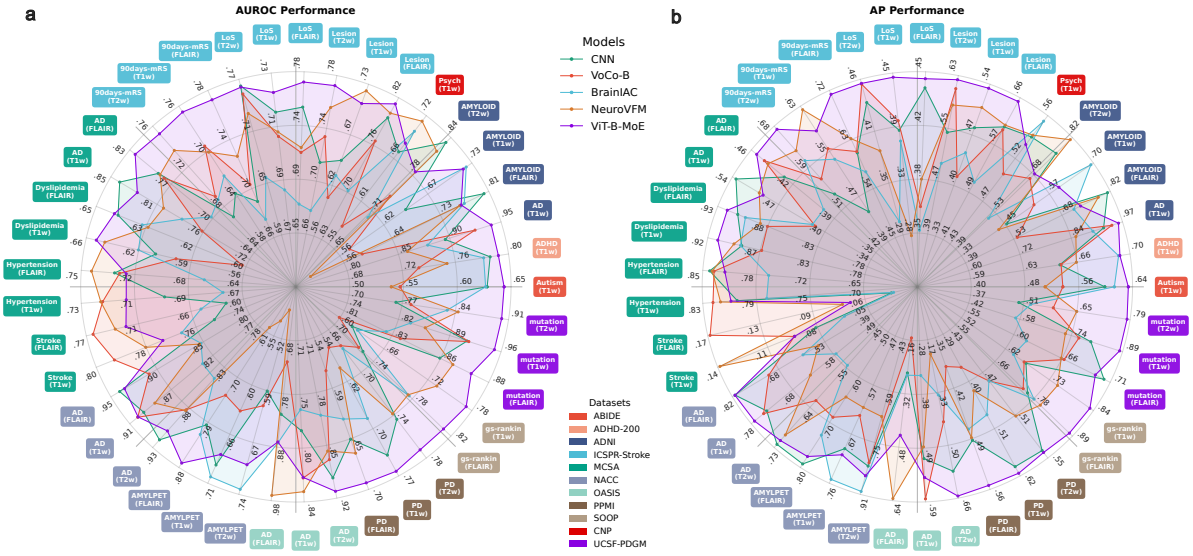
(b) Brain Age Prediction ($n = 757$)

Model	R^2 ↑	MAE↓ (yr)	RMSE↓ (yr)
VoCo-B	0.111 [0.065, 0.165]	6.22 [5.46, 6.94]	12.03 [10.54, 13.36]
BrainIAC	0.522 [0.447, 0.591]	5.42 [4.93, 5.95]	8.82 [7.77, 9.78]
NeuroVFM	0.673 [0.611, 0.735]	4.36 [3.91, 4.77]	7.29 [6.35, 8.12]
CNN	0.858 [0.827, 0.883]	3.25 [2.99, 3.50]	4.81 [4.39, 5.21]
Neuro-JEPA	**0.894** [0.860, 0.917]	**2.78** [2.56, 3.02]	**4.15** [3.69, 4.70]

H Additional Validation Analyses

H.1 Comparison with Simple CNN Baseline

To assess whether foundation-model pretraining provides practical value beyond conventional simple models, we benchmarked a specific designed CNN baseline for neuroimaging [15] (a wide 4-layers CNN with 32 million parameters) trained directly for the downstream task from scratch. All evaluations follow hyperparameters sweep detailed in Supplementary Section B.7. This comparison provides an important calibration on if pre-existing foundation models bring any benefits beyond simple methods. The evaluation is performed on the 41 combinations from 12 public datasets and age prediction. Across these evaluations, existing foundation models did not consistently surpass the CNN baseline, whereas Neuro-JEPA achieved consistent gains over the CNN on both averaged AUROC (3.7% improvement) and AUPRC (4.5% improvement) (Supplementary Table 12; Supplementary Figure 36). On age prediction, Neuro-JEPA presents the only foundation model performs better with Quasi-RAW scans (+2.8 on R^2, -0.37 on *MAE* and -0.50 on *RMSE*). The results indicate that Neuro-JEPA represents first successful foundation model on surpassing simple CNN baseline with proper scaling and algorithmic design.



Supplementary Figure 36. All Evaluated Models vs. Simple CNN Baseline. AUROC and AUPRC for 41 different combinations on public datasets. **a,** AUROC comparison across tasks and modalities. **b,** AUPRC comparison across tasks and modalities.

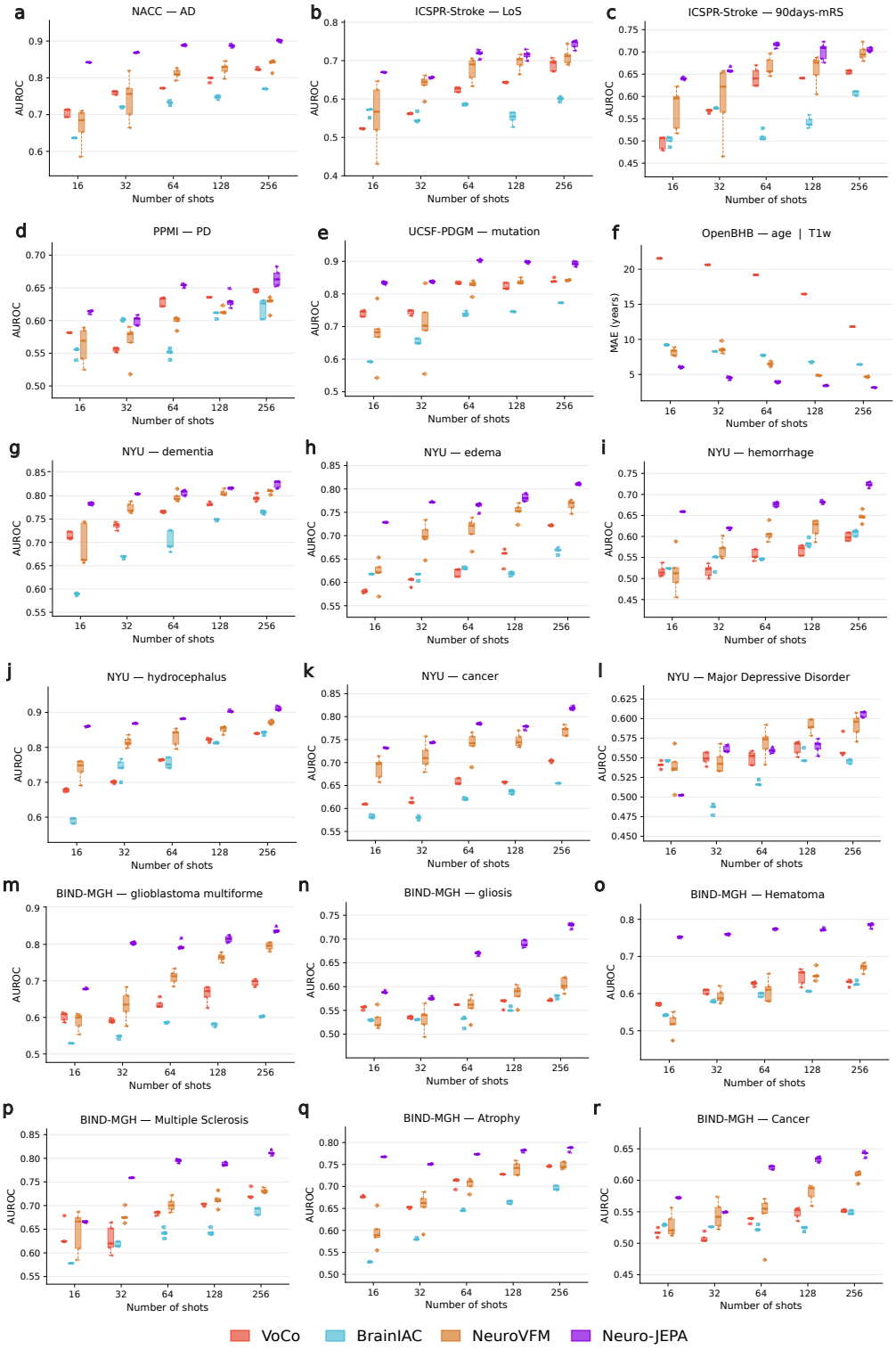
H.2 Few-shot Analysis

H.2.1 AUROC and AUPRC on Few-shot Performance Averaged Across Modalities

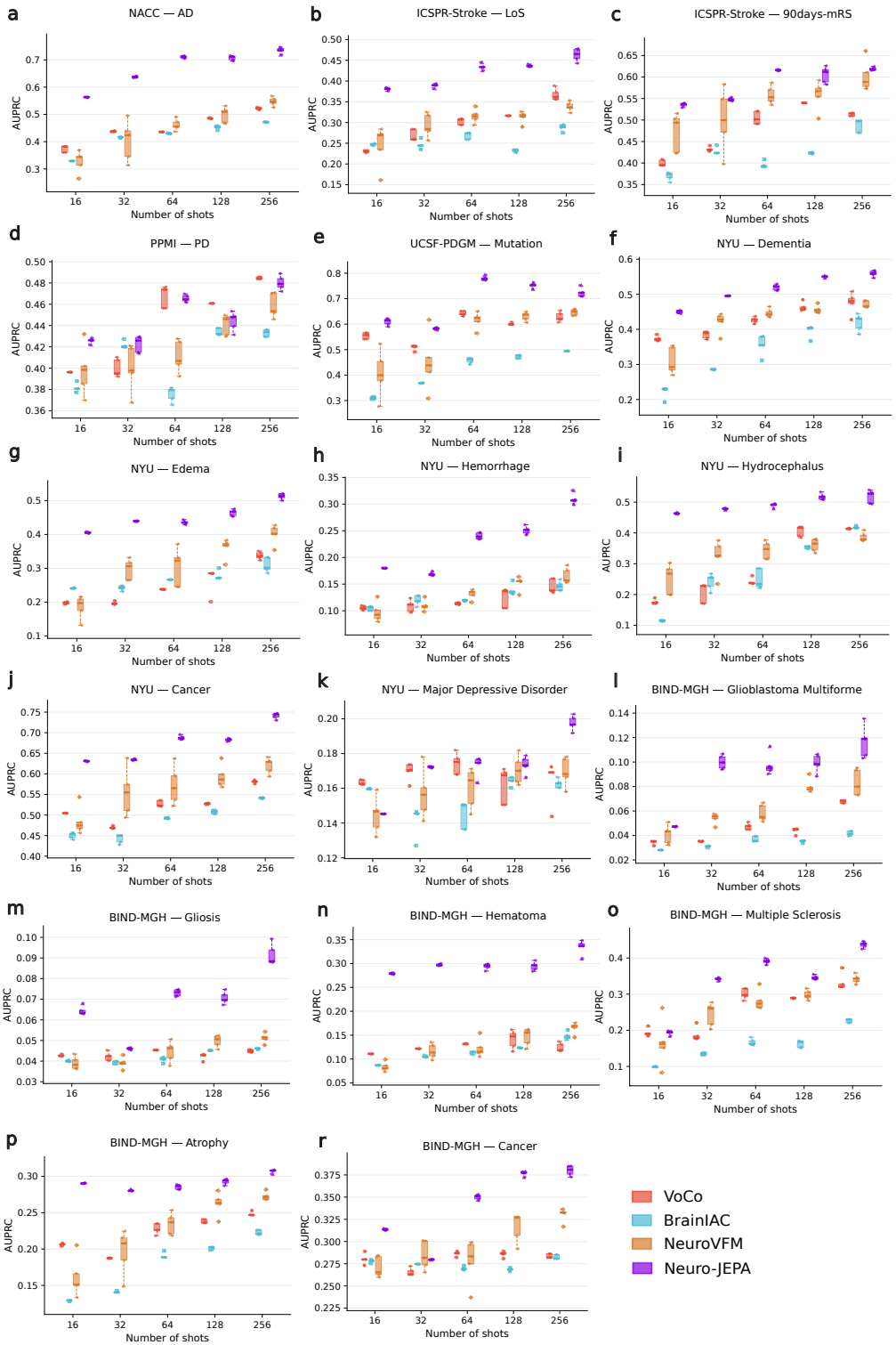
We report few-shot model performance on AUROC, AUPRC and MAE for age prediction regression averaged across differnt modalities on more diverse tasks in Supplementary Figures 37 and 38. Consistently with the finding in the main article Figure 5. We found consistently improved few-shot performance for Neuro-JEPA across majority of tasks on both AUROC and AUPRC.

