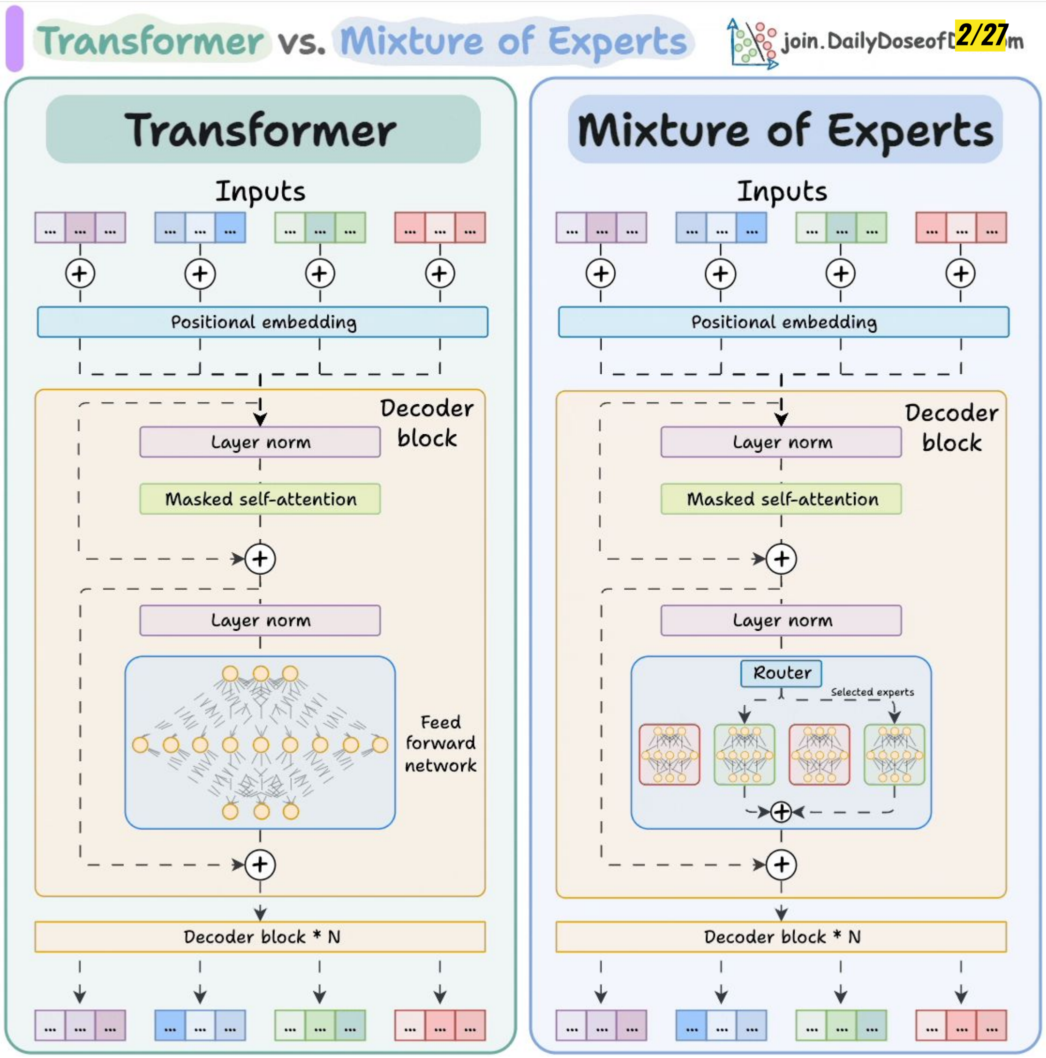
Flex-MoE: Modeling Arbitrary Modality Combination via the Flexible Mixture-of-Experts

Sukwon Yun[1], Inyoung Choi[2] , Jie Peng[3] , Yangfan Wu[3] , Jingxuan Bao[2] , Qiyiwen Zhang[2] , Jiayi Xin[2] , Qi Long[2] , Tianlong Chen[1]

[1]University of North Carolina at Chapel Hill
[2]University of Pennsylvania
[3]University of Science and Technology of China

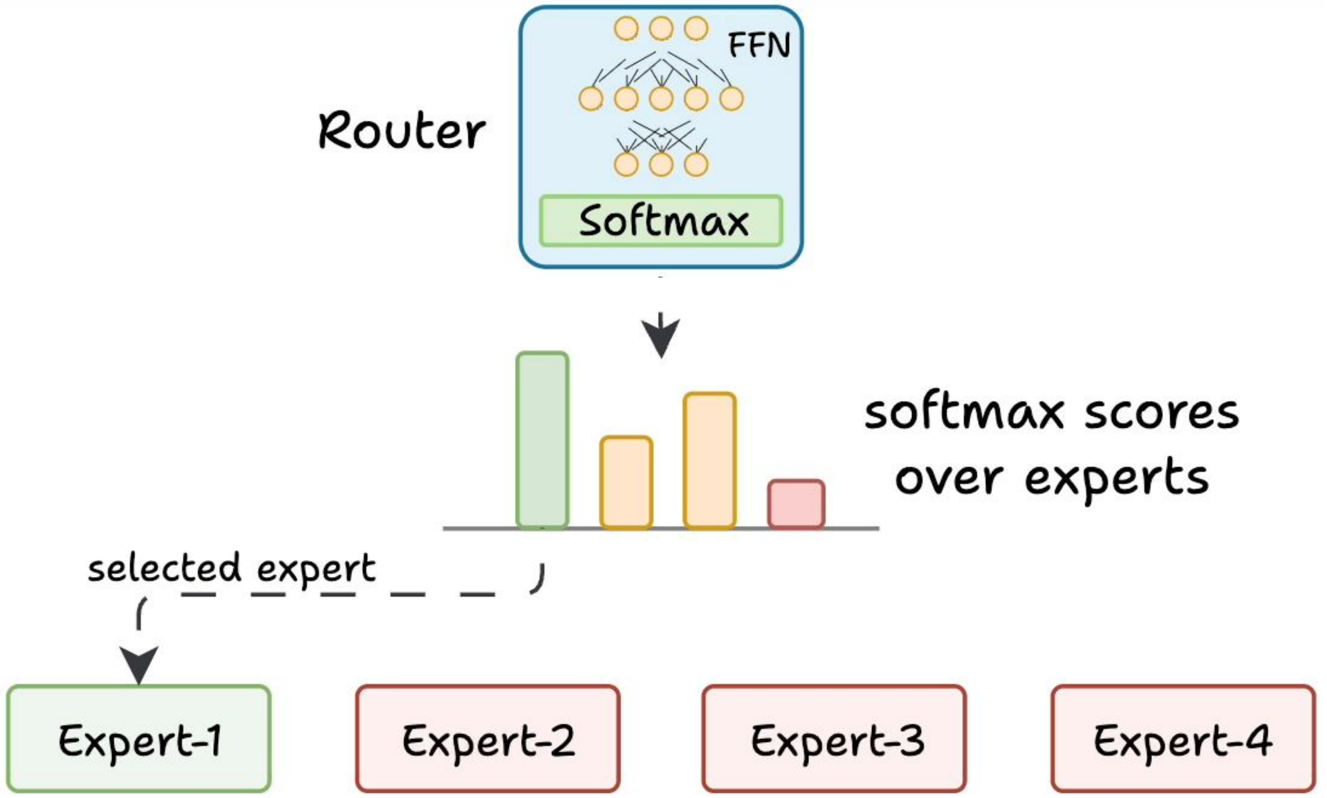
MoE vs Transformer

- **MoE** - model architecture where many specialized sub-models (experts) exist.
- **Transformer** - sequence modeling architecture that relies on self-attention mechanisms.
- MoE often applied to transformer architecture due to their success/scalability.



