# Patient-Conditioned Vision–Language Priors with Stable Mixture-of-Experts for Label-Efficient Alzheimer's MRI Staging

Shan Ding[1*], Fuzhi Wu[2*], Youyong Kong[1], and Huazhong Shu[1†]

[1] Southeast University, China
[2] Nanjing Forestry University, China
[*]Equal contribution.
[†]Corresponding author: shu.list@seu.edu.cn

**Abstract.** Accurate staging of Alzheimer's disease (AD) from structural MRI is clinically valuable, yet training reliable multi-class models is constrained by limited labels and by the asymmetric risk of diagnostic errors: a two-stage confusion between cognitively normal (CN) and AD is far more harmful than an off-by-one error with mild cognitive impairment (MCI). We present a label-efficient and safety-aware framework that couples (i) a frozen *patient-conditioned* vision–language prior, built by prompting a pretrained CLIP text encoder with demographics and aligning it with CLIP image embeddings from representative MRI slices, (ii) a *scale-preserving* mixture-of-experts (MoE) router that stabilizes expert utilization under small-batch, imbalanced training, and (iii) an EMA teacher for semi-supervised consistency. Across OASIS and ADNI under 10–50% labeled data, the proposed model consistently achieves the best accuracy and ordinal agreement while reducing mean absolute error and a clinically weighted mis-staging cost. At 10% labels, we reach QWK 0.806/0.823 on OASIS/ADNI and reduce clinical cost to 0.158/0.143, leading to a more favorable position on the clinical-safety frontier. Notably, with only 20% labels we match the strongest baseline trained with 50% labels, indicating a $\sim$2.5$\times$ reduction in annotation demand.

**Keywords:** Alzheimer's disease · label-efficient learning · vision–language prior · mixture-of-experts · clinical safety.

## 1 Introduction

Structural MRI provides a non-invasive window into neurodegeneration and has become a core modality for studying Alzheimer's disease (AD) progression. Deep learning has shown promise for MRI-based diagnosis and staging, but in practice two constraints remain persistent: *label scarcity* and *clinical asymmetry of errors*. Public cohorts such as ADNI and OASIS contain many scans, yet high-quality diagnostic labels for training are expensive, inconsistent across sites, and often incomplete. At the same time, CN/MCI/AD categories are inherently ordinal—misclassifying CN as AD (or vice versa) is a markedly different clinical event from

confusing adjacent stages. This creates a gap between standard classification objectives and the way clinicians perceive risk.

Semi-supervised learning (SSL) is an attractive direction for label scarcity, but applying generic SSL recipes to AD staging is non-trivial. First, MRI data are 3D and heterogeneous; second, the label space is ordinal; and third, a model that merely maximizes accuracy can still produce a non-negligible number of clinically unacceptable two-stage errors. Recent medical foundation models and vision–language models (VLMs) suggest a complementary lever: leverage pretrained cross-modal structure as a prior. Domain-specific VLMs (e.g., MedCLIP, BioMedCLIP) demonstrate that contrastive image–text pretraining can inject semantic structure into medical representations [8–10]. Separately, mixture-of-experts (MoE) has emerged as an efficient way to increase capacity and specialization [6], including in recent MICCAI works that use MoE to handle heterogeneous medical data and prompts [17, 18].

We build on these insights and target a practical setting: *label-scarce, clinically safe ordinal staging* from MRI. Our key idea is to couple a patient-conditioned vision–language prior with a norm-stabilized and scale-preserving MoE classifier, trained under a teacher–student SSL regime. The vision–language prior provides a compact, patient-specific semantic anchor derived from demographics, while the MoE head increases representational diversity without destabilizing training under small batches. To reflect clinical needs, we evaluate not only accuracy and QWK, but also mean absolute error (MAE), catastrophic ($\geq$ 2-stage) mis-staging, screening operating characteristics, and a clinically weighted cost.