H.2.2 Per-modality AUROC and AUPRC Few-short Performance

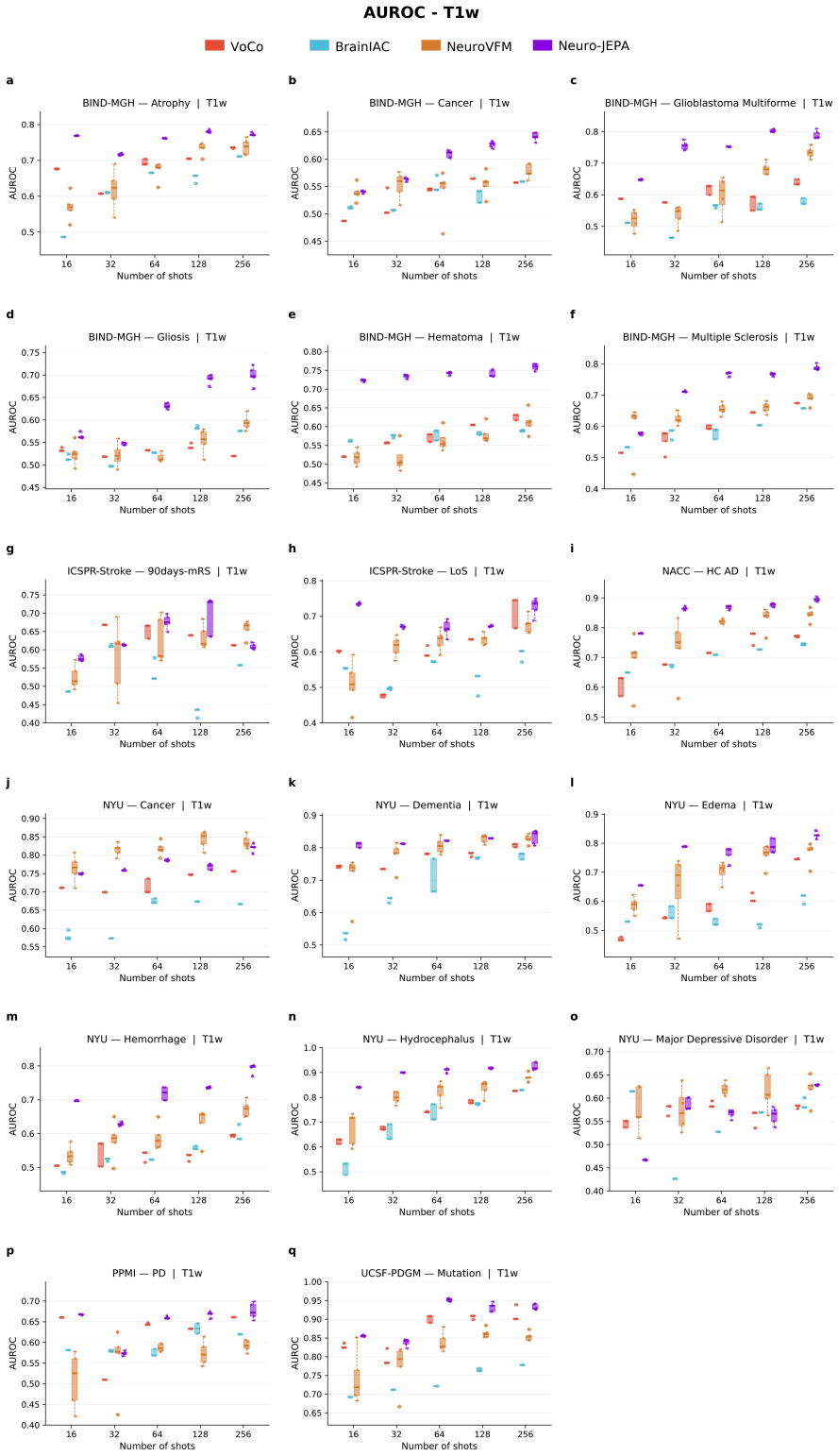
We report few-shot model performance on AUROC, AUPRC and MAE for age prediction regression per modality (T1w, T2w, FLAIR) in Supplementary Figures 39 to 44. Per modality performance is consistent with the averaged performance where Neuro-JEPA improved in comparison to other evaluated models.



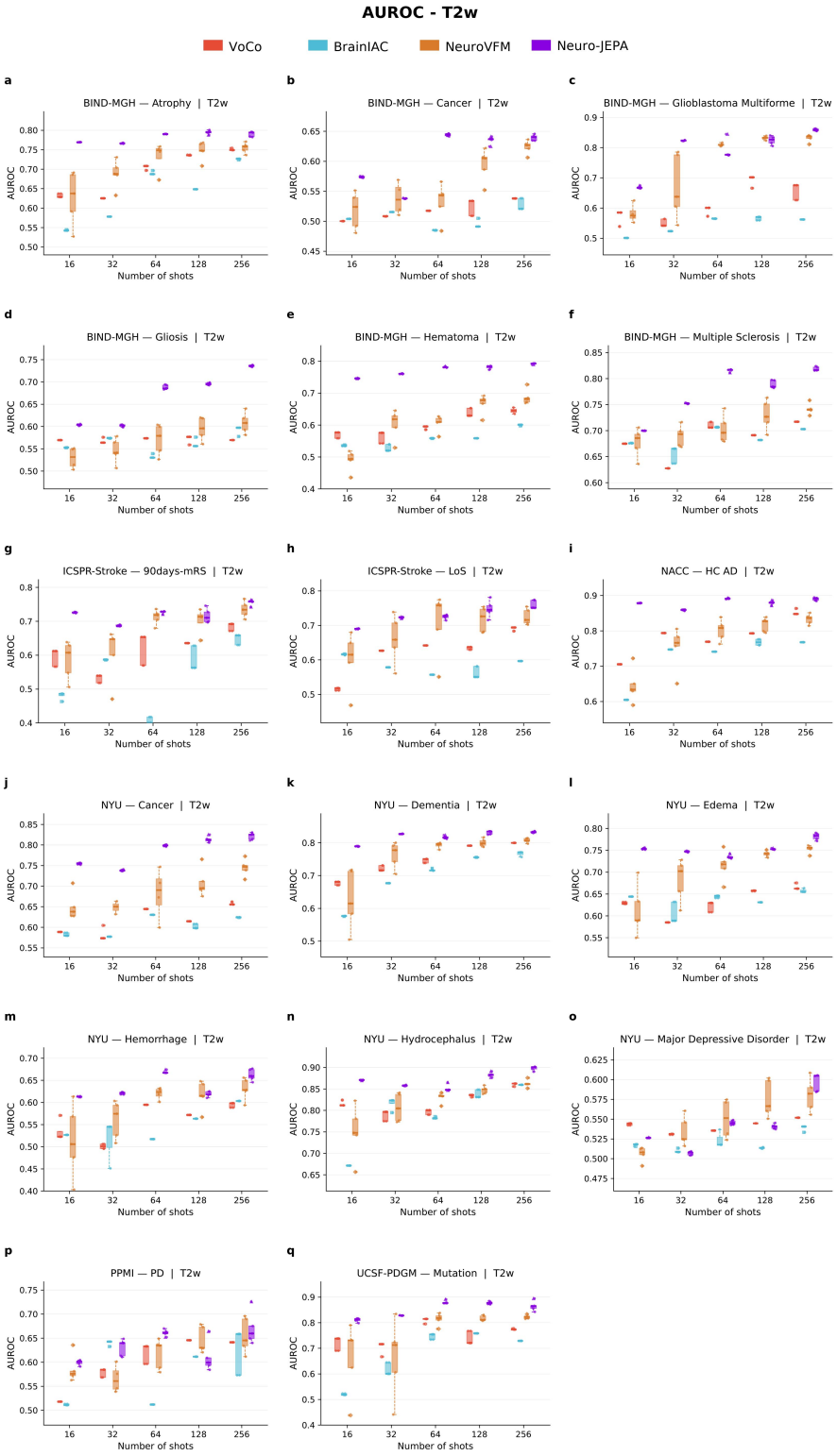
Supplementary Figure 37. Few-shot Analysis - we examine the evaluated models label efficiency when only $k = \{16, 32, 64, 128, 256\}$ positive samples are provided on more diverse selected tasks. The performance in reported in AUROC for classification and MAE for regression. **a-f,** Few-shot performance on selected tasks from public datasets. All result is reported as averaged performance across all available modalities for each task **g-l,** Few-shot performance on selected tasks from NYU-Langone dataset. **m-r,** Few-shot performance on selected tasks from BIND-MGH dataset.



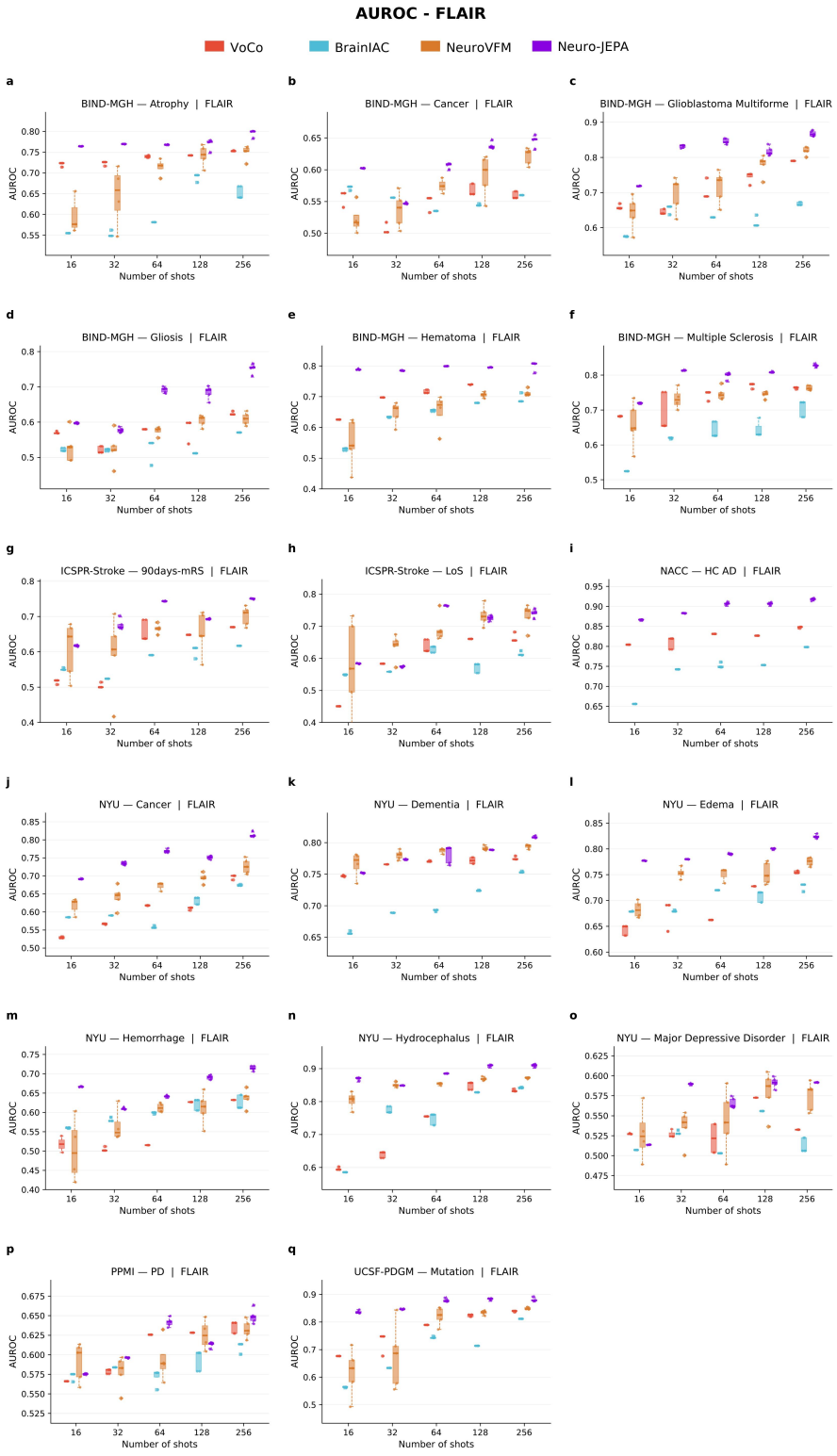
Supplementary Figure 38. Few-shot Analysis - we examine the evaluated models label efficiency when only $k = \{16, 32, 64, 128, 256\}$ positive samples are provided on more diverse selected tasks. The performance in reported in AUPRC for classification. **a-e,** Few-shot performance on selected tasks from public datasets. All result is reported as averaged performance across all available modalities for each task **f-k,** Few-shot performance on selected tasks from NYU-Langone dataset. **l-r,** Few-shot performance on selected tasks from BIND-MGH dataset.



