QA-MoE: Quality-Aware and Stable Multimodal Mixture-of-Experts for Robust Clinical Prediction in Noisy and Missing-Modal Settings

Linpeng Sun[1] , **Victor S. Sheng**[1]
[1]Department of Computer Science, Texas Tech University
{linsun, victor.sheng}@ttu.edu

Abstract

Clinical prediction increasingly relies on multimodal inputs, where reliability and efficiency are crucial for real-world deployment. However, mainstream fusion and MoE gating typically treat all available modalities as uniformly beneficial and allow noisy or weakly informative modalities to perturb routing, leading to instability, routing collapse, and miscalibrated confidence under missingness and shift. We propose **QA-MoE**, a **Q**uality-**A**ware and stable multimodal **M**ixture-**o**f-**E**xperts that decouples reliability estimation from routing to enable robust, sparse, and interpretable fusion. QA-MoE adopts a modular architecture where each modality is initially encoded into a shared embedding space. To deal with structurally missing data, we employ a completion pathway that maintains a consistent interface. Unlike standard approaches, QA-MoE separates reliability estimation from the routing process. We propose an Evidential Quality Scorer to measure epistemic uncertainty, which then guides a Stability-Enhanced Subset Selector to filter out noisy modalities on the fly. Additionally, we include a Ternary Expert Aggregation mechanism acting as a specialized branch to stabilize predictions when data missingness is severe. Evaluations on clinical benchmarks (ADNI for Alzheimer's staging and MIMIC-IV for Length-of-Stay) demonstrate that QA-MoE outperforms strong multimodal baselines, improving reliability while cutting down unnecessary computation. This indicates that QA-MoE offers a robust solution for multimodal decision support, especially in clinical settings prone to noise and missing data.

1 Introduction

Multimodal perception and reasoning is a key bottleneck on the path toward reliable, deployable AI systems, and the challenge is particularly acute in healthcare [Le *et al.*, 2025; Zhang *et al.*, 2025]. Modern clinical decision support must integrate diverse signals. These inputs range from structured EHR (Electronic Health Record) variables to free-text reports. Imaging and lab measurements are also essential [Chang *et al.*, 2025]. However, these data types differ drastically. They vary significantly in availability. Their resolution and reliability also differ [Hao *et al.*, 2025]. Critically, real-world patients rarely present with "complete" multimodal evidence: some modalities are missing by design due to cost, workflow, and contraindications, while others are present but noisy or weakly informative [Dörrich *et al.*, 2025; Wang *et al.*, 2025b]. There is a pressing need for multimodal models that can reason about which modalities are actually useful for a given patient, rather than assuming that using more data is always better.

Recent work advanced multimodal clinical modeling and scalable MoE routing [Yun *et al.*, 2024; Fedus *et al.*, 2022], but two limitations remain: (i) most fusion pipelines treat all available modalities as uniformly helpful [Yao *et al.*, 2024], and (ii) missing-modality completion does not guarantee task-relevant information [Kim and Kim, 2024]. Consequently, robust trade-offs between performance and reliability under heterogeneous quality are still challenging.

To address these limitations, we propose a quality-aware and stable multimodal MoE framework (QA-MoE) that explicitly learns *modality-subset selection*: for each patient, the model predicts a task-adaptive combination of modalities, instead of always using all available modalities of inputs. The core idea is to decouple reliability estimation from routing. Concretely, QA-MoE (i) assigns each modality a calibrated reliability score via evidential modeling [Sensoy *et al.*, 2018], (ii) performs stability-enhanced gating to select an informative subset of modalities under noise and missingness, and (iii) routes the selected representation through a ternary MoE router that supports positive, zero, and negative expert contributions to mitigate interference.

Our main contributions are three-fold:

- We introduce a **quality-aware multimodal MoE** that learns instance-wise modality-subset selection for clinical prediction, reflecting the practical observation that strong performance can often be achieved with only a small, high-signal subset of modalities.
- We propose a **stable selection-and-routing pipeline** that couples evidential reliability estimation with bootstrap-consensus gating and ternary expert fusion, improving robustness to heterogeneous modality quality.
- We demonstrate consistent gains on two clinical benchmarks,