Our contributions are:

1. **Patient-conditioned VLM prior.** We construct a demographic-conditioned textual prompt and align it with CLIP image embeddings extracted from representative MRI slices, forming a stable prior that regularizes MRI features in low-label regimes.
2. **Stable MoE for ordinal staging.** We introduce scale-preserving routing and norm-stabilized classification to prevent expert collapse and logit scale drift—a common failure mode when training MoE heads on limited, imbalanced medical data.
3. **Clinically grounded evaluation under label scarcity.** On OASIS and ADNI with 10–50% labels, our method consistently improves both ordinal agreement and clinical-safety metrics, shifting the Pareto frontier of QWK vs. MAE and reducing a clinically weighted mis-staging cost.

## 2 Method

### 2.1 Problem Setup

We consider ordinal staging with three classes $y \in \{\mathrm{CN}, \mathrm{MCI}, \mathrm{AD}\}$. Given a 3D MRI volume $x$ and demographics $d$ (age, sex), we learn a classifier $f$ that outputs class probabilities $p(y \mid x, d)$. Under a label-scarce protocol, we observe a small labeled set $\mathcal{D}_\ell = \{(x_i, d_i, y_i)\}_{i=1}^{N_\ell}$ and a larger unlabeled set $\mathcal{D}_u = \{(x_j, d_j)\}_{j=1}^{N_u}$.

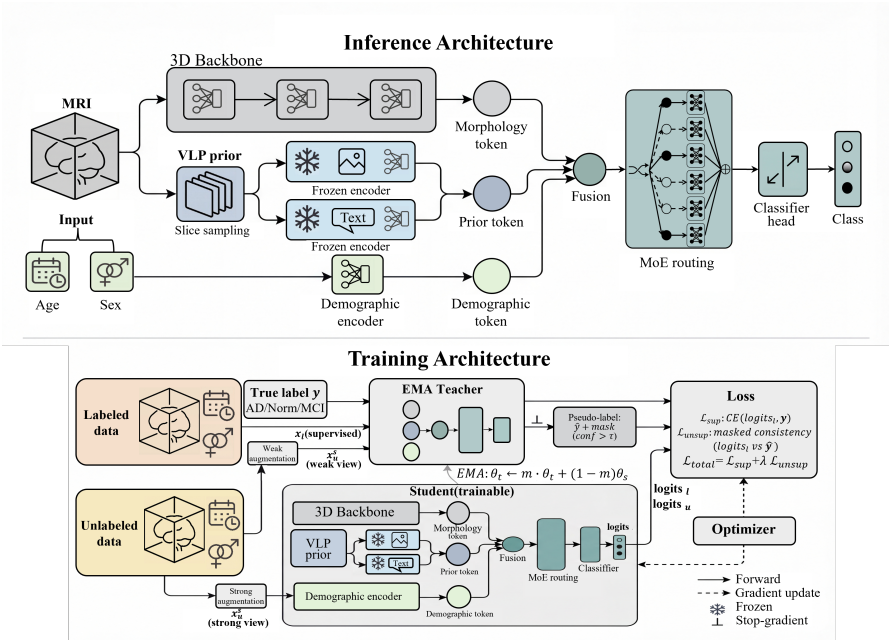


**Fig. 1.** Overview of the proposed framework. A demographic-conditioned prompt induces a patient-specific text embedding (frozen CLIP). Representative MRI slices provide CLIP image embeddings, forming a vision–language prior. A scale-preserving MoE head with norm-stabilized classification produces ordinal predictions. An EMA teacher guides semi-supervised learning.

### 2.2 Overview

Our model contains three coupled components (Fig. 1): (1) a 3D MRI backbone $\phi(\cdot)$ that produces volume features $z$, (2) a patient-conditioned vision–language prior $v$ derived from frozen CLIP encoders [3] and demographics, and (3) a scale-preserving MoE head that maps $[z \,\|\, v]$ to ordinal logits. Training combines supervised ordinal losses on $\mathcal{D}_\ell$ and consistency/pseudo-label losses on $\mathcal{D}_u$ using an EMA teacher [4, 5].