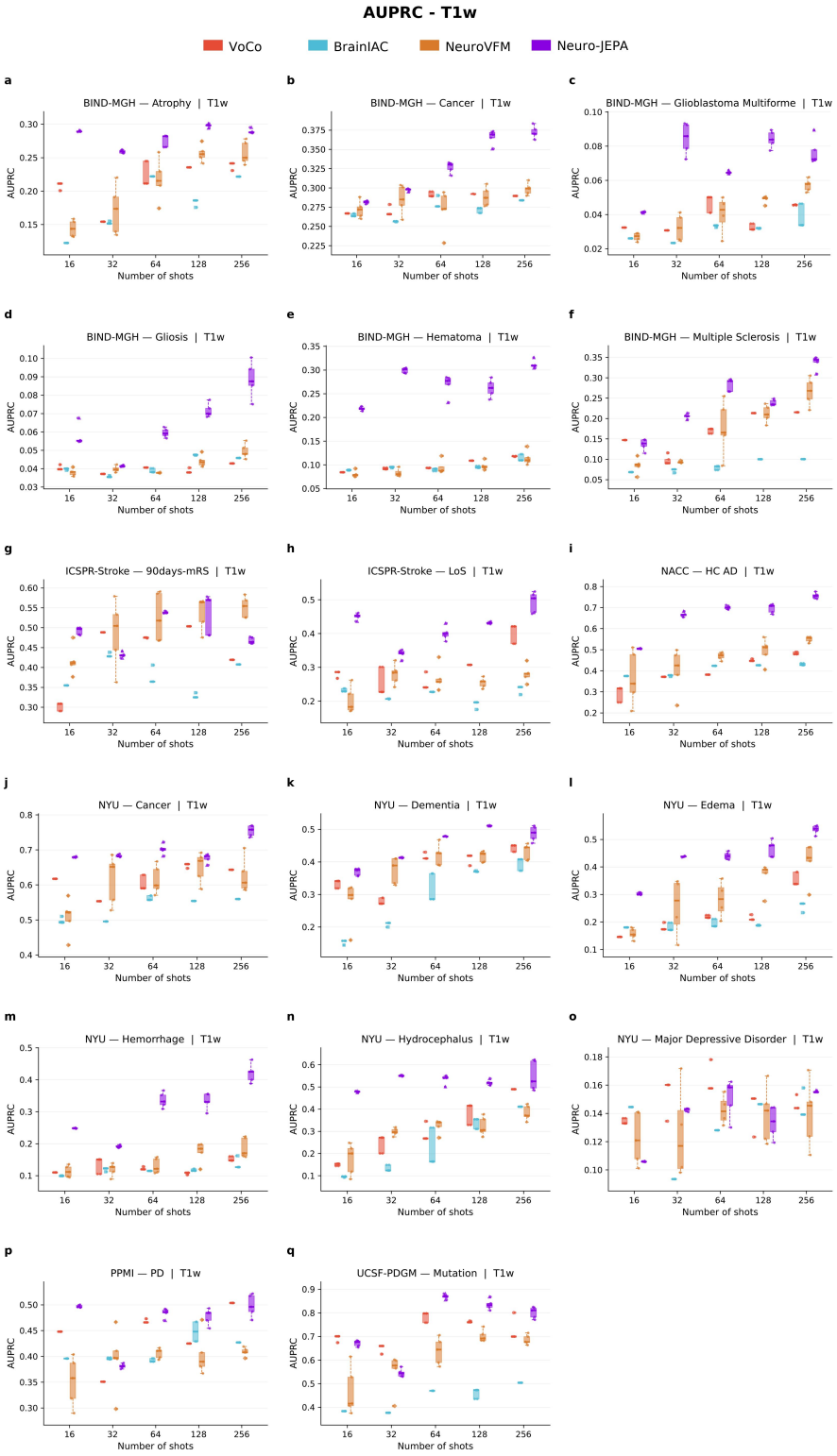
Supplementary Figure 39. AUROC Few-shot Analysis on T1w - few-shot performance across tasks for T1w.



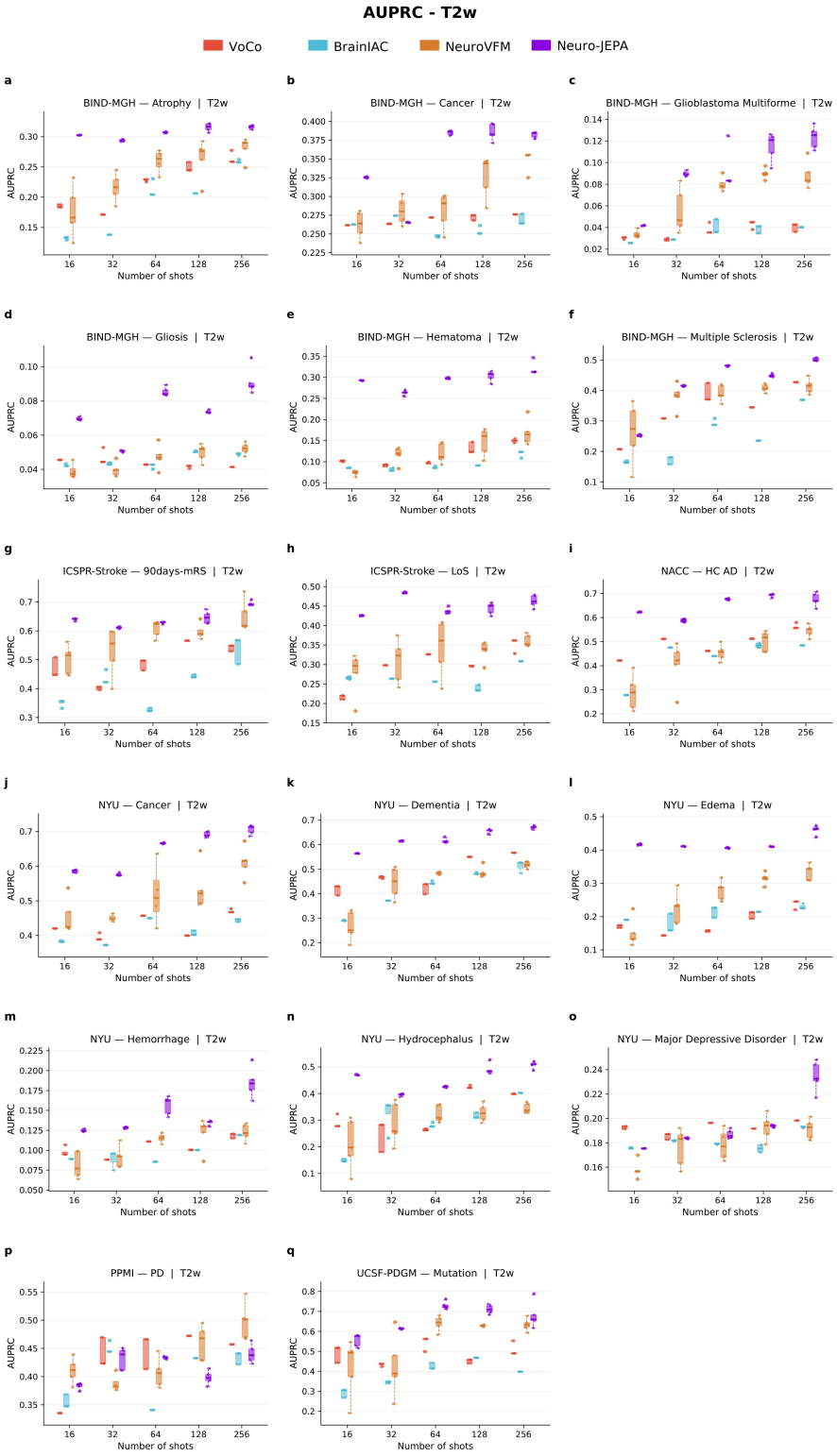
Supplementary Figure 40. AUROC Few-shot Analysis on T2w - few-shot performance across tasks for T2w.



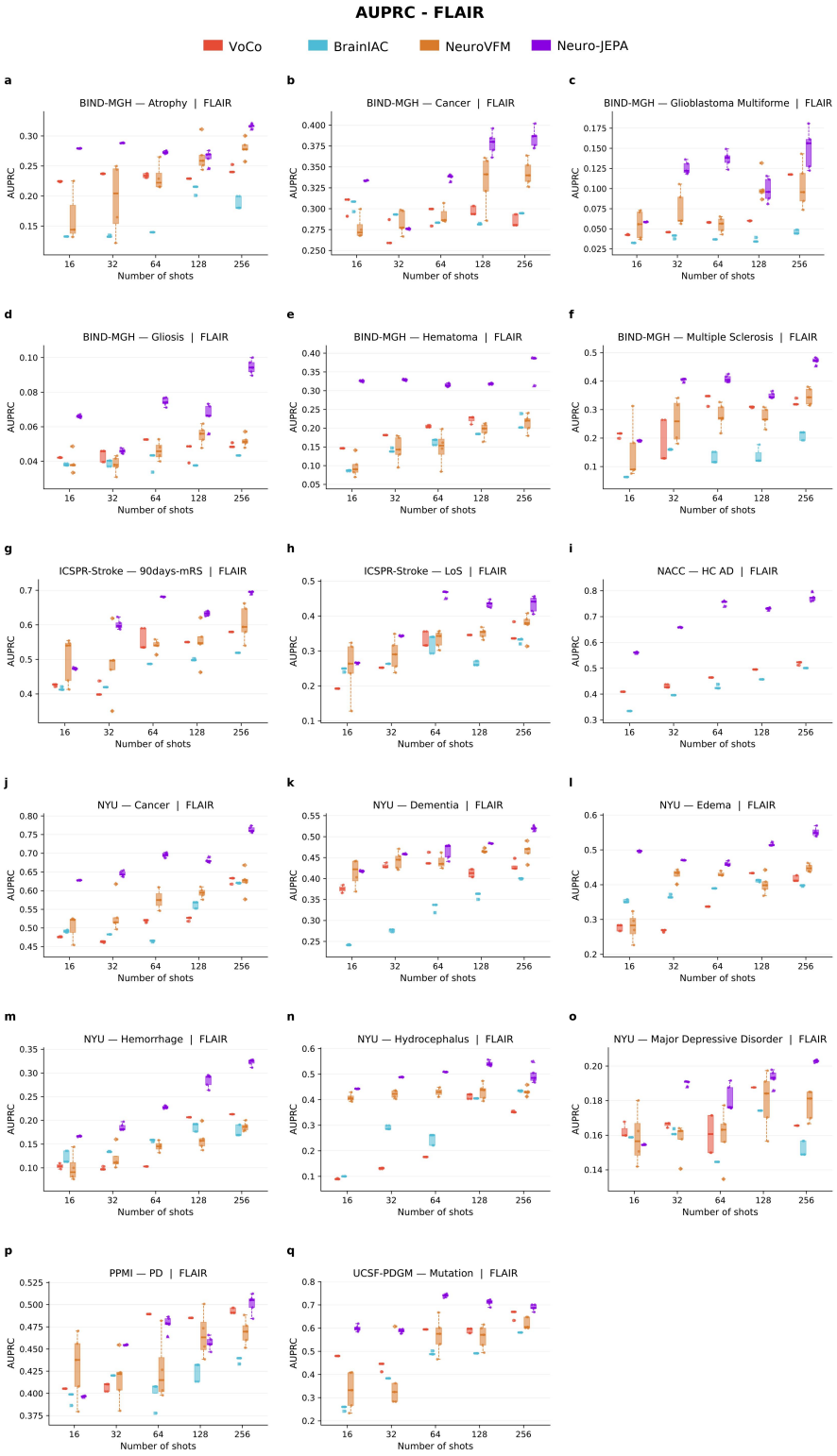
Supplementary Figure 41. AUROC Few-shot Analysis on FLAIR - few-shot performance across tasks for FLAIR.