Img Src: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts

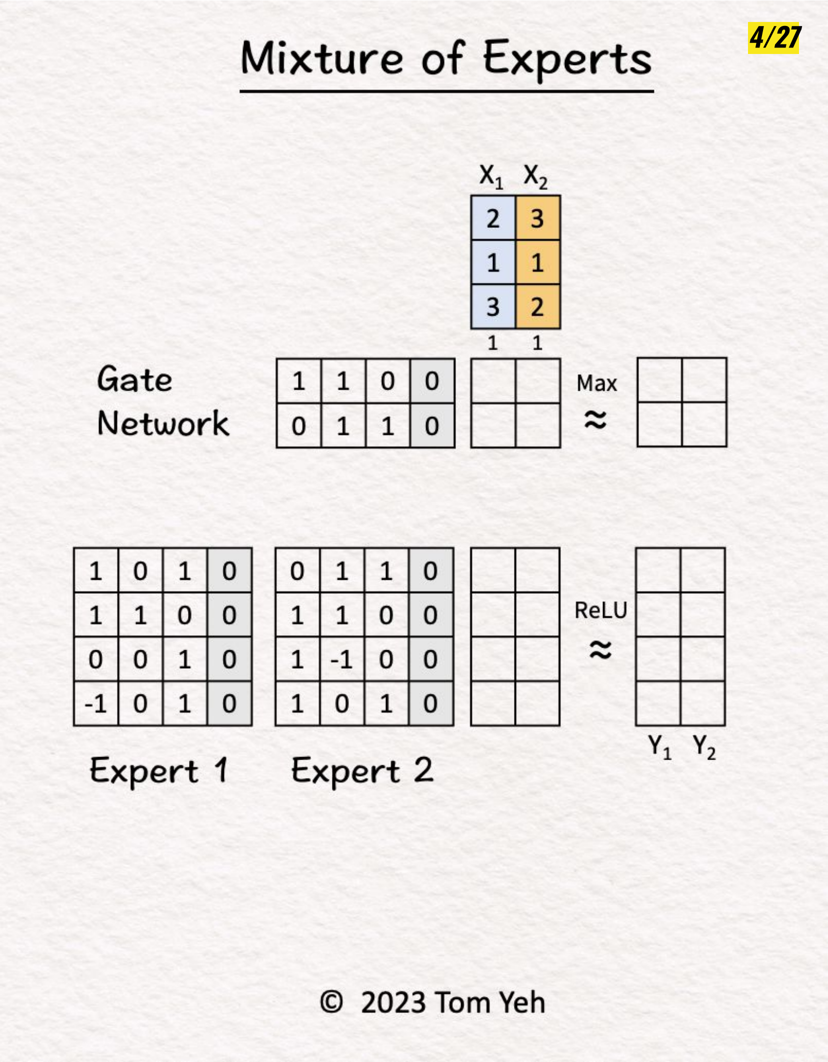
Routers (or Gates)



Img Src: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts

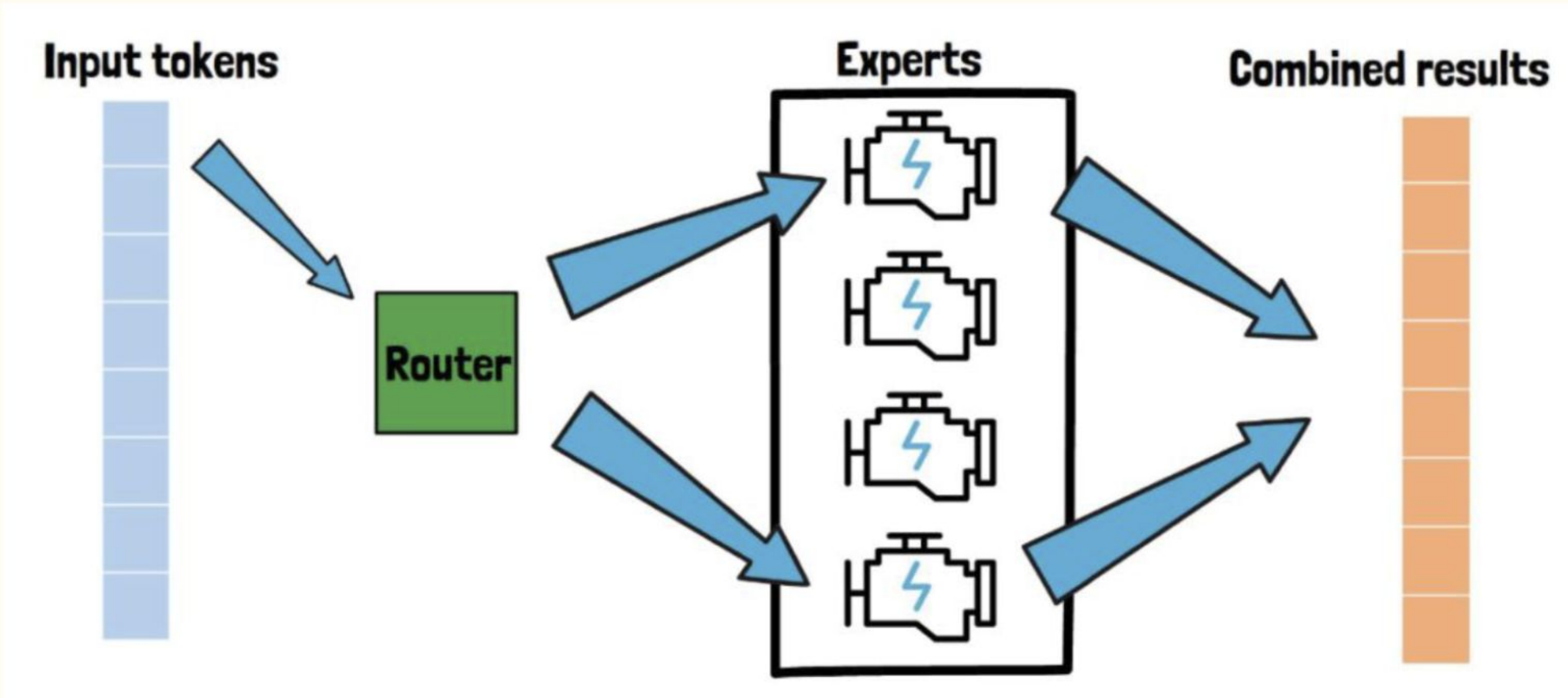
Experts

- An expert is a subset of the original network that is independently parameterized.
- Each has its own weights not shared with other experts.
- Inputs are processed by the gate network which decides which expert to use.



