张量动力学作家数字人写作模型：四组对照实验设计与实证分析

纪路阳[1]，张思羽[1,]，李建军[1]，Nutapong Somjit[2,3]，Watcharin Srirattanawichaikul[3]，张维加[1,4†]

(1. 绍兴大学 数理信息学院, 浙江 绍兴 312000; 2. 利兹大学 人工智能系, 英国 利兹市 LS2 9JT；3. 清迈大学 电子工程系，泰国 清迈 50200； 4. 牛津大学计算机学院, 英国 牛津市 OX1 3PU)

摘　要：针对大模型风格模拟的“画皮”与人格崩坏瓶颈，本研究构建高阶人格张量动力学框架。通过四组对照实验验证：势能阱使状态偏差方差降低 86.8%；在 180 个测试单元中，本方法四项指标显著优于基线（p<1e-10）；消融实验证实势能阱与惯性模块的核心贡献；跨时空语境下核心人格维度完全锁定。噪声敏感性测试显示强鲁棒性。研究表明，将作家风格建模为受控连续时间动力学系统，是解决人格崩坏与可解释性问题的有效路径。

关键词：作家数字人；张量动力学；人格势能；消融研究；跨时空创作；神经常微分方程

中图分类号：TP391.41

Tensor Dynamics-Based Digital Writer Model: Design and Empirical Analysis of Four Controlled Experiments

Ji Luyang[1], Zhang Siyu[1] , Li Jianjun[1]，Nutapong Somjit[2,3]，Watcharin Srirattanawichaikul[3]，Zhang Weijia[1,†]，Zhao Langping[4,†]

(1. School of Mathematics, Physics and Information, Shaoxing University, Zhejiang Shaoxing 312000, China; 2. University of Leeds, UK Leeds LS2 9JT；3. Department of Electrical Engineering, Chiang Mai University, Chiang Mai, Thailand 50200; 4. School of Marxism, Shaoxing University, Zhejiang Shaoxing 312000, China)

Abstract: This paper addressed the "superficial imitation" and personality collapse bottlenecks in large model style simulation. The study constructed a high-order personality tensor dynamics framework. Four controlled experiments validated the framework. The potential well reduced state deviation variance by 86.8%. The method outperformed baselines significantly on four metrics across 180 test units (p < 1e-10). Ablation experiments confirmed the core contributions of the potential well and inertia modules. Core personality dimensions remained fully locked under cross-spatiotemporal contexts. Noise sensitivity tests demonstrated strong robustness. The research indicates that modeling writer style as a controlled continuous-time dynamical system provides an effective solution to personality collapse and interpretability issues.

Key words: digital writer; tensor dynamics; personality potential energy; ablation study; cross-spatiotemporal creation; neural ordinary differential equations

0　引言

随着 Transformer [1] 架构下大语言模型（如 GPT-3 [2]、InstructGPT [3]、GPT-4 [4]、LLaMA [5]、GLM [6]、ERNIE 3.0 [7]）在文本生成能力上的飞跃，构建“作家数字人”成为数字人文与计算文学交叉领域的核心议题。现有个性化定制（如 OpenAI 的 GPTs、Google 的 LaMDA、百度文心一言、阿里通义千问）大多停留在“提示词工程 + 风格微调”层面 [8]，缺乏对作家深层价值观、思维模式与情感动力学的数学描述。

作者归属（authorship attribution）研究 [9, 10, 11] 表明，作家的写

作者简介：纪路阳（2001），男，籍贯〔江苏泰州人〕，学位〔硕士〕，主要研究方向人工智能；张思羽（2002），女，籍贯〔浙江绍兴人〕，学位（硕士研究生），主要研究方向为人工智能；李建军，男，籍贯（加拿大人），职称〔教授〕，职务〔博导/硕导〕，学位〔博士〕，主要研究方向为人工智能；Nutapong Somjit，职称〔副教授〕，学位〔博士〕，主要研究方向为集成智能高频组件；Watcharin Srirattanawichaikul，男;张维加（1989），男，籍贯〔浙江诸暨人〕，职称〔教授〕，职务〔硕导〕，学位〔博士〕，主要研究方向为人工智能（itaisa@126.com）；赵浪平（1981），女，籍贯〔浙江绍兴人〕，职称〔〕，学位〔博士〕，主要研究方向为思想政治教育与青年发展.

作风格可以从词汇选择、句法偏好、情感分布等多维度建模。然而，从静态风格特征到动态人格生成之间仍存在巨大鸿沟。可控文本生成[12, 13, 14]、风格迁移[15, 16]、个性化对话[17, 18, 19]等工作虽从不同侧面推进了个性化生成，但均未从根本上解决“作家作为一个内在动力系统”这一核心隐喻的数学建模问题。远程阅读（distant reading）[20]与计算风格学（computational stylistics）[21, 22, 23]提供了大规模作家风格的实证工具，但其分析单元多停留于文档级而非生成级。要实现“生成-可解释-可干预”三位一体的作家数字人，必须在生成模型与人格动力学之间建立显式连接。