Supplementary Figure 42. AUPRC Few-shot Analysis on T1w - few-shot performance across tasks for T1w.



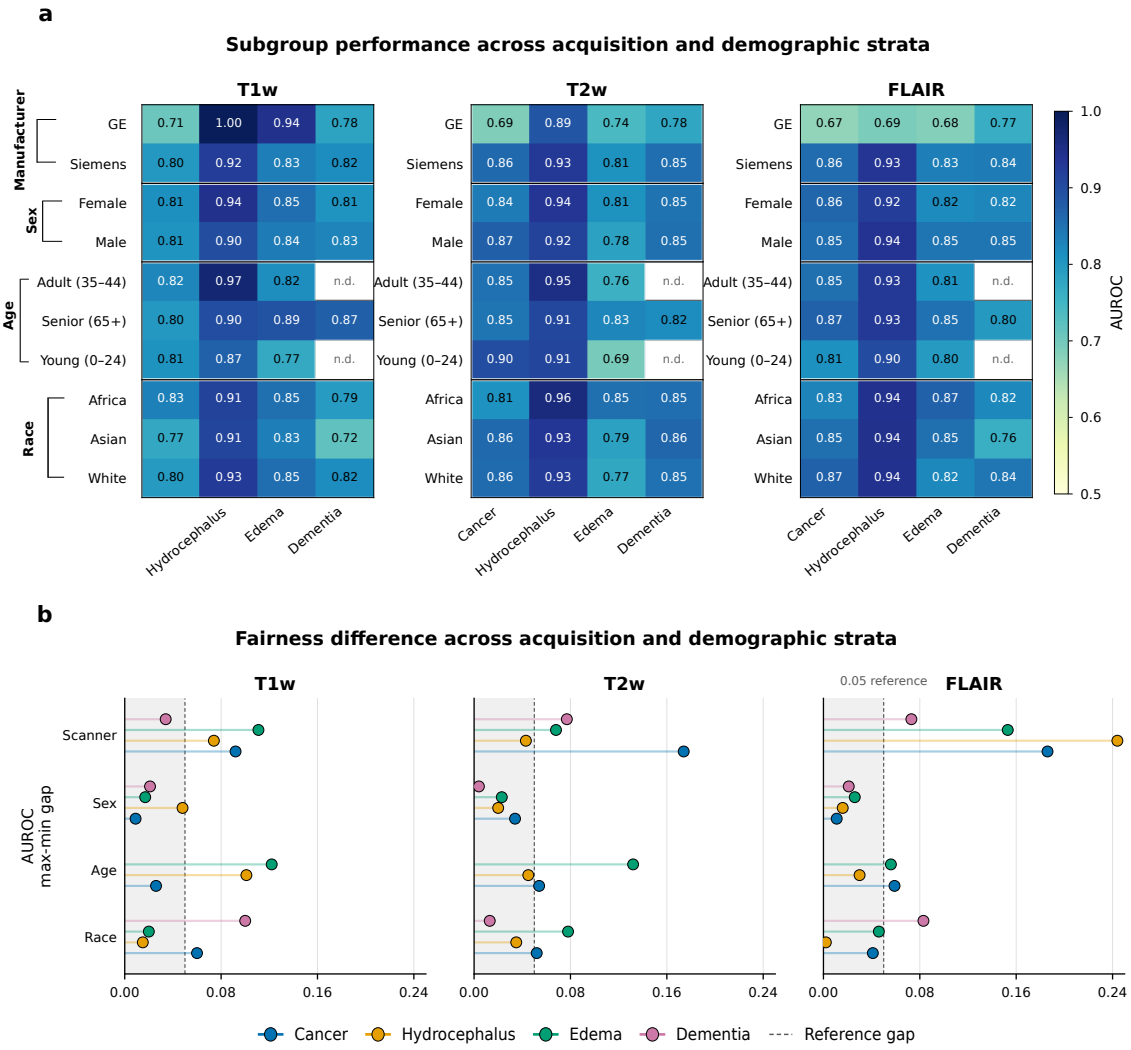
Supplementary Figure 43. AUPRC Few-shot Analysis on T2w - few-shot performance across tasks for T2w.



Supplementary Figure 44. AUPRC Few-shot Analysis on FLAIR - few-shot performance across tasks for FLAIR.

H.3 Fairness Analysis

We performed a fairness analysis on the curated evaluation dataset across subgroups defined by age, sex, race and scanner manufacturer (Supplementary Figure 45). For each subgroup, we report AUROC and quantify the fairness gap as the difference between the maximum and minimum AUROC across subgroup categories. Overall, performance was broadly consistent across subgroups, with limited evidence of large disparities in most settings. The main exceptions were scanner-associated differences for T2w and FLAIR scans. The pattern is consistent with the underlying data distribution: Siemens scanners were predominant in the pretraining cohort. Thus, the observed subgroup gaps likely reflect realistic sources of distributional heterogeneity rather than systematic model failure.

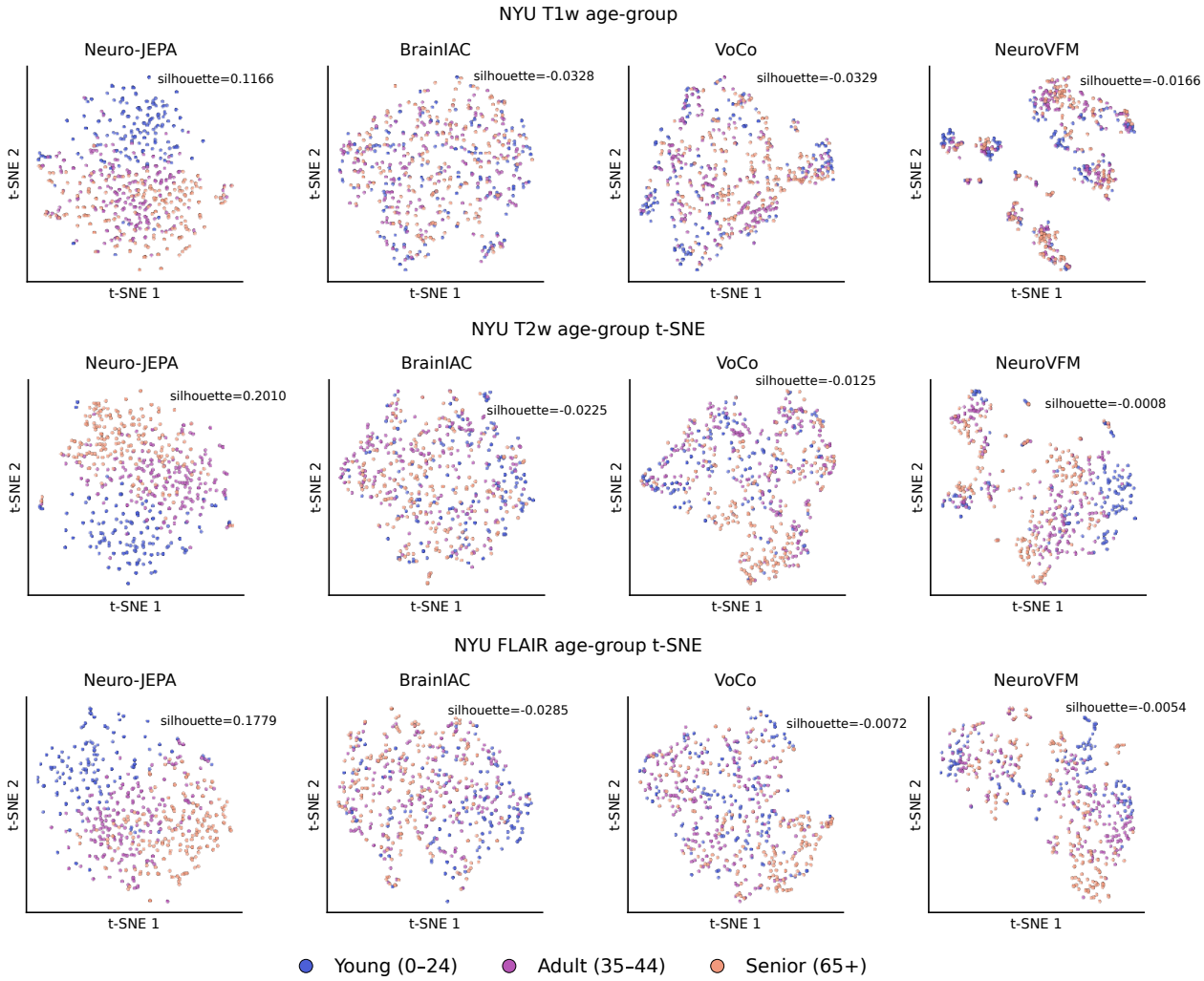


Supplementary Figure 45. Fairness Analysis across sub-cohorts. Fairness comparison on representative diseases (Cancer, Hydrocephalus, Edema, Dementia) from NYU-Langone dataset across different sub-cohorts by attentive probing. **a,** AUROC on difference sub-cohorts and diseases. **b,** Fairness gap on maximum AUROC minus minimum AUROC for each sub-cohort and disease

H.4 TSNE Visualization on Age Groups

We evaluated the separability of age-related representations across pretrained models using t-SNE projections of pooled embeddings, obtained by averaging token representations without further fine-tuning. Patients were grouped into three age cohorts: Young (0–24 years), Adult (35–44 years), and Senior (65+ years). The analysis was performed on a subset of NYU Langone patients with available age records who were excluded from the pretraining dataset. For each modality, 600 samples were analyzed, with 200 samples per

age cohort. Results are shown in Supplementary Fig. Figure 46. Both visualization and quantitative silhouette scores indicate that Neuro-JEPA achieves the best age-group separability among the evaluated models.



Supplementary Figure 46. Age TSNE Across Different Pretrained Models. The plot present TSNE visualization on different age subgroups (Young (0-24), Adult (35-44) and Senior (65+)) and modalities (T1w, T2w, T2-FLAIR) for NYU Langone Dataset. The silhouette scores and visualization present that Neuro-JEPA shows best separation on age subgroups.

I Additional Result for MoE Routing Analysis and Visualization

I.1 MoE Foreground vs. Background Routing

We present result for different model size foreground vs. background routing on NYU-Langone dataset with ViT-Base-MoE and ViT-Large-MoE on all layers in Supplementary Section I.5 and figure 47. The routing distribution shows that experts are able to separate between foreground and background especially on early layers where all foreground/background tokens routed to different experts.

I.2 MoE Different Modalities Routing

We present result for different model size MoE routing on different modalities in Supplementary Figures 48 and 49. While a majority of experts have balanced routing across modalities, few experts present different routing behaviors across modalities such as Expert 16 on Layer 5 and Expert 7 on Layer 7. Given that different neuroimaging modalities can have very similar anatomies with shared information, this indicate MoE learns when to deviate from shared computation, assigning only a subset of experts to modality-sensitive contrast information.

