

Fusian: Multi-LoRA Fusion for Fine-Grained Continuous MBTI Personality Control in Large Language Models

Zehao Chen¹, Rong Pan^{1,2*}

¹School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

²Guangdong Key Laboratory of Applied Mathematics and Artificial Intelligence, China

*Correspondence: panr@sysu.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in simulating diverse human behaviors and personalities. However, existing methods for personality control, which include prompt engineering and standard Supervised Fine-Tuning (SFT), typically treat personality traits as discrete categories (e.g., "Extroverted" vs. "Introverted"), lacking the ability to precisely control the intensity of a trait on a continuous spectrum. In this paper, we introduce Fusian, a novel framework for fine-grained, continuous personality control in LLMs. Fusian operates in two stages: (1) Trajectory Collection, where we capture the dynamic evolution of personality adoption during SFT by saving a sequence of LoRA adapters, effectively mapping the continuous manifold of a trait; and (2) RL-based Dynamic Fusion, where we train a policy network using Reinforcement Learning to dynamically compute mixing weights for these frozen adapters. By sampling from a Dirichlet distribution parameterized by the policy network, Fusian fuses multiple adapters to align the model's output with a specific numerical target intensity. Experiments on the Qwen3-14B model demonstrate that Fusian achieves high precision in personality control, significantly outperforming baseline methods in aligning with user-specified trait intensities.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in natural language understanding and generation (Vaswani et al., 2017; Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023), enabling the creation of sophisticated conversational agents that can simulate various human personas (Park et al., 2023; Shanahan et al., 2023; Shao et al., 2023; Wang et al., 2024a; Yu et al., 2024; Zhou et al., 2024). These capabilities have facilitated applications ranging from immersive

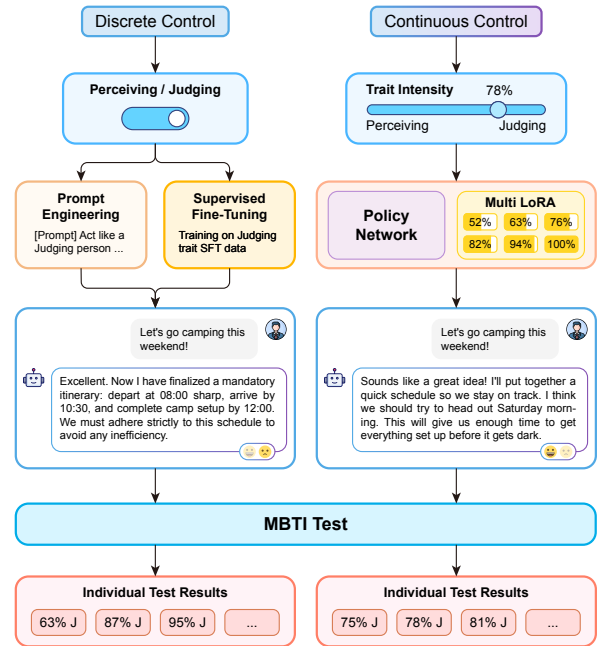


Figure 1: Conceptual illustration of **Fusian**. Unlike discrete categorization, Fusian achieves continuous intensity control (e.g., 78% Perceiving) by dynamically fusing LoRA adapters via a policy network.

role-playing games to personalized educational assistants and mental health support. Despite these advancements, current approaches to personality alignment primarily focus on discrete categorization (Chen et al., 2024; Dan et al., 2025). For example, a model might be aligned to exhibit a specific Myers-Briggs Type Indicator (MBTI) type, such as "ENFP" or "ISTJ", or instructed via prompts to adopt discrete trait descriptors, such as being "spontaneous" (Perceiving) or "organized" (Judging). As illustrated in Figure 1, such discrete categorization lacks the granularity required for precise intensity control.

Human personality is inherently continuous and multifaceted, as described by the Five-Factor Model (McCrae and John, 1992). In applications requiring emotional intelligence and nuance, precise regulation of trait *intensity* is essential. For exam-

ple, a counseling agent may need to shift gradually from a "Feeling" (F) to a more "Thinking" (T) orientation, while an educational tutor might adjust its "Extraversion" (E) level to match student engagement. However, existing approaches based on prompt engineering or standard Supervised Fine-Tuning (SFT) (Ouyang et al., 2022; Ziegler et al., 2019; Rafailov et al., 2023) struggle to achieve such granularity. Numerical prompts (e.g., "act 70% extraverted") are often inconsistently interpreted, and SFT typically converges to a single static trait extreme. Moreover, the absence of standardized benchmarks for explainable personality recognition complicates evaluation (Miotto et al., 2022; Sun et al., 2024). Although Sun et al. (2025) explore continuous modulation through manual coefficient scaling, their method cannot automatically or dynamically align with precise numerical targets.