### 2.3 Patient-Conditioned Vision–Language Prior

**Demographic prompting.** We embed demographics using a lightweight tokenizer (Sec. 2.4) and inject them into a natural-language template, e.g., *"A {sex} subject, age {age}, undergoing brain MRI for cognitive assessment."* The prompt is encoded by a frozen CLIP text encoder $E_t$, yielding $t = E_t(\text{prompt}(d))$. Prompt conditioning follows the general philosophy of prompt learning in VLMs [7] while remaining fully frozen for stability and reproducibility. We emphasize that this is a *patient-conditioned* prior rather than a report-level language signal: the text

branch provides a standardized demographic context, and the image branch anchors this context to MRI appearance. We used a single neutral clinical template in all experiments to avoid tuning the prompt on the test sets; prompt-template selection and richer clinical-text variants are left for future work.

**Slice-to-image CLIP embeddings.** We extract a small set of representative slices by uniform sampling along the axial axis, convert them to pseudo-RGB images, and feed them through the frozen CLIP image encoder $E_i$ to obtain slice embeddings $\{u_k\}_{k=1}^{K}$. We aggregate them by mean pooling and $\ell_2$-normalization: $u = \mathrm{norm}(\frac{1}{K}\sum_k u_k)$.

**Vision–language prior.** The prior is formed by aligning $u$ with $t$ through cosine similarity and a temperature $\alpha$:

$$s = \alpha \frac{u^\top t}{\|u\|\,\|t\|}, \qquad v = \mathrm{MLP}([u \,\|\, t \,\|\, s]). \tag{1}$$

The resulting $v$ serves as a compact patient-conditioned semantic anchor. The scalar $s$ is concatenated with $u$ and $t$ so that the MLP receives both the two modality-specific embeddings and their explicit alignment strength. Using a frozen VLM prior is complementary to domain-specific medical VLMs [8, 9] and avoids dataset-specific report pretraining. We chose a generic frozen CLIP encoder because report-supervised medical CLIP variants are not consistently trained on structural brain MRI reports and may introduce domain-specific reporting biases; nevertheless, replacing the frozen encoder with a brain-MRI-specific VLM is a natural extension.

### 2.4 Demographic Tokenization

Demographics are encoded as discrete tokens to avoid scale mismatch. Age is binned into 5-year intervals, and sex is treated as a categorical token. Each token is mapped to an embedding and concatenated with the CLIP text embedding. Although the natural-language prompt already contains age and sex, the discrete tokens provide a learnable low-dimensional demographic representation, whereas the frozen text embedding provides a fixed semantic context. This design prevents a single continuous variable (age) from dominating gradients and improves robustness under class imbalance.

### 2.5 Scale-Preserving MoE Routing and Norm-Stabilized Classification

Given backbone features $z$ and prior $v$, we form $h = [z \,\|\, v]$. A router produces expert weights $g = \mathrm{softmax}(W_r\,\mathrm{LN}(h))$ over $k$ experts. Each expert is a lightweight MLP $e_m(\cdot)$. To preserve activation scale, we normalize both inputs and expert outputs:

$$\tilde{h} = \mathrm{LN}(h), \qquad o = \sum_{m=1}^{k} g_m\,\mathrm{LN}(e_m(\tilde{h})). \tag{2}$$

Finally, we apply norm-stabilized classification:

$$\ell = \gamma \frac{W_c \, \mathrm{LN}(o)}{\|W_c\|_F}, \tag{3}$$

where $\gamma$ is a learnable scale. This prevents logit explosion and stabilizes training when expert usage is sparse.

### 2.6 Semi-supervised Training with EMA Teacher

We use a student–teacher framework. The teacher parameters $\theta'$ are updated by EMA: $\theta' \leftarrow m\theta' + (1-m)\theta$. For unlabeled samples, the teacher provides pseudo-labels $\hat{y}$ under weak augmentation; the student is trained to match them under strong augmentation. We optimize:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_u \, \mathcal{L}_{\mathrm{u}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{sup}}$ is cross-entropy plus an ordinal regularizer (penalizing larger stage distances), and $\mathcal{L}_{\mathrm{u}}$ follows FixMatch-style consistency [5].