I.3 MoE Heatmap

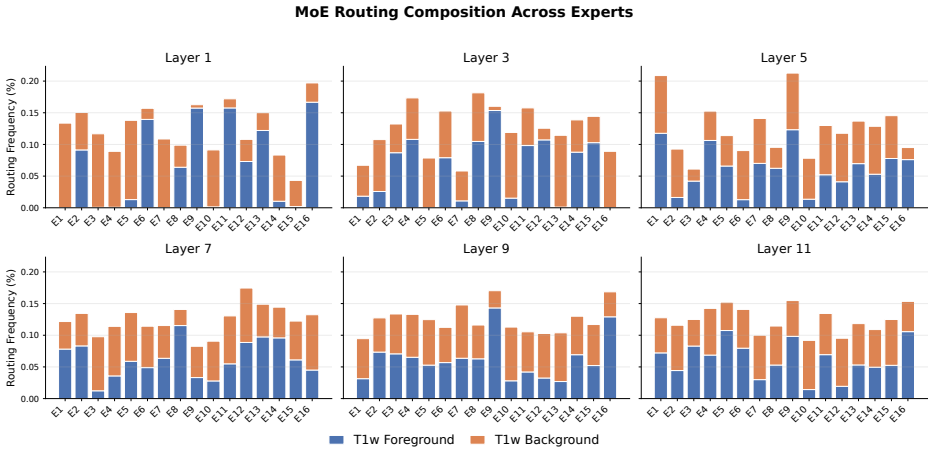
We present heatmap result for different model size MoE routing on examining if different experts assign different distribution on different tokens in Supplementary Figures 50 to 55. The result shows different experts can focus on different tokens across layers, indicating MoE learns on routing based on different anatomy structures.

I.4 MoE Visualization

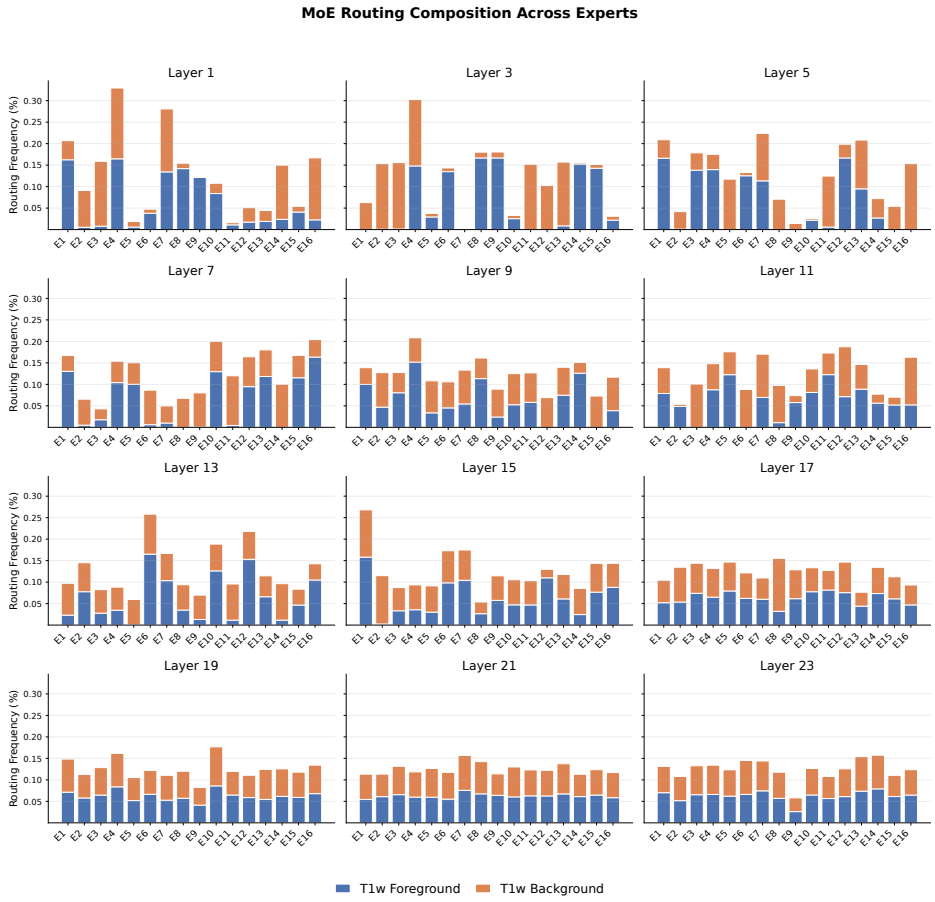
Supplementary Figures 57 to 60 present different patch size MoE routing visualization on selected slices on NYU-Langone and BIND-MGH datasets with T1w and T2w template. The result shows that different experts focus on different anatomical structure for MoE. Additionally, both NYU-Langone and BIND-MGH present very similar routing behaviors, indicating the revealed anatomical structure focus is not coincident.

I.5 Attention Map Visualization

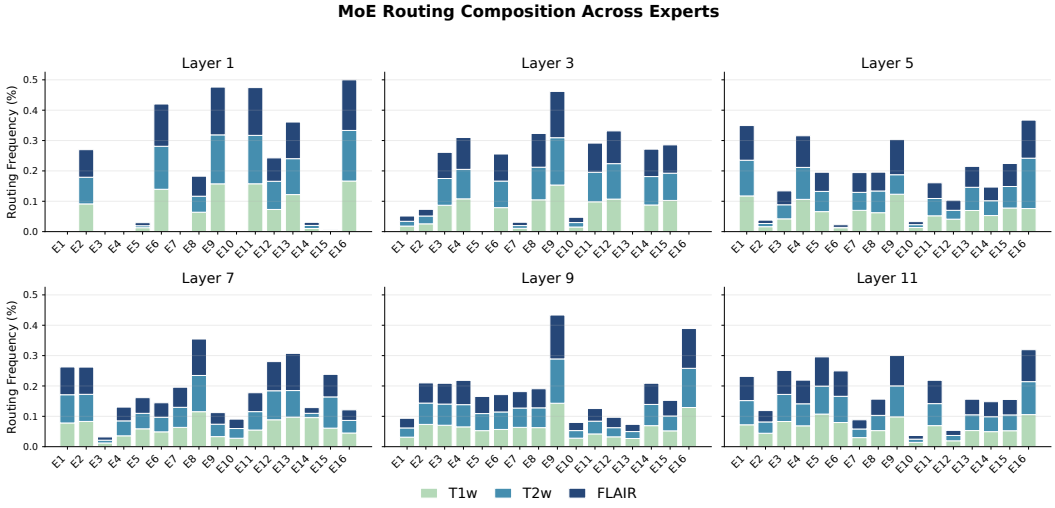
Extended Fig 8-10 present attention map visualization for stroke, multiple sclerosis and gliobastoma lesions. The visualization presents that the pretrained model is able to attend to different pathology without further adapting.



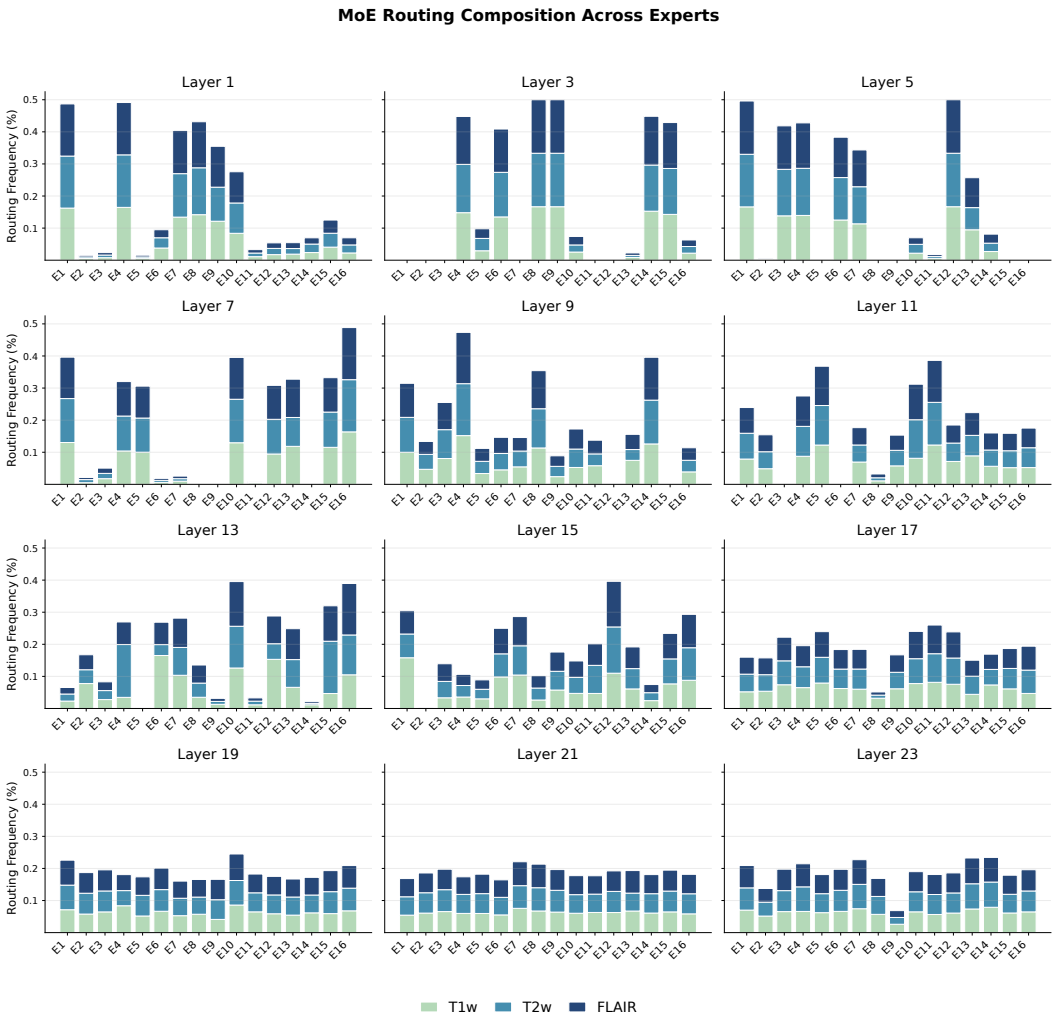
Supplementary Figure 47. Supplementary MoE Routing FG vs BG – T1w NYU - ViT-B



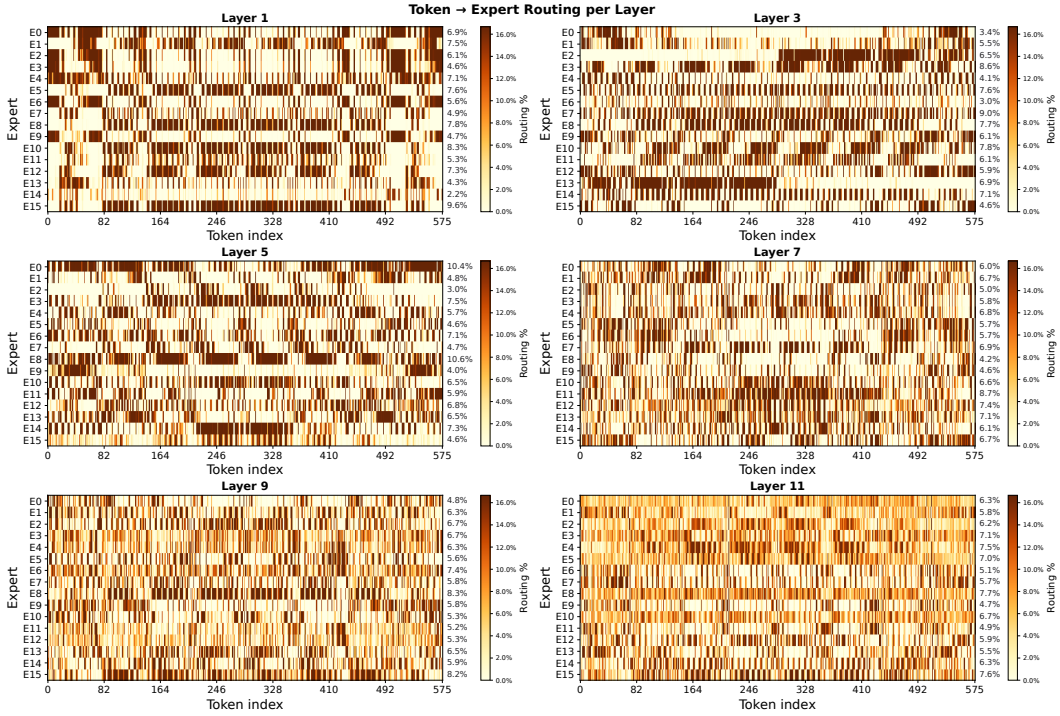
Supplementary Figure 48. Supplementary MoE Routing FG vs BG – T1w MGH - ViT-L