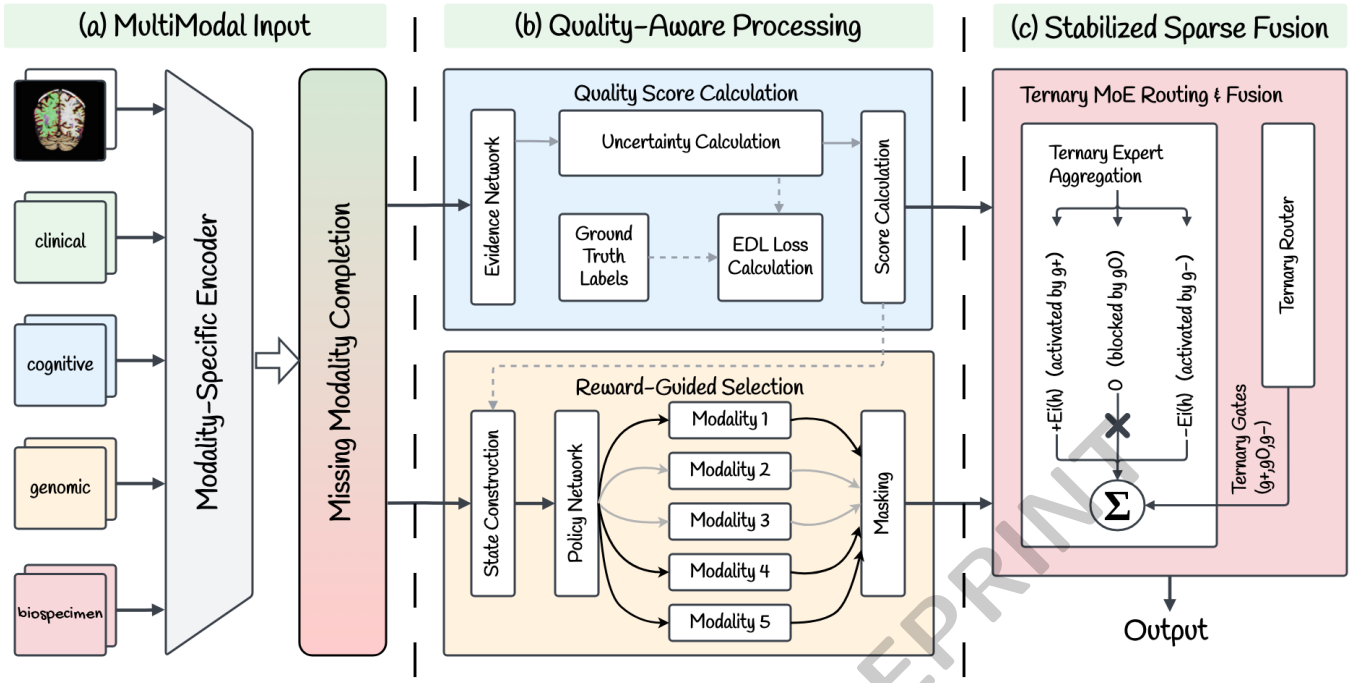

Figure 1: QA-MoE architecture overview. The framework consists of three main stages: (a) **MultiModal Input**: Heterogeneous clinical modalities (MRI, clinical scores, cognitive tests, genomics, biospecimens) are encoded through modality-specific encoders and processed via missing-modality completion. (b) **Quality-Aware Processing**: The evidential quality scorer computes calibrated reliability estimates for each modality using Dirichlet evidence, while the reward-guided subset selector performs stability-enhanced modality gating with bootstrap consensus. (c) **Stabilized Sparse Fusion**: Filtered embeddings are routed through a ternary MoE system where experts can contribute positively ($+E_i(h)$), abstain (0), or provide negative contributions ($-E_i(h)$), enabling robust fusion under modality disagreement.

ADNI for Alzheimer's staging [Weiner *et al.*, 2010] and MIMIC-IV for LOS prediction [Johnson *et al.*, 2023], while achieving favorable calibration and lower computational cost. Notably, while validated here on clinical tasks, our framework is model-agnostic and readily adaptable to other safety-critical scenarios requiring high reliability.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces QA-MoE and its components. Section 4 presents experiments and analyses on ADNI and MIMIC-IV. Section 5 concludes with limitations and future directions. Our code is publicly available at https://github.com/Illidon/QA-MoE.

2 Related Work

Multimodal Mixture-of-Experts MoE-based multimodal architectures have gained attention due to their ability to activate modality-specific pathways. FuseMoE [Han *et al.*, 2024] introduces flexible routing for missing modalities, and Flex-MoE [Yun *et al.*, 2024] further extends this idea by offering expert libraries specialized for observed subsets, enabling partial-modality composition in heterogeneous clinical inputs. MoE-Health [Wang and Yang, 2025] explores EHR–image fusion using sparse expert activation for clinical classification. Although these systems improve flexibility, routing still relies on softmax-based gating, which is highly sensitive to inter-modal noise and lacks explicit reliability modeling. Recent vision-language MoEs reinforce similar observations: dynamic gating is fragile to modality imbalance and quality heterogeneity.

Quality Estimation and Multimodal Uncertainty Reliable uncertainty estimation is critical in clinical settings. Classical approaches including Monte Carlo dropout, deep ensembles, and approximate variational inference are computationally expensive or unstable under sparse activation [Abboud *et al.*, 2024; Tocchetti and Mancini, 2024]. Evidential Deep Learning provides a deterministic alternative by predicting Dirichlet evidence that encodes both class probabilities and epistemic uncertainty [Chen *et al.*, 2024]. Although EDL has been applied in perception and medical diagnosis [Fan *et al.*, 2024; Liu *et al.*, 2025], its integration into multimodal MoE routing has not been explored. Existing multimodal uncertainty techniques often rely on heuristic confidence measures that are not calibrated under distribution shift.

Positioning w.r.t. confidence-weighted fusion and MoE gating. A key source of conceptual overlap in this space is where uncertainty is injected. Confidence-weighted fusion methods typically inject reliability after modality encoding, by weighting modality embeddings or logits during fusion [Gao *et al.*, 2024; Sidheekh *et al.*, 2024], which improves robustness but does not necessarily prevent unreliable modalities from perturbing the routing signal. In contrast, missing-modality completion methods focus on recovering representations when modalities are absent, but do not explicitly decide whether a recovered modality is useful

for the downstream task. TC-MoE-style [Yan *et al.*, 2025; Wang *et al.*, 2025a] routers improve expert stability through ternary activation, yet do not model modality reliability. QA-MoE is distinguished by injecting quality before subset selection and routing: reliability scores guide a stable instance-wise modality subset, and the selected representation is then routed through a ternary MoE. This separation clarifies that our gains arise from the interaction between calibrated quality estimation, sparsity-inducing subset selection, and stable ternary routing, rather than any single component alone. Unlike prior works that address modality missingness or expert routing in isolation, QA-MoE contributes a tightly coupled pipeline where evidential uncertainty directly stabilizes the selection and gating mechanisms under severe structural missingness.

