R^2MoE: Representation and Expert Selection Dual-Regularized Mixture-of-Experts for Multimodal Clinical Data

Supplementary Material

1. Expert Activation Patterns

As discussed in Sec.3.3.2, our key idea is that although the fused representation $\mathbf{z}_{\text{fuse}}$ encodes the semantic information derived from a patient's multimodal inputs, it does not fully describe how the Mixture-of-Experts (MoE) layer processes that patient. The routing weights of the gate carry complementary information about which Specialized and Flex Experts are engaged, and this routing behaviour can reveal stable, subtype-specific computation patterns. To capture both aspects, R^2MoE integrates three complementary components: (i) expert selection regularization, (ii) balanced secondary expert utilization, and (iii) a routing-based descriptor, the Gating Weight Fingerprint (GWF).

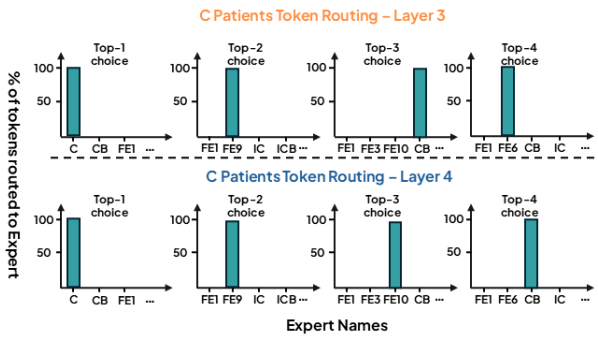
Top-1 specialization and $\mathcal{L}_{\text{SE}}$. We specifically anchor the Top-1 expert, as it receives the highest routing weight and thus contributes most significantly to the fused representation. Anchoring secondary selections (e.g., Top-2 or beyond) would reduce routing flexibility and require additional specialized experts, increasing computational cost while limiting the model's ability to learn shared cross-modality patterns. By constraining only the primary expert and leaving secondary routing unconstrained, the model achieves a balance between stable specialization and flexible expert collaboration. Fig. 1 visualizes the Top-1 to Top-4 expert selection histograms for all six observed modality combinations (C, CB, IC, GCB, ICB, IGCB) at layers 3 and 4. Each panel reports, for a given combination, the percentage of tokens assigned to each expert. The Top-1 bars confirm that tokens from each modality combination are consistently routed to the correct Specialized Expert (SE), e.g., CB tokens to the CB expert, IGCB tokens to the IGCB expert, and so on. This is further corroborated by Table 1, which quantifies this behaviour: each modality combination is routed to its designated SE (at top-1) more than 97% of the time, demonstrating highly stable and modality-combination-consistent Top-1 specialization. This behaviour directly reflects the effect of the expert-selection loss $\mathcal{L}_{\text{SE}}$, which anchors the top-1 choice to the designated SE for that modality combination. Without this constraint, we observed that the gate often sends the same combination to multiple SEs, preventing each expert to learn a coherent function, thus weakening specialization and leading to inconsistent outputs. The concentration of Top-1 mass on the correct SE in Fig. 1 empirically validates that $\mathcal{L}_{\text{SE}}$ successfully enforces the intended specialization.

Secondary routes and $\mathcal{L}_{\text{balance}}$. Beyond the primary SE choice, the Top-2–Top-4 bars in Fig. 1 show that each modality combination relies on only a *small* subset of secondary experts. For example, IC patients at layer 4 predominantly use FE-4, FE-5, and FE-7 as secondary routes, while other combinations activate different expert subsets. These combination-specific co-activation patterns are desirable because they reflect meaningful structure within and across modality combinations, but they also pose a risk: without additional regularization, the gate could overuse a few experts and ignore the rest, leading to expert under-utilization and reduced robustness. To counteract this, we introduced $\mathcal{L}_{\text{balance}}$, which explicitly encourages the Top-2–4 traffic to be spread across the secondary experts. Table 2 reports the activation percentages for all experts at secondary choice, showing that, after training all the samples with $\mathcal{L}_{\text{balance}}$, the average utilization of experts is well-distributed rather than collapsing onto a single dominant expert. Thus, $\mathcal{L}_{\text{SE}}$ and $\mathcal{L}_{\text{balance}}$ work together: the former fixes which SE owns each modality combination (Top-1), while the latter keeps the shared pool healthy and diverse for secondary routing.