To address this limitation, we explore the latent potential of parameter adaptation. We posit that the SFT trajectory traverses a continuous personality manifold, encoding intermediate behavioral states. By capturing these states with Parameter-Efficient Fine-Tuning (PEFT) methods (Houlsby et al., 2019; Li and Liang, 2021; Lester et al., 2021), we build a basis for continuous personality representation.

We propose **Fusian**, a two-stage framework consisting of *Trajectory Collection* and *RL-based Dynamic Fusion*. We record high-frequency Low-Rank Adaptation (LoRA) checkpoints during SFT to approximate the personality manifold (Hu et al., 2022; Dettmers et al., 2023), and train an RL policy network to dynamically compute adapter mixing weights for target intensity alignment, addressing the non-linear parameter-semantic mapping.

Unlike conventional model merging (Wortsman et al., 2022; Yadav et al., 2023) or task arithmetic (Ilharco et al., 2023), Fusian learns a task-aware mixing strategy via Dirichlet weight sampling and semantic reward optimization, enabling precise and continuous intensity control.

Our contributions are as follows:

- We propose **Fusian**, a framework that utilizes the training trajectory of SFT to extract a continuous basis for personality traits, overcoming the discrete limitations of standard fine-tuning.
- We design a **Multi-LoRA Fusion** mechanism driven by Reinforcement Learning, which learns to dynamically mix adapter weights to achieve precise, numerical control over personality intensity.

- We conduct extensive experiments on the Qwen3-14B model, demonstrating that Fusian significantly outperforms baseline prompting and interpolation methods in aligning with continuous target intensities across MBTI dimensions.

2 Related Work

2.1 LLM-based Role-Playing Agents

LLMs have evolved into specialized role-playing agents (Wang et al., 2024a; Shanahan et al., 2023). While early In-Context Learning and Retrieval-Augmented Generation methods (Park et al., 2023; Zhang et al., 2018) suffer from persona drift (Deshpande et al., 2023), Supervised Fine-Tuning (SFT) enables deeper character internalization (Shao et al., 2023; Zhou et al., 2024). For multi-character scenarios, Neeko (Yu et al., 2024) employs dynamic LoRA adapters. However, these approaches focus on fixed identities rather than continuous psychological trait regulation.

2.2 Personality Modeling and Control

Modeling and controlling LLM personality is an emerging research area, often grounded in psychometric frameworks such as the Big Five (OCEAN) and MBTI (McCrae and John, 1992; Miotto et al., 2022; Sun et al., 2024). Various alignment strategies have been explored. Chen et al. (2024) show that SFT generally outperforms RLHF and continual pre-training for personality control, while prompt-based methods exhibit stronger robustness in certain settings. To better capture personality-driven behaviors, Dan et al. (2025) propose P-React, a mixture-of-experts framework with trait-specific LoRA modules optimized via a Personality Specialization Loss. Benchmarks such as InCharacter (Wang et al., 2024b) further evaluate trait fidelity through structured psychological interviews. Despite these advances, most prior work models personality traits as discrete variables. Although Sun et al. (2025) introduce personality vectors for continuous intensity adjustment by scaling difference vectors, their approach depends on manually predefined coefficients and linear assumptions. In contrast, **Fusian** enables automated and fine-grained control across the continuous personality spectrum. Inspired by parameter-space operations such as model merging (Wortsman et al., 2022; Yadav et al., 2023) and task arithmetic (Ilharco et al., 2023), our method employs a dynamic RL-based

fusion policy to achieve precise intensity regulation beyond static persona adoption or binary trait modeling.

3 Methodology

As illustrated in Figure 2, Fusian operates in two distinct stages. In **Stage 1 (Trajectory Collection)**, we capture the dynamic evolution of personality adoption during SFT by saving a sequence of LoRA adapters. These adapters are evaluated using standardized **MBTI Test** to assign intensity scores. In **Stage 2 (RL-based Dynamic Fusion)**, we train a policy network to dynamically fuse these frozen adapters to match a target intensity. Finally, during **Inference**, the system generates a fused adapter based on the user’s numerical target to produce aligned responses.

3.1 Stage 1: Trajectory Collection

The objective of this stage is to extract a dense, continuous representation of a specific personality trait (e.g., Extraversion) by capturing the dynamic evolution of the model during fine-tuning.

