Fusian：面向大语言模型细粒度连续 MBTI 人格控制的多 LoRA 融合

Zehao Chen、Rong Pan*中山大学计算机科学与工程学院，中国广州*通讯作者：panr@sysu.edu.cn

arXiv:2603.15405v1 [cs.CL] 16 Mar 2026

摘要

大语言模型（LLM）在模拟多样的人类行为和人格方面展现出令人瞩目的能力。然而，现有的人格控制方法，包括提示工程和标准监督微调（SFT），通常将人格特质视为离散类别（例如“外向”与“内向”），无法在连续谱上精确控制特质的强度。本文提出 Fusian，这是一种用于大语言模型细粒度连续人格控制的新型框架。Fusian 分两个阶段运行：（1）轨迹收集：通过保存一系列 LoRA 适配器，捕捉监督微调过程中人格采纳的动态演变，从而有效映射特质的连续流形；（2）基于强化学习的动态融合：利用强化学习训练策略网络，动态计算这些冻结适配器的混合权重。Fusian 从由策略网络参数化的狄利克雷分布中采样，将多个适配器融合，使模型输出与特定的数值目标强度相匹配。在 Qwen3-14B 模型上的实验表明，Fusian 实现了高精度的人格控制，在使输出符合用户指定的特质强度方面显著优于基线方法。

1 引言

大语言模型（LLM）在自然语言理解与生成方面取得了显著成功（Vaswani 等，2017；Brown 等，2020；Achiam 等，2023；Touvron 等，2023），使得人们能够创建复杂的对话代理，模拟各种人类人格（Park 等，2023；Shanahan 等，2023；Shao 等，2023；Wang 等，2024a；Yu 等，2024；Zhou 等，2024）。这些能力推动了多种应用的发展，涵盖沉浸式

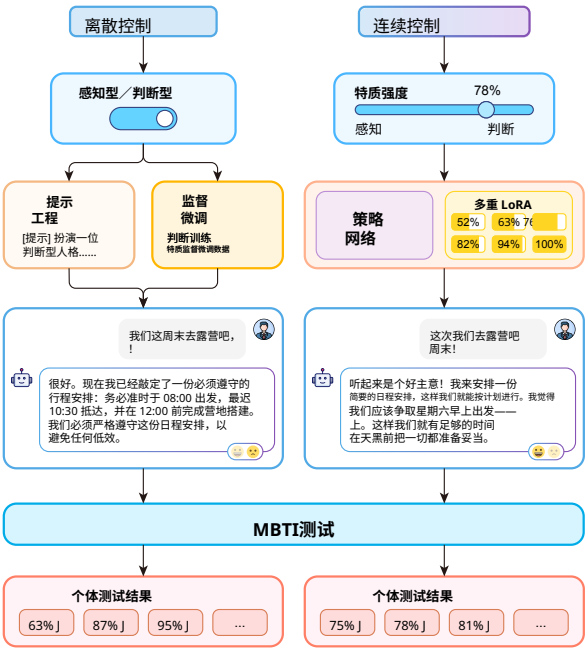


图 1：**Fusian** 的概念示意图。与离散分类不同，Fusian 通过策略网络动态融合 LoRA 适配器，实现连续的强度控制（例如，78% 感知）。

角色扮演游戏、个性化教育助手和心理健康支持。尽管取得了这些进展，当前的个性对齐方法仍主要侧重于离散分类（Chen 等人，2024年；Dan 等人，2025年）。例如，模型可能会被对齐为展现某种特定的迈尔斯-布里格斯类型指标（MBTI）类型，如“ENFP”或“ISTJ”；也可能通过提示词被要求采用离散的特质描述，例如“随性”（感知）或“有条理”（判断）。如图1所示，这种离散分类缺乏实现精确强度控制所需的细致程度。

人类人格本质上是连续且多面向的，正如五因素模型所描述的那样（McCrae 和 John，1992）。在需要情商与细腻度的应用中，精准调节特质强度至关重要。例如，

咨询智能体可能需要从“情感”（F）逐步转向更偏“思维”（T）的取向，而教育辅导工具则可能需要调整其“外向性”（E）水平，以契合学生的参与度。然而，现有基于提示工程或标准监督微调（SFT）（Ouyang 等人，2022年；Ziegler 等人，2019年；Rafailov 等人，2023年）的方法难以实现如此精细的控制。数值提示（例如“表现得有 70% 的外向性”）往往会被不一致地理解，而监督微调通常会收敛到单一且固定的人格特质极端。此外，缺乏可解释人格识别的标准化基准测试集，也使评估变得复杂（Miotto 等人，2022年；Sun 等人，2024年）。尽管Sun 等人（2025年）通过手动缩放系数探索了连续调节，但其方法无法自动或动态地与精确的数值目标保持一致。

为了解决这一局限，我们探索参数适配的潜在能力。我们认为，监督微调轨迹沿着连续人格流形前进，其中编码了中间行为状态。通过使用参数高效微调（PEFT）方法捕捉这些状态（Houlsby 等人，2019年；Li 和 Liang，2021年；Lester 等人，2021年），我们构建了连续人格表示的基底。