Stabilizing MoE Routing MoE routers suffer from load imbalance, expert collapse, and high-variance gradient flows. MoE-CE [Li *et al.*, 2025] introduces auxiliary-free load balancing via dynamic bias shifts, SAFEx mitigates gating variance through bootstrap masking, and TC-MoE proposes ternary expert activation to reduce negative interference. Concurrent work in large-scale sparse models emphasizes the importance of expert frequency regularization and router debiasing for stable training [Wang *et al.*, 2025c]. While these methods improve routing behavior, they do not incorporate modality quality awareness and therefore remain vulnerable to heterogeneous or incomplete multimodal inputs. QA-MoE unifies uncertainty estimation and routing stabilization within a multimodal framework.

Missing-Modality and Noisy-Modality Fusion In medical applications, missingness is structural and cannot be treated as random dropout. Methods such as Modality Dropout, confidence-aware gating, or imputation-based fusion address missing or corrupted signals, but do not scale well to high-dimensional clinical modalities [Dai *et al.*, 2024; Gliozzo *et al.*, 2025]. Recent work in RAG (Retrieval-Augmented Generation) proposes a principled approach for missing-modality completion by selecting appropriate partial-modality experts. We extend this mechanism by employing a flexible completion strategy as a fallback pathway: when a modality is structurally missing, QA-MoE selects a partial-modality embedding template before scoring and routing, ensuring consistency in the downstream MoE pipeline.

3 Method

QA-MoE comprises four components: (i) modality encoders that generate continuous embeddings from heterogeneous clinical data; (ii) an Evidential Quality Scorer that computes calibrated reliability scores; (iii) a Stability-Enhanced Subset Selector that determines which modalities to retain; and (iv) a ternary MoE router with adaptive load balancing and expert fusion. The design is fully modular, enabling each component to operate independently or to be ablated individually.

3.1 Modality Encoders

Each modality x_m is processed through a modality-specific encoder to produce an embedding $e_m \in \mathbb{R}^D$. For images, lightweight convolutional encoders are used; for clinical and genomic features, multilayer perceptrons map structured data to the shared embedding space; for text or free-text clinical narratives, a projection layer maps high-dimensional contextual embeddings to D. Each encoder includes a learned noise parameter that models intrinsic signal variability in medical measurements. When a modality is structurally missing, a partial-modality embedding template is substituted in place of the missing encoder output, ensuring consistent downstream quality scoring and routing.

3.2 Evidential Quality Scoring

The Evidential Quality Scorer provides calibrated reliability estimates for each modality by leveraging evidential deep learning principles. For each modality embedding $e_m \in \mathbb{R}^D$, a lightweight MLP ϕ_{evid} maps e_m to evidence parameters $f_m \in \mathbb{R}^K$, where K denotes the number of classes:

$$f_m = \phi_{\text{evid}}(e_m), \quad f_m \in [-10, 10],$$

where the evidence parameters are clamped to prevent numerical instability. These parameters are then transformed into Dirichlet concentration parameters $\alpha_m \in \mathbb{R}^K$:

$$\alpha_{m,k} = \exp(f_{m,k}) + 1, \quad \forall k \in \{1, \ldots, K\}.$$

The Dirichlet distribution $\text{Dir}(\alpha_m)$ models the epistemic uncertainty over class predictions. The total evidence strength is quantified as $S_m = \sum_{k=1}^{K} \alpha_{m,k}$, and the epistemic uncertainty is defined as:

$$u_m = \frac{K}{S_m}.$$

The reliability score $q_m \in [0, 1]$ is then computed as:

$$q_m = 1 - u_m = 1 - \frac{K}{\sum_{k=1}^{K} \alpha_{m,k}}.$$

This formulation ensures that modalities with high evidence (large S_m) receive high reliability scores, while noisy or uncertain modalities yield low scores. The uncertainty regularizer $\mathcal{L}_{\text{uncert}}$ encourages calibration by penalizing overconfident predictions with low evidence:

$$\mathcal{L}_{\text{uncert}} = \sum_{m=1}^{M} \left[\sum_{k=1}^{K} (y_k - \hat{p}_{m,k})^2 + u_m(1 - u_m) \right],$$

where $\hat{p}_{m,k} = \alpha_{m,k}/S_m$ is the predicted class probability and y is the one-hot encoded ground truth. The first term encourages accurate predictions, while the second term regularizes uncertainty estimates to prevent overconfidence. This scoring mechanism is fully differentiable and provides explicit quality estimates that directly inform the subset selection process.

Remark on the evidential regularizer Our uncertainty penalty is a lightweight surrogate designed for stability in a sparse, multi-component optimization setting. It deviates from standard EDL objectives. In this submission, we utilize it as a practical calibration-oriented regularizer, leaving more rigorous theoretical justifications and alternative evidential formulations to future work. Although this heuristic trades strict theoretical guarantees for empirical stability in sparse routing, our ablations confirm its practical necessity.

Usefulness vs. confidence We emphasize that q_m measures predictive confidence under the learned representation, which is not identical to task usefulness: a modality can be confident but irrelevant or spurious for the target. QA-MoE mitigates this gap by coupling quality scoring with *task-driven* selection: encoders and the evidential scorer are trained jointly with the supervised objective, and the subset selector is optimized to maximize downstream performance under a sparsity budget. Nevertheless, confident-but-misleading modalities remain a possible failure mode; we discuss this limitation and suggested diagnostics in Section 5.

3.3 Stability-Enhanced Subset Selection

The subset selector produces an instance-wise modality mask from quality scores and embeddings. Given $\{q_m\}_{m=1}^M$ and $\{e_m\}_{m=1}^M$, we form a state vector

$$s = [q_1, \ldots, q_M, \text{flatten}(e_1), \ldots, \text{flatten}(e_M)].$$

A lightweight actor π_θ outputs continuous selection probabilities:

$$p_m = \sigma(\pi_\theta(s)_m), \quad p_m \in (0, 1),$$

which are used as soft masks in training; at inference we threshold at $\tau = 0.5$:

$$m_m = \begin{cases} 1 & \text{if } p_m > \tau \\ 0 & \text{otherwise} \end{cases}.$$

The filtered embeddings are then computed as $\tilde{e}_m = m_m \cdot e_m$.