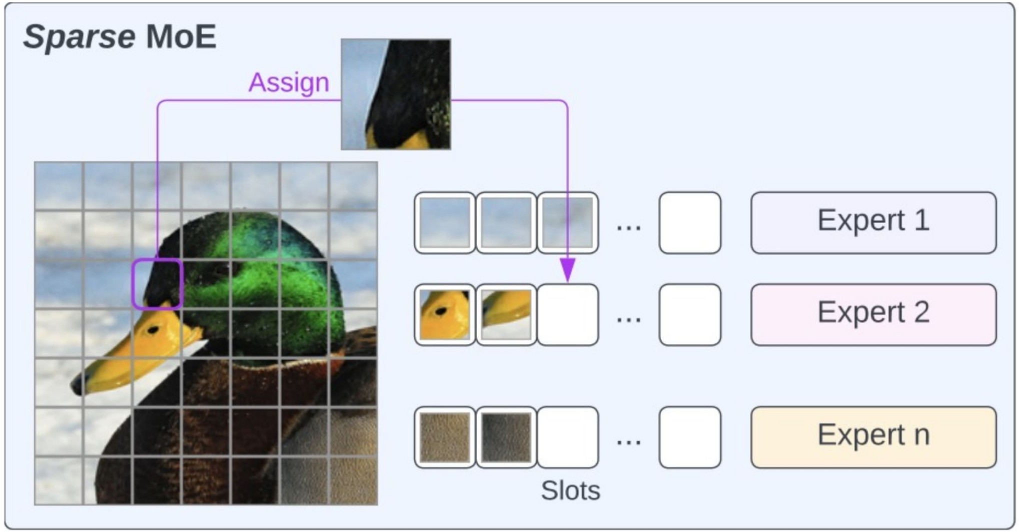
ImgSrc:https://www.linkedin.com/posts/tom-yeh_deeplearning-generatieveai-llms-activity-7141461533112381441-J35v?utm_source=li_share&utm_content=feedcontent&utm_medium=g_dt_web&utm_campaign=copy

Combination Function



Sparse MoE – Passing a token via the router and chosen experts

Sparse Mixture-of-Experts (SMoE)



ImgSrc:https://sh-tsang.medium.com/review-scaling-vision-with-sparse-mixture-of-experts-dd4de8ad27fa

- The model learns which experts are best with particular tokens
- Only a set of experts are activated for each input rather than all experts
- Makes computation more efficient, with possible trade-offs in accuracy

Rise of Mixture of Experts (MoE)

Mixture of Experts
1991

Jacobs et al.
Combined local experts for vowel classification

Top-K Routing
2017

Shazeer et al.
Sparsely activated experts

Switch Transformer
2022

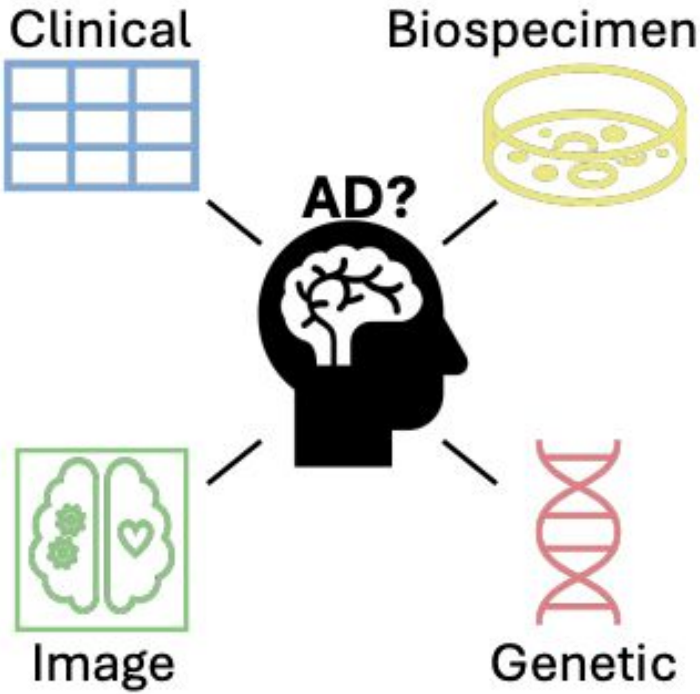
Fedus et al.
Routed to one expert

Dropless MoE
2022+

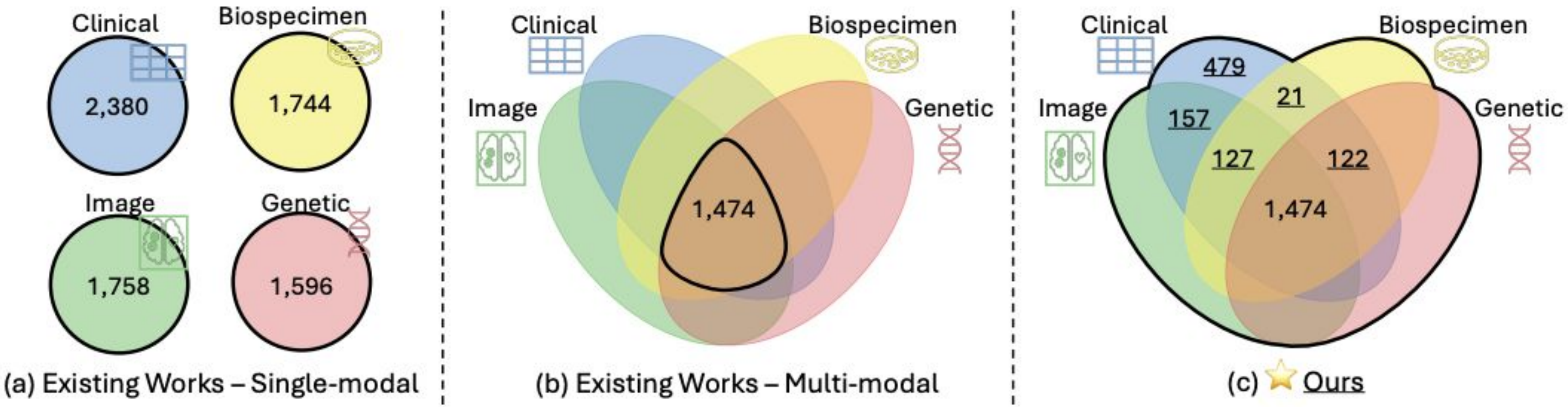
Gale et al.
Block-sparse matrix mult. “Megablocks”; also Soft MoE

Multimodal Alzheimer's Disease (AD)

- AD pathologies involve many hypotheses spanning various modalities.
- Modalities might be missing for certain patients that could be useful in understanding their progression.
- Most models expect all modalities to be present.
- Number of modalities are growing along with possible combinations.