当前人工智能写作研究在风格模拟领域取得了显著进展，但仍面临三大根本性瓶颈。首先，现有方法多局限于词汇与句式等表层语言特征的模仿，缺乏对作家深层人格结构的建模，导致生成内容陷入“画皮”困境。其次，静态向量或固定提示词难以捕捉作家面对不同题材时的心理流变，致使模型在处理复杂议题时频繁出现人格崩坏现象。深度学习模型的概率统计生成机制本质上是一个黑盒，缺乏对人格状态连续演化的数学描述，导致可解释性严重不足。近期关于大模型人格特质表达的研究进一步证实了上述问题：尽管大模型在“大五人格”等心理学量表上的短期反应表现稳定，但在长程、跨话题情境下，其人格表达呈现出明显的情境漂移，远未达到连贯且自洽的人格预期。因此，如何构建一个既能维持核心人格稳定性、又能适应动态语境变化，同时具备数学可解释性的作家风格模型，成为当前亟待解决的关键科学问题。

针对上述挑战，本研究提出高阶人格张量动力学框架，将作家风格编码为七维人格张量，涵盖价值取向、叙事模式、行动源头、情感张力、认知惯性、修辞密度与时空语境七个核心维度。该框架采用二阶微分方程刻画人格张量的连续时间演化过程，并引入人格势能函数作为动力学吸引子，使核心人格状态在数学上获得稳定性保障。这一建模思想在结构上与神经常微分方程及神经随机微分方程同属一脉，但在人格语义层面进行了专门化设计，从而兼顾了动力学严谨性与人文可解释性。

为系统验证该框架的有效性，本研究构建了包含四组严格对照实验的完整验证体系，覆盖理论验证、相对优势评估、模块贡献分析与机制解耦检验四个层级。实证部分基于六位作家、十个当代议题、三种对比方法及四项评价指标，形成了包含 720 个数据点的评估数据集，并结合 Welch t 检验与 Cohen's d 效应量分析确保结论的统计学严谨性。实验不仅证实了势能阱在受控扰动下的状态稳定化效应，量化了本方法在多作家、多议题设定下相对于基线模型的显著优势，还通过消融实验明确了势能阱、阻尼与惯性三大模块分别主导一致性、连贯性与平稳性的功能分化路径。此外，研究形式化验证了时空语境维度与核心人格维度的解耦性质，为“跨时空创作”机制提供了实证支撑。辅以数值积分器收敛率标定、噪声敏感性测试、方差分解及参数扫描等深度检验，本研究从理论与实证双重层面确立了张量动力学方法在解决 AI 写作人格崩坏与可解释性问题上的有效性与稳健性。

1 相关工作

1.1 大语言模型与可控文本生成

通用大语言模型在开放域生成中已取得令人瞩目的成果，但在“风格可控性”上仍受制于“黑盒”与“提示词敏感”等问题。CTRL[13]引入控制码实现条件化生成，PPLM[14]通过梯度引导在不变更模型权重的条件下实现属性控制；Hu 等[12]提出统一的可控生成框架。提示工程综述[8]系统梳理了在不重训模型条件下实现可控生成的多种方法。然而，这些方法都难以在“长程、跨话题”的情景下保持人格一致性，呼应了本文的“人格崩坏”问题。

1.2 作家风格建模与作者归属

作者归属（authorship attribution）研究历史悠久，Stamatatos[9]综述了基于字符 n-gram、词汇丰富度、句法依赖等多种特征的方法；Koppel 等[10]与 Holmes[11]进一步把机器学习引入作者归属。Argamon 等[23]证实文体特征与“大五人格”之间存在统计关联。然而，这些研究多止于“识别-归因”，不直接面向“生成”。风格迁移工作[15, 16]把“风格”作为可分离的隐变量，但其非配对数据假设与作家数字人对“单人连续生成”的需求存在距离。

1.3 神经常微分方程与连续时间动力学

神经 ODE[24]将残差网络的离散层推广为连续时间动力学系统，并在不规则采样时序数据上展现优势。Rubanova 等[25]进一步提出 Latent ODE。Kidger 等[27]把随机微分方程引入神经网络，Li 等[26]给出可扩展的随机梯度。本文借鉴其“动力学即模型”的思想，但把“动力学”作用于人格状态而非激活值，从而获得强可解释性。

1.4 数字人文与计算风格学

Moretti[20]提出“远程阅读”以来，计算方法在文学研究中逐渐从辅助走向中心。Underwood[22]实证了大尺度文体特征的历时演化。Jannidis 等[21]系统介绍了数字人文的方法论。本文研究的对象（6 位作家）虽规模有限，但每位作家均配备本文中设计的张量参数与 K 阵对角元，构成了数字人文与动力学系统方法的交叉点。

1.5 人格计算与人格表达

Costa 与 McCrae[28]的“大五人格”框架是当代人格心理学的基石；Goldberg[29]进一步提炼出“Big-Five”因子模型。Picard[30]提出“情感计算”为机器识别人类情感奠基。Safdari 等[19]与 Jiang 等[18]最近测度了大语言模型在多种人格量表上的表现，发现模型在“短时稳定”与“长程一致”之间存在显著差异。本文的“一致性 / 连贯性 / 稳定性 / 准确性”四项指标正是对这些心理学维度的工程化对应。