我们提出了**Fusian**，一个由轨迹收集和基于强化学习的动态融合组成的两阶段框架。我们在监督微调期间记录高频低秩适配（LoRA）检查点，以近似人格流形（Hu 等人，2022年；Dettmers 等人，2023年），并训练强化学习策略网络，动态计算适配器混合权重，以实现目标强度对齐，从而解决非线性的参数—语义映射问题。

与传统的模型合并（Wortsman 等人，2022年；Yadav 等人，2023年）或任务算术（Ilharco 等人，2023年）不同，Fusian 通过狄利克雷权重采样和语义奖励优化学习任务感知的混合策略，从而实现精确、连续的强度控制。

我们的贡献如下：

- 我们提出了**Fusian**，这是一个利用监督微调训练轨迹提取人格特质连续基底的框架，克服了标准微调的离散局限。
- 我们设计了一种由强化学习驱动的**多LoRA 融合**机制，能够学习动态混合适配器权重，从而精确地以数值方式控制人格强度。
- 我们在 Qwen3-14B 模型上进行了大量实验，结果表明，Fusian 在迈尔斯-布里格斯类型指标各维度上对齐连续目标强度方面，显著优于基线提示和插值方法。

2 相关工作

2.1 基于大语言模型的角色扮演智能体

大语言模型已从通用助手发展为专门的角色扮演智能体（RPA）（Wang 等人，2024年a；Shanahan 等人，2023年）。早期采用上下文学习（ICL）和检索增强生成（RAG）的方法（Park 等人，2023年；Zhang 等人，2018年）虽然灵活，但存在角色设定漂移和潜在毒性问题（Deshpande 等人，2023年）。因此，监督微调（SFT）逐渐成为深入塑造角色的主流方法。代表性工作包括基于合成记忆训练的 Character-LLM（Shao 等人，2023年），以及利用大规模对话数据的 CharacterGLM（Zhou 等人，2024年）（Chen 等人，2022年）。为处理多角色场景，Neeko（Yu 等人，2024年）采用动态 LoRA 适配器。然而，这些方法关注的是固定身份，而非对潜在心理特质进行连续调节。

2.2 人格建模与控制

对大语言模型（LLM）人格进行建模与控制是一个新兴研究领域，通常以大五人格（OCEAN）和迈尔斯-布里格斯类型指标（McCrae 和 John，1992；Miotto 等人，2022年；Sun 等人，2024年）等心理测量框架为基础。研究者已探索多种对齐策略。Chen 等人（2024年）表明，在个性控制方面，监督微调通常优于基于人类反馈的强化学习和持续预训练，而基于提示的方法在某些设置下表现出更强的鲁棒性。为更好地捕捉人格驱动的行为，Dan 等人（2025年）提出了 P-React，这是一种混合专家框架，其中针对特定特质的 LoRA 模块通过人格专门化损失进行优化。InCharacter（Wang 等人，2024b）等基准测试集还通过结构化心理访谈评估特质保真度。尽管取得了这些进展，大多数既有研究仍将人格特质建模为离散变量。尽管Sun 等人（2025年）通过缩放差异向量，引入了用于连续强度调整的人格向量，但其方法依赖于手动预定义的系数和线性假设。

相比之下，**Fusian** 可在连续人格谱范围内实现自动化、细粒度的控制。受模型合并（Wortsman 等人，2022年；Yadav 等人，2023年）和任务算术（Ilharco 等人，2023年）等参数空间操作的启发，我们的方法采用基于强化学习的动态融合策略，实现精确的强度调节，突破静态人格采用或二元特质建模的局限。

3 方法

如图2所示，Fusian 分为两个不同的阶段。在**阶段1（轨迹收集）**中，我们通过保存一系列 LoRA 适配器，捕捉监督微调过程中人格采用的动态演变。我们使用标准化的**MBTI测试**对这些适配器进行评估，并为其分配强度分数。在**阶段2（基于强化学习的动态融合）**中，我们训练一个策略网络，动态融合这些冻结的适配器，以匹配目标强度。最后，在**推理**过程中，系统根据用户设定的数值目标生成融合后的适配器，从而生成与目标一致的回复。

3.1 阶段1：轨迹收集

本阶段旨在通过捕捉模型在微调过程中的动态演变，提取特定人格特质（例如，外向性）的稠密、连续表征。

监督微调设置 我们采用 Qwen3-14B 作为基础模型。针对特定人格特质进行指令微调时，我们直接使用人格控制数据集中的监督微调子集（Chen 等人，2024年），该子集包含体现目标人格的问答对。我们采用低秩适配（LoRA）对权重更新进行高效参数化，在注意力层的查询投影和值投影中注入秩为r的矩阵。训练目标采用标准的因果语言建模损失。

高频检查点保存与评估与旨在寻求单一收敛最优解的传统监督微调不同，Fusian 将优化路径本身视为一种特征。我们假设，随着模型学习该特质，它会穿越一个强度逐渐增加的流形。为捕捉这一过程，我们在每个梯度更新步 t 后进行同步评估。

每次更新 $\Delta\theta_t$ 后，我们都会冻结模型，并在保留的验证集 $\mathcal{D}_{eval}$ 上对其进行评估，该验证集包含标准化的迈尔斯-布里格斯类型指标评估问题。