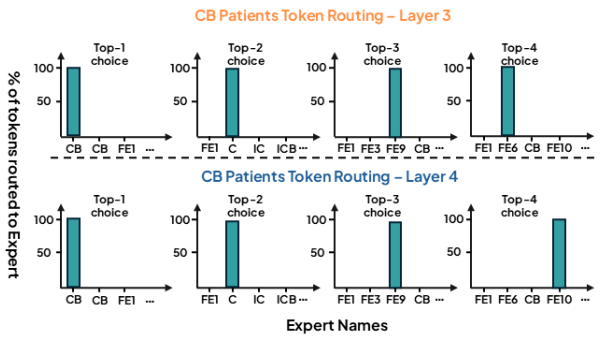
Table 1. **Specialized Top-1 routing.** Percentage of tokens routed to the correct Specialized Expert (SE) at Top-1 for each modality combination.

Modality Combination	SE Top-1 Activation (%)
C	98.8
CB	97.7
IC	99.4
GCB	99.4
ICB	99.6
IGCB	97.0

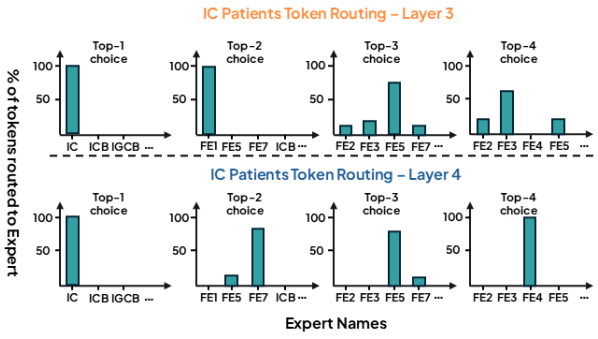
From secondary patterns to GWF. The same histograms in Fig. 1 also motivated our use of the Gating Weight Fingerprint (GWF). Within each modality combination, the Top-2–4 distributions are not random; instead, they form stable co-activation profiles that are consistent across patients and tokens of the same combination. In other words, patients with the same combination (e.g., IC or GCB) tend to reuse a characteristic subset of experts with similar relative weights, and different combinations exhibit different signatures. These secondary patterns are precisely the high-order routing structure that $\mathbf{z}_{\text{fuse}}$ alone cannot capture. We therefore aggregate the normalized Top-2–4 gating weights from the last MoE layers into a compact vector, the GWF,



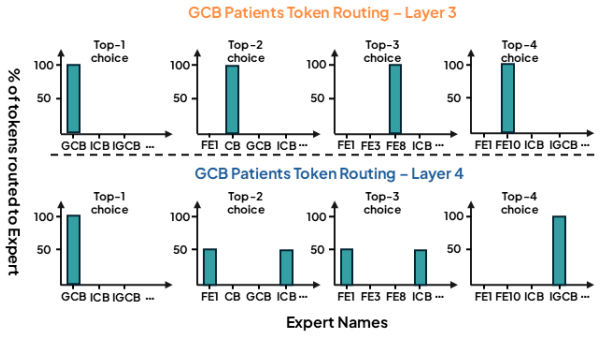
(a) C patients



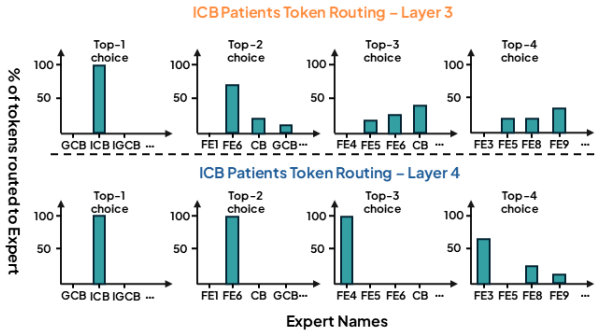
(b) CB patients



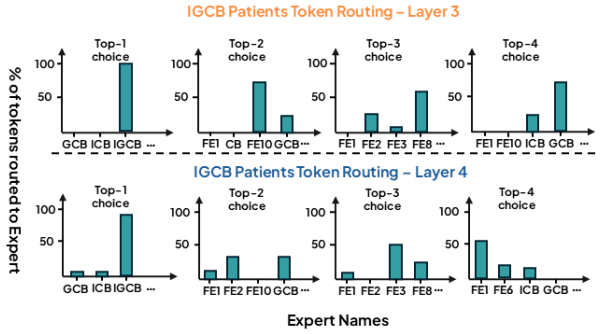
(c) IC patients



(d) GCB patients



(e) ICB patients



(f) IGCB patients

Figure 1. **Top-k expert selection patterns across all modality combinations.** For each modality combination (C, CB, IC, GCB, ICB, IGCB), we show the token-level expert selection histograms at the 3^{rd} and 4^{th} MoE layers for Top-1 to Top-4 choices. Across all cases, the Top-1 bar confirms that the correct Specialized Expert (SE) is consistently selected for each corresponding modality combination, while the secondary choices (Top-2–Top-4) exhibit stable, combination-specific co-activation patterns that underline the Gating Weight Fingerprint.