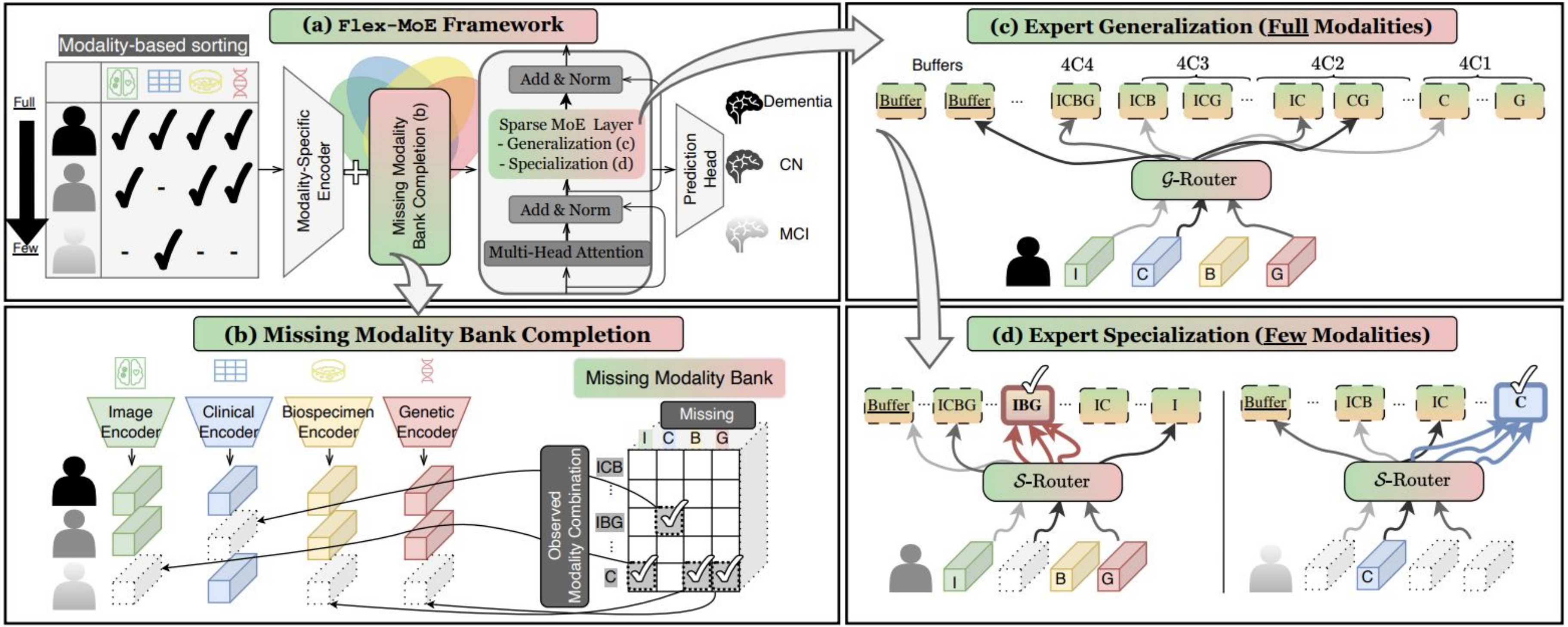


Missing Modality Problem



Flex-MoE

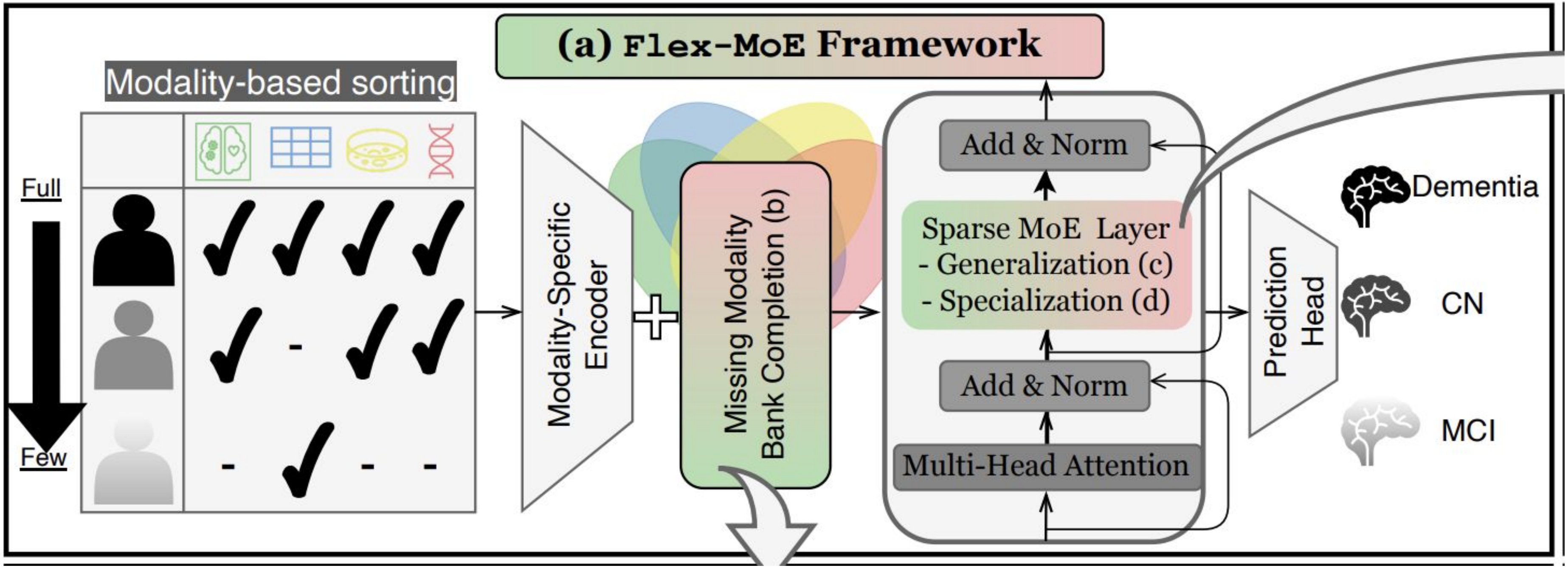
Methods Overview



B = biospecimen, C = clinical, I = image, G = genetic