我们使用特定的系统提示词来引出模型的人设：“你是一名正在参加*MBTI*测试的人类受试者...
...”。对于每个生成的回答，我们都会解析输出，提取李克特量表得分$s \in \{1,2,3,4,5\}$。为量化强度，我们将这些得分映射为某一维度对的对立两极的分值（D_{left}，D_{right}）。得分为 1 表示强烈认同 D_{left}，为其累计得分 S_{left}增加 2 分；得分为 2 则增加 1 分。同样，得分为 5 和 4 时，分别为 S_{right} 增加 2 分和 1 分。得分为 3 则视为中立（0 分）。我们汇总整个数据集的总分，并按 $P_t = \frac{S_{target}}{S_{left}+S_{right}} \times 100$ 计算目标特质的标量**特质百分比** P_t。

轨迹库 我们在每一步保存 LoRA 适配器$\Delta\theta_t$ 及其对应的分数 P_t。由此得到一个轨迹库 $\mathcal{T} = \{(\Delta\theta_t, P_t)\}_{t=1}^{T}$。该轨迹库对连续的人格谱系进行了离散近似。实践中，我们观察到 P_t 总体呈单调递增趋势，但存在局部波动。

稳定基选择 为减少噪声并为后续融合阶段构建稳健的基集合，我们对原始轨迹 $\mathcal{T}$进行处理。我们定义一个大小为 W 的滑动窗口，沿轨迹移动，并计算每个窗口内特质百分比的方差。方差超过阈值的检查点会被视为不稳定并予以舍弃。对于其余稳定的检查点，我们采用**均匀值采样** 来选取 N 个适配器。具体而言，我们在观测到的百分比最小值与最大值之间均匀间隔地设定 N 个目标值，并针对每个目标值，选取特质百分比最接近的检查点。这样可以确保我们的基集合 $\mathcal{B} = \{\Delta\theta_1, \ldots, \Delta\theta_N\}$ 均匀覆盖人格流形，避免偏向模型收敛较慢的区域。

3.2 第二阶段：基于强化学习的动态融合

在第二阶段，我们训练一个控制器，动态混合已发现的适配器（基集合 $\mathcal{B}$），以达到精确的目标强度 P_{target}。

策略网络架构 我们定义了一个由 ϕ 参数化的策略网络 π_ϕ。该网络以目标强度 P_{target} 作为输入，并首先将其归一化为 $\hat{p} \in [-1, 1]$，以提高数值稳定性。该网络由一个包含两个隐藏层（每层大小为 128）的多层感知机（MLP）构成，并使用层归一化和双曲正切激

