arXiv:2601.01060v1 [cs.CL] 3 Jan 2026

面向可控强度的无监督文本风格迁移

ShuhuanGu[1], **Wenbiao Tao**[1],**XinchenMa**[1],**KangkangHe**[2],**YeGuo**[2], **XiangLi**[1],**Yunshi Lan**[1]*[1]华东师范大学，中国上海[2]宝武集团，中国上海treegu2003@gmail.comyslan@dase.ecnu.edu.cn

摘要

无监督文本风格迁移（UTST）旨在构建一个系统，在没有平行文本对的情况下，迁移给定文本的风格属性。与不同风格极性之间的文本迁移相比，面向可控强度的UTST更具挑战性，这是因为不同强度等级之间的风格特征存在细微差异。针对缺乏平行数据以及相邻强度等级难以区分所带来的挑战，我们提出了一种先监督微调后近端策略优化范式，用于微调大语言模型。我们首先使用合成平行数据对大语言模型进行微调。随后，我们进一步使用近端策略优化训练大语言模型，其中精心设计奖励，以分层级别区分风格强度。我们综合考虑全局和局部风格特征来构建奖励函数。在两个无监督文本风格迁移基准上的实验表明，两种奖励各具优势，将其用于大语言模型微调能够基于各种评估指标有效提升大语言模型骨干的性能。即使对于相近的强度级别，我们仍然可以观察到生成文本之间明显的风格差异。

1 引言

文本风格迁移（TST）旨在自动修改给定文本的风格属性，例如可读性（North 等人，2023年）、情感（Zhang 等人，2015年）或正式程度（Briakou 等人，2021年），同时保留其原有的语义内容。这项将原文转换为所需风格的任务，在教育（Tanet al.，2024年）、法律（Li 等人，2021年）和社交媒体（Prabhumoye 等人，2018年）等领域有广泛的应用。除了在风格极性之间进行迁移外，近期研究也开始评估不同强度尺度上的风格迁移。由于不同强度级别的风格属性和特征差异细微，这一任务更具挑战性（Mir 等人，2019年；Mukherjee等人，2025年）。此外，由于缺乏大规模、高质量的平行语料库，无监督文本风格迁移（UTST）也面临新的挑战。图1展示了一个可控强度无监督文本风格迁移的示例。

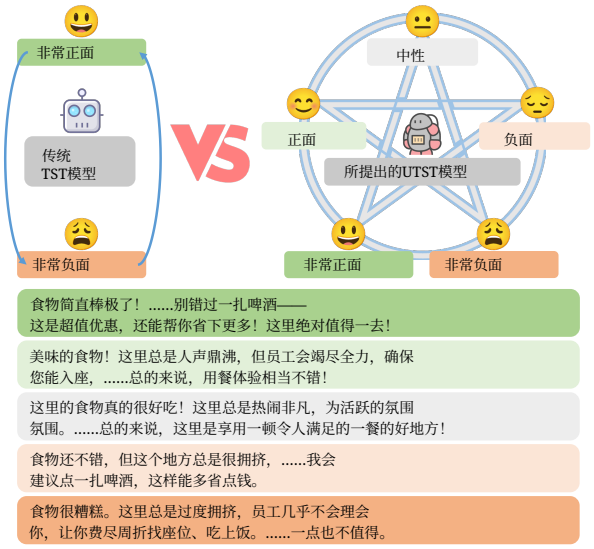


图1：风格极性文本迁移与跨强度级别文本迁移的比较。

现有研究已经取得了显著进展。我们将其分为两类：(1)属性替换。这些方法会将文本中的特定风格短语替换为目标风格中的相应短语。因此，额外的语料库（Li 等人，2018年；Sudhakar 等人，2019年）、分类器模型（Lee，2020年；Fu 等人，2018年）和映射模型（Luo 等人，2019b年）被引入迁移模型中。然而，由于直接替换属性，迁移后的句子可能会出现语义不连贯。(2)句子生成。得益于大语言模型的出现，文本风格可以逐个词元进行迁移。提示工程（Reif 等人，2022年；Suzgun等人，2022年）、指令微调（Mukherjee 和 Dušek，2023年；Zong 等人，

* 通讯作者

2024年）和强化学习（RL）（Deng 等人，2022年；Liu 等人，2024年）被应用于语言模型，使其学会生成目标风格文本。然而，面对可控强度生成的需求时，这些方法经常存在风格缺陷。