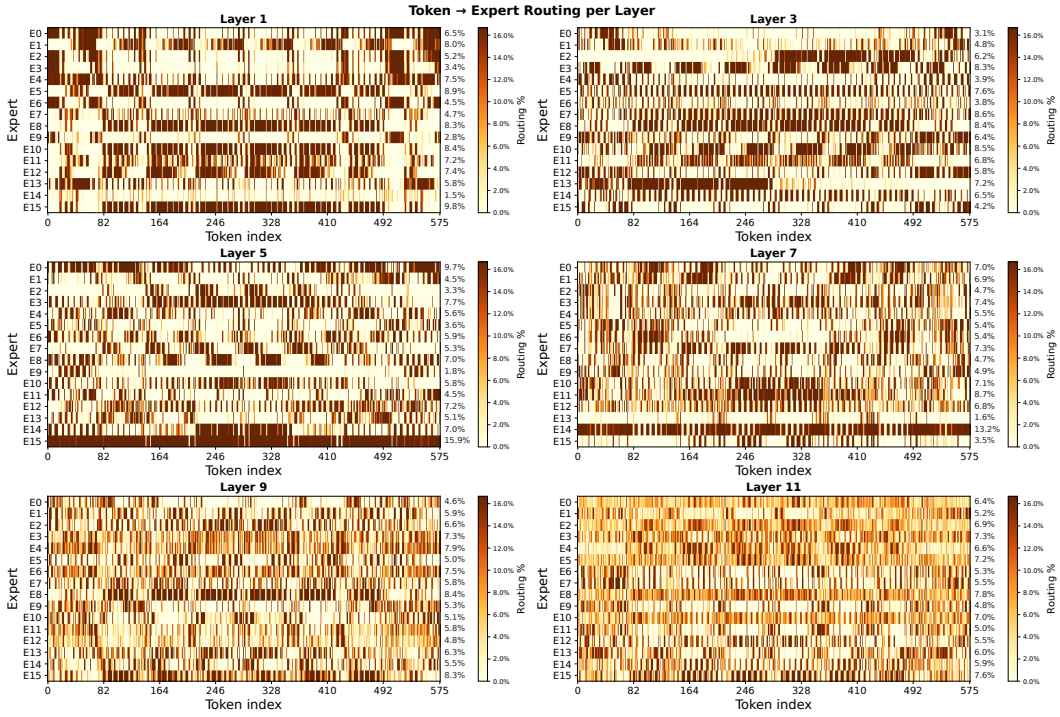
Supplementary Figure 49. Supplementary MoE Routing on Different Modalities – NYU - ViT-B



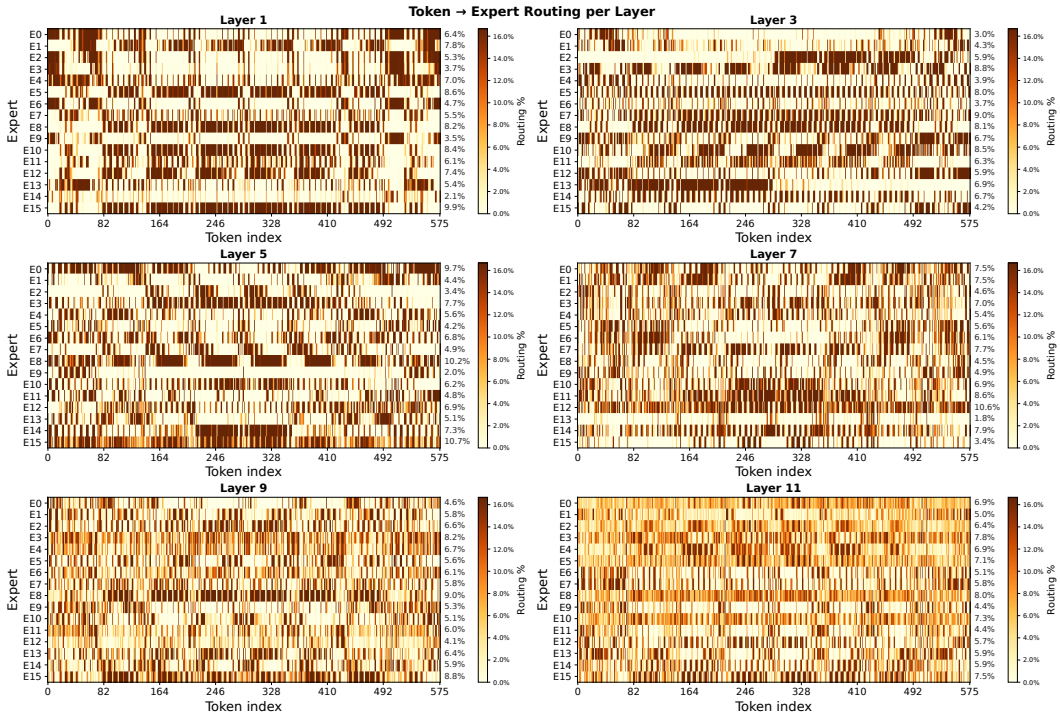
Supplementary Figure 50. Supplementary MoE Routing on Different Modalities – NYU - ViT-L



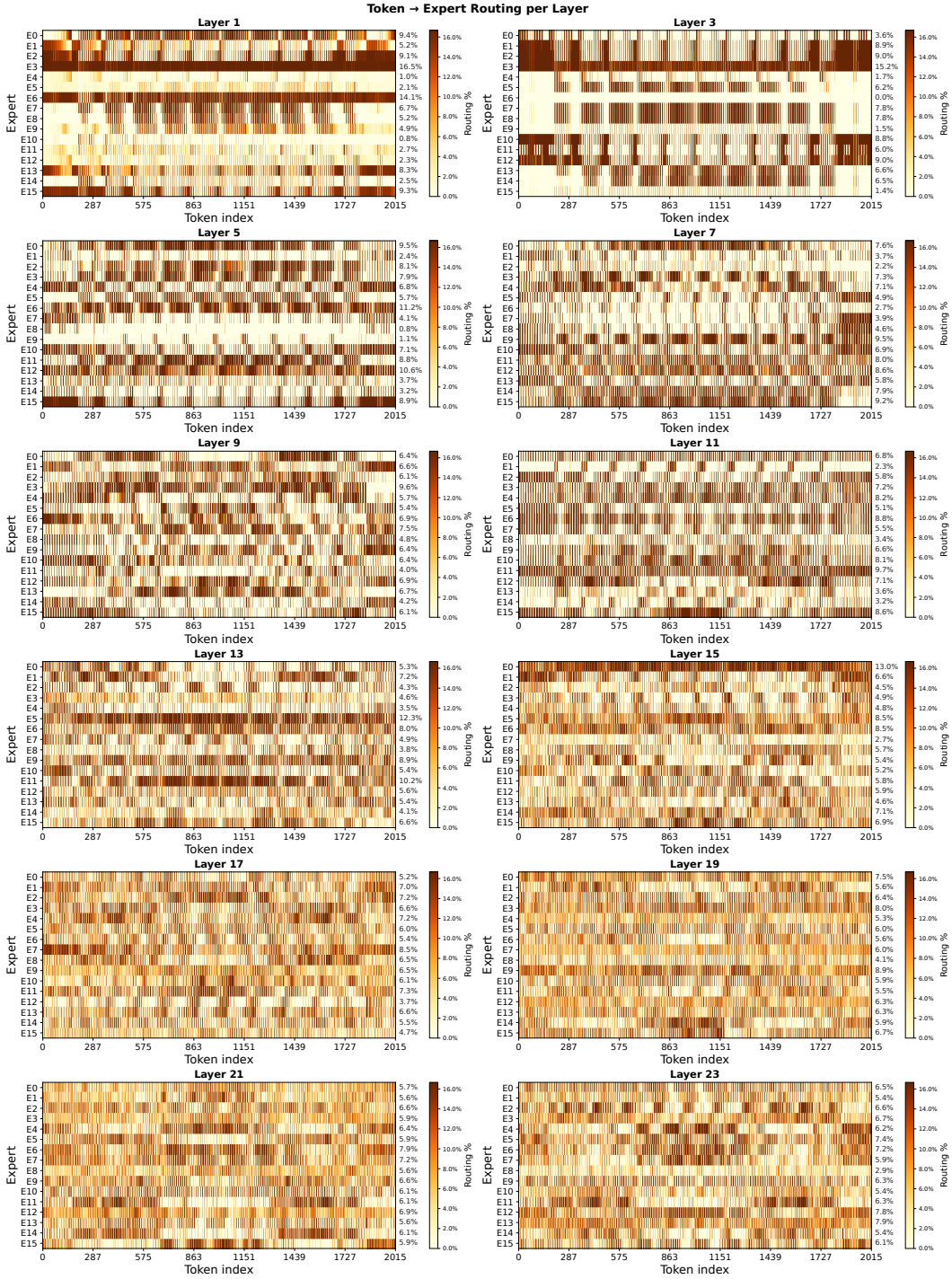
Supplementary Figure 51. Supplementary MoE Routing Heatmaps – NYU T1w - ViT-B