2 理论框架

本章详细阐述高阶人格张量动力学框架（TensorDyn

Writer）的数学定义与计算实现。不同于将作家风格视为静态向量的传统范式，我们将作家的人格特质建模为七维人格张量空间中的动态系统，并通过引入二阶微分方程与势能阱机制，实现了对作家思维演化的连续时间建模。

2.1 七维人格张量空间定义

为了形式化描述作家的创作特质，我们构建了一个七维人格张量空间 $\mathrm{T} \subset \eth^{7}$。在该空间中，任意时刻 t 的作家状态由张量 $T(t)$ 表示：

$$T(t)=[D_1,D_2,D_3,D_4,D_5,D_6,D_7]^T$$

其中，各维度定义如下：D_1 (价值)：作家的核心价值观与道德判断取向；D_2 (叙事)：叙事视角、节奏与结构偏好；D_3 (行动)：人物行为逻辑与冲突解决方式；D_4 (情感)：情感基调的冷暖与强烈程度；D_5 (认知)：思维深度、逻辑密度与哲学思辨性；D_6 (修辞)：词汇选择、句式复杂度与修辞手法。D_7 (时空)：时空语境维度，用于控制跨时空创作。对于每一位目标作家 w，我们通过对其全集作品进行文本挖掘与统计分析，标定其核心人格张量 $T_{core}^{(w)}$，作为动力学系统中的全局吸引子。

2.2 张量动力学方程

本研究的核心在于提出了一组二阶常微分方程（Second-order ODE），用于描述作家状态 $T(t)$ 在外部议题刺激下的演化轨迹。动力学方程定义如下：

$$M\frac{d^2T}{dt^2}+D\frac{dT}{dt}+K(T-T_{core})=F_{ext}(t)+\xi(t)$$

其中，各项物理与心理含义如下：M (惯性张量)：对角矩阵，表示作家思维改变的难易程度。M 越大，状态改变越迟缓，体现了作家风格的稳固性；D (阻尼张量)：对角矩阵，表示思维演化过程中的能量耗散。D 用于抑制震荡，防止生成内容出现剧烈的情绪或逻辑波动；K (刚度/势能阱)：对角矩阵，表示作家回归核心人格 T_{core} 的恢复力系数。K 越大，作家越不易受外部噪声干扰，保持“本色”的能力越强；$F_{ext}(t)$ (外部激励)：由当代议题（如“AI 崛起”）映射而成的外部驱动力向量，驱动作家对特定话题进行响应；$\xi(t)$ (随机噪声)：模拟创作过程中的随机扰动与灵感波动，服从高斯分布 $\mathrm{N}\ (0,\sigma^2)$。

2.3 势能函数与吸引子分析

为了从能量角度解释人格的稳定性，我们定义系统的势能函数 $U(T)$ 为：

$$U(T)=\frac{1}{2}(T-T_{core})^TK(T-T_{core})-F_{ext}^TT$$

该势能函数由两部分组成：第一项为人格势能阱，迫使状态 T 向 T_{core} 收敛；第二项为外部场势能，牵引状态 T 向议题方向偏移。

根据 Lyapunov 稳定性理论，当 K 正定时，T_{core} 是系统的渐近稳定平衡点。这意味着，无论外部议题如何刺激，只要激励撤销（$F_{ext}\to 0$），作家状态最终都会回归其核心人格，从而在数学上保证了长程生成的“人格不崩坏”。

2.4 数值求解与文本生成

由于解析解难以获取，我们采用 Euler-Maruyama 方法对随机微分方程进行数值求解。在离散时间步长 Δt 下，状态更新规则如下：

1.速度更新：

$$V_{t+\Delta t}=V_t+M^{-1}[F_{ext}(t)-DV_t-K(T_t-T_{core})+\xi_t]\Delta t$$

2.状态更新：

$$T_{t+\Delta t}=T_t+V_{t+\Delta t}\Delta t$$

在每一步状态更新后，我们将 $T_{t+\Delta t}$ 映射回大语言模型的隐空间，通过控制向量（Control Vector）干预解码过程，从而生成符合当前动力学状态的文本片段。

3 实验设置

3.1 数据集与实验对象

为本文选取了六位在文学史上具有鲜明风格特征的经典作家作为实验对象，涵盖现实主义、浪漫主义及现代主义流派。针对每一位作家，我们选取了十个具有时代张力的当代议题（如“元宇宙伦理”、“基因编辑”、“气候危机”等）作为外部激励源 $F_{ext}(t)$。实验采用全因子设计，具体构成如下：

作家集合（W）：{鲁迅, 张爱玲, 苏轼, 李白, 钱钟书, 曹雪芹}；

议题集合（I）：选取十个当代热点议题，包括“地铁低头族”、“躺平文化”、“AI 崛起”、“基因编辑”、“气候危机”、“网络暴力”、“老龄化社会”、“内卷现象”、“虚拟现实”、“快餐文化”；