To stabilize selection against noise, we adopt a SAFEx-inspired bootstrap consensus and quality gate. We treat this mechanism as a non-differentiable regularizer. Consequently, gradients flow primarily through the continuous-mask path.

The selector is trained using a reward signal that balances task accuracy and modality sparsity:

$$R = \text{ACC} + \eta \left(1 - \frac{\sum_{m=1}^M m_m}{M}\right) - \lambda_{\text{reg}} \cdot \text{Var}(\{m_m\}),$$

where $\eta = 0.2$ encourages selecting fewer modalities, and λ_{reg} penalizes high variance in selection patterns.

Selector objective. We use a differentiable reward-weighted term (no environment interaction loop):

$$\mathcal{L}_{\text{sel}} \;=\; -\mathbb{E}\Big[(R - \bar{R}) \cdot \frac{1}{M} \sum_{m=1}^{M} \log(p_m + \epsilon)\Big],$$

where $\bar{R}$ is a batch baseline and ϵ is a small constant.

3.4 Ternary MoE Routing and Fusion

The Ternary MoE Router extends standard MoE gating to enable experts to contribute positively, abstain, or provide negative contributions. Filtered embeddings $\{\tilde{e}_m\}_{m=1}^M$ are first aggregated into a routing token:

$$h = \frac{1}{M} \sum_{m=1}^{M} \tilde{e}_m \in \mathbb{R}^D.$$

A linear gate $\mathbf{W}_g \in \mathbb{R}^{D \times N}$ produces base logits for N experts:

$$\ell_i^{\text{base}} = \mathbf{W}_g[i] \cdot h, \quad i \in \{1, \ldots, N\}.$$

For each expert i, QA-MoE generates three logits corresponding to positive ($+1$), zero (0), and negative (-1) contributions:

$$\ell_i^+ = \ell_i^{\text{base}} + b_i^+, \tag{1}$$

$$\ell_i^0 = \ell_i^{\text{base}} - 10 + b_i^0, \tag{2}$$

$$\ell_i^- = \ell_i^{\text{base}} + b_i^-, \tag{3}$$

where $\{b_i^+, b_i^0, b_i^-\}$ are learnable per-expert biases. The large negative offset (-10) in ℓ_i^0 initially biases experts toward abstention. A per-expert softmax is applied to obtain ternary gate values:

$$[g_i^+, g_i^0, g_i^-] = \text{softmax}([\ell_i^+, \ell_i^0, \ell_i^-]),$$

ensuring $g_i^+ + g_i^0 + g_i^- = 1$ for each expert. This ternary formulation prevents global competition between experts and allows experts to abstain when input signals are unreliable, reducing interference from noisy modalities. An explicit g_i^0 provides a learned abstention option (instead of manual thresholds).

We apply ALFLB (Auxiliary-Loss-Free Load Balancing) [Fedus *et al.*, 2022] bias updates to reduce persistent expert imbalance (hyperparameters in Section 3.5).

Fusion is performed by combining expert outputs according to their ternary gate values. Each expert E_i computes a base output $E_i(h) \in \mathbb{R}^D$. The final fused representation is:

$$h_{\text{fused}} = \sum_{i=1}^{N} \left[g_i^+ \cdot E_i(h) + g_i^0 \cdot \mathbf{0} + g_i^- \cdot (-E_i(h))\right].$$

This formulation enables experts to: (i) amplify their responses when g_i^+ is high, (ii) abstain when g_i^0 dominates, and (iii) invert their contributions when g_i^- is high, modeling antagonistic modality interactions. The final prediction is obtained via a linear classifier:

$$\hat{y} = \text{softmax}(\mathbf{W}_{\text{out}} \cdot h_{\text{fused}} + \mathbf{b}_{\text{out}}),$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{K \times D}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^K$ are learnable parameters. This ternary fusion mechanism enhances robustness when clinical modalities disagree, as experts can explicitly model conflicting signals through negative contributions.

3.5 Training Objective

All components are optimized end-to-end with a joint objective that combines task supervision, quality calibration, selection regularization, and MoE balancing:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} \;+\; \alpha\, \mathcal{L}_{\text{uncert}} \;+\; \beta\, \mathcal{L}_{\text{sel}} \;+\; \gamma\, \mathcal{L}_{\text{bal}},$$

where $\mathcal{L}_{\text{task}}$ is the cross-entropy loss, $\mathcal{L}_{\text{uncert}}$ is the evidential uncertainty regularizer (Eq. (9)), $\mathcal{L}_{\text{bal}}$ encourages balanced expert utilization, and $\mathcal{L}_{\text{sel}}$ is a reward-guided differentiable term that biases the selector toward subsets that improve downstream accuracy while reducing average subset size (as in Eq. (18)). This formulation clarifies that improvements are not attributed to a single mechanism: calibration shapes q_m, selection turns q_m into patient-specific sparsity, and ternary routing stabilizes expert fusion under disagreement. Beyond simply reducing noise, this adaptive sparsity lowers the computational burden at inference time. Since it skips uninformative branches, QA-MoE represents an efficient solution for clinical environments that lack high-end hardware but require rapid triage.