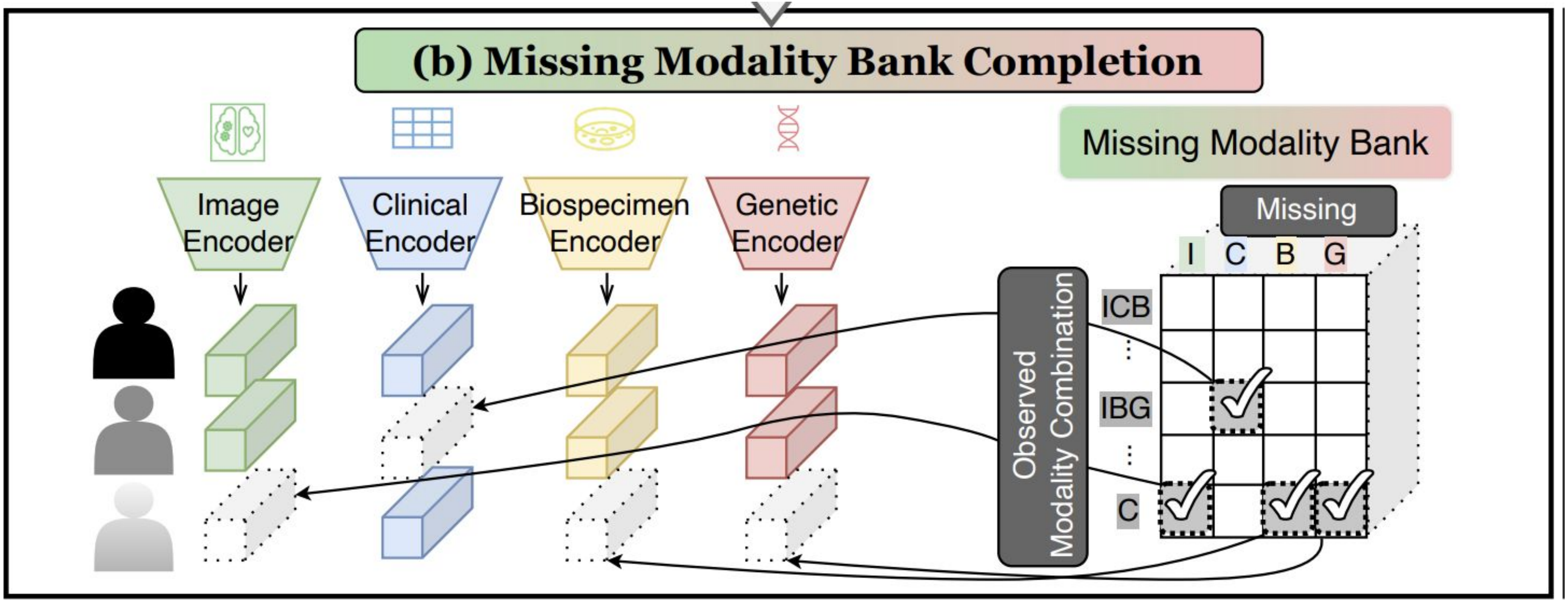
CN = normal cognitive aging, MCI = mild cognitive impairment

Flex-MoE Framework



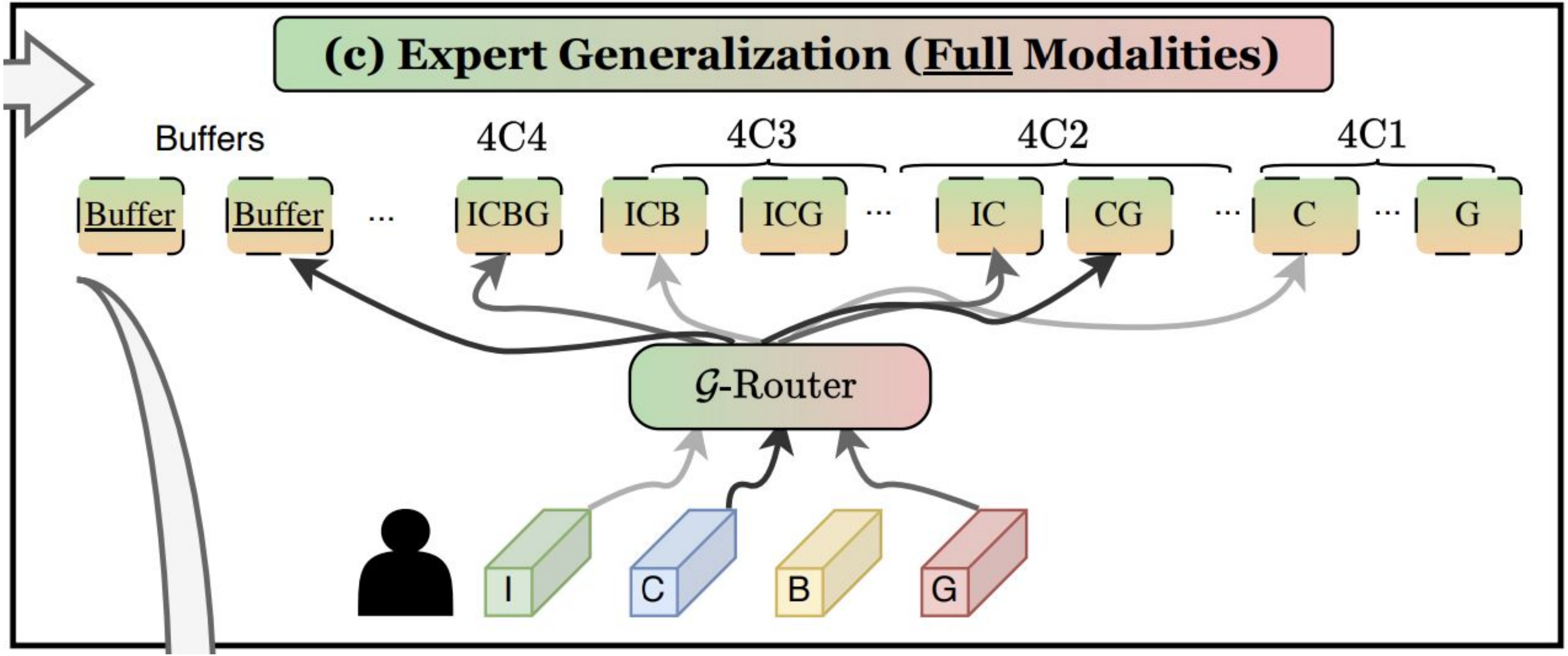
CN = normal cognitive aging, MCI = mild cognitive impairment