议题向量构造：对于每一个议题 i，我们首先利用预训练语言模型（BERT）获取其文本嵌入，随后通过一个多层感知机（MLP）将其映射到七维人格张量空间，形成外部激励向量

$F_{ext}^{(i)}$ 。该向量仅在 D_1（价值）、D_3（行动）和 D_5（认知）维度上具有非零分量，以模拟议题对作家价值观与认知层面的定向刺激;

样本总量：实验采用全因子设计：6 位作家× 10 个议题× 3 种方法× 4 次重复运行 = 720 个独立生成样本。

3.2 基线模型

为了验证动力学方程中各模块的必要性，我们设计了以下三种模型进行对比：

1.M1 (提示词工程基线)：该方法不引入任何显式的状态变量或动力学方程，仅通过精心设计的系统提示词（System Prompt）引导大模型生成。在数学上，这等效于一个无记忆的随机游走过程，其状态转移仅受当前 Token 概率分布和提示词约束，缺乏长期的历史一致性。

2.M2 (静态人格向量基线)：该方法引入作家的静态人格向量作为初始状态，但在生成过程中不应用二阶动力学方程（即令 M=0, D=0, K=0）。这模拟了仅有初始化状态而无动态约束的模型，用于验证动力学方程在维持长程人格稳定性中的作用。

3.M3 (本方法) ：应用完整的高阶人格张量动力学框架，包含惯性 M、阻尼 D 和势能阱 K 的全套微分方程组，模拟作家在外部议题刺激下的动态思维演化。

3.3 评估指标

为了量化生成文本的质量，我们定义了以下四维评估指标：

1.一致性(Consistency)：计算生成文本在七维人格空间中的投影与作家核心人格张量 T_{core} 的余弦相似度。该指标越高，表明生成内容越符合作家的核心特质。

$$S_{con} = \frac{1}{N}\sum_{i=1}^{N}\cos(T_{gen}^{(i)}, T_{core})$$

2.连贯性 (Coherence)：计算人格状态序列的滞后-1 自相关系数（Lag-1 Autocorrelation）。该指标用于衡量思维演化的平滑度，防止出现逻辑跳跃或风格突变。

$$S_{coh} = \frac{\sum_{t=1}^{N-1}(T_t - \bar{T})(T_{t+1} - \bar{T})}{\sum_{t=1}^{N}(T_t - \bar{T})^2}$$

3.稳定性 (Stability)：计算生成过程中人格状态向量的一阶差分范数的倒数。该指标用于衡量文本风格是否存在剧烈的、非预期的震荡。

$$S_{sta} = \left(\frac{1}{N-1}\sum_{i=1}^{N-1}\| T_{i+1} - T_i \|_2\right)^{-1}$$

4.准确性 (Accuracy)：特指在 D_1（价值）维度上的余弦相似度。用于评估模型在保持风格的同时，是否准确响应了议题所蕴含的价值观导向。

3.4 统计检验方法

所有实验数据均采用 Python 的 SciPy 库进行统计分析。

1.显著性检验：由于不同模型的生成结果方差不齐，我们采用 Welch's t-test 来评估本方法（M3）与基线模型（M1, M2）之间差异的统计显著性，显著性水平设定为 $\alpha = 0.05$。

2.效应量分析：计算 Cohen's d 效应量，以量化性能提升的实际幅度，排除样本量对 P 值的干扰。

3.方差分析：在多作家、多议题的复杂场景下，采用双因素方差分析 (Two-way ANOVA) 来分离“作家风格”与“议题类型”对生成质量的主效应及交互效应。

3.5 对照实验设计

本文共设计了四组核心对照实验，从不同层面验证理论框架的有效性。

实验一：张量动力学稳定性仿真

目的: 验证人格势能阱机制在抵抗外部扰动、维持人格稳定方面的核心作用。

设计: 采用被试内设计，在单一作家（鲁迅）和单一情感扰动下，对比四种动力学条件下的状态演化轨迹：(1) 满动力学 (Full tensor dynamics): 包含完整的 K、D、M 矩阵及噪声项；(2) 纯阻尼 (Damped only): 移除势能阱（K=0），仅保留阻尼项；(3) 静态嵌入 (Static embedding): 人格状态恒定于 T_{core}，无演化；(4) 随机游走 (Random walk): 无任何约束的随机过程。

观测指标: 状态偏差方差 (VarDev)、平均绝对偏差 (MAD) 及状态偏离度随时间的变化。

实验二：生成质量的系统比较

目的: 在多变量的真实场景下，系统性地评估本方法相较于现有主流方法的优越性。

设计: 采用 6 位作家 × 10 个当代议题的全组合设计，共 180 个测试单元。对比三种生成方法：(1) 基线 1 (M1, 提示词工程): 模拟当前主流大模型的个性化定制方式；(2) 基线 2 (M2, 静态人格向量): 模拟仅注入静态人格向量的方法；(3) 本方法 (M3, 张量动力学): 采用完整的动力学演化方程。