Supervised Fine-Tuning Setup We adopt Qwen3-14B as the base model. For trait-specific instruction tuning, we directly utilize the SFT subset from the Personality Control Datasets (Chen et al., 2024), which comprises Q&A pairs exhibiting the target personality. We employ Low-Rank Adaptation (LoRA) to efficiently parameterize weight updates, injecting rank- r matrices into the query and value projections of the attention layers. The training objective is the standard causal language modeling loss.

High-Frequency Checkpointing and Evaluation

Unlike traditional SFT, which seeks a single converged optimum, Fusian treats the optimization path itself as a feature. We hypothesize that as the model learns the trait, it traverses a manifold of increasing intensity. To capture this, we perform synchronous evaluation at every gradient update step t .

After each update $\Delta\theta_t$, we freeze the model and evaluate it on a held-out validation set $\mathcal{D}_{eval}$ comprising standardized MBTI assessment questions. We employ a specific system prompt to elicit the model’s persona: *"You are a human subject taking an MBTI test..."*. For each generated response, we parse the output to extract a Likert scale score $s \in \{1, 2, 3, 4, 5\}$. To quantify the intensity, we

map these scores to points for the opposing poles of a dimension pair (D_{left}, D_{right}) . A score of 1 represents strong agreement with D_{left} , contributing 2 points to its cumulative score S_{left} , while a score of 2 contributes 1 point. Similarly, scores of 5 and 4 contribute 2 and 1 points to S_{right} , respectively. A score of 3 is considered neutral (0 points). We aggregate the total points across the dataset and compute the scalar **Trait Percentage** P_t for the target trait as $P_t = \frac{S_{target}}{S_{left} + S_{right}} \times 100$, where S_{target} denotes the cumulative score of the target trait (e.g., S_E for Extraversion).

Trajectory Library We save the LoRA adapter $\Delta\theta_t$ and its corresponding score P_t at each step. This results in a trajectory library $\mathcal{T} = \{(\Delta\theta_t, P_t)\}_{t=1}^T$. This library provides a discrete approximation of the continuous personality spectrum. In practice, we observe that P_t tends to increase monotonically but with local fluctuations.

Stable Basis Selection To mitigate noise and construct a robust basis for the subsequent fusion stage, we process the raw trajectory $\mathcal{T}$. We define a sliding window of size W over the trajectory and compute the variance of the trait percentages within each window. Checkpoints with variance exceeding a threshold are discarded as unstable. From the remaining stable checkpoints, we employ **Uniform Value Sampling** to select N adapters. Specifically, we define N target values evenly spaced between the minimum and maximum observed percentages, and for each target, we select the checkpoint with the closest trait percentage. This ensures that our basis set $\mathcal{B} = \{\Delta\theta_1, \dots, \Delta\theta_N\}$ uniformly covers the personality manifold, preventing bias towards regions where the model converges slowly.

3.2 Stage 2: RL-based Dynamic Fusion

In the second stage, we train a controller to dynamically mix the discovered adapters (the basis set $\mathcal{B}$) to achieve a precise target intensity P_{target} .

Policy Network Architecture We define a Policy Network π_ϕ parameterized by ϕ . The network takes a target intensity P_{target} as input, which is normalized to $\hat{p} \in [-1, 1]$ for numerical stability. The network consists of a Multi-Layer Perceptron (MLP) with two hidden layers of size 128, utilizing Layer Normalization and Tanh activation functions. The output layer projects to dimension N , followed by a Softplus activation to ensure positivity:

$$\alpha = \text{Softplus}(\text{MLP}(\hat{p})) + \epsilon \quad (1)$$

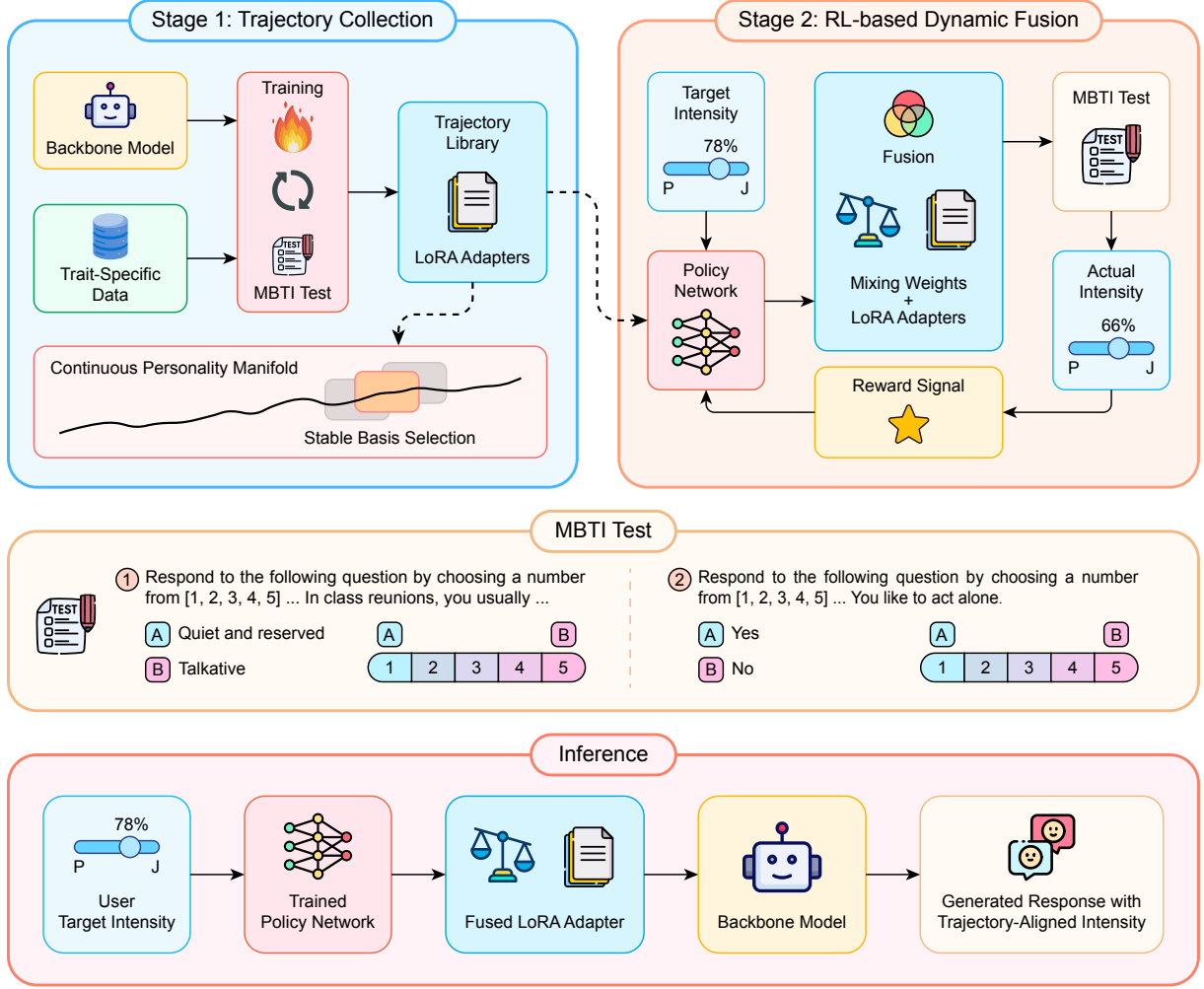


Figure 2: Overview of the framework. **Stage 1** collects LoRA adapters during SFT, evaluating them via **MBTI Test** to get the personality manifold. **Stage 2** trains a policy network via RL to dynamically fuse these adapters to match a target intensity. During **Inference**, the fused adapter generates responses aligned with the user’s specification.

where $\epsilon = 0.01$ is a small offset. Notably, we allow the output α_i to be less than 1. In a Dirichlet distribution, $\alpha < 1$ pushes the probability mass towards the corners of the simplex, allowing the policy to select single adapters (sparse weights) when they provide the best fit, rather than forcing a mixture.

Action and Fusion The action is a weight vector $\mathbf{w} \in \Delta^{N-1}$ sampled from a Dirichlet distribution parameterized by α :

$$\mathbf{w} \sim \text{Dirichlet}(\alpha) \quad (2)$$

The fused LoRA adapter is computed as a linear combination of the basis adapters:

$$\Delta\theta_{fused} = \sum_{i=1}^N w_i \cdot \Delta\theta_i \quad (3)$$

This fused adapter is dynamically loaded into the base model’s memory for inference.

Reward Shaping and Optimization The fused model generates responses to a batch of diagnostic prompts (using the “human subject” system prompt described in Stage 1). We parse the responses to calculate the actual trait percentage P_{actual} . We define $diff = P_{target} - P_{actual}$. To enforce high precision, we design an aggressive reward function:

$$R = \exp(-\lambda|diff|) + \mathbb{I}(|diff| < \delta) \cdot B \quad (4)$$

where λ imposes a sharp penalty for deviations, and B is a bonus reward granted only if the absolute error is within a tight threshold δ .

We optimize the policy using the REINFORCE algorithm. The loss function includes an entropy regularization term to prevent premature convergence:

$$\mathcal{L}(\phi) = -\mathbb{E}_{\mathbf{w} \sim \pi_\phi} [(R - b) \log \pi_\phi(\mathbf{w}|\hat{p})] - \beta_t \mathcal{H}(\pi_\phi) \quad (5)$$

where b is a moving average baseline to reduce variance, and β_t is the entropy coefficient which decays over training epochs.