Missing Modality Bank



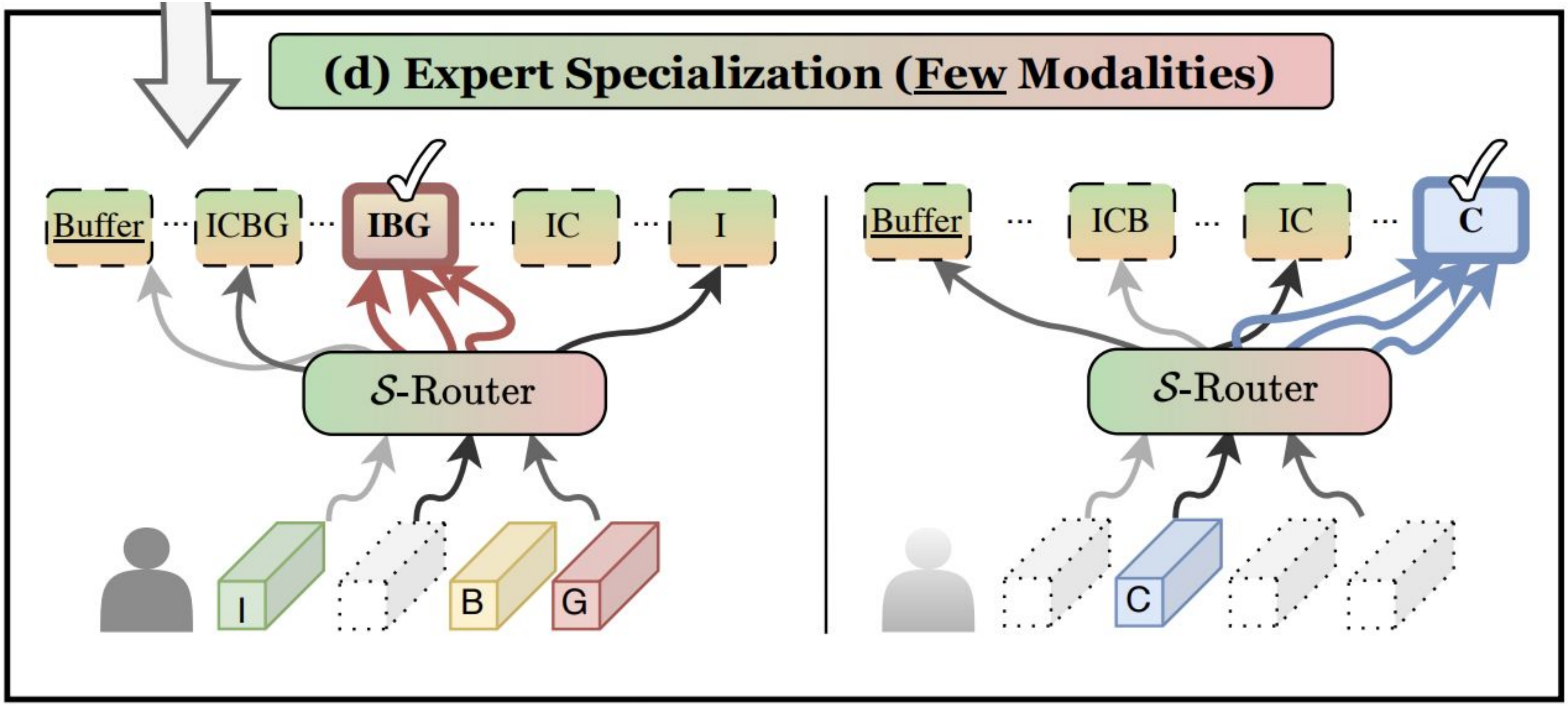
B = biospecimen, C = clinical, I = image, G = genetic

Expert Generalization



B = biospecimen, C = clinical, I = image, G = genetic

Expert Specialization



B = biospecimen, C = clinical, I = image, G = genetic

Generalization (SMoE)

- Train first layer of SMoE
  - Easiest examples first
  - All modalities fully observed
- More challenging examples appear later
  - Follows vanilla SMoE equation
  - Input tokens only consist of full modality combinations

$$\mathbf{y} = \sum_{i=1}^{|E|} \mathcal{R}(\mathbf{x})_i \cdot f_i(\mathbf{x}),$$

$$\mathcal{R}(\mathbf{x}) = \text{Top-K}(\text{softmax}(g(\mathbf{x})), k),$$

$$\text{TopK}(\mathbf{v}, k) = \begin{cases} \mathbf{v}, & \text{if } \mathbf{v} \text{ is in the top } k, \\ 0, & \text{otherwise.} \end{cases}$$

- y - final output of MoE layer
- |E| - number of experts
- f_i(x) - output of the ith expert, given x
- R(x)$_i$ - routing weight assigned to expert i, given x
- TopK(v, k) - keeps only top k probabilities
  - Only most relevant experts preserved
  - Masks others to zero

Specialization

Once the experts are initially trained, they use a special routing mechanism (S-Router) to target specific experts with specific modality combinations.

This is achieved by the following loss function

$$\mathcal{L}_{ce} = -\sum_{j=1}^{n} \mathcal{MC}(\mathbf{x}_j) \log(\max(\mathcal{S}\text{-Router}(\mathbf{x}_j)))$$

- **MC** - one hot vector indicating which combination to target
- **S-router** - outputs probability distribution over experts for input x_j
- Accumulates the loss over all inputs in the batch
- Penalizes S-router when selected top-1 expert doesn't match modality combination

Datasets

	ADNI	MIMIC-IV
Type	Alzheimer's multimodal dataset	ICU clinical dataset
Data	MRI, PET, genetics, clinical, biospecimens	ICD-9, clinical text, labs/vitals
Patients	Alzheimer's cases across stages	Adults with ≥ 2 visits
Size	~2,000 subjects (varies by modality)	~50,000 patients (subset of full MIMIC-IV)
Task	Multi-class prediction of AD stage- Dementia, CN, or MCI	One-year mortality binary classification
Prep	Mean imputation for missing data	Drop death-time visits, use last visit only
Access	Multi-center, open-access	Single center, de-identified

CN = normal cognitive aging, MCI = mild cognitive impairment, AD = Alzheimer’s Disease

Experimental Design

	ADNI	MIMIC-IV
Dataset focus	AD prediction	One-year patient mortality prediction
Classification task	3-class classification: CN, MCI, Dementia	Binary classification: 1-year mortality (yes/no)
Modalities Used	MRI, PET, Genetic (APOE, SNPs), Clinical, Biospecimen (CSF, blood, urine)	ICD-9 codes, Clinical Text, Labs & Vitals
Baselines	3D CNN, VGG, ResNet-18, ResNet-34, Autoencoders, GRU, ShaeSpec, mmFormer, MAG, MulT, TF	Same multimodal models: FuseMoE, MulT, MAG, TF, LIMoE
Fusion Strategy	For baselines lacking imputation/fusion: zero-padding used during batch training	Same

Baselines

Modality	Model/Method	Description
Image-only	3D CNN	Processes 3D MRI scans.
Image-only	3D CNN + 3D CLSTM	Combines 3D CNN with convolutional LSTM for temporal features.
Image-only	2D VGG	Pretrained VGG with layer-wise transfer learning on 2D MRI slices.
Image-only	Modified ResNet-18	Adapted for 2D MRI scans.
Genetic-only	ResNet-34	Handles high-dimensional genetic data.
Multimodal (ADNI)	Autoencoder + 3D CNN	Integrates imaging, genetic, and clinical data.
Multimodal (ADNI)	GRU-based Architecture	Incorporates imaging, genetic, clinical, and biospecimen data.
Multimodal (ADNI)	ShaeSpec	Spectral attention mechanism across modalities.
Multimodal (ADNI)	mmFormer	Transformer-based multimodal fusion with attention.
Multimodal (ADNI & MIMIC-IV)	FuseMOE	Mixture-of-experts strategy for direct multimodal integration.
Multimodal (ADNI & MIMIC-IV)	MulT	Cross-attention for cross-modal interaction.
Multimodal (ADNI & MIMIC-IV)	MAG	Multimodal fusion via adaptation vector mapping.
Multimodal (ADNI & MIMIC-IV)	TF	Combines embedding sub-networks and a tensor fusion layer.
Multimodal (ADNI & MIMIC-IV)	LIMoE	Uses entropy regularization for stable multimodal learning with contrastive learning.