and concatenate it with $\mathbf{z}_{\text{fuse}}$ to form the final representation. Empirically, this routing-aware descriptor improves discrimination and calibration, especially under missing-modality settings, by encoding how the model chooses to process each patient rather than only what is encoded in the fused feature space.

In summary, Fig. 1 and Table 1 and 2 jointly illustrate the motivations behind our design: $\mathcal{L}_{\text{SE}}$ secures a consistent mapping from modality combination to SE, $\mathcal{L}_{\text{balance}}$ maintains healthy utilization of experts for secondary routes, and GWF turns the resulting stable co-activation patterns into explicit, subtype-aware signal improving the representation.

2. Additional Analysis

We report the computational cost of R^{2}MoE compared to prior multimodal models in Table 3. The reported mean time corresponds to per-iteration training time, including routing overhead. Importantly, R^{2}MoE does not introduce additional routing stages beyond standard MoE architectures. R^{2}MoE achieves competitive computational efficiency while maintaining improved routing stability and predictive performance. Despite introducing additional regularization, the model incurs only a modest increase in runtime compared to FlexMoE, while remaining significantly

Table 2. **Top-2–Top-4** activations across all experts (specialized + flexible), illustrating expert diversity promoted by $\mathcal{L}_{\text{balance}}$.

Secondary Expert	Top-2–Top-4 Activation (%)
C	6.25
CB	4.66
GCB	8.73
IC	5.35
ICB	4.66
IGCB	5.92
FE1	9.24
FE2	6.49
FE3	9.13
FE4	7.22
FE5	8.45
FE6	8.27
FE7	6.42
FE8	7.75
FE9	8.73
FE10	9.10

Table 3. Compute cost comparison of Mean Time, GFLOPs, and # of parameters.

Metric	MulT	MAG	LIMOE	FuseMoE	Flex-MoE	R^{2}MoE
Mean Time (s)	38.70	16.04	17.96	20.71	16.00	19.37
GFLOPs	60.12	59.06	59.41	59.76	59.07	59.24
# Params	46.4M	36.5M	38.6M	340.9M	36.9M	38.6M

more efficient than large-capacity models such as FuseMoE.

We also analyze the sensitivity of R^{2}MoE to key architectural hyperparameters, including the number of experts, the number of MoE layers, and the Top-k routing parameter.

As shown in Fig. 2, we found out that while increasing the number of experts initially improves model capacity, excessive experts (e.g., 32) can lead to degraded performance due to increased complexity and routing instability. The best performance is achieved with 16, showing balance between capacity and specialization.

For the number of MoE layers, performance exhibits a non-monotonic trend. While increasing depth from 2 to 4 layers improves accuracy, shallow configurations (e.g., 2 layers) underperform due to insufficient modeling capacity. The best performance is achieved with deeper configurations, indicating that hierarchical expert routing can improve representation learning when appropriately scaled.

Finally, the Top-k parameter controls the number of active experts per token. We observe that performance improves as Top-k increases up to 4, after which it declines. This suggests that moderate sparsity (Top-$k = 4$) provides the best trade-off between expressivity and specialization. Lower values restrict model capacity, while higher values dilute expert specialization.

Overall, R^{2}MoE achieves optimal performance under moderate model capacity, sufficient depth, and controlled sparsity, and remains robust across a range of hyperparameter settings.

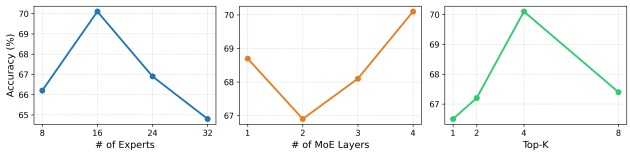


Figure 2. Sensitivity analysis of R^{2}MoE with respect to key hyperparameters. From left to right: number of experts, number of MoE layers, and Top-k routing.