文本转换的可控强度要求模型充分理解不同层次的风格特征并进行风格迁移。总而言之，我们需要重点关注两个挑战：

- **（C1）缺乏高质量文本风格迁移平行数据**。配对文本风格迁移数据能够帮助模型学习精确的迁移模式。而在缺乏高质量平行数据的情况下，模型的知识来源受到限制，从而影响可控强度的文本风格迁移。

- **（C2）跨越难以区分的强度进行迁移**。广为人知的两极风格转换具有易于识别的鲜明特征。然而，阶梯式强度文本之间存在更为细微的差异，如图1所示，这清晰地展示了 UTST 任务。

为应对这些挑战，我们提出了一种用于 UTST 的先 SFT 后 PPO 范式。我们的方法首先使用大语言模型将源文本重写为不同强度等级的目标风格，从而构建伪平行语料库。随后，使用预训练风格分类器对生成结果进行筛选，以确保其与所需的风格强度保持一致，最终得到高质量伪平行语料库。在合成数据的基础上，我们对现有大语言模型进行微调，以建立基础的文本风格迁移能力。为进一步提升风格一致性，并实现超越监督学习的细粒度控制，我们引入带有层次化奖励的近端策略优化。具体而言，我们设计了句子级奖励以捕捉全局风格对齐，并设计了词典级奖励以指导细粒度的词汇选择。这种双层奖励结构使模型能够生成风格强度更加准确且一致的输出。

总而言之，我们的贡献如下：

- 我们提出一种强度可控的可控强度文本风格迁移范式：首先通过监督微调对模型进行微调，以实现基础风格迁移；随后利用强化学习进一步优化模型，使其能够处理不同的风格强度。

- 我们构建了伪平行语料库，以解决TST模型微调所需平行数据匮乏的问题；此外，我们还设计了句子级、词典级和语义级奖励，使模型能够从多个粒度适应可控强度的UTST。

- 我们为这项任务构建了两个基准测试。在可读性和情感风格迁移方面的实验结果表明，即使强度级别相近，我们的方法也能有效生成风格差异明显的文本。

2 相关工作

文本风格迁移。文本风格迁移通过属性替换或句子生成来改变文本风格，同时保留其含义。

属性替换 方法利用外部语料库替换具有特定风格的短语（Li 等人，2018年；Sudhakar 等人，2019年）、分类器（Lee，2020年；Fu 等人，2018年）或映射模型（Luo 等人，2019b年）。这些方法在粗粒度的情感极性迁移方面效果显著，但往往存在语义不连贯的问题，且无法处理细粒度强度——直接替换短语（例如“good”→ “bad”）无法捕捉强度差异（例如“great”和“amazing”）（Mir 等人，2019年）。

句子生成方法可分为基于大语言模型的非解耦方法和解耦方法。非解耦方法（例如提示工程（Reif 等人，2022年；Suzgun等人，2022年）、指令微调（Mukherjee 和 Dušek，2023年；Zong 等人，2024年）采用交织的风格—内容表征，这简化了训练过程，但缺乏调节细微风格强度的能力。强化学习扩展（Deng 等人，2022年；Liu 等人，2024年）仅优化宽泛的风格对齐，而非风格的细微梯度。解耦方法（Han 等人，2024年）通过辅助任务将风格与内容解耦，以实现精确控制，但需要大量人工标注。