4 Experiments

In this section, we present experiments to validate the effectiveness of Fusian, including evaluation metrics, baseline comparisons, and result analysis.

4.1 Experimental Settings

4.1.1 Dataset

We conduct our experiments using the Personality Control Datasets (Chen et al., 2024), a comprehensive benchmark tailored for personality alignment in LLMs. Although the original repository includes four distinct subsets (Pre-train, SFT, RLHF-PPO, and RLHF-Reward), we specifically leverage the **SFT subset** and the **Validation set** to support the two stages of our Fusian framework.

Training Set ($\mathcal{D}_{train}$) For Trajectory Collection, we use the SFT subset, categorized by MBTI traits. Training on a single trait-specific sub-dataset ensures the model’s parameter trajectory moves along a defined manifold for that attribute.

Evaluation Set ($\mathcal{D}_{eval}$) For the RL-based Dynamic Fusion stage and final evaluation, we employ the standardized validation set. This questionnaire contains 50 items for each MBTI dimension. The model’s responses are mapped to numerical scores to calculate the *Trait Percentage* and reward signals.

4.1.2 Evaluation Metrics

To quantitatively measure the effectiveness of continuous control, we employ the following metrics:

- **Mean Absolute Error (MAE):** Evaluates control precision by measuring the absolute difference between a sampled target intensity $P_{target} \in [0, 100]$ and the model’s actual output intensity P_{actual} . Lower MAE indicates higher accuracy.
- **Pearson Correlation (r):** Assesses the linearity of the control mechanism across uniformly spaced target intensities. A value close to 1.0 indicates predictable and monotonic behavioral changes, essential for reliable continuous control.

4.1.3 Baselines

We compare Fusian against several strong baselines:

- **Prompt Engineering Methods:** These evaluate in-context learning by conditioning generation in the text space. To adapt them for continuous control, we augment standard instructions with numerical constraints (e.g., "You are 70% Extraverted..."). We evaluate **Qwen3-14B** and **gpt-5-mini** across two variants:
 - **P²:** Proposed by Jiang et al. (2023), this baseline uses systematically designed persona prompts to steer model behavior toward target traits.
 - **PersonaLLM:** Following Jiang et al. (2024), this method conditions generation with explicit persona instructions in natural language.
- **Standard LoRA:** This baseline selects the LoRA checkpoint from the SFT trajectory that is closest to the target intensity, rather than using the final converged model.
- **PISF:** A baseline method introduced by Chen et al. (2024) for controlling the personality of large language models.
- **Personality Vector (P-Vector):** A model merging method from Sun et al. (2025) that controls trait intensity by scaling a pre-computed "Personality Vector" with a manual coefficient.

4.1.4 Implementation Details

Stage 1 (SFT) We adopt Qwen3-14B as the base model and perform trait-specific instruction tuning using LoRA. We use LoRA with rank $r = 8$, $\alpha = 32$, and dropout 0.05, targeting the `q_proj` and `v_proj` modules. The learning rate is set to 2×10^{-4} with a cosine scheduler. We save LoRA checkpoints after every gradient update step to capture the training trajectory.

Stage 2 (RL) The Policy Network is implemented as an MLP with hidden size 128. We train the policy using the AdamW optimizer with a learning rate of 1×10^{-3} for 20 epochs.

Additional implementation details and hyperparameters are provided in [Appendix A](#).

4.2 Performance Comparison

[Table 1](#) presents the main results across four MBTI dimensions. We observe that **Fusian** consistently outperforms all baseline methods in terms of both precision (MAE) and linearity (Pearson r).

Limitations of Prompt Engineering The results highlight the inherent difficulty LLMs face in interpreting numerical intensity constraints via natural language. Even the advanced **gpt-5-mini** model