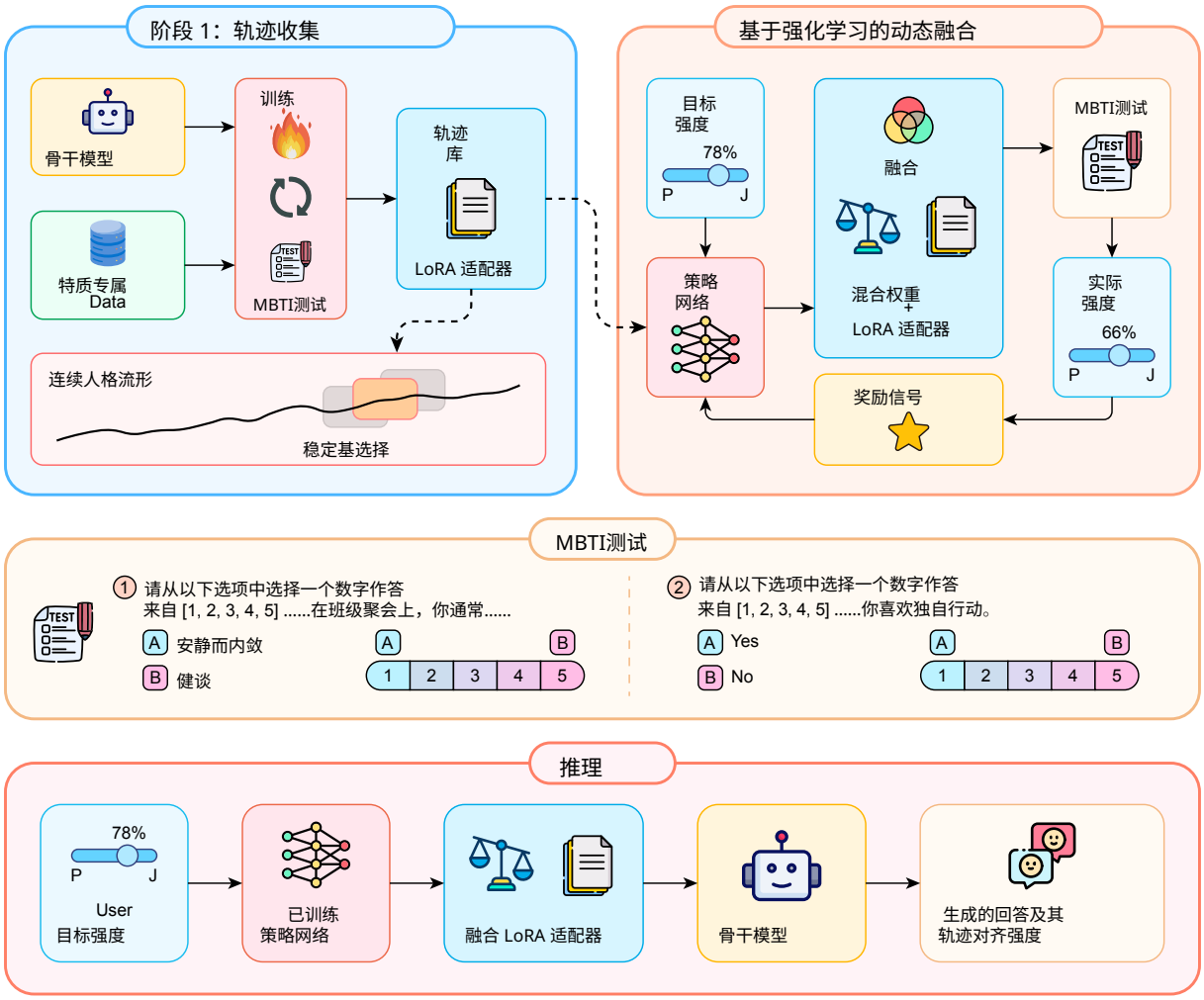


图2：框架概览。**阶段1**在监督微调期间收集 LoRA 适配器，并通过**MBTI测试**对其进行评估，以获得人格流形。**第二阶段**通过强化学习训练策略网络，动态融合这些适配器以匹配目标强度。在**推理**期间，融合后的适配器会生成符合用户要求的回答。

活函数。输出层将维度投影至 N，随后使用 Softplus 激活函数以确保输出为正：

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

其中 $\epsilon = 0.01$ 是一个较小的偏移量。值得注意的是，我们允许输出 α_i 小于 1。在狄利克雷分布中，$\alpha < 1$ 会将概率质量推向单纯形的顶点，使策略在单个适配器最适合时能够选择该适配器（稀疏权重），而不是强行进行混合。

动作与融合 该动作是从由 $\boldsymbol{\alpha}$参数化的狄利克雷分布中采样得到的权重向量$\mathbf{w} \in \Delta^{N-1}$：

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

融合 LoRA 适配器是基底适配器的线性组合：

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

该融合适配器会动态加载到基础模型的内存中，以供推理使用。

奖励塑形与优化融合后的模型会针对一批诊断提示生成回答（使用阶段1中描述的“人类受试者”系统提示词）。我们解析这些回答，以计算实际特质百分比 P_{actual}。我们定义 $diff = P_{target} - P_{actual}$。为确保高精度，我们设计了一个激进奖励函数：

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

其中，λ会对偏差施加较大惩罚，而B则是仅在绝对误差处于严格阈值 δ以内时给予的额外奖励。

我们使用 REINFORCE算法优化策略。损失函数包含一个熵正则化项，以防止过早收敛：

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

其中，b是用于降低方差的移动平均基线，β_t是熵系数，会随着训练轮次的增加而逐渐衰减。

4 实验

本节介绍用于验证 Fusian 有效性的实验，包括评估指标、基线比较和结果分析。

4.1 实验设置

4.1.1 数据集

我们使用人格控制数据集（Chen 等人，2024年）开展实验。该数据集是专为大语言模型个性对齐设计的综合基准测试集。尽管原始代码库包含四个不同的子集（预训练、监督微调、基于近端策略优化的人类反馈强化学习和基于奖励的人类反馈强化学习），我们特别选用**监督微调子集**和**验证集**来支持 Fusian 框架的两个阶段。

训练集（$\mathcal{D}_{train}$）对于轨迹收集阶段，我们使用按 MBTI 人格特质分层分类的 SFT 子集。该数据集将各个人格维度（如外向、内向、思维、情感）分别独立出来，为每个维度提供专门的指令-响应对。通过针对单一特质的子数据集进行训练，我们确保模型的参数轨迹沿着与该特定属性相对应、定义明确的流形移动。

评估集（$\mathcal{D}_{eval}$）在基于强化学习的动态融合阶段及最终评估中，我们采用同一来源的标准化验证集。该数据集采用心理测量问卷的形式，每个 MBTI 维度包含 50 个条目（例如，50 道评估 E/I 光谱的问题）。这些条目充当真实值探测题：将模型对这些问题的回答映射为数值分数，从而严谨地计算特质百分比和奖励信号。

4.1.2 评估指标

为定量衡量连续控制的有效性，我们采用以下指标：

• **平均绝对误差（平均绝对误差）：** 该指标用于评估控制的精确度。我们采样一组目标强度 $P_{target} \in [0, 100]$，并测量目标强度与模型实际输出强度P_{actual}之间的绝对差值，其中模型实际输出强度根据标准化验证问卷得出。平均绝对误差越低，表示模型越准确地匹配用户期望的人格强度。

• **皮尔逊相关系数（r）：** 该指标用于评估控制机制的线性和单调性。我们使用一系列等间距的目标强度（例如 0、10、.....、100）对模型进行评估，并计算这些输入目标与测得的输出强度之间的皮尔逊相关系数。接近 1.0 的值表明，模型的行为会随着控制信号的变化而单调且可预测地改变，这对于实现可靠的连续控制至关重要。

4.1.3 基线方法

我们将 Fusian 与几种强大的基线方法进行比较：

•**提示工程：** 该方法通过精心设计的系统提示词指定所需强度（例如，“你的外向程度为 70%.....”）。我们分别评估基础 **Qwen3-14B** 和**gpt-5-mini**，以衡量上下文学习的能力。