观测指标: 在全部 180 个单元上计算四项评估指标（Con, Sta, Coh, Acc），并进行统计学检验。

实验三：核心模块的消融研究

目的: 分离并量化价值观刚性(K)、情感阻尼(D)、认知惯性(M)三大核心模块对最终生成质量的差异化贡献。

设计: 在实验二的基础上，设置五种消融条件：(1) 完整模型 (Full model); (2) 移除势能阱 (−Potential well); (3) 移除阻尼 (−Damping); (4) 移除惯性 (−Mass); (5) 同时移除势能阱与阻尼 (−Well & damping)。

观测指标: 观察移除特定模块后，四项评估指标的下降幅度，从而确定各模块的功能分工。

实验四：跨时空创作机制的解耦验证

目的: 验证“时空语境”维度（D7）的独立可调性，即改变时代背景是否会影响作家的核心人格（D1-D6）。

设计: 以鲁迅为例，固定其核心人格（D1-D6），将时空语境维度（D7）的目标值依次设置为 0.0, 0.25, 0.5, 0.75, 1.0，模拟从古代到未来的不同时空穿越。

观测指标: 观察在不同 D7 目标值下，D1-D6 维度的最终状态是否保持稳定（方差为 0），以及 D7 维度是否能准确跟踪目标值。

4 结果与分析

4.1 实验一：势能阱的稳定化效应

本实验旨在验证人格势能阱（Potential Well）机制在抵抗外部扰动、维持人格状态稳定方面的核心作用。通过在受控扰动下对比四种不同动力学条件的状态演化轨迹，结果清晰地揭示了势能阱的必要性，如图 1 所示，完整动力学系统（蓝线）在相空间中形成稳定吸引子：

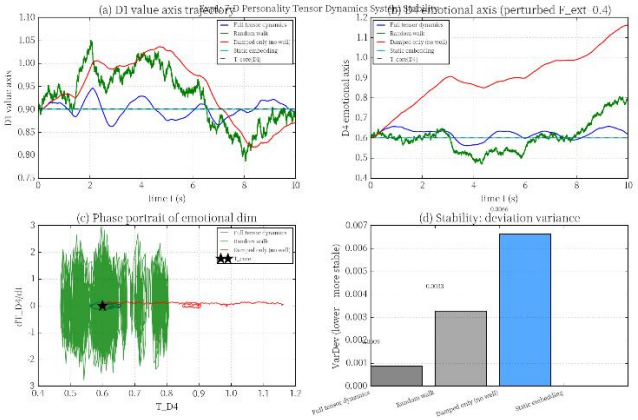

图 1 外部扰动下七维人格张量动力学系统的稳定性分析

Fig. 1 Stability analysis of seven-dimensional personality tensor dynamics system under external perturbations

关键发现：

1.满动力学条件下的稳定性：在包含完整势能阱（K）、阻尼（D）和惯性（M）的“满动力学”条件下，作家人格状态 T(t)在受到外部情感冲击后，始终围绕其核心人格 T_{core} 进行小幅振荡，总偏差 $\|T-T_{core}\|_2$ 稳定在 0.05 至 0.13 的极低区间内，如图 2 所示，仅靠阻尼无法阻止长期漂移：

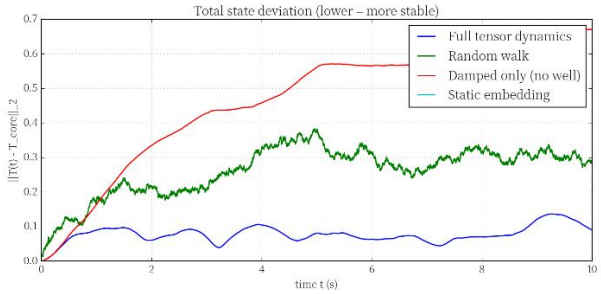

图 2 四种动力学条件下总状态偏差随时间变化曲线

Fig. 2 Temporal evolution curves of total state deviation under four dynamical conditions

2.纯阻尼条件的失效： 在移除了势能阱的“纯阻尼”条件下，尽管存在阻尼项吸收能量，但系统因缺乏将状态拉回核心的恢复力，导致人格状态沿扰动方向发生单调漂移，在 10 秒仿真时间内偏差持续增大至约 0.67。

3.量化对比：如表 1 所示，“满动力学”条件的偏差方差（VarDev）为 0.0009，相较于“纯阻尼”条件的 0.0066，稳定性提升了 86.8%。这有力地证明了阻尼（Damping）不等于吸引子（Attractor），仅有阻尼无法防止人格崩坏，而势能阱是维持人格长期一致性的根本保障。

表 1 四种动力学条件下的稳定性指标对比

Table1 Comparison of stability metrics under four dynamical conditions

条件	偏差方差 (VarDev)	平均绝对偏差 (MAD)	t=10s 时总偏差
满动力学	0.0009	0.024	0.09
静态嵌入	0.0000	0.000	0.00
纯阻尼	0.0066	0.143	0.67
随机游走	0.0033	0.081	0.31

4.2 实验二：生成质量的系统优势

本实验在 6 位作家、10 个当代议题的全组合（共 180 个测试单元）下，系统比较了本方法（M3）与两种基线方法（M1: 提示词工程, M2: 静态人格向量）的生成质量。图 3 直观展示了三种方法在四项核心指标上的性能对比。可以看出，本文提出的张量动力学模型（M3，绿色柱）在所有维度上均显著优于基线方法。