Dataset	Modality	Available	Missing	Missing Rate
ADNI	Clinical	3,746	1,119	23.0%
	Genomic	757	4,108	84.4%
	Image	3,231	1,634	33.6%
	Biospecimen	2,639	2,226	45.8%
	Cognitive	4,861	4	0.1%
MIMIC-IV	Lab	447,690	98,082	18.0%
	Diagnosis	545,497	275	0.1%
	Demographic	218,171	327,725	60.0%

Table 1: Modality Missing Rates for ADNI and MIMIC-IV Datasets.

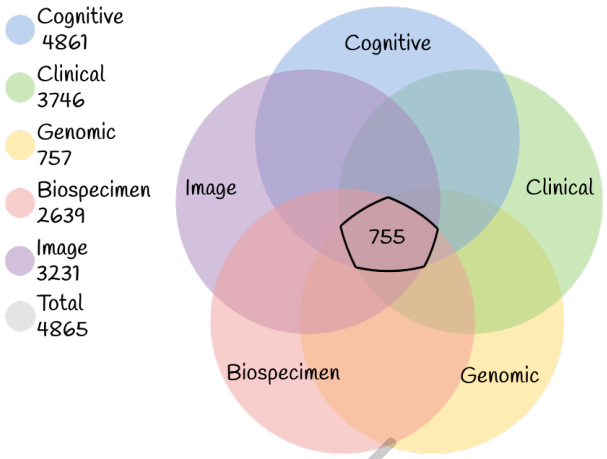

Figure 2: ADNI dataset modality availability statistics. The figure shows the number of patients with each modality and pairwise overlaps. Clinical and Cognitive modalities have the highest availability, while Genomic data is sparse. Only a small fraction of patients have all five modalities available, demonstrating the challenge of multimodal fusion in clinical settings.

4 Experiments

We evaluate QA-MoE on two complementary medical benchmarks: (i) ADNI for neurodegenerative disease staging, and (ii) MIMIC-IV for hospital Length-of-Stay (LOS) prediction. These datasets represent two distinct challenges. The first is heterogeneous multimodal quality in neuroimaging, while the second is missing or noisy clinical variables in EHR data.

4.1 ADNI: Alzheimer's Disease Progression (3-way Classification)

Task Setup We construct a uniquely comprehensive ADNI benchmark spanning five distinct modalities: MRI scans, cognitive assessments, clinical features, and biospecimen biomarkers. This multi-view selection captures a holistic profile of disease progression. We formulate a three-class classification task over CN (Cognitively Normal), MCI (Mild Cognitive Impairment), and AD (Alzheimer's disease).

Modalities We use the following pre-processing pipeline to encode each modality in the ADNI dataset:

- **Structural MRI**: 3D volume, encoded via 3D-ResNet18 $\rightarrow$ GAP
- **Clinical scores**: 50-d, encoded via MLP
- **Genomic SNPs**: 1000-d, encoded via MLP
- **Cognitive tests**: 30-d, encoded via MLP
- **CSF biomarkers**: 100-d, encoded via MLP

Dataset Statistics Our curated subset addresses the limitations of prior works by significantly expanding both modality coverage and patient count. Figure 2 and Table 1 summarize the data availability. Unlike standard subsets used in previous MoE studies, our pipeline successfully integrates nearly 80% more patients, providing a significantly more rigorous testbed for multimodal reliability. Figure 3 visualizes the learned class-dependent selection patterns.

4.2 MIMIC-IV: LOS Prediction (6-class)

Task Setup Unlike prior benchmarks that focus on binary mortality prediction within 24 hours of admission which often lacking practical utility for long-term resource planning, we address the more challenging and granular task of LOS prediction. We retain only the final admission per patient, excluding deceased and invalid entries. The continuous stay duration is discretised into six clinically relevant categories: $<$1 day (Class 0), 1–3 days (Class 1), 3–7 days (Class 2), 7–14 days (Class 3), 14–30 days (Class 4), and $>$30 days (Class 5). This formulation requires the model to distinguish between complex LOS ranges with varying class distributions, providing a rigorous test for multimodal reasoning.

Modalities

- **Lab tests**: 100-d, mean-aggregated 100 frequent tests
- **Diagnosis codes (ICD)**: 200-d, multi-hot of top 200 ICD codes
- **Demographics**: 10-d, expanded age + gender

Modalities are highly incomplete: lab panels are frequently missing ($\approx$ 30%), and ICD codes vary widely by case. While demographics exist at the patient level, our LOS setup operates at the admission level and requires joining multiple tables; consequently, demographic features are not available for all admissions after preprocessing (Table 1), and missing demographics are represented by a zero vector.

4.3 Baselines

We compare QA-MoE with a compact yet representative set of state-of-the-art multimodal baselines to evaluate the necessity of quality-aware expert routing under noisy and missing-modality conditions. We include dense multimodal Transformers, MMT [Xu *et al.*, 2023] (Multimodal Transformer) and Perceiver-IO [Jaegle *et al.*, 2021], as strong non-sparse baselines that rely on attention-based fusion and serve to assess whether MoE-style sparsity provides additional benefits. To directly compare against existing multimodal MoE designs, we evaluate FuseMoE [Han *et al.*, 2024], which performs modality-aware routing under missing inputs, Flex-MoE [Yun *et al.*, 2024], which extends MoE routing with partial-modality expert libraries and is particularly relevant for missing-modality completion, and TC-MoE [Yan *et al.*, 2025], which introduces ternary expert activation to stabilize MoE training. Finally, we consider robustness-oriented and