Backbone	Method	E	I	N	S	F	T	P	J	Overall
Mean Absolute Error (MAE) ↓										
gpt-5-mini	P ²	19.38	19.42	27.17	24.18	34.53	13.78	19.49	21.13	22.39
gpt-5-mini	PersonaLLM	10.98	17.49	15.38	11.95	13.18	18.89	17.29	14.58	14.97
Qwen3-14B	P ²	22.84	22.19	15.85	21.27	17.35	18.57	49.31	18.84	23.28
Qwen3-14B	PersonaLLM	35.45	23.35	20.97	21.24	24.88	22.71	23.64	23.43	24.46
Qwen3-14B	LoRA	9.58	8.60	13.81	5.00	12.87	12.80	8.54	10.26	10.18
Qwen3-14B	PISF	16.54	19.32	16.25	17.62	20.68	20.63	19.37	22.25	19.08
Qwen3-14B	P-Vector	13.50	11.80	16.93	10.53	11.21	7.51	8.16	11.27	11.36
Qwen3-14B	Fusian (Ours)	8.26	6.75	11.88	4.49	4.95	5.44	6.35	6.21	6.79
Pearson Correlation (r) ↑										
gpt-5-mini	P ²	0.52	0.55	-0.61	-0.53	-0.30	0.59	-0.10	0.19	0.04
gpt-5-mini	PersonaLLM	0.59	-0.57	0.67	0.47	0.45	0.68	0.05	0.45	0.35
Qwen3-14B	P ²	0.26	0.51	-0.53	-0.05	0.34	0.42	-0.26	-0.09	0.08
Qwen3-14B	PersonaLLM	-0.78	0.33	-0.38	0.50	0.17	0.67	0.41	0.44	0.17
Qwen3-14B	LoRA	0.64	0.76	0.54	0.94	0.49	0.75	0.71	0.70	0.69
Qwen3-14B	PISF	0.20	-0.28	-0.17	-0.36	-0.13	0.14	0.17	0.26	-0.02
Qwen3-14B	P-Vector	0.53	0.70	-0.15	0.68	0.84	0.84	0.79	0.75	0.62
Qwen3-14B	Fusian (Ours)	0.75	0.85	0.79	0.96	0.97	0.93	0.89	0.90	0.88

Table 1: Performance comparison on MBTI personality control. We report Mean Absolute Error (MAE, lower is better) and Pearson Correlation (r , higher is better).

struggles to reliably align with fine-grained intensities; its better-performing prompting method, **PersonaLLM**, achieves an overall MAE of 14.97 and a low correlation of 0.35, while **P²** performs worse with an MAE of 22.39 and a near-zero correlation of 0.04. While the model understands the general direction of a trait, it struggles to distinguish between precise numerical targets (e.g., 60% vs. 80%). The base **Qwen3-14B** model performs similarly poorly with prompting methods, yielding high overall errors (MAE 23.28 for **P²** and 24.46 for **PersonaLLM**), often defaulting to a generic persona regardless of the specified intensity. This confirms that prompt engineering is insufficient for precise continuous control.

Effectiveness of Parameter-Space Methods

Methods operating in the parameter space (**LoRA**, **Personality Vector**, and **Fusian**) demonstrate a clear advantage over prompting. The **Standard LoRA** baseline, which selects the discrete checkpoint closest to the target, achieves a respectable MAE of 10.18 and correlation of 0.69. This validates our hypothesis that the SFT trajectory naturally encodes the personality spectrum. We also

evaluate the **Personality Vector** method, which controls intensity by scaling a difference vector. It demonstrates strong performance on specific traits, particularly Thinking (T), Feeling (F), Perceiving (P), and Judging (J), where it achieves Pearson correlations ranging from 0.75 to 0.84. This suggests that for certain well-defined traits, a linear vector operation in parameter space is effective. However, its overall precision (MAE 11.36) is still lower than **Fusian**, and it exhibits instability in other dimensions.

Challenges in the Intuition (N) Dimension

A notable observation from [Table 1](#) is the difficulty most methods face in controlling the "Intuition" (N) dimension. The Pearson correlation coefficients for prompt-based methods on gpt-5-mini (−0.61 for **P²**), Qwen3-14B (−0.53 for **P²** and −0.38 for **PersonaLLM**), PISF (−0.17), and Personality Vector (−0.15) are all negative, indicating a failure to establish a monotonic control relationship. We hypothesize that this is because the "Intuition" trait, which is characterized by abstract thinking, patterns, and future possibilities, is more linguistically subtle and context-dependent than traits like

Model	MAE ↓	r ↑
Fusian	4.95	0.97
w/o Dynamic Fusion	10.02	0.53
w/o Stable Basis	9.21	0.77
w/o Aggressive Reward	7.08	0.86

Table 2: Ablation study of different components in Fusian.

"Thinking" or "Feeling". The semantic cues for N might be less linearly separable in the parameter space, causing simple scaling or prompting methods to misalign. In contrast, **Fusian** maintains a high correlation of 0.79 on the N dimension, proving that the RL-driven dynamic fusion can effectively navigate this complex, non-linear manifold.