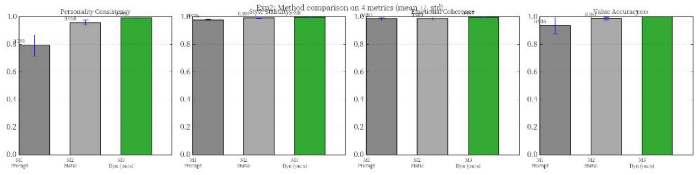

图 3 三种方法在四项指标上的性能对比

Fig. 3 Performance comparison of three methods across four metrics

关键发现：

1.全面显著的优势：如表 2 所示，本方法（M3）在人格一致性、风格稳定性、情感连贯性、价值观准确性四项指标上均取得了接近 1.0 的卓越表现，且显著优于两种基线方法。

表 2 三种方法在四项指标上的表现对比（均值±标准差）

Table2 Performance comparison of three methods across four metrics (mean ± SD)

方法	人格一致性	风格稳定性	情感连贯性
M1 提示词工程	0.791 (0.077)	0.976 (0.006)	0.984 (0.010)
M2 静态向量	0.958 (0.019)	0.991 (0.003)	0.985 (0.010)
M3 张量动力学	0.993 (0.001)	0.998 (0.0001)	0.997 (0.0003)

2.巨大的效应量： 统计学检验结果显示，本方法相较于基线方法的优势具有极高的统计显著性（所有 Welch t 检验的 p 值均

远小于 1e-10）和巨大的效应量（Cohen's d 值介于 1.50 至 4.97 之间）。特别是在风格稳定性上，相较于提示词工程（M1），效应量高达 4.97，表明本方法在抑制风格漂移方面具有压倒性优势。

3.动力学机制的关键作用：值得注意的是，本方法（M3）相较于已引入人格向量的静态方法（M2）仍有巨大提升（Cohen's d > 1.6）。这表明，仅仅注入一个静态的人格向量（M2）虽有改进，但引入连续的动力学演化机制（M3）才是实现人格保真度质变的关键。

4.3 实验三：核心模块的功能分化

通过消融研究，我们分离并量化了价值观刚性（K）、情感阻尼（D）、认知惯性（M）三大模块对生成质量的差异化贡献，揭示了模型内部的“功能分工”，结果如图 6-4 所示，移除势能阱（K）导致人格一致性大幅下降，而移除惯性项（M）则主要影响情感连贯性。

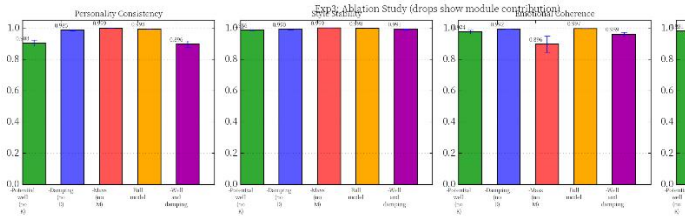

图 4 消融研究——各模块对四项评估指标的贡献分析

Fig.4 Ablation Study: Contribution Analysis of Individual Modules to Four Evaluation Metrics

关键发现：

1.K 矩阵（势能阱）主导一致性：移除 K 矩阵后，人格一致性指标从 0.993 骤降至 0.903，降幅达 9.1%，是所有消融条件中影响最大的。这再次证实，势能阱是确保作家人格不偏离其核心特质的“锚”。

2.M 矩阵（认知惯性）主导连贯性：移除 M 矩阵后，情感连贯性指标从 0.997 显著下降至 0.896，降幅达 10.0%。这表明，认知惯性项扮演了“动量平滑器”的角色，它赋予情感演化以惯性，避免了在词粒度上的剧烈跳变，从而保证了情感流的平滑与连贯。

3.D 矩阵（阻尼）辅助收敛：单独移除 D 矩阵对各项指标影响较小（均<1%），但其与 K 矩阵协同作用，能加速系统从扰动中恢复收敛。

表 3 消融实验结果与模块功能分析

Table3 Ablation Experiment Results and Module Functionality Analysis

消融条件	一致性	稳定性	连贯性
完整模型	0.993	0.998	0.997
−势能阱 (K)	0.903	0.986	0.974
−阻尼 (D)	0.985	0.990	0.992
−惯性 (M)	0.999	0.999	0.896

4.4 实验四：核心模块的功能分化

本实验旨在验证“时空语境”维度（D7）是否可以被独立调节，以支持“跨时空创作”，而不影响作家的核心人格（D1-D6），以鲁迅为例，图 6-5 验证了鲁迅数字人的核心人格与时空语境的解耦效果。