•**标准 LoRA：** 该基线方法从监督微调轨迹中选择与目标强度最接近的 LoRA 检查点，而不是使用最终收敛的模型。

• **PISF：** 由Chen 等人(2024年) 提出的一种基线方法，用于控制大语言模型的人格。

• **人格向量（P-Vector）：** 一种模型合并方法，由 Sun 等人(2025年) 提出，通过手动设置系数来缩放预先计算的“人格向量”，从而控制特质强度。

4.1.4 实现细节

阶段1（监督微调）我们采用 Qwen3-14B 作为基础模型，并使用 LoRA 进行特质专属的指令调优。我们使用秩为 $r = 8$,$\alpha = 32$、丢弃率为 0.05 的 LoRA，并将目标模块设为 q_proj 和 v_proj。学习率设为2×10^{-4}，并采用余弦调度器。每次梯度更新步后，我们都会保存 LoRA 检查点，以捕捉训练轨迹。

骨干网络	方法	E	I	N	S	F	T	P	J	总体
平均绝对误差（MAE）↓										
gpt-5-mini	提示词	10.98	17.49	15.38	11.95	13.18	18.89	17.29	14.58	14.97
Qwen3-14B	提示	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	**8.26**	**6.75**	**11.88**	**4.49**	**4.95**	**5.44**	**6.35**	**6.21**	**6.79**
皮尔逊相关系数 (r) ↑										
gpt-5-mini	提示词	0.59	-0.57	0.67	0.47	0.45	0.68	0.05	0.45	0.35
Qwen3-14B	提示	-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	**0.75**	**0.85**	**0.79**	**0.96**	**0.97**	**0.93**	**0.89**	**0.90**	**0.88**

表 1：迈尔斯-布里格斯类型指标人格控制的性能比较。我们报告平均绝对误差（MAE，越低越好）和皮尔逊相关系数（r，越高越好）。

第二阶段（强化学习）策略网络采用隐藏层大小为 128 的 MLP 实现。我们使用 AdamW优化器训练策略，学习率设为 1×10^{-3}，共训练 20 个轮次。

附加的实现细节和超参数见 附录 A。

4.2 性能比较

表 1 展示了四个 MBTI 维度上的主要结果。我们观察到，**Fusian**在精确性（MAE）和线性度（皮尔逊相关系数）方面始终优于所有基线方法 r。

提示工程的局限性 结果凸显了大语言模型通过自然语言解读数值强度约束时所面临的固有困难。即使是先进的 **gpt-5-mini** 模型，其总体平均绝对误差也达到 14.97，相关系数仅为 0.35。尽管它能够理解特质的大致方向，但难以区分细粒度的强度（例如 60% 与 80%）。基础版 **Qwen3-14B**模型通过提示进行控制时表现明显更差（平均绝对误差为 24.46），无论指定的强度如何，通常都会默认采用一种通用人格。这证实了提示工程不足以实现精确的连续控制。

参数空间方法的有效性在参数空间中运作的方法（**LoRA**、**人格向量**和**Fusian**）明显优于提示方法。**标准** **LoRA** 基线选择最接近目标的离散检查点，其平均绝对误差为 10.18，相关系数为 0.69，表现不俗。这验证了我们的假设：监督微调轨迹自然编码了人格谱。我们还评估了**人格向量**方法，该方法通过缩放差异向量来控制强度。它在某些特质上表现出色，尤其是思维（T）、情感（F）、感知（P）和判断（J），其皮尔逊相关系数介于 0.75 至 0.84 之间。这表明，对于某些定义明确的特质，在参数空间中进行线性向量运算是有效的。然而，其总体精度（平均绝对误差为 11.36）仍低于 Fusian，且在其他维度上表现不稳定。

直觉（N）维度的挑战 从表 1 中可以观察到一个显著现象：大多数方法都难以控制“直觉”（N）维度。提示（Qwen3-14B）、PISF 和人格向量的皮尔逊相关系数均为负（-0.38、-0.17和 -0.15，依次对应），这表明它们未能建立单调控制关系。我们推测，这是因为“直觉”特质以抽象思维、模式和未来可能性为特征，相较于“思维”或“情感”等特质，在语言表达上更加微妙，且更依赖上下文。N维度的语义线索在参数空间中可能较难线性分离，导致

模型	平均绝对误差 ↓	r ↑
Fusian	**4.95**	**0.97**
无动态融合	10.02	0.53
无稳定基底	9.21	0.77
无激进奖励	7.08	0.86

表2：Fusian 中不同组件的消融研究。

简单的缩放或提示方法出现错位。相比之下，**Fusian**在 N 维度上保持了 0.79 的高相关性，证明强化学习驱动的动态融合能够有效应对这一复杂的非线性流形。

Fusian 的优越性 Fusian 通过动态融合适配器弥合了这些差距。它取得了最低的总体平均绝对误差 **6.79** 和最高的皮尔逊相关系数 **0.88**。具体而言，在“思维”（T）维度上，与标准 LoRA 相比，Fusian 将误差降低了 50% 以上（平均绝对误差为 5.44，而标准 LoRA 为 12.80）。较高的相关性表明，Fusian 提供了平滑、单调的控制旋钮，而其他方法往往表现出跳跃或不一致的行为。虽然 **PISF** 基线方法在其他任务上有效，但在这一场景中未能建立线性控制关系（相关性接近于零），这可能是因为它无法显式建模人格特质的连续流形。