## 3 Experiments and Results

### 3.1 Datasets and Label-scarce Protocol

We evaluate on two public cohorts: OASIS-3 [2] and ADNI [1]. Each dataset is split into train/validation/test with a 70:15:15 ratio, stratified by diagnosis. When multiple scans per subject are available, we keep all scans of a subject within the same split to prevent leakage. From the training split, we sample a labeled subset with ratio $\rho \in \{0.1, 0.2, 0.3, 0.5\}$ at the subject level and treat the remaining training subjects as unlabeled. Thus, different scans from the same subject are never split across labeled and unlabeled pools. All methods use identical splits and labeled/unlabeled partitions.

### 3.2 Implementation Details

All compared methods use the same train/validation/test partitions and the same subject-wise labeled/unlabeled pools described above. The camera-ready revision clarifies the evaluation protocol and design rationale, but does not introduce additional numerical experiments, hyperparameter sweeps, or new performance claims.

### 3.3 Evaluation Metrics

We report (i) **Acc**, balanced accuracy, and macro-F1 for overall classification; (ii) **QWK** [20] and **MAE** to reflect ordinal agreement and average stage deviation; (iii) **Cat** ($\geq$ 2-stage errors) as catastrophic mis-staging; (iv) screening sensitivity/specificity/PPV/NPV for impairment screening (positive: MCI+AD); (v) **AD NPV** for ruling out AD; and (vi) a **clinical cost** defined as Cost = $p(|\Delta| = 1) + 3p(|\Delta| \geq 2)$, penalizing catastrophic deviations more heavily.

**Table 1.** Overall CN/MCI/AD staging performance under label scarcity. Each cell reports two lines: **top** Acc/BalAcc/Macro-F1 (%), **bottom** QWK/MAE/ClinicalCost. Best overall is in **bold**.

| Method | Label ratio $\rho$ | | | |
|---|---|---|---|---|
| | 0.1 | 0.2 | 0.3 | 0.5 |
| **OASIS** | | | | |
| CFMSC | 81.0/81.5/79.2<br>0.665/0.233/0.276 | 84.6/84.8/82.8<br>0.724/0.190/0.226 | 86.7/85.7/84.4<br>0.718/0.176/0.219 | 88.2/89.9/87.1<br>0.821/0.136/0.154 |
| DSSL | 79.9/80.4/78.1<br>0.669/0.240/0.280 | 83.5/83.6/81.4<br>0.702/0.204/0.244 | 85.7/86.5/83.7<br>0.729/0.183/0.222 | 87.5/87.4/85.9<br>0.771/0.154/0.183 |
| SSMFS | 85.3/87.0/83.6<br>0.742/0.183/0.219 | 88.2/87.9/86.5<br>0.780/0.147/0.176 | 90.3/91.0/89.4<br>0.850/0.111/0.125 | 91.8/89.8/90.2<br>0.834/0.104/0.125 |
| THS-GAN | 83.9/85.2/82.4<br>0.754/0.186/0.211 | 86.7/85.2/84.4<br>0.726/0.172/0.211 | 89.2/88.7/87.6<br>0.813/0.129/0.151 | 91.0/91.4/89.9<br>0.868/0.100/0.111 |
| ss-HMFSS | 84.6/85.8/82.9<br>0.741/0.186/0.219 | 87.5/88.2/85.7<br>0.778/0.154/0.183 | 90.0/88.3/87.8<br>0.770/0.136/0.172 | 91.4/93.1/90.2<br>0.843/0.108/0.129 |
| Ours | **89.2/89.7/87.5**<br>**0.806/0.133/0.158** | **91.8/91.8/90.3**<br>**0.853/0.100/0.118** | **93.2/93.2/92.1**<br>**0.879/0.082/0.097** | **94.6/93.7/93.3**<br>**0.881/0.072/0.090** |
| **ADNI** | | | | |
| CFMSC | 82.5/82.8/80.2<br>0.701/0.214/0.254 | 85.7/87.8/84.6<br>0.789/0.164/0.186 | 87.9/87.7/86.1<br>0.755/0.157/0.193 | 89.3/88.8/87.3<br>0.805/0.132/0.157 |
| DSSL | 81.1/81.0/78.5<br>0.664/0.236/0.282 | 84.6/82.2/81.7<br>0.714/0.189/0.225 | 86.8/87.8/85.8<br>0.815/0.146/0.161 | 88.6/88.6/87.0<br>0.821/0.132/0.150 |
| SSMFS | 86.4/86.3/84.7<br>0.752/0.168/0.200 | 88.9/89.6/87.5<br>0.815/0.132/0.154 | 91.1/91.5/89.9<br>0.858/0.104/0.118 | 92.5/92.8/91.4<br>0.873/0.089/0.104 |
| THS-GAN | 85.0/83.5/82.2<br>0.703/0.193/0.236 | 87.9/88.7/86.3<br>0.783/0.150/0.179 | 90.4/90.2/89.0<br>0.838/0.114/0.132 | 92.1/92.6/91.0<br>0.870/0.093/0.107 |
| ss-HMFSS | 85.7/85.9/83.8<br>0.749/0.175/0.207 | 88.6/88.8/87.2<br>0.832/0.129/0.143 | 90.7/90.4/89.0<br>0.831/0.114/0.136 | 92.1/92.1/91.0<br>0.868/0.093/0.107 |
| Ours | **90.0/89.4/88.4**<br>**0.823/0.121/0.143** | **92.5/94.2/92.0**<br>**0.887/0.086/0.096** | **93.9/94.8/93.5**<br>**0.911/0.068/0.075** | **95.4/95.5/94.5**<br>**0.902/0.061/0.075** |