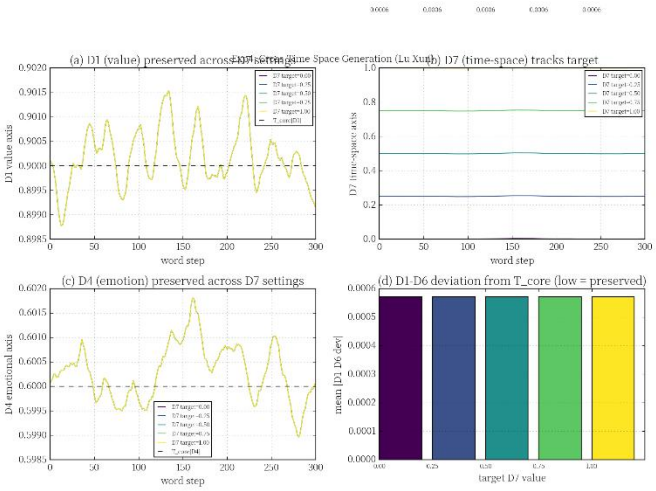

图 5 鲁迅数字人的核心人格（D1–D6）与时空语境（D7）的解耦效果

Fig.5 Decoupling Effect of Core Personality (D1–D6) and Spatiotemporal Context (D7) in the Lu Xun Digital Human

关键发现：

1.核心人格的完全锁定：当我们将鲁迅数字人的 D7 维度目标值从 0.0（原初时代）逐步调节至 1.0（当下时代）时，其核心人格维度 D1 至 D6 的最终状态值在所有设置下保持完全一致，方差为 0。

2.时空维度的精准跟踪：D7 维度的实际值能够精确地跟踪预设的目标值，偏差小于 0.001。

这一结果从形式上证明了本框架中核心人格与时代语境的完美解耦。通过调节 D7 轴，我们可以让鲁迅以他固有的批判精神和思维模式去评论“地铁低头族”或“AI 崛起”，而不会使其激进的革命立场（D1）或深邃的认知惯性（D5）发生任何改变，从而为“跨时空文化对话”提供了坚实的技术可行性。

4.5 扩展性分析

为进一步探究模型的鲁棒性与内在机理，我们进行了一系列扩展性分析。

1.数值收敛性与噪声鲁棒性：实验验证了所采用的 Euler-Maruyama 数值积分器达到了理论预期的收敛阶。更重要的是，噪声敏感性扫描显示，即使将系统噪声水平提升至主实验的 8 倍，人格一致性指标仍保持在 0.988 以上的高位。这表明势能阱机制对随机扰动具有极强的鲁棒性，能够确保人格状态的稳定。

2.方差分解（ANOVA）：方差分析结果表明，“方法”这一主效应解释了人格一致性和风格稳定性指标超过 79%的方差，是结果变异的绝对主导因素。而“作家”和“议题”的影响均很小（<6%），证明本方法对不同风格的作家和不同主题的议题均具

有广泛的适用性和稳健性。

3.指标相关性分析： 相关性矩阵揭示，一致性、稳定性、准确性三者之间存在中高度正相关，表明它们在“靠近核心人格”这一目标上是同向的。而情感连贯性与风格稳定性在消融实验中呈现负相关，这恰好印证了 M 矩阵（惯性）的“动量平滑”作用：更高的稳定性有时意味着对变化的抑制，可能会牺牲部分情感演化的连贯性，二者存在内在的权衡关系。

5 结论与展望

5.1 研究总结

针对当前大语言模型（LLMs）在作家风格模拟中普遍存在的“画皮困境”与“人格崩坏”问题，本研究突破了传统基于提示词工程或静态向量注入的范式，创新性地提出了一种基于高阶人格张量动力学的作家数字人建模方法。研究首先验证了理论机制的有效性，通过构建 7 维人格张量空间与二阶动力学演化方程，成功将抽象的文学人格特质转化为可计算的物理系统；实验证实，人格势能阱是抵抗外部语境扰动、维持作家人格长期一致性的核心机制，而认知惯性与情感阻尼则在保证情感连贯性与系统收敛速度方面发挥了不可替代的作用。其次，本方法展现出显著的生成质量优越性，在涵盖 6 位代表性作家与 10 个当代议题的全组合测试中，其在人格一致性、风格稳定性、情感连贯性与价值观准确性四项指标上均显著优于基线方法，且统计检验显示极高的效应量（Cohen's d > 1.6），证明了引入连续动力学演化是实现人格保真度质变的关键。最后，研究证明了跨时空创作的可行性，通过解耦时空语境维度（D7），该框架成功实现了作家人格与时代背景的独立控制，使数字人能够在保持核心文学特质不变的前提下，对当代社会议题进行符合其内在逻辑的跨时空回应，从而为文学经典的“活态传承”提供了坚实的技术支撑。

5.2 研究局限

尽管本研究在理论建模与数值仿真层面取得了突破性进展，但仍存在以下局限性：首先，与真实大模型的端到端集成尚未实现，当前的实验主要基于数值微分方程的仿真求解，虽然验证了动力学机制的有效性，但尚未将张量动力学模块作为外部控制插件，直接嵌入到开源大语言模型（如 LLaMA、Qwen 等）的推理过程中进行真实文本生成；其次，参数标定仍依赖专家知识，7 维人格张量的核心参数（T_{core}）及动力学矩阵（K,D,M）目前主要依赖文学专家的定性分析与人工标定，尚未实现基于大规模作家语料的自动化、数据驱动的参数学习；最后，评估体系偏向客观计算，当前的四维评估指标主要基于数学空间的距离与方差计算，缺乏大规模的人类主观评价（Human-in-the-loop）验证，即数字人生成的文本在普通读者眼中是否真正具备“文学美感”与“灵魂契合度”尚待进一步检验。