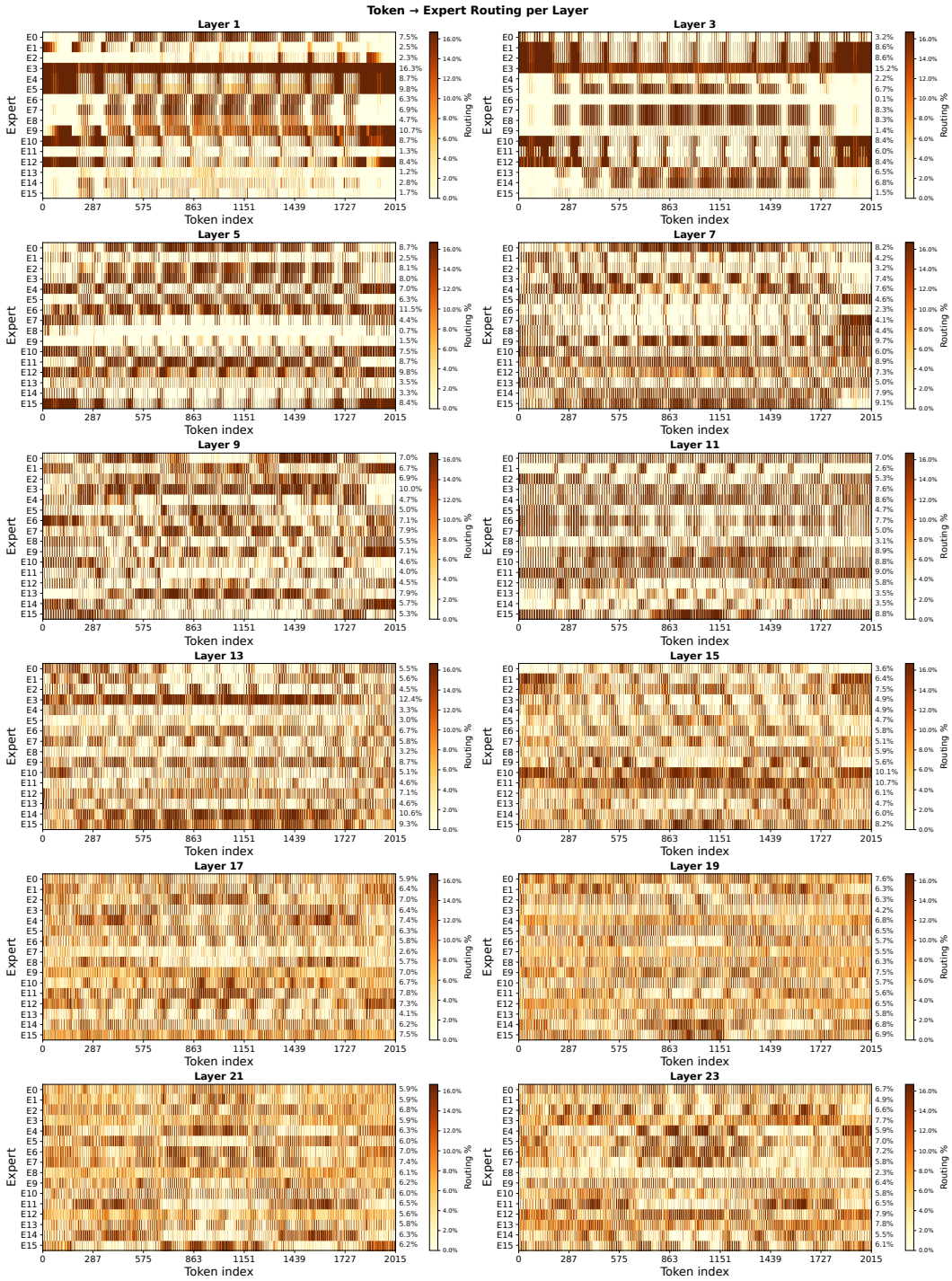
Supplementary Figure 52. Supplementary MoE Routing Heatmaps – NYU T2w - ViT-B



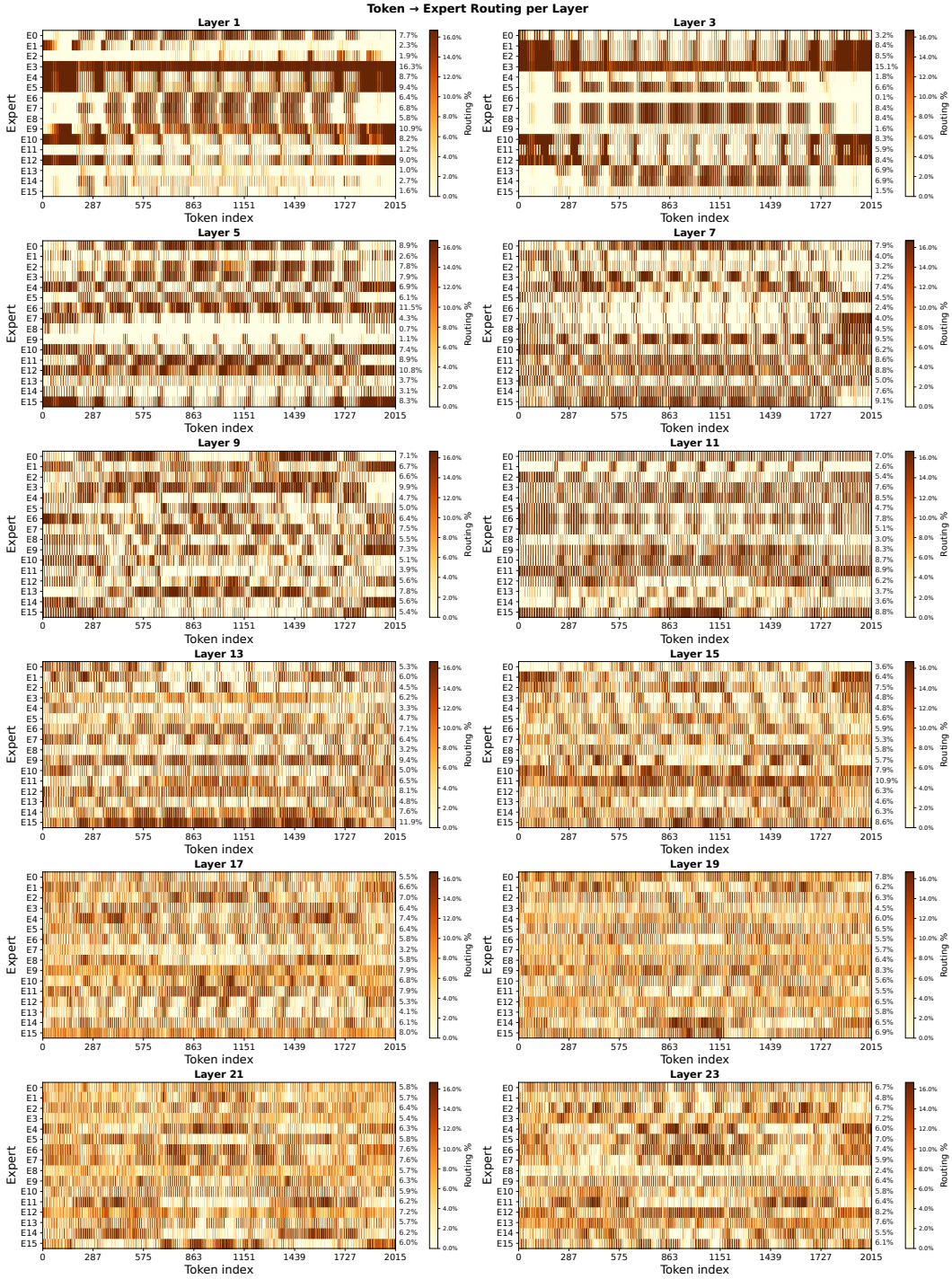
Supplementary Figure 53. Supplementary MoE Routing Heatmaps – NYU FLAIR - ViT-B



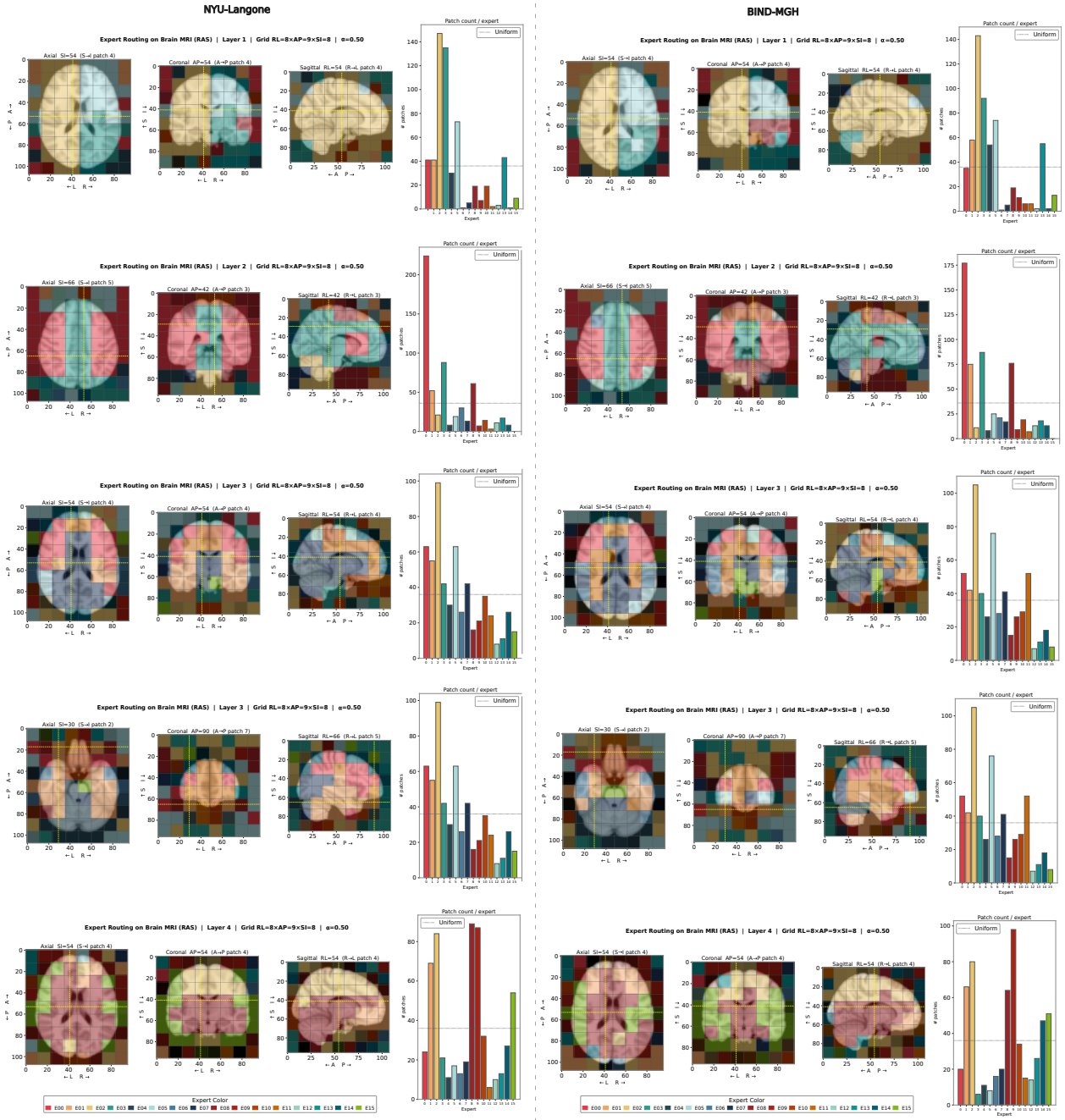
Supplementary Figure 54. Supplementary MoE Routing Heatmaps – NYU T1w - ViT-L



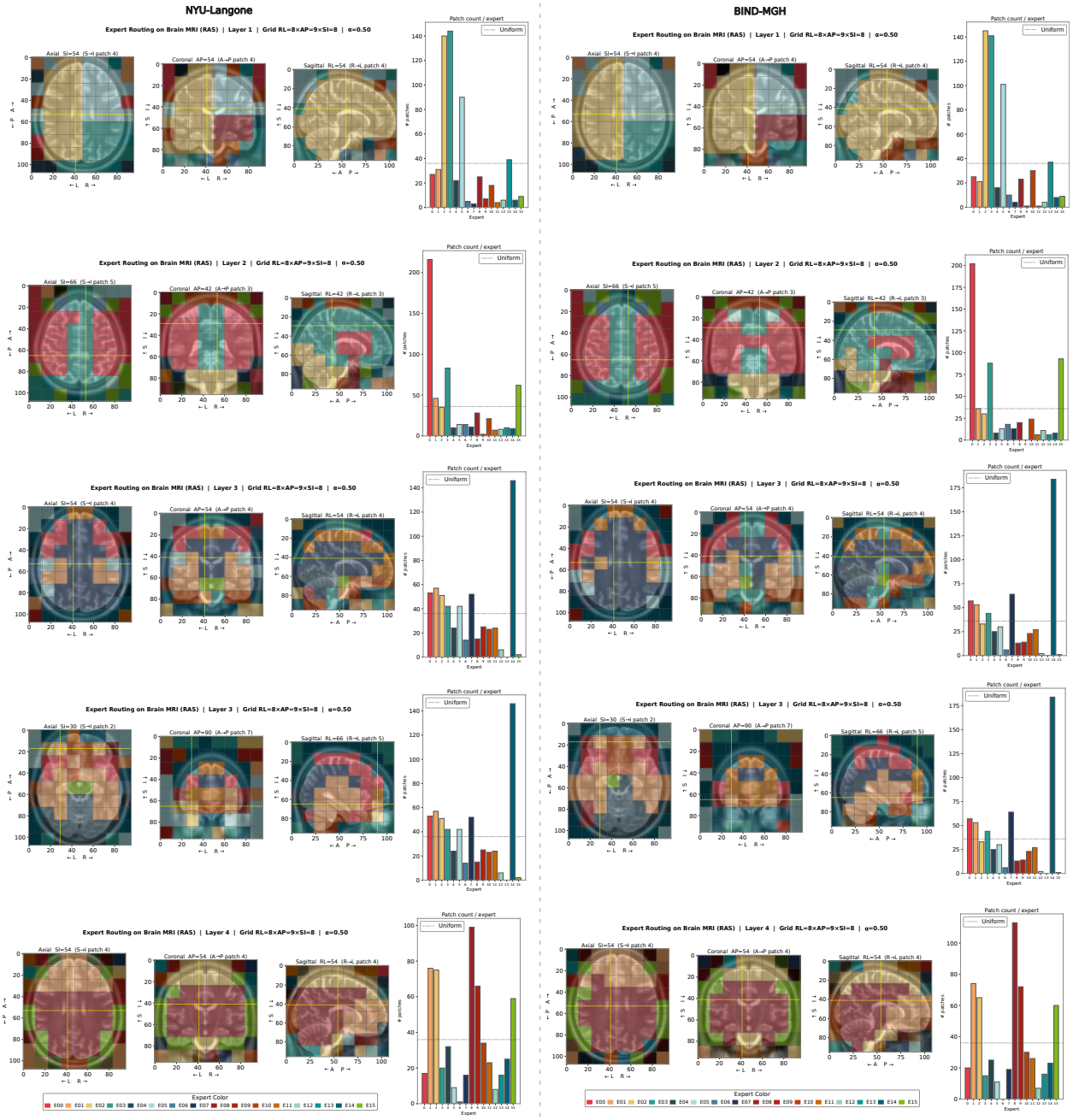
Supplementary Figure 55. Supplementary MoE Routing Heatmaps – NYU T2w - ViT-L