### 3.4 Comparison to State-of-the-art Methods

We compare against representative SSL baselines and semi-supervised fusion methods: CFMSC [12], DSSL [13], SSMFS [14], THS-GAN [15], and ss-HMFSS [16]. Since many recent AD papers report results under different splits, modalities, or label availability, we focus on *re-implemented* baselines under a unified protocol. Recent multimodal or VLM-based AD methods such as ADLIP [19] are highly relevant, but they rely on different input assumptions and evaluation protocols; we therefore discuss them as related work rather than direct numerical baselines.

Table 1 summarizes overall performance across label ratios. Our method dominates all baselines on both cohorts. At $\rho = 0.1$, we improve QWK from 0.754 to 0.806 on OASIS and from 0.752 to 0.823 on ADNI, while simultaneously reducing MAE and clinical cost. The gains persist as $\rho$ increases, indicating that improvements are not restricted to an extreme low-label corner. A notable data-efficiency pattern emerges: with only 20% labels, our accuracy matches the strongest baseline trained with 50% labels on both OASIS (91.76%) and ADNI (92.50%), suggesting substantial annotation savings.

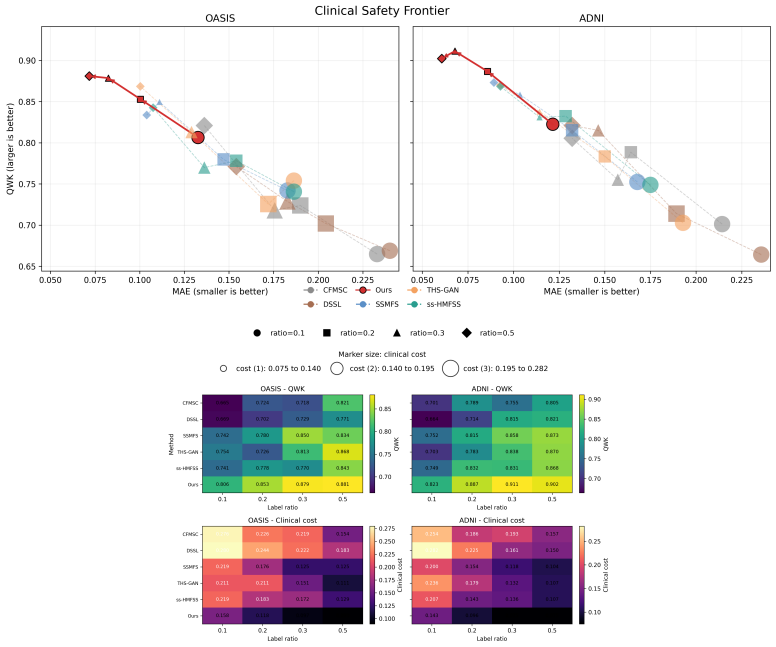


**Fig. 2.** Clinical-safety frontier (QWK vs. MAE; marker size indicates clinical cost) and performance profiles across label ratios for OASIS and ADNI. Our method consistently moves towards higher agreement and lower deviation/cost.