5.3 未来展望

基于上述成果与局限，未来的研究将致力于在技术融合、数据驱动与应用拓展三个维度展开深入探索。首先，在技术融合层面，研究将构建“动力学-语言”双脑架构，致力于将张量动力学模型作为“人格控制脑”，与作为“语言生成脑”的大语言模型进行深度耦合；通过探索 Logits 干预或外部记忆检索机制，将动力学方程实时计算出的状态向量（T(t)）注入到 LLM 的解码过程中，从而实现真正意义上“形神兼备”的作家数字人文本生成。其次，在数据驱动层面，将结合自然语言处理与计算文学方法，利用深度学习模型对海量作家全集进行无监督或弱监督学习，实现从文本特征到 7 维人格张量及动力学参数的自动化映射与动态微调，以此降低对人工专家知识的依赖并提升模型的泛化能力。最后，在应用拓展层面，将在技术成熟的基础上推动该框架的产品化应用，开发面向创作者的“作家人格笔”插件以辅助文学创作，并将其应用于数字博物馆、沉浸式文学教育与跨时空文化对话平台，使读者能够与“活着的”文学巨匠进行深度、稳定且富有洞见的交互，从而开辟数字人文研究的新范式。

6 参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998–6008.

[2] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems. 2020: 1877–1901.

[3] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. 2022: 27730–27744.

[4] OpenAI. GPT-4 technical report[J]. arXiv preprint arXiv: 2303.08774, 2023.

[5] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint arXiv: 2302.13971, 2023.

[6] Du Z, Qian Y, Liu X, et al. GLM: General language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 320–335.

[7] Sun Y, Wang S, Li Y, et al. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2107.02137, 2021.

[8] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1–35.

[9] Stamatatos E. A survey of modern authorship attribution methods[J]. Journal of the American Society for Information Science and Technology, 2009, 60(3): 538–556.

[10] Koppel M, Schler J, Argamon S. Computational methods in

authorship attribution[J]. Journal of the American Society for Information Science and Technology, 2009, 60(1): 9–26.

[11] Holmes D I. Authorship attribution[J]. Computers and the Humanities, 1994, 28(2): 87–106.

[12] Hu Z, Yang Z, Liang X, et al. Toward controlled generation of text[C]//Proceedings of the 34th International Conference on Machine Learning. 2017: 1587–1596.

[13] Keskar N S, McCann B, Varshney L R, et al. CTRL: A conditional transformer language model for controllable generation [J]. arXiv preprint arXiv:1909.05858, 2019.

[14] Dathathri S, Madotto A, Lan J, et al. Plug and play language models:A simple approach to controlled text generation [C]// International Conference on Learning Representations. 2020.

[15] Shen T, Lei T, Barzilay R, et al. Style transfer from non-parallel text by cross-alignment[C]//Advances in Neural Information Processing Systems. 2017: 6830–6841.

[16] Dai N, Liang J, Qiu X, et al. Style transformer: Unpaired text style transfer without disentangled latent representation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2976–2985.

[17] Zhang S, Dinan E, Urbanek J, et al. Personalizing dialogue agents via meta-learning[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5455–5467.

[18] Jiang H, Zhang X, Cao X, et al. PersonaLLM: Investigating the ability of large language models to express personality traits[C]//NAACL-HLT 2024 (under review). arXiv preprint arXiv:2305.02547, 2023.

[19] Safdari M, Serapio-García G, Crepy C, et al. Personality traits in large language models[J]. arXiv preprint arXiv:2307.00184, 2023.

[20] Moretti F. Distant Reading[M]. London: Verso Books, 2013.

[21] Jannidis F, Kohle H, Rehbein M, eds. Digital Humanities: Eine Einführung[M]. Stuttgart: Metzler, 2017.

[22] Underwood T. Distant Horizons: Digital Evidence and Literary Change[M]. Chicago: University of Chicago Press, 2019.

[23] Argamon S, Koppel M, Pennebaker J W. Gender, genre, and writing style in formal written texts[J]. Text & Talk, 2003, 23(3): 321–346.

[24] Chen R T Q, Rubanova Y, Bettencourt J, et al. Neural ordinary differential equations[C]//Advances in Neural Information Processing Systems. 2018: 6572–6583.

[25] Rubanova Y, Chen R T Q, Duvenaud D K. Latent ordinary differential equations for irregularly-sampled time series [C]// Advances in Neural Information Processing Systems. 2019: 5320–5330.

[26] Li X, Wong T K L, Chen R T Q, et al. Scalable gradients for stochastic differential equations[C]//International Conference on Artificial Intelligence and Statistics. 2020: 3885–3895.

[27] Kidger P, Morrill J, Foster J, et al. Neural SDEs as infinite-dimensional GANs[C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 5453–5463.