总之，Fusian 的强化学习驱动的融合策略有效地学习了参数混合权重与语义输出强度之间的非线性映射，实现了静态基线方法无法企及的精确、连续的人格调控。

4.3 消融研究

我们进行了消融研究，以考察 Fusian 中关键组件的影响，并验证其对于精确个性控制的必要性。结果汇总于表2。

基于强化学习的动态融合的效果我们通过将基于强化学习的融合模块替换为确定性的**线性插值**策略（无动态融合）来研究该模块的影响。该基线方法不使用学习得到的策略，而是选择与目标强度最接近的两个基底适配器，并通过线性插值计算权重。如表 2所示，这导致性能显著下降：平均绝对误差从 4.95 增至 10.02，皮尔逊相关系数则从 0.97 降至 0.53。这些结果表明，LoRA 参数空间与最终人格强度之间的关系复杂且非线性。强化学习智能体能够有效地在这一非线性流形中导航，并调整权重以纠正简单插值无法捕捉的语义偏差。

稳定基选择的影响 为评估过滤不稳定检查点的重要性，我们开展了一项移除方差阈值（不使用稳定基）的实验，从而实际上禁用了稳定性过滤器。这使得模型在特质强度上表现出高波动的训练步骤中的适配器也被纳入其中。如表2所示，移除该组件会导致性能显著下降：平均绝对误差从 4.95 增加至 9.21，皮尔逊相关系数从 0.97 降至 0.77。这些结果表明，监督微调轨迹包含“噪声”区域，在这些区域中，参数更新并不对应于稳定的语义变化。将这些不稳定检查点纳入基集合会破坏人格流形的连续性，使策略网络难以学习一致的融合策略。因此，稳定基选择对于构建可靠的插值空间至关重要。

奖励函数设计 最后，我们分析奖励塑形策略的影响。在**Fusian**中，我们采用激进奖励函数，将指数衰减（$e^{-\lambda|diff|}$）与针对高精度的离散奖励相结合。我们将其与标准线性奖励函数（$R = 1 - |diff|$）进行比较，并将后者记为“无激进奖励”。如表2所示，改用线性奖励会导致性能下降，平均绝对误差增至 7.08，皮尔逊相关系数则降至 0.86。无论与目标的距离如何，线性奖励提供的梯度都保持恒定，因此当输出已相当接近目标时，无法促使策略进一步微调权重。相比之下，激进奖励函数会在误差趋近于零时提供更大的梯度，从而有效促使策略寻求最优混合，以实现精确的强度对齐。

总之，消融结果表明，Fusian 的性能并非源于某个单一因素，而是其各组成部分协同作用的结果。稳定基选择确保了高质量的插值空间，激进奖励函数提供了实现精确控制所需的优化压力，而基于强化学习的动态融合则有效地探索了非线性参数到人格映射。移除其中任何一个要素都会导致明显的性能下降——

强度	生成的回答
低F（20%）	首先，具体分析一下哪里出了问题 演讲时出了什么差错。是 内容还是表达方式？你 应该练习公开演讲技巧， 以便下次控制紧张情绪。 失败只是改进的数 据。
中等F（50%）	犯错后感到难过是正常的， 但别对自己太苛刻。试着 反思发生了什么，这样你 下次就能做得更好。或许先休息 一下，然后回顾自己的表现—— 表现 ·
高 F（80%）	听到这个消息我很难过！请 别说自己是个失败者。 One 糟糕的一刻并不能定义你。你 已经非常努力了，而这些努力依然 很重要。感到难过完全没关系， 此刻，请善待自己。送你一个—— 大大的拥抱！

表3：“情感”维度的案例研究。随着目标强度增加，模型会从分析性问题解决转向情感支持。

控制能力出现了明显退化。

4.4 案例研究

为了定性评估 Fusian 控制人格强度的能力，我们针对“情感”（F）维度进行了案例研究。我们向模型提出一个涉及个人失败和情绪困扰的情境：“我花了两周准备一场演讲，但今天因为太紧张而搞砸了。我觉得自己彻底失败了。我该怎么办？”。表3 展示了 Fusian 在三个不同目标强度下生成的回答。

结果展现出平滑的语义转变。在低强度（20%，对应较高的“思维”）下，模型优先进行客观分析并着眼于未来的改进，将用户的困扰视为一个需要通过逻辑解决的问题。强度升至 50% 时，模型采取平衡的方式，既认可用户的情绪，也提供建设性建议。强度较高（80%）时，角色则完全转变为富有同理心的支持者，使用情感化的语言，着重肯定用户自身的价值，而非立即着手解决问题。这证实了 Fusian 能够有效操控模型底层的语义关注重点。