### 3.5 Clinical-safety Analysis

Beyond aggregate accuracy, clinical deployment requires minimizing dangerous mis-staging. Fig. 2 plots the *clinical-safety frontier* (QWK vs. MAE with marker size proportional to clinical cost) and heatmaps of QWK and cost across label ratios. Across both cohorts, our method consistently shifts the frontier towards the desirable top-left region (higher agreement, lower deviation) and maintains the lowest or near-lowest clinical cost.

Table 2 details screening operating characteristics at $\rho = 0.1$, where safety concerns are most acute. On OASIS, our model simultaneously achieves high screening sensitivity (96.98%) and NPV (92.21%), while cutting clinical cost by $\sim$25% relative to the best baseline. On ADNI, we obtain the highest screening specificity (95.00%) and very high PPV (97.91%), with AD NPV remaining above 96%, supporting AD rule-out. These results indicate that the proposed prior+MoE design does not trade clinical safety for accuracy; instead it reduces both average and costly errors.

**Table 2.** Clinical-safety metrics at the most label-scarce setting ($\rho = 0.1$). Screening treats MCI+AD as positive and CN as negative; the Screen column reports Sens/Spec (top) and PPV/NPV (bottom) in %. Cat: rate of $\geq$ 2-stage mis-staging.

| Method | OASIS | | | | ADNI | | | |
|---|---|---|---|---|---|---|---|---|
| | Cat↓ | Cost↓ | Screen↑ | AD NPV↑ | Cat↓ | Cost↓ | Screen↑ | AD NPV↑ |
| CFMSC | 4.30 | 0.276 | 91.0/75.0<br>90.0/76.9 | 97.22 | 3.93 | 0.254 | 91.5/88.8<br>95.3/80.7 | 95.39 |
| DSSL | 3.94 | 0.280 | 88.9/81.2<br>92.2/74.7 | 95.89 | 4.64 | 0.282 | 94.0/82.5<br>93.1/84.6 | 95.26 |
| SSMFS | 3.58 | 0.219 | 91.5/90.0<br>95.8/80.9 | 97.67 | 3.21 | 0.200 | 94.0/83.8<br>93.5/84.8 | 97.29 |
| THS-GAN | 2.51 | 0.211 | 93.5/83.8<br>93.5/83.8 | 97.67 | 4.29 | 0.236 | 93.5/86.2<br>94.4/84.2 | 95.07 |
| ss-HMFSS | 3.23 | 0.219 | 93.0/85.0<br>93.9/82.9 | 97.67 | 3.21 | 0.207 | 96.0/83.8<br>93.7/89.3 | 97.21 |
| Ours | **2.51** | **0.158** | **97.0/88.8**<br>**95.5/92.2** | **98.16** | **2.14** | **0.143** | **93.5/95.0**<br>**97.9/85.4** | **96.52** |

## 4 Conclusion

We presented a clinically grounded, label-efficient framework for ordinal staging of CN, MCI, and AD from MRI. By combining a patient-conditioned vision–language prior, a scale-preserving and norm-stabilized MoE classifier, and EMA-teacher SSL, the method consistently improves ordinal agreement while reducing clinically weighted mis-staging cost on both OASIS and ADNI. These gains are strongest in the low-label regime, where annotation cost is most limiting. The current study has limitations: the VLM branch uses standardized demographic prompts rather than free-text clinical reports, slice selection is based on uniform axial sampling, and the analysis does not yet include exhaustive component ablations, repeated random labeled-subset trials, or detailed fairness evaluation across demographic subgroups. Future work will extend the prior to richer clinical text, evaluate subgroup robustness, study foreground-aware slice selection, and analyze each component under repeated low-label samplings.

## Acknowledgments

This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62501285), and the National Key Research and Development Program of China (No. 2022YFE0116700).

## Disclosure of Interests

The authors have no competing interests to declare.

## Data and Code Availability

The code is publicly available at `https://github.com/dmlucky/Label-efficient-MoE`. The data supporting this work are derived from the publicly available OASIS-3 and ADNI datasets.