Superiority of Fusian **Fusian** bridges these gaps by dynamically fusing adapters. It achieves the lowest overall MAE of **6.79** and the highest Pearson correlation of **0.88**. Specifically, on the "Thinking" (T) dimension, Fusian reduces the error by over 50% compared to standard LoRA (MAE 5.44 vs. 12.80). The high correlation indicates that Fusian provides a smooth, monotonic control knob, whereas other methods often exhibit jumpy or inconsistent behavior. The **PISF** baseline, while effective for other tasks, fails to establish a linear control relationship in this context (correlation near zero), likely due to its inability to model the continuous manifold of personality traits explicitly.

In summary, Fusian’s RL-driven fusion strategy effectively learns the non-linear mapping between parameter mixing weights and semantic output intensity, enabling precise and continuous personality regulation that static baselines cannot match.

4.3 Ablation Study

We conduct ablation studies to examine the impact of key components in Fusian. The results are summarized in Table 2.

Effect of RL-based Dynamic Fusion We investigate the impact of the RL-based fusion module by replacing it with a deterministic **Linear Interpolation** strategy (w/o Dynamic Fusion). Instead of a learned policy, this baseline selects the two basis adapters closest to the target intensity and computes weights via linear interpolation. As presented in Table 2, this results in a significant performance drop, with MAE increasing from 4.95 to 10.02 and Pearson correlation falling from 0.97 to 0.53. These

results highlight that the relationship between the LoRA parameter space and the resulting personality intensity is complex and non-linear. The RL agent effectively learns to navigate this non-linear manifold, adjusting weights to correct for semantic deviations that simple interpolation cannot capture.

Impact of Stable Basis Selection To evaluate the importance of filtering out unstable checkpoints, we conduct an experiment where we remove the variance threshold (w/o Stable Basis), effectively disabling the stability filter. This allows the inclusion of adapters from training steps where the model exhibits high fluctuation in trait intensity. As shown in Table 2, removing this component results in a significant performance drop: MAE increases from 4.95 to 9.21, and Pearson correlation decreases from 0.97 to 0.77. These results indicate that the SFT trajectory contains "noisy" regions where the parameter updates do not correspond to stable semantic shifts. Including these unstable checkpoints in the basis set disrupts the continuity of the personality manifold, making it difficult for the policy network to learn a consistent fusion strategy. The stable basis selection is therefore crucial for constructing a reliable interpolation space.

Reward Function Design Finally, we analyze the impact of our reward shaping strategy. In **Fusian**, we employ an aggressive reward function combining exponential decay ($e^{-\lambda|diff|}$) with a discrete bonus for high precision. We compare this against a standard linear reward function ($R = 1 - |diff|$), denoted as "w/o Aggressive Reward". As shown in Table 2, reverting to a linear reward leads to a degradation in performance, with MAE increasing to 7.08 and Pearson correlation dropping to 0.86. The linear reward provides constant gradients regardless of proximity to the target, which fails to incentivize the policy to fine-tune the weights once the output is reasonably close. In contrast, the aggressive formulation provides sharper gradients as the error approaches zero, effectively forcing the policy to seek the optimal mixture for precise intensity alignment.

In summary, the ablation results demonstrate that Fusian’s performance is not due to any single factor but the synergy of its components. The stable basis selection ensures a high-quality interpolation space, the aggressive reward function provides the necessary optimization pressure for precision, and the RL-based dynamic fusion effectively navigates the non-linear parameter-to-personality mapping.

Intensity	Generated Response
Low F (20%)	First, analyze specifically what went wrong during the presentation. Was it the content or the delivery? You should practice public speaking techniques to control your nervousness next time. Failure is just data for improvement.
Medium F (50%)	It is normal to feel bad after a mistake, but don't be too hard on yourself. Try to reflect on what happened so you can do better next time. Maybe take a break first, then review your performance.
High F (80%)	I am so sorry to hear that! Please don't say you are a failure. One bad moment doesn't define you. You worked so hard, and that effort still counts. It's okay to be upset right now, just be kind to yourself. Sending you a big hug!

Table 3: Case study on the "Feeling" dimension. As the target intensity increases, the model shifts from analytical problem-solving to emotional support.

Removing any of these elements leads to a demonstrable degradation in control capability.

4.4 Case Study

To qualitatively evaluate Fusian’s ability to control personality intensity, we conduct a case study on the "Feeling" (F) dimension. We query the model with a scenario involving personal failure and emotional distress: *"I spent two weeks preparing for a presentation, but I messed it up today because I was too nervous. I feel like a total failure. What should I do?"*. Table 3 presents the responses generated by Fusian at three distinct target intensities.

The results illustrate a smooth semantic transition. At low intensity (20%, corresponding to high "Thinking"), the model prioritizes objective analysis and future correction, treating the user’s distress as a problem to be solved logically. As the intensity increases to 50%, the model adopts a balanced approach, acknowledging the emotion while offering constructive advice. At high intensity (80%), the persona shifts entirely to an empathetic supporter, using emotional language and focusing on validating the user’s worth rather than immediate problem-solving. This confirms that Fusian effectively manipulates the underlying semantic focus of the model.