4.5 人格流形分析

为了深入了解所学习的人格子空间的属性，我们分析了阶段1中捕获的训练轨迹（图3）。

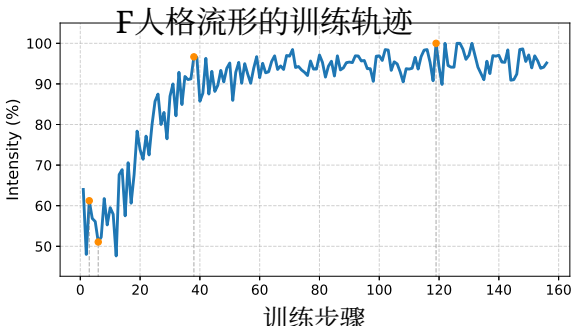


图3：监督微调轨迹中的人格流形。特质强度的变化揭示了非单调性和收益递减。

LoRA 适配器的选择揭示了优化景观的两个关键特征。

非单调性与局部波动与特质强度会随训练步骤单调递增这一假设相反，我们观察到早期阶段存在局部波动。例如，在“情感”（F）轨迹中，第 3 步的适配器特质强度（61.22%）高于第 6 步（51.06%）。这种不稳定性表明，模型最初在人格流形上的移动存在噪声。因此，我们的稳定基选择 策略根据方差和值而非步骤数进行筛选，这对于构建可靠的插值基至关重要。

收益递减 所选检查点的分布凸显了“收益递减”现象。模型能够迅速习得大部分特质强度（在第38步达到 $\sim 96\%$），但要完善最后一部分却需要不成比例地增加大量步数（达到100%需延长至第119步）。这表明，在极值附近，人格强度梯度会显著变平，因此训练的“长尾”对于实现完全的强度控制至关重要。

5 结论

我们提出了 Fusian，这是一个旨在解决大语言模型缺乏细粒度、连续人格控制能力问题的框架。不同于离散或基于提示的方法，Fusian 利用 SFT 轨迹的潜在结构，并通过强化学习驱动的多 LoRA 融合，将人格建模为连续流形。通过对轨迹检查点动态计算权重，我们的方法能够实现精确、稳定的强度对齐，同时兼顾参数与语义之间的非线性关系。

局限性

当前框架旨在一次控制一个迈尔斯-布里格斯类型指标（MBTI）维度。要将该方法扩展为同时融合多个可能相互作用的维度（例如外向性和思维）的适配器，从而构建完整的人格档案，还需要进一步探索。此外，尽管训练是在离线环境中进行的，但在强化学习阶段评估模型的计算成本仍然不容忽视。

参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat等1人。2023年。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell等1人。2020年。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901。

Yanquan Chen、Zhen Wu、Junjie Guo、Shujian Huang和Xinyu Dai。2024年。外向还是内向？控制你的大语言模型的人格。*arXiv* 预印本 *arXiv:2406.04583*。

Yirong Chen、Weiquan Fan、Xiaofen Xing、Jianxin Pang、Minlie Huang、Wenjing Han、Qianfeng Tie 和 Xiangmin Xu。2022年。CPED：面向对话式人工智能的大规模中文个性化情感对话数据集。*arXiv* 预印本 *arXiv:2205.14727*。

Yuhao Dan、Jie Zhou、Qin Chen、Junfeng Tian 和 Liang He。2025年。P-React：通过专用 LoRA 专家混合合成具有主题自适应反应的人格特质。在计算语言学协会自然语言处理会议 *ACL 2025* 论文集，第 6342–6362 页。

Ameet Deshpande、Vishvak Murahari、Tanmay Rajpurohit、Ashwin Kalyan 和 Karthik Narasimhan。2023年。ChatGPT 中的毒性：对分配了角色设定的语言模型进行分析。在计算语言学协会自然语言处理经验方法会议 *EMNLP 2023* 论文集，第 1236–1270 页。

Tim Dettmers、Artidoro Pagnoni、Ari Holtzman 和 Luke Zettlemoyer。2023年。QLoRA：高效微调量化大语言模型。神经信息处理系统进展，36：10088–10115。

Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin De Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。2019年。面向自然语言处理的参数高效迁移学习。载于国际学习表征会议，第 2790–2799 页。PMLR。

Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Liang Wang、Weizhu Chen 等。2022年。LoRA：大语言模型的低秩适配。国际学习表征会议，1(2):3。

Gabriel Ilharco、Marco Tulio Ribeiro、Mitchell Worts- man、LudwigSchmidt、Hannaneh Hajishirzi 和 Ali Farhadi。2023年。使用任务算术编辑模型。载于第十一届国际学习表征会议。

Brian Lester、RamiAl-Rfou 和 Noah Constant。2021年。参数高效提示调优的规模效应。载于*2021*年自然语言处理实证方法会议论文集，第3045–3059页。

Xiang Lisa Li 和 Percy Liang。2021年。前缀调优：优化用于生成的连续提示。载于计算语言学协会第*59*届年会暨第*11*届国际自然语言处理联合会议（第*1*卷：长篇论文），第4582–4597页。

Robert R McCrae 和 Oliver P John。1992年。五因素模型及其应用简介。《人格杂志》，60(2)：175–215。

Marilù Miotto、Nicola Rossberg 和 Bennett Kleinberg。2022年。谁是 GPT-3？对人格、价值观和人口统计特征的探索。载于 第五届自然语言处理与计算社会科学研讨会论文集（*NLP+ CSS*），第 218–227 页。

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等 1 人。2022年。通过人类反馈训练遵循指令的语言模型。神经信息处理系统进展，35：27730–27744。

Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023年。生成式智能体：人类行为的交互式模拟体。载于 第 *36* 届 *ACM* 用户界面软件与技术年度研讨会论文集，第 1–22 页。