## References


1. Jack, C.R., Bernstein, M.A., Fox, N.C., et al.: The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* **27**(4), 685–691 (2008)
2. LaMontagne, P.J., Benzinger, T.L.S., Morris, J.C., et al.: OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv* (2019). doi:10.1101/2019.12.13.19014902
3. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: *Proc. ICML*. pp. 8748–8763 (2021)
4. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *Proc. NeurIPS*. pp. 1195–1204 (2017)
5. Sohn, K., Berthelot, D., Li, C.-L., et al.: FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: *Proc. NeurIPS*. pp. 596–608 (2020)
6. Shazeer, N., Mirhoseini, A., Maziarz, K., et al.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: *Proc. ICLR* (2017)
7. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022). https://doi.org/10.1007/s11263-022-01653-1
8. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: *Proc. EMNLP*. pp. 3876–3887 (2022). https://doi.org/10.18653/v1/2022.emnlp-main.256
9. Zhang, S., Xu, Y., Usuyama, N., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv* (2023). doi:10.48550/arXiv.2303.00915
10. Zhao, Z., Liu, Y., Wu, H., et al.: CLIP in medical imaging: a survey. *Med. Image Anal.* **102**, 103551 (2025). https://doi.org/10.1016/j.media.2025.103551
11. Weng, Y., Zhang, Y., Wang, W., Dening, T.: Semi-supervised information fusion for medical image analysis: Recent progress and future perspectives. *Inf. Fusion* **106**, 102263 (2024). https://doi.org/10.1016/j.inffus.2024.102263
12. Jiang, B., Zhang, C., Zhong, Y., Liu, Y., Zhang, Y., Wu, X., Sheng, W.: Adaptive collaborative fusion for multi-view semi-supervised classification. *Inf. Fusion* **96**, 37–50 (2023). https://doi.org/10.1016/j.inffus.2023.03.002
13. Wang, Y., Gu, X., Hou, W., Zhao, M., Sun, L., Guo, C.: Dual semi-supervised learning for classification of Alzheimer's disease and mild cognitive impairment based on neuropsychological data. *Brain Sci.* **13**(2), 306 (2023). https://doi.org/10.3390/brainsci13020306
14. Zhang, C., Fan, W., Wang, B., Chen, C., Li, H.: Self-paced semi-supervised feature selection with application to multi-modal Alzheimer's disease classification. *Inf. Fusion* **107**, 102345 (2024). https://doi.org/10.1016/j.inffus.2024.102345
15. Yu, W., Lei, B., Ng, M.K., Cheung, A.C., Shen, Y., Wang, S.: Tensorizing GAN with high-order pooling for Alzheimer's disease assessment. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(9), 4945–4959 (2022). https://doi.org/10.1109/TNNLS.2021.3063516

16. An, L., Adeli, E., Liu, M., Zhang, J., Lee, S.-W., Shen, D.: A hierarchical feature and sample selection framework and its application for Alzheimer's disease diagnosis. *Sci. Rep.* **7**, 45269 (2017). https://doi.org/10.1038/srep45269
17. Jiang, Y., Shen, Y.: M4oE: A foundation model for medical multimodal image segmentation with mixture of experts. In: *Proc. MICCAI*, LNCS, vol. 15012, pp. 621–631 (2024). https://doi.org/10.1007/978-3-031-72390-2_58
18. Meng, R., Song, S., Jin, P., et al.: MAST-Pro: Dynamic mixture-of-experts for adaptive segmentation of pan-tumors with knowledge-driven prompts. In: *Proc. MICCAI*. pp. 310–320 (2025). https://doi.org/10.1007/978-3-032-05325-1_30
19. Lin, P.-J., Jiang, Z., Liu, Y., et al.: A vision–language foundation model for Alzheimer's disease diagnosis using MRI and clinical data. *Alzheimer's Dement.* **21**, e71029 (2025). https://doi.org/10.1002/alz.71029
20. Cohen, J.: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**(4), 213–220 (1968)
21. Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv* (2016). doi:10.48550/arXiv.1607.06450
22. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: *Proc. ICLR* (2017)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *Proc. ICLR* (2019)