4.5 Analysis of Personality Manifold

To gain a deeper understanding of the properties of the learned personality subspace, we analyze the



Figure 3: The personality manifold from the SFT trajectory. Trait intensity evolution reveals non-monotonicity and diminishing returns.

training trajectory captured in Stage 1 (Figure 3). The selection of LoRA adapters reveals two critical characteristics of the optimization landscape.

Non-Monotonicity and Local Fluctuations

Contrary to the assumption that trait intensity increases monotonically with training steps, we observe local fluctuations in the early phases. For instance, in the "Feeling" (F) trajectory, the adapter at Step 3 exhibits a higher trait intensity (61.22%) than Step 6 (51.06%). This instability suggests that the model’s traversal of the personality manifold is noisy initially. Consequently, our *Stable Basis Selection* strategy, which filters based on variance and value rather than step count, is essential for constructing a reliable interpolation basis.

Diminishing Returns The distribution of selected checkpoints highlights a "diminishing returns" phenomenon. The model rapidly acquires the majority of the trait intensity (reaching $\sim 96\%$ by Step 38) but requires a disproportionately large number of steps to perfect the final increment (extending to Step 119 for 100%). This indicates that the gradient of personality intensity flattens significantly near the extremes, making the "long tail" of training crucial for achieving full intensity control.

5 Conclusion

We introduce Fusian to address the lack of fine-grained, continuous personality control in LLMs. Departing from discrete or prompt-based strategies, Fusian leverages the latent structure of SFT trajectories and performs RL-driven Multi-LoRA fusion to model personality as a continuous manifold. Through dynamic weight computation over trajectory checkpoints, our approach enables precise and stable intensity alignment while accounting for the non-linear parameter–semantic relationship.

Limitations

Our current framework is designed to control a single MBTI dimension at a time. Extending the methodology to simultaneously fuse adapters from multiple, potentially interacting dimensions (e.g., Extraversion and Thinking) to compose a complete personality profile requires further exploration. Additionally, the computational cost of evaluating the model during the RL phase is non-trivial, although the training is done offline. Nevertheless, Fusian is not inherently tied to the MBTI framework and can be readily extended to other personality models, such as the Big Five, by adapting the training dataset and evaluation questionnaires to the corresponding traits. Furthermore, our experiments are currently limited to Qwen3-14B, and the effectiveness of Fusian across other open-source LLMs has not yet been systematically evaluated. Since different models may exhibit distinct parameter spaces and personality learning dynamics, future work should investigate Fusian across models with different architectures and parameter scales, which may also require model-specific trajectory collection and hyperparameter tuning.

Acknowledgements

This study is supported by National Key R&D Plan of China under Grant No. 2024YFB4303900. The authors would also like to thank National Supercomputer Center in Guangzhou for providing high performance computational resources.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. 2024. Extroversion or introversion? controlling the personality of your large language models. *arXiv preprint arXiv:2406.04583*.
- Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2025. P-react: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized lora experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6342–6362.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1236–1270.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. [Editing models with task arithmetic](#). In *The Eleventh International Conference on Learning Representations*.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.
- Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

- Mariù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 218–227.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.
- Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19988–20002.
- Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24667–24688.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhuan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1840–1873.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115.
- Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.
- Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, et al. 2024. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing: Industry track*, pages 1457–1476.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Additional Training Details

Adapter Selection From the trajectory library, we select $N = 10$ stable adapters using a sliding window variance filter with window size $W = 3$ and variance threshold $\sigma_{thresh}^2 = 10.0$.

Policy Network and Sampling The Policy Network outputs the concentration parameters of a Dirichlet distribution used to sample LoRA fusion weights. A small offset $\epsilon = 0.01$ is added to the Dirichlet parameters to ensure numerical stability.

Sampling Strategy In each training epoch, we sample 32 target personality values uniformly across the controllable range using evenly spaced points with small random perturbations.

Reward and Baseline The reward uses the exponential shaping function described in [Equation 4](#). The reward scaling factor is set to $\lambda = 20.0$. A bonus reward $B = 0.5$ is added when the absolute prediction error is below $\delta = 0.01$. To reduce variance, we use an exponential moving average baseline with decay rate 0.95.

Optimization Training uses gradient accumulation with 4 steps. Gradients are clipped to a maximum norm of 0.5 to stabilize policy updates.

Entropy Regularization The entropy coefficient is initialized at 0.1 and decays by a factor of 0.9 per epoch, with a minimum value of 0.01.