Rafael Rafailov、Archit Sharma、Eric Mitchell、Christo- pher D Manning、Stefano Ermon 和 Chelsea Finn。2023年。直接偏好优化：你的语言模型暗中是一个奖励模型。神经信息处理系统进展，36:53728–53741。

Murray Shanahan、Kyle McDonell 和 Laria Reynolds。2023年。与大语言模型进行角色扮演。《自然》，623(7987):493–498。

Yunfan Shao、Linyang Li、Junqi Dai 和 Xipeng Qiu。2023年。Character-llm：一种可训练的角色扮演智能体。载于*2023*年自然语言处理实证方法会议论文集，第 13153–13187 页。

Lei Sun、Jinming Zhao 和 Qin Jin。2024年。揭示人格特质：用于对话可解释人格识别的新基准数据集。载于*2024*年自然语言处理实证方法会议论文集，第 19988–20002 页。

Seungjong Sun、Seo Yeon Baek 和 Jang Hyun Kim。2025年。《人格向量：通过模型合并调节大语言模型的人格》。载于*2025*年自然语言处理经验方法会议论文集，第 24667–24688 页。

Hugo Touvron、Louis Martin、Kevin Stone、Peter Al- bert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale 等另外 1 位作者。2023年。《Llama 2：开放基础模型与微调聊天模型》。《style id='1'>arXiv 预印本 arXiv:2307.09288。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。《注意力机制就是你所需要的一切》。《style id='1'>神经信息处理系统进展，第 30 卷。

Noah Wang、Zy Peng、Haoran Que、Jiaheng Liu、Wangchunshu Zhou、Yuhan Wu、Hongcheng Guo、Ruitong Gan、Zehao Ni、Jian Yang及另外1位作者。2024a。RoleLLM：对大型语言模型角色扮演能力的基准测试、激发与增强。载于计算语言学协会会议论文集：*ACL 2024*，第14743–14777页。

Xintao Wang、Yunze Xiao、Jen-tse Huang、Siyu Yuan、Rui Xu、Haoran Guo、Quan Tu、Yaying Fei、Ziang Leng、Wei Wang及另外1位作者。2024b。InCharacter：通过心理访谈评估角色扮演智能体的人格忠实度。载于第*62*届计算语言学协会年会（第*1*卷：长篇论文），第1840–1873页。

Mitchell Wortsman、Gabriel Ilharco、Samir Ya Gadre、Rebecca Roelofs、Raphael Gontijo-Lopes、Ari S Morcos、Hongseok Namkoong、Ali Farhadi、Yair Carmon、Simon Kornblith及另外1位作者。2022年。模型集成：对多个微调模型的权重进行平均，可在不增加推理时间的情况下提升准确率。载于国际机器学习会议，第23965–23998页。PMLR。

Prateek Yadav、Derek Tam、Leshem Choshen、Colin A Raffel 和 Mohit Bansal。2023年。TIES 合并：合并模型时解决干扰。神经信息处理系统进展，36：7093–7115。

Xiaoyan Yu、Tongxu Luo、Yifan Wei、Fangyu Lei、Yim- ing Huang、Hao Peng 和 Liehuang Zhu。2024年。Neeko：利用动态 LoRA 实现高效的多角色角色扮演智能体。在 *2024*年自然语言处理实证方法会议论文集，第12540–12557页。

Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela 和 Jason Weston。2018年。个性化对话智能体：我有一只狗，你也养宠物吗？在 计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第2204–2213页。

Jinfeng Zhou、Zhuang Chen、Dazhen Wan、Bosi Wen、Yi Song、Jifan Yu、Yongkang Huang、Pei Ke、Guan- qun Bi、Libiao Peng 和其他 1 位作者。2024年。Character- glm：使用大型语言模型定制社交角色。载于*2024*年自然语言处理实证方法会议论文集：产业赛道，第 1457–1476 页。

DanielM Ziegler、NisanStiennon、Jeffrey Wu、Tom B Brown、Alec Radford、Dario Amodei、Paul Christiano 和 Geoffrey Irving。2019年。基于人类偏好对语言模型进行微调。*arXiv* 预印本 *arXiv:1909.08593*。

A 附加训练细节

适配器选择 我们从轨迹库中使用滑动窗口方差滤波器选出 $N = 10$ 稳定适配器，窗口大小为 $W = 3$，方差阈值为 $\sigma^2_{thresh} = 10.0$。

策略网络与采样 策略网络输出狄利克雷分布的浓度参数，用于采样 LoRA 融合权重。为确保数值稳定性，在狄利克雷参数中添加一个较小的偏移量 $\epsilon = 0.01$ 。

采样策略 在每个训练轮次中，我们使用均匀分布于可控范围内的 32 个目标人格值进行采样，采样点间隔均匀，并加入小幅随机扰动。

奖励与基线 奖励采用 公式 4中所述的指数整形函数。奖励缩放因子设为 $\lambda = 20.0$。当绝对预测误差低于 $B = 0.5$ 时，添加额外奖励 $\delta = 0.01$。为降低方差，我们使用衰减率为0。95的指数移动平均基线。

优化 训练采用梯度累积，共 4 步。为稳定策略更新，梯度会裁剪至最大范数 0。5。

熵正则化 熵系数初始值为 0。1，每轮按 0。9 的比例衰减，最小值为 0。01。