Category	Model	ACC↑	F1↑	AUC↑	ECE↓	FLOPs↓
MoE Routing	TC-MoE	0.6969 ± 0.0081	0.6990 ± 0.0068	0.8160 ± 0.0050	0.1620 ± 0.0036	0.3507G
Multimodal MoE	FuseMoE	0.6696 ± 0.0269	0.6721 ± 0.0254	0.7932 ± 0.0216	0.1358 ± 0.06025	0.3523G
	FlexMoE	0.6903 ± 0.0269	0.6929 ± 0.0254	0.8177 ± 0.0216	0.1288 ± 0.0593	0.3518G
Non-MoE	MMT	0.6947 ± 0.0095	0.6892 ± 0.0025	0.8337 ± 0.0066	0.1045 ± 0.0070	0.4172G
	MADDi	0.7040 ± 0.0038	0.7020 ± 0.0060	0.8390 ± 0.0010	0.0600 ± 0.0115	**0.0207G**
	Perceiver-IO	0.7051 ± 0.0036	0.7032 ± 0.0063	0.8403 ± 0.0008	0.0592 ± 0.0123	0.9508G
Quality	EDL	0.6321 ± 0.1220	0.5680 ± 0.1970	0.7195 ± 0.1559	0.0760 ± 0.0221	0.3431G
	QA-MoE	$\mathbf{0.8912 \pm 0.0140}$	$\mathbf{0.8914 \pm 0.0139}$	$\mathbf{0.9555 \pm 0.0088}$	$\mathbf{0.0224 \pm 0.0055}$	0.1742G

Table 2: Performance Comparison on ADNI Dataset.

Category	Model	ACC↑	Weighted-F1↑	AUC↑	ECE↓	FLOPs↓
MoE Routing	TC-MoE	0.4420 ± 0.0052	0.3920 ± 0.0205	0.8140 ± 0.0026	0.0600 ± 0.0088	0.0099G
Multimodal MoE	FuseMoE	0.3133 ± 0.0001	0.0795 ± 0.0001	0.5000 ± 0.0001	0.0981 ± 0.0006	0.0100G
	FlexMoE	0.4380 ± 0.0050	0.3880 ± 0.0190	0.8120 ± 0.0028	0.0640 ± 0.0085	0.0099G
Non-MoE	MMT	0.4850 ± 0.0058	0.4020 ± 0.0264	0.8180 ± 0.0019	$\mathbf{0.0477 \pm 0.0121}$	0.0111G
	MADDi	0.4744 ± 0.0017	0.3873 ± 0.0052	0.8182 ± 0.0011	0.0574 ± 0.0075	**0.0003G**
	Perceiver-IO	0.4780 ± 0.0042	0.3980 ± 0.0220	0.8150 ± 0.0020	0.0520 ± 0.0100	0.0253G
Quality	EDL	0.4400 ± 0.0055	0.3900 ± 0.0200	0.8150 ± 0.0025	0.0620 ± 0.0090	0.0019G
	QA-MoE	$\mathbf{0.5210 \pm 0.0139}$	$\mathbf{0.4550 \pm 0.0123}$	$\mathbf{0.8350 \pm 0.0048}$	0.0505 ± 0.0059	0.0034G

Table 3: Performance Comparison on MIMIC-IV LOS Prediction.

uncertainty-oriented fusion methods that do not employ expert routing, including MADDi [Golovanevsky *et al.*, 2022] for noise- and missingness-robust multimodal learning, EDL-based fusion [Sensoy *et al.*, 2018] for uncertainty-weighted integration, and Modality Dropout as a strong and widely adopted robustness baseline. Together, these baselines cover dense fusion, sparse MoE routing, and uncertainty-aware integration.

4.4 Implementation Details

Experiments were performed on an NVIDIA RTX 3090 GPU with a data split of 70% training, 15% validation, and 15% testing. We report the average results over three runs with different random seeds. Baselines are trained using the official settings from their respective papers to ensure fair comparison. regarding our proposed method, for ALFLB we use τ_{over}=2.0, τ_{under}=0.8, and step size γ_{bias} (ADNI: 0.02; MIMIC-IV LOS: 0.003) with $\tanh$ bias clipping. For the selector/SES, we set bootstrap samples B (ADNI: 100; MIMIC-IV LOS: 50), SES trigger probability (ADNI: 0.3; MIMIC-IV LOS: 0.1), Gaussian perturbation scale σ_{noise}=0.05, consensus threshold 80%, and inference threshold τ=0.5. Full parameter settings are available in the Supplementary Material.

4.5 Main Results

Performance on ADNI. As shown in Table 2, QA-MoE establishes a new state-of-the-art on the ADNI benchmark, achieving an accuracy of 89.12% and an AUROC of 95.55%. This represents a substantial margin over existing methods, outperforming the strongest MoE baseline (TC-MoE) by $+19.4\%$ in accuracy and the best non-MoE baseline (Perceiver-IO) by $+18.6\%$. Notably, QA-MoE achieves this superior performance with remarkable efficiency: its computational cost is approximately half that of TC-MoE and less than one-fifth of Perceiver-IO. Furthermore, our method demonstrates exceptional reliability, achieving the lowest ECE, indicating that its confidence estimates are highly aligned with true probabilities. The pronounced margin on ADNI occurs because extreme structural missingness degrades standard baselines, whereas QA-MoE explicitly prunes noisy paths.

The dramatic performance leap on ADNI highlights the efficacy of our quality-aware routing in high-dimensional, heterogeneous feature spaces. Unlike standard MoE approaches which treat all modalities as potentially useful, QA-MoE effectively filters out noise through its quality scoring mechanism. The learned subsets allow the model to focus computational resources on the most diagnostic signals while suppressing less informative or noisy modalities. This selective activation not only boosts predictive accuracy but also significantly reduces floating-point operations by bypassing irrelevant expert pathways.