Experimental Settings

Setting	Values
LR	1e-3, 1e-4, 1e-5
Hidden Dim	64, 128, 256
Batch Size	8, 16
# Experts	16, 32
Top-k	2, 3, 4
Loss Coeff.	0.01
Data Split	70% train, 15% val, 15% test
Modality Handling	Intersection for val/test; zero-pad if needed
Runs	3 seeds, averaged
Hardware	NVIDIA A100 GPUs

Results - ADNI

	Modalities				Dataset: ADNI / Metric: ACC									
$\mathcal{MC}$					Transformer-based [59]	GRU-based [33]	ShaSpec	mmFormer	TF	MulT	MAG	LIMoE	FuseMoE	Flex-MoE
$\mathcal{I},\mathcal{G}$	•	•			54.81 ±1.45	53.59 ±2.98	48.09 ±0.66	49.85 ±4.92	59.94 ±0.40	60.32 ±0.95	59.94 ±1.00	59.29 ±0.95	60.41 ±0.87	**61.08** ±0.78
$\mathcal{I},\mathcal{C}$	•		•		44.35 ±1.99	**57.15** ±1.58	47.62 ±1.81	51.96 ±4.23	54.53 ±0.66	50.14 ±1.05	52.19 ±2.90	52.38 ±3.46	53.13 ±1.97	56.49 ±2.55
$\mathcal{I},\mathcal{B}$	•			•	40.80 ±2.94	57.61 ±1.86	50.98 ±2.09	51.45 ±3.53	52.57 ±2.06	51.17 ±2.88	52.47 ±4.11	53.87 ±2.75	49.67 ±1.97	**60.41** ±0.26
$\mathcal{G},\mathcal{C}$		•	•		51.91 ±1.39	52.85 ±2.47	52.85 ±2.65	49.58 ±4.45	38.38 ±3.03	46.03 ±5.42	40.34 ±6.11	35.76 ±6.24	38.84 ±2.42	**60.60** ±0.26
$\mathcal{G},\mathcal{B}$		•		•	45.01 ±1.30	52.66 ±3.63	58.54 ±2.97	48.45 ±4.56	42.20 ±1.78	39.40 ±2.91	40.52 ±2.52	36.88 ±5.04	37.91 ±0.80	**63.59** ±1.04
$\mathcal{C},\mathcal{B}$			•	•	44.63 ±0.92	**63.68** ±0.48	59.10 ±2.69	47.71 ±4.49	39.68 ±2.38	44.54 ±0.82	40.15 ±2.58	43.98 ±0.00	37.91 ±0.80	60.50 ±0.82
$\mathcal{I},\mathcal{G},\mathcal{C}$	•	•	•		55.12 ±2.38	54.72 ±0.28	49.30 ±3.17	46.49 ±3.57	54.06 ±1.98	60.97 ±0.95	61.34 ±0.61	53.50 ±2.25	60.97 ±1.32	**63.21** ±1.73
$\mathcal{I},\mathcal{G},\mathcal{B}$	•	•		•	56.12 ±3.44	55.28 ±3.44	52.85 ±0.53	47.15 ±6.43	54.44 ±2.26	53.03 ±1.95	54.15 ±1.06	53.97 ±1.08	52.85 ±1.00	**62.28** ±2.75
$\mathcal{I},\mathcal{C},\mathcal{B}$	•		•	•	43.79 ±0.69	60.97 ±2.60	52.85 ±3.30	47.18 ±4.68	52.29 ±1.47	49.86 ±1.50	53.24 ±0.50	54.97 ±0.00	49.67 ±1.00	**64.05** ±1.78
$\mathcal{G},\mathcal{C},\mathcal{B}$		•	•	•	45.28 ±1.85	53.87 ±3.35	62.09 ±3.27	46.38 ±4.24	43.33 ±4.43	43.32 ±6.74	37.25 ±1.99	40.99 ±2.62	34.64 ±1.95	**65.36** ±1.38
$\mathcal{I},\mathcal{G},\mathcal{C},\mathcal{B}$	•	•	•	•	52.10 ±0.99	55.64 ±1.86	52.84 ±0.53	58.92 ±6.58	57.24 ±3.05	58.82 ±0.82	61.44 ±1.61	55.18 ±4.22	59.52 ±1.00	**66.11** ±1.14

3-class classification of AD stage- Dementia, CN, or MCI

CN = normal cognitive aging, MCI = mild cognitive impairment, AD = Alzheimer's Disease

Image ($\mathcal{I}$,) Clinical ($\mathcal{C}$,) Biospecimen ($\mathcal{B}$,) Genetic ($\mathcal{G}$,)

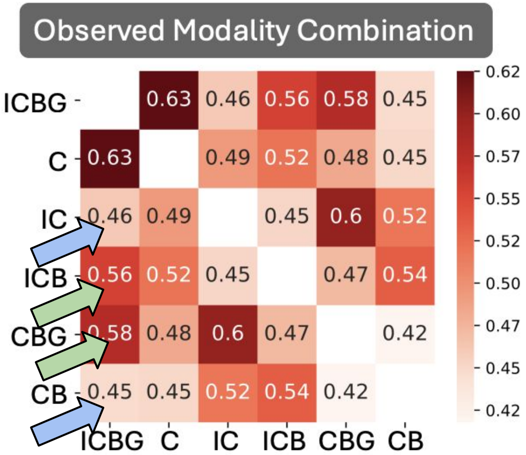
Results - MIMIC-IV

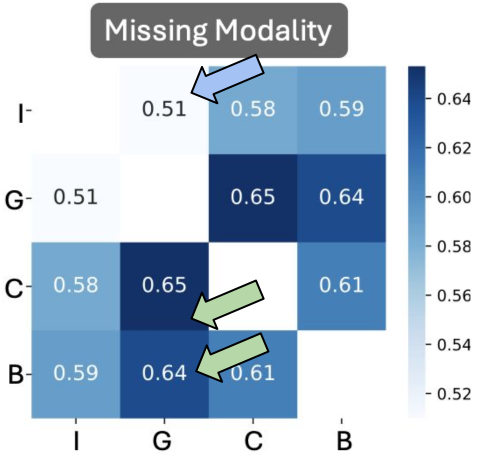
	Modalities			Dataset: MIMIC-IV / Metric: ACC					
$\mathcal{MC}$				TF	MulT	MAG	LIMoE	FuseMoE	Flex-MoE
$\mathcal{L},\mathcal{N}$	•	•		60.05 ±1.96	57.96 ±7.25	62.72 ±2.36	63.80 ±1.99	60.50 ±3.82	**76.14** ±0.73
$\mathcal{L},\mathcal{C}$	•		•	64.13 ±3.39	62.47 ±2.01	60.13 ±1.97	64.89 ±1.46	63.31 ±3.21	**75.15** ±0.55
$\mathcal{N},\mathcal{C}$		•	•	60.97 ±2.36	62.23 ±2.81	59.41 ±4.15	64.27 ±4.05	64.77 ±3.05	**74.96** ±1.59
$\mathcal{L},\mathcal{N},\mathcal{C}$	•	•	•	63.11 ±2.17	64.62 ±0.44	62.87 ±2.50	61.61 ±2.37	63.90 ±1.72	**76.81** ±0.90

