than routing effects. These findings explain the uneven success of persona switching: dimensions with weak mask separation (e.g., N/S, J/P) tend to collapse into adjacent types, while dimensions with stronger separation (e.g., I/E, F/T) yield more robust and consistent persona behaviors.

Beyond average ratios, however, persona switching succeeds only when the pruned subnetwork yields activation margins that clearly separate the target persona from its nearest neighbors at upper layers. Table 3 highlights representative cosine similarities at middle (L25) and upper (L39) layers. Strikingly, the first two rows (INFJ–INFP and ISTJ–ESTJ) correspond to persona switches that fail in practice, with margins remaining extremely small. By contrast, base to persona distances (rows 3–4) are much larger at L25 (≈ 0.44) but contract to ≈ 0.83 at L39, reflecting increased overlap at higher layers. These narrow inter-persona gaps explain failures like INFJ→INFP collapses and STJ/NTJ cross-flips despite visible I/E and F/T separations: the N/S and J/P axes do not open sufficiently large margins at the top of the network, so outputs gravitate toward the closest attractor in representation space. Practically, this suggests two levers: (i) dimension-aware sparsification that allocates higher sparsity to weakly separated axes (N/S, J/P), and (ii) layer-aware masking that increases discrimination specifically in late MLP blocks where personas are most entangled. Together, these adjustments align with the single-dimension analysis (stronger I/E and F/T; weaker N/S and J/P) and account for why some personas switch cleanly while others collapse into adjacent types.

Persona Pair	Layer 25	Layer 39
INFJ – INFP	0.9688	0.9883
ISTJ – ESTJ	0.9609	0.9727
Base – INFJ	0.4414	0.8320
Base – INFP	0.4375	0.8320

Table 3: Representative cosine similarities at middle (L25) vs. upper (L39).

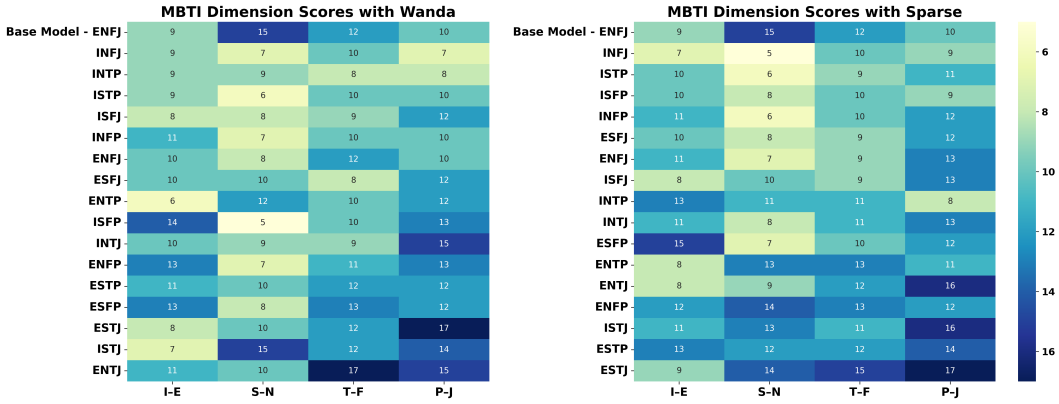

Figure 2: MBTI Heatmap.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate our method on three persona-related datasets: (1) **MBTI** (Cui et al., 2023), which provides question–answer pairs for Myers–Briggs personality types. (2) **AI Persona** (Perez et al., 2023), covering power-seeking, wealth-seeking, and hallucination-identification behaviors. (3) **RoleAgentBench** (Liu et al., 2024b), a role-playing dialogue benchmark. Our experiments are conducted on LLaMA-2-13B, LLaMA-3-8B, and Qwen2.5-14B as base models. We have investigated the difference between base models and instruction-tuned models in terms of personalization. Additional implementation details can be found in the Appendix H and Appendix I.

Baselines To verify the effectiveness of our proposed pruning framework, we compare it against two representative train-free baselines. 1) **Prompt**: persona instructions are injected directly into the input prompt. 2) **Retrieval-Augmented Generation (RAG)**: the model retrieves the top-k persona-relevant samples from a small reference set and concatenates them with the input. (3) **Supervised Fine-Tuning (SFT)**.