Performance on MIMIC-IV. On the MIMIC-IV LOS prediction task (Table 3), QA-MoE consistently outperforms all baselines, achieving the highest accuracy, weighted-F1, and AUROC. Compared to the competitive MMT baseline, QA-MoE yields a solid improvement of $+3.6\%$ in accuracy and $+5.3\%$ in weighted-F1. In terms of efficiency, QA-MoE remains extremely lightweight, surpassing the efficiency of both Perceiver-IO and TC-MoE while maintaining superior predictive capability.

The MIMIC-IV results underscore the robustness of QA-

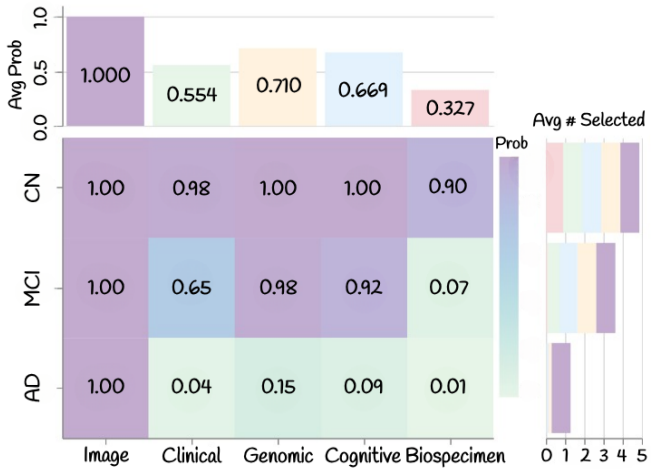

Figure 3: QA-MoE modality selection patterns on ADNI dataset. The visualization consists of three components: (top) average selection probability for each modality, (center) heatmap showing modality selection probability for each class (CN, MCI, AD), and (right) stacked bar chart showing the average number of modalities selected per class.

MoE in scenarios characterized by severe missingness and class imbalance. By explicitly modeling the quality of available modalities, QA-MoE stabilizes the routing process, ensuring that decision-making is driven by the most reliable clinical indicators. Notably, regarding calibration, while MMT achieves a marginally lower mean ECE, QA-MoE exhibits significantly superior stability, with a standard deviation less than half that of MMT. This indicates that our quality-driven approach prevents the large fluctuations in reliability estimates often observed in baselines, offering a more consistent and reproducible solution for clinical deployment.

4.6 Ablation Studies

Impact of Key Components. To systematically verify the contribution of each design module, we conduct an ablation study on the ADNI dataset, as summarized in Table 4. We sequentially remove the *Subset Selection* mechanism, the *Quality Scorer*, the *Ternary Routing* extension, and the auxiliary training objectives (*ALFLB* and *Stability Masking*).

The results clearly indicate that *Subset Selection* is the most critical component, with its removal leading to the largest performance drop. This validates our core hypothesis: in high-dimensional clinical spaces, strictly limiting the input scope to a coherent subset of modalities is more effective than soft weighting. The *Quality Scorer* is the second most influential factor, confirming that dynamically estimating modality reliability is essential for robust expert routing. Furthermore, removing the *ALFLB* regularization and *Ternary Extension* results in notable degradation, demonstrating that stable, mathematically grounded routing objectives are necessary to prevent the collapse often seen in sparse MoE training. Collectively, these ablations show that QA-MoE's performance is not driven by a single trick, but by the synergistic interaction between quality estimation and discrete, reliable expert selection.

Method	ACC↑	ΔACC
QA-MoE (full)	$\mathbf{0.8912 \pm 0.0140}$	–
– EDL Quality Scorer	0.8475 ± 0.018	↓4.9
– Subset Selection	0.8279 ± 0.019	↓7.1
– Ternary Extension	0.8582 ± 0.017	↓3.7
– ALFLB + Reward Loss	0.8519 ± 0.018	↓4.4
– SES Stability Masking	0.8733 ± 0.016	↓2.0

Table 4: Ablation study on ADNI dataset.

4.7 Modality Selection Analysis

Figure 3 visualizes the learned class-dependent selection policies. We observe a striking inverse correlation between disease severity and modality usage. For CN subjects, the model adopts a "holistic validation" strategy, activating nearly all available modalities to rigorously rule out pathology. In contrast, for AD, the router converges to a highly sparse, MRI-dominant regime. This suggests the model has learned that structural atrophy visible in MRI is sufficiently diagnostic for advanced disease, rendering noisy auxiliary signals redundant. Notably, MRI acts as a universal anchor, while weaker modalities are selectively pruned to maximize the signal-to-noise ratio.

5 Discussion

QA-MoE improves accuracy and calibration on ADNI and MIMIC-IV by better stabilizing multimodal fusion. Instead of relying on standard methods, we combine quality-aware scoring with ternary routing to separate useful diagnostic signals from noisy inputs. This leads to direct efficiency gains: because the model dynamically skips non-informative experts, QA-MoE beats state-of-the-art dense architectures with much fewer FLOPs. This suggests our approach is valid for resource-constrained clinical settings.

Despite these results, limitations remain. Although we achieved a theoretical reduction in FLOPs, we still need to measure end-to-end latency to understand the real-world hardware overhead of dynamic routing. Also, we only tested reliability on in-distribution data; future work needs to verify performance under out-of-distribution (OoD) shifts. Finally, distinguishing between a modality's reliability and its semantic usefulness is a complex problem that requires better interpretability tools.

6 Conclusion

In this work, we propose QA-MoE, a multimodal Mixture-of-Experts framework built to tackle reliability and scalability issues in medical data. Instead of standard fusion methods, we combine evidential quality scoring with a stability-enhanced routing mechanism. This creates a model that is both interpretable and robust, even when data is severely missing. Our evaluation on ADNI and MIMIC-IV shows that QA-MoE outperforms dense Transformers and uncertainty-based baselines in accuracy, while significantly cutting down computational costs. Given this balance of efficiency and performance, we believe QA-MoE is a viable option for deployment in clinical decision support systems.