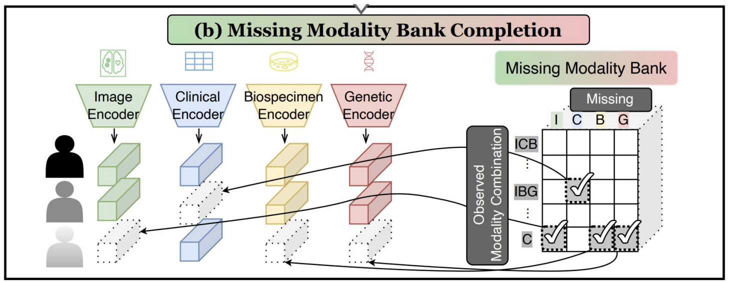
Binary classification on 1-year mortality

Clinical Notes ($\mathcal{N}$,) ICD-9 Codes ($\mathcal{C}$,) Lab and Vital values ($\mathcal{L}$,),

Results - Missing Modality Bank



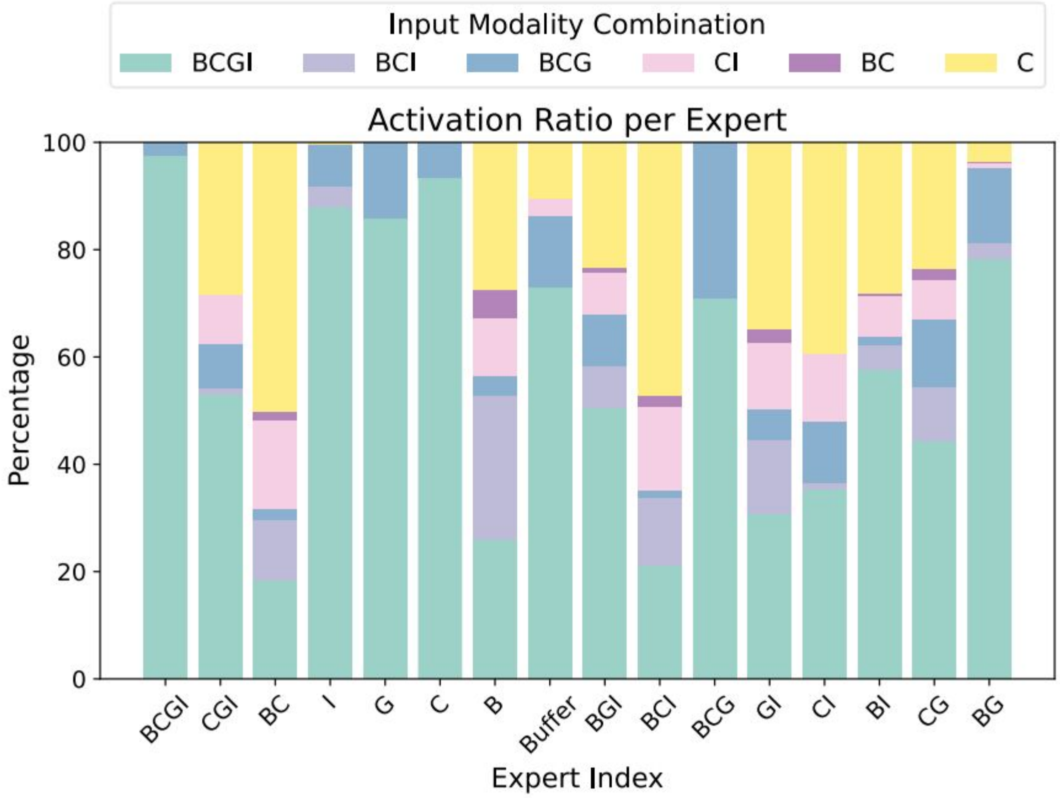




B = biospecimen, C = clinical, I = image, G = genetic

- (LEFT) Cosine similarity between observed modalities and their bank representation
    - More overlapping combinations share similar embedding information
- (RIGHT) Cosine similarity between missing modalities and retrieved bank representation
    - Certain missing modalities are handled more similarly by the model than others

Results - Modality Combination Activation Ratio



1. Generalized knowledge (BCGI) is distributed across all experts
   a. Due to expert generalization
2. Each expert is able to acquire specialized knowledge
   a. Due to expert specialization

B = biospecimen, C = clinical, I = image, G = genetic

Ablation Test

Table 3: Ablation study of Flex-MoE.

	ACC	F1
Flex-MoE	**66.11**	**64.73**
w/o ES	62.75	60.79
w/o {ES + EG}	62.49	60.07
w/o embedding bank	63.87	62.48
w/o sorting - random	62.65	60.70
w/o sorting - ascending	63.87	62.22

1. When ES/GS removed, accuracy dropped lowest
2. Embedding bank also important, accuracy dropped when removed
3. Ascending order sorting shown less performant than descending order sorting

Sensitivity Analysis

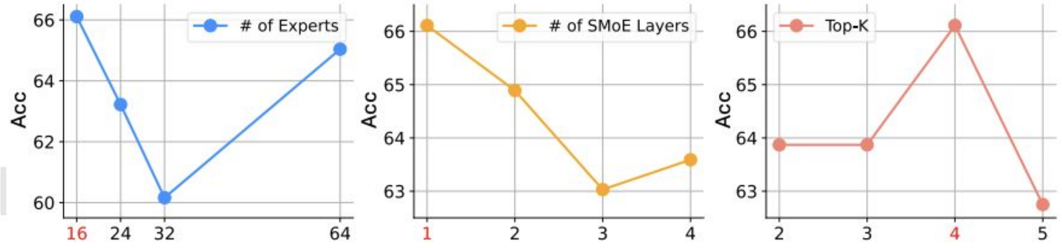


Figure 6: Sensitivity analysis of `Flex-MoE`. The hyper-parameters include the number of experts, the number of SMoE layers and Top-k expert selection. For the experiment, ADNI dataset with full modalities is used.

Tested 3 Hyperparameters:

- # Experts
  - More isn't always better
- # SMoE Layers
  - Using a single layer most effective
- Top-K
  - Found 4 to be the best choice

References/Resources

- Visual guide to MoEs
  - https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts
- Calculating an MoE by hand
  - https://www.linkedin.com/posts/tom-yeh_deeplearning-generatieveai-llms-activity-7141461533112381441-J35v?utm_source=share&utm_medium=member_desktop
- Review - Scaling vision with sparse mixture of experts
  - https://sh-tsang.medium.com/review-scaling-visio ... a

1. Mixture of Experts (MoE)

TOM YEH
DEC 15, 2023

...sformer vs. Mixture of Experts in LLMs

...explained visually.

AVI CHAWLA
FEB 27, 2025

Review — Scaling Vision with Sparse Mixture of Experts

V-MoE, up to 24 MoE Layers, 32 Experts Per Layer, Almost 15B Parameters

Sik-Ho Tsang · Follow
6 min read · Oct 4, 2022