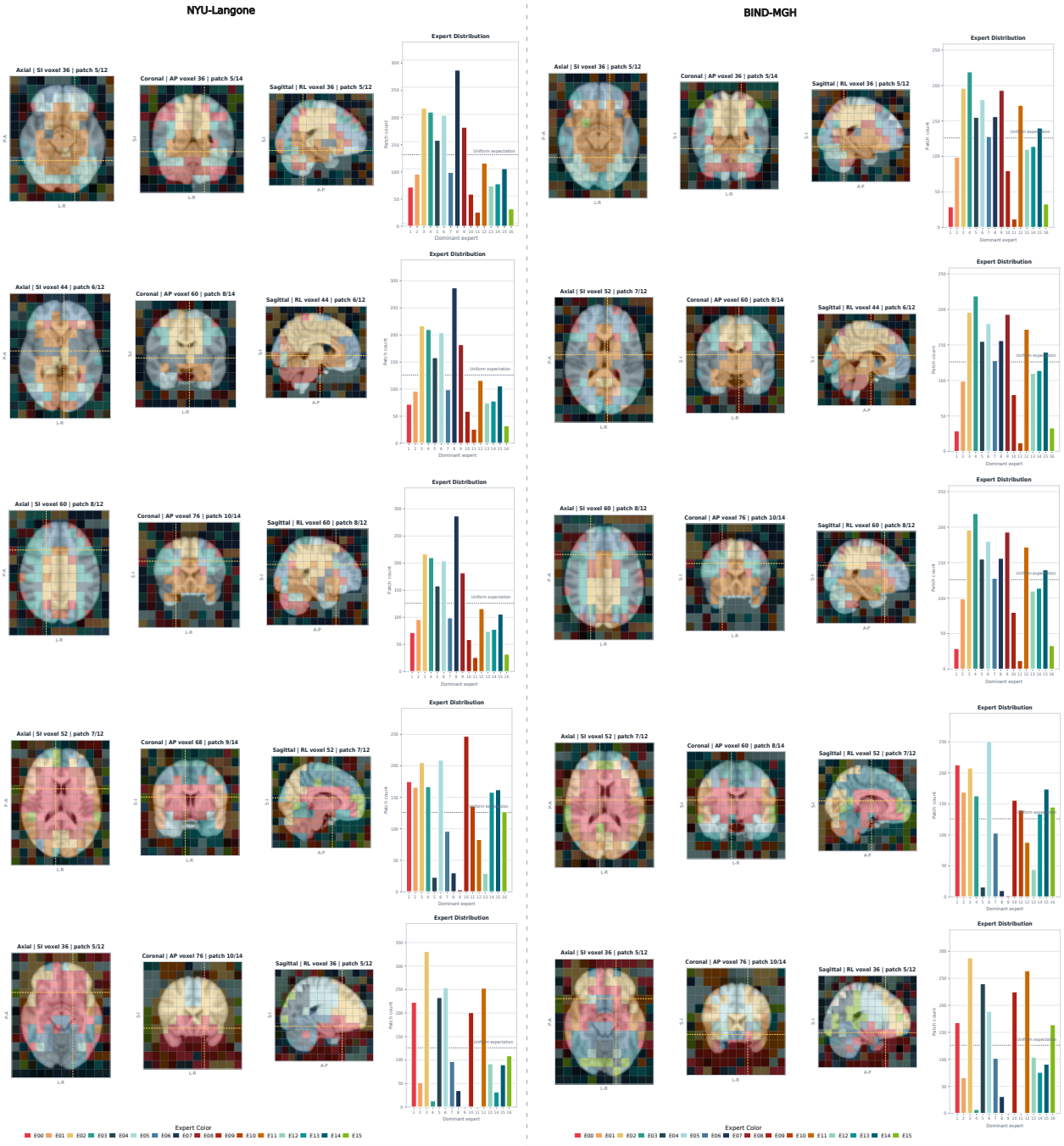
Supplementary Figure 56. Supplementary MoE Routing Heatmaps – NYU FLAIR - ViT-L



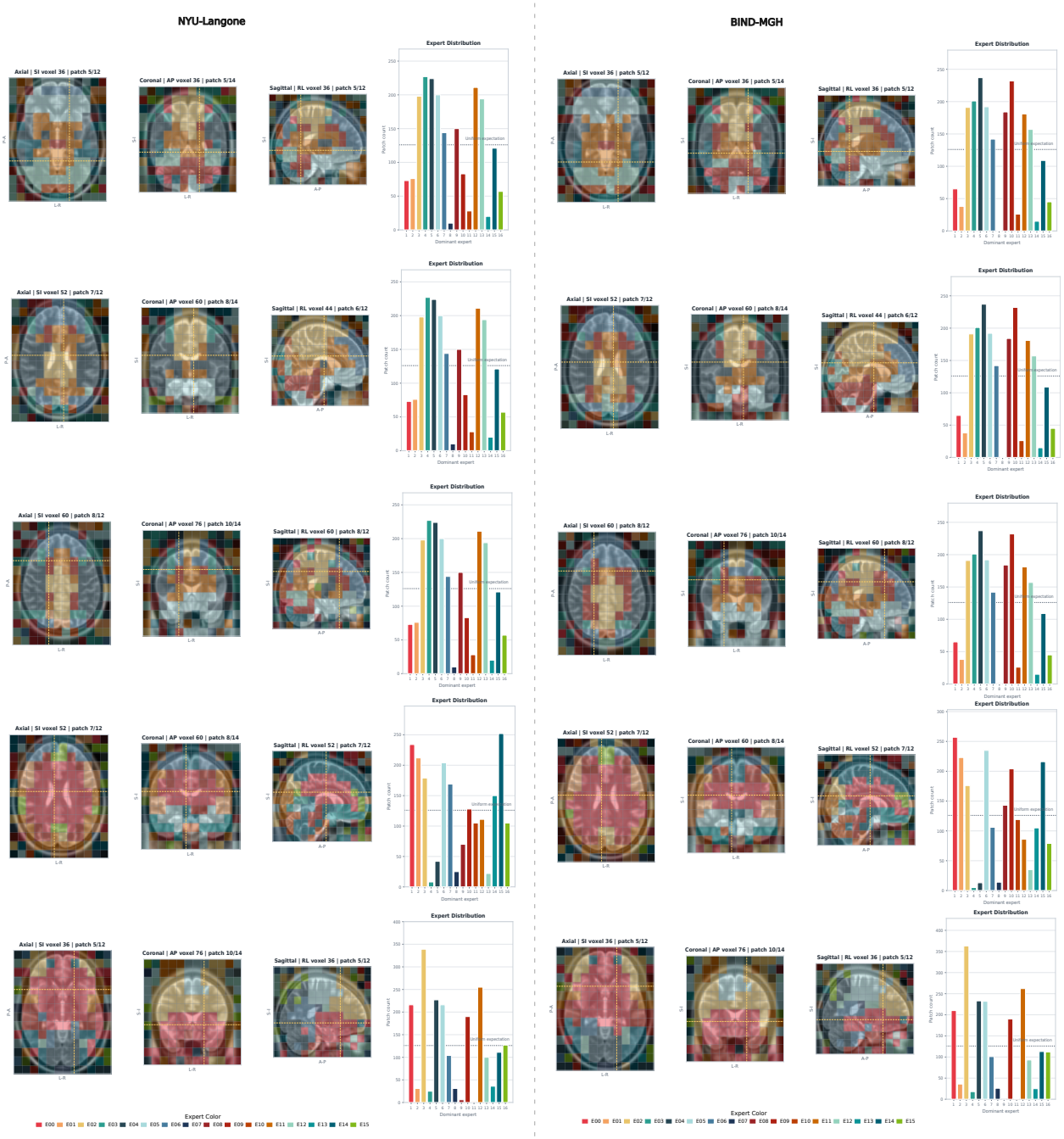
Supplementary Figure 57. MoE Visualization on T1w template - Patch Size 12.



Supplementary Figure 58. MoE Visualization on T2w template - Patch Size 12.



Supplementary Figure 59. MoE Visualization on T1w template - Patch Size 8.



Supplementary Figure 60. MoE Visualization on T2w template - Patch Size 8.

References

1. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Mach. Learn. Res.* (2024). Featured Certification.
2. Siméoni, O. *et al.* Dinov3 (2025). 2508.10104.
3. Kondepudi, A. *et al.* Health system learning enables generalist neuroimaging models. *Nat. Medicine* DOI: 10.1038/s41591-026-04497-1 (2026).
4. Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**, 1771–1800, DOI: 10.1162/089976602760128018 (2002).
5. Madaan, D., Makino, T., Chopra, S. & Cho, K. Jointly modeling inter- & intra-modality dependencies for multi-modal learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).
6. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211, DOI: 10.1038/s41592-020-01008-z (2021).
7. Isensee, F. *et al.* nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In Linguraru, M. G. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 488–498 (Springer Nature Switzerland, Cham, 2024).
8. Ma, J. *et al.* Segment anything in medical images. *Nat. Commun.* **15**, 654, DOI: 10.1038/s41467-024-44824-z (2024).
9. Zhao, T. *et al.* A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nat. Methods* **22**, 166–176, DOI: 10.1038/s41592-024-02499-w (2025).
10. He, Y. *et al.* Vista3d: A unified segmentation foundation model for 3d medical imaging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20863–20873, DOI: 10.1109/CVPR52734.2025.01943 (2025).
11. Asadi, M. *et al.* Mirage: The illusion of visual understanding (2026). 2603.21687.
12. Madaan, D., Muhunthan, V., Cho, K. & Chopra, S. Multi-modal data spectrum: Multi-modal datasets are multi-dimensional. In *The Fourteenth International Conference on Learning Representations* (2026).
13. Gu, Y. *et al.* Evaluating the robustness and readiness of large frontier models in health ai applications. *Nat. Medicine* DOI: 10.1038/s41591-026-04501-8 (2026).
14. Assran, M. *et al.* V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* (2025).
15. Liu, S. *et al.* Generalizable deep learning model for early alzheimer's disease detection from structural mris. *Sci. Reports* **12**, 17106, DOI: 10.1038/s41598-022-20674-x (2022).