References

[Abboud *et al.*, 2024] Zeinab Abboud, Herve Lombaert, and Samuel Kadoury. Sparse bayesian networks: Efficient uncertainty quantification in medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 675–684. Springer, 2024.

[Chang *et al.*, 2025] Jee Suk Chang, Hyunwook Kim, Eun Sil Baek, Jeong Eun Choi, Joon Seok Lim, Jin Sung Kim, and Sang Joon Shin. Continuous multimodal data supply chain and expandable clinical decision support for oncology. *npj Digital Medicine*, 8(1):128, 2025.

[Chen *et al.*, 2024] Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[Dai *et al.*, 2024] Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27445–27455, 2024.

[Dörrich *et al.*, 2025] Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hamed Mirbagheri, David J Fischer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, et al. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16(1):7163, 2025.

[Fan *et al.*, 2024] Lei Fan, Mingfu Liang, Yunxuan Li, Gang Hua, and Ying Wu. Evidential active recognition: Intelligent and prudent open-world embodied perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16351–16361, 2024.

[Fedus *et al.*, 2022] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 2022.

[Gao *et al.*, 2024] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26876–26885, 2024.

[Gliozzo *et al.*, 2025] Jessica Gliozzo, Mauricio A Soto Gomez, Arturo Bonometti, Alex Patak, Elena Casiraghi, and Giorgio Valentini. miss-snf: a multimodal patient similarity network integration approach to handle completely missing data sources. *Bioinformatics*, 41(4):btaf150, 2025.

[Golovanevsky *et al.*, 2022] Michal Golovanevsky, Carsten Eickhoff, and Ritambhara Singh. Multimodal attention-based deep learning for alzheimer's disease diagnosis. *Journal of the American Medical Informatics Association*, 29(12):2014–2022, 2022.

[Han *et al.*, 2024] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37:67850–67900, 2024.

[Hao *et al.*, 2025] Yan Hao, Chao Cheng, Juanjuan Li, Hongwen Li, Xingsi Di, Xiaoxia Zeng, Shoumei Jin, Xiaodong Han, Chongsong Liu, Qianqian Wang, et al. Multimodal integration in health care: Development with applications in disease management. *Journal of medical Internet research*, 27:e76557, 2025.

[Jaegle *et al.*, 2021] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[Johnson *et al.*, 2023] Alistair E. W. Johnson, Tom J. Pollard, Roger G. Mark, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 2023.

[Kim and Kim, 2024] Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pages 171–187. Springer, 2024.

[Le *et al.*, 2025] Lien P Le, Thu Nguyen, Michael A Riegler, Pål Halvorsen, and Binh T Nguyen. Multimodal missing data in healthcare: A comprehensive review and future directions. *Computer Science Review*, 56:100720, 2025.

[Li *et al.*, 2025] Tianyu Li, Yan Xin, et al. Moe-ce: Enhancing generalization for deep learning based channel estimation via a mixture-of-experts framework. *arXiv preprint arXiv:2509.15964*, 2025.

[Liu *et al.*, 2025] Yuanye Liu, Zheyao Gao, Nannan Shi, Fuping Wu, Yuxin Shi, Qingchao Chen, and Xiahai Zhuang. Merit: Multi-view evidential learning for reliable and interpretable liver fibrosis staging. *Medical Image Analysis*, 102:103507, 2025.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Mehmet Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[Sidheekh *et al.*, 2024] Sahil Sidheekh, Pranuthi Tenali, Saurabh Mathur, Erik Blasch, and Sriraam Natarajan. On the robustness and reliability of late multi-modal fusion using probabilistic circuits. In *2024 27th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2024.

[Tocchetti and Mancini, 2024] Tommaso T Tocchetti and Massimiliano Mancini. Credal deep ensembles for uncertainty quantification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[Wang and Yang, 2025] Xiaoyang Wang and Christopher Yang. Moe-health: A mixture of experts framework for robust multimodal healthcare prediction. In *Proceedings of the 16th ACM International Conference on Bioinformatics,*

Computational Biology, and Health Informatics, pages 1–9, 2025.

[Wang *et al.*, 2025a] Hongyu Wang, Jiayu Xu, Ruiping Wang, Yan Feng, Yitao Zhai, Peng Pei, Xunliang Cai, and Xilin Chen. Mote: Mixture of ternary experts for memory-efficient large multimodal models. *arXiv preprint arXiv:2506.14435*, 2025.

[Wang *et al.*, 2025b] Muyu Wang, Shiyu Fan, Yichen Li, Zhongrang Xie, and Hui Chen. Missing-modality enabled multi-modal fusion architecture for medical data. *Journal of Biomedical Informatics*, 164:104796, 2025.

[Wang *et al.*, 2025c] Ziteng Wang, Jun Zhu, and Jianfei Chen. Remoe: Fully differentiable mixture-of-experts with relu routing. 2025.

[Weiner *et al.*, 2010] Michael W. Weiner, Danielle P. Veitch, Paul S. Aisen, et al. The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 2010.

[Xu *et al.*, 2023] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[Yan *et al.*, 2025] Shen Yan, Xingyan Bin, Sijun Zhang, Yisen Wang, and Zhouchen Lin. Tc-moe: Augmenting mixture of experts with ternary expert choice. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Yao *et al.*, 2024] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16416–16424, 2024.

[Yun *et al.*, 2024] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

[Zhang *et al.*, 2025] Baoyi Zhang, Zhuoya Wan, Yige Luo, Xi Zhao, Josue Samayoa, Weilong Zhao, and Si Wu. Multimodal integration strategies for clinical application in oncology. *Frontiers in Pharmacology*, 16:1609079, 2025.