This enables rapid persona switching with minimal overhead. We optionally support a soft gating mechanism

$$G = \mathbf{M}^p + \gamma(1 - \mathbf{M}^p), \quad \gamma \in [0, 1), \tag{7}$$

where $\gamma = 0$ corresponds to standard hard masking. This enables efficient persona switching without modifying model parameters or caches.

Contrastive Pruning Many personas naturally form semantic oppositions (e.g., power-seeking vs. power-rejecting). Standard pruning methods may yield overlapping masks that fail to capture contrasting behaviors, as they optimize importance scores independently for each persona. We therefore introduce a contrastive pruning strategy that explicitly maximizes parameter separation between opposing personas (p_+, p_-) by leveraging differential activation patterns.

For opposing persona pairs, we collect activation statistics $\mu_{ij}^{p_+}, \mu_{ij}^{p_-}$ and variances $\sigma_{ij}^{p_+}, \sigma_{ij}^{p_-}$ across their respective calibration data. In the contrastive Wanda variant, importance is defined by scaling weight magnitudes with standardized activation differences:

$$S_{ij}^p = |w_{ij}| \cdot \phi\left(\frac{\mu_{ij}^{p_+} - \mu_{ij}^{p_-}}{\sqrt{\sigma_{ij}^{p_+} + \sigma_{ij}^{p_-}} + \varepsilon}\right), \tag{8}$$

where ϕ is a monotonic activation function (e.g., ReLU or softplus) and ε is a small constant for numerical stability. In the contrastive-Sparse variant, we normalize importance scores column-wise and compute contrastive importance as:

$$C_{ij} = \left|\tilde{S}_{ij}^{p_+} - \tilde{S}_{ij}^{p_-}\right|, \quad \tilde{S}_{ij}^p = \frac{S_{ij}^p}{\sum_k S_{ik}^p}, \tag{9}$$

Parameters are then ranked according to their persona-specific scores and assigned to disjoint masks $\mathbf{M}^{p_+}, \mathbf{M}^{p_-}$, with each parameter allocated to the persona exhibiting the larger score, encouraging the distribution divergence of local neurons that could result in personality separation. This procedure can be interpreted as minimizing an overlap regularizer $\Omega(\mathbf{M}^{p_+}, \mathbf{M}^{p_-})$ alongside the calibration loss. By explicitly modeling opposition rather than treating personas independently, both variants encourage divergent subnetworks that achieve superior specialization for contrasting behaviors while maintaining computational efficiency through training-free mask construction.

Note that our local contrastive constraint that encourages the pruning procedure to make explicit trade-offs between competing activations, assigning each high-importance parameter to the persona for which it is most informative. Crucially, this does not imply that the underlying activation distributions are non-overlapping or that the entire resulting subnetwork structure is orthogonal, especially since shared components like the LM head remain unpruned. For example, if two different personality subnetworks have a sparsity of 40%, there will be overlap between the two subnetworks. We have analyzed the similarity and overlap between different personas in Section 3.4 and the Appendix G.

3.4 Mask Analysis

While the pruning method successfully elicits distinct personas on Llama 2-13B, we observe a notable variance in success rates: some personas are clearly differentiated, while others remain entangled or fail to emerge. To investigate the underlying mechanisms, we analyze the pruned subnetworks along two axes: (i) mask-level separation and (ii) layer-wise representation similarity. Table 2 provides the average differential ratios across MBTI dimensions, serving as a measure of mask-level separation. We observe that I/E and F/T pairs exhibit substantially higher divergence than N/S and J/P, suggesting that certain dimensions are more strongly encoded in the model's internal representations. Moreover, across all dimensions, differences are consistently larger in MLP blocks than in attention layers, indicating that persona separation primarily relies on feed-forward transformations rather

Dimension Pair	Avg. Diff. (%)	Attn	MLP
I vs. E	1.34	1.28	1.44
N vs. S	0.75	0.75	0.76
F vs. T	1.08	1.03	1.14
J vs. P	0.76	0.73	0.79

Table 2: Average differential mask ratios across MBTI dimensions.