我们的范式沿袭非解耦方法的路线，通过大语言模型生成伪平行数据，并采用基于分层 PPO 的强化学习实现细粒度风格控制，从而无需平行数据即可进行稳健的 UTST。

可控文本生成。可控文本生成（可控文本生成）能够生成具有所需属

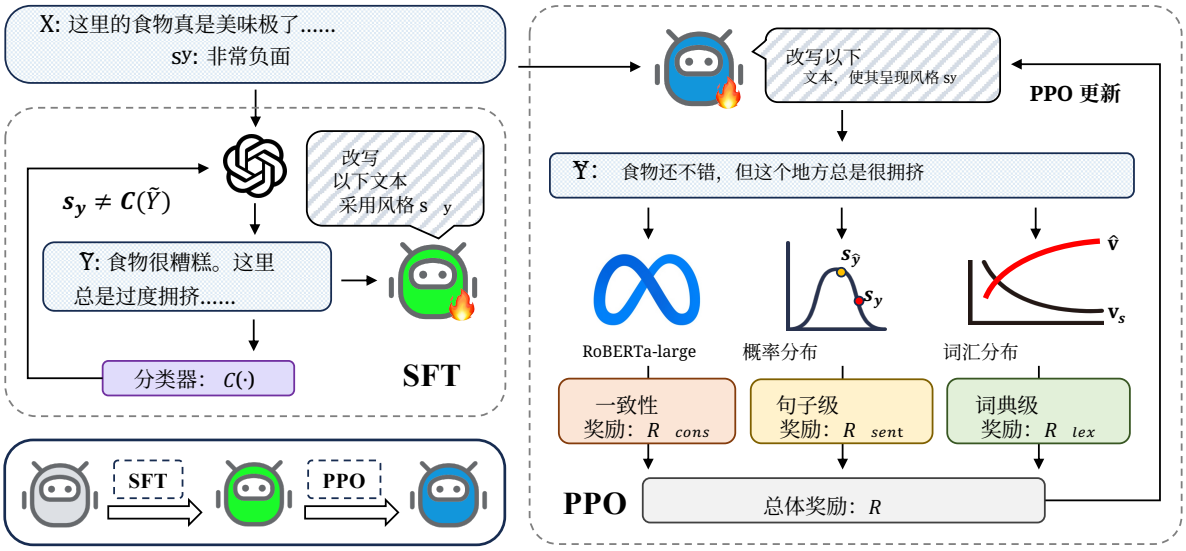


图2：采用 SFT 后接 PPO 范式的 UTST。

性的文本。基于 Transformer 的大语言模型通过基于提示的方法推动了可控文本生成的发展（Suzgun等人，2022年），使用户能够通过自然语言指令灵活地进行控制。微调（Mukherjee 和 Dušek，2023年）和强化学习（RL）（Deng 等人，2022年；Liu 等人，2024年）进一步提升了模型在情感修改（Zhang 等人，2015年）和可读性调整（North等人，2023年）等任务中的适应能力。然而，由于缺乏平行数据，且难以捕捉细微的风格差异，具有可控强度的 UTST 仍然面临挑战（Mukherjee等人，2025年）。我们的方法通过结合伪平行数据生成与近端策略优化（PPO）方法中的层次化奖励，弥补了这些不足，从而实现细粒度的风格控制，并在无需监督的情况下取得出色的 UTST 性能。

3 任务陈述

UTST 是一项自然语言处理任务，旨在不借助平行文本对的情况下转换给定文本的风格。具体而言，给定 k 组非平行语料库，每组都可以表示为 $\mathcal{D}^s = \{X_i\}_{i=1}^{|\mathcal{D}^s|}$。其中，$s$ 标注了预定义的风格强度集合 $\mathcal{S} = \{s_1, s_2, .., s_k\}$，该集合是风格迁移的类别尺度。给定一段风格强度为 $s_x \in \mathcal{S}$ 的源文本 $X = (x_1, x_2, ..., x_n)$，UTST 的关键在于开发一种模型，使其能够将文本转换为风格强度为$s_y \in \mathcal{S}$ 的目标文本$Y = (y_1, y_2, ..., y_m)$，同时保留原始内容。

关于可控强度的 UTST，生成过程可以建模为：

$$\hat{Y} \leftarrow \mathcal{G}(X, s_y),$$

ˆ 其中，$\mathcal{G}(\cdot)$ 是一个 UTST 模型，而 Y 应是风格强度为 s_y 的迁移文本。

4 UTST 的范式

为了训练这样的模型 $\mathcal{G}(\cdot)$ 在没有平行语料库的情况下，我们充分利用大语言模型进行可控强度的 UTST，并提出一种先 SFT 后 PPO 的范式，通过层级奖励训练大语言模型。具体而言，我们首先使用基于大语言模型的流水线生成高质量的伪平行语料库。然后，使用平行风格化文本对大语言模型进行微调，将其训练为 $\mathcal{G}(\cdot)$。接着，我们进一步使用近端策略优化（PPO）训练该大语言模型，其中的奖励经过精心设计，用于区分细粒度风格强度。我们的方法概览如图 2所示。我们将详细讨论获取 $\mathcal{G}(\cdot)$ 的步骤。

4.1 使用合成平行文本进行监督微调

监督微调对于增强大语言模型的文本风格迁移能力至关重要。尽管大语言模型在 TST 任务中展现出令人印象深刻的零样本和少样本性能，但其输出通常缺乏遵循特定风格约束的一致性，例如目标可读性等级或情感语调（Suz- gun等人，2022年；Hu等人，2020年）。监督微调通过将大语言模型的输出准确地与目标文本的属性对齐来解决这一问题，尤其是在风格标注有细粒度强度时。为此

，我们采用指令微调（Zhang 等人，2023年），在这一过程中，大语言模型基于精心设计的提示词进行微调，这些提示词明确指定了目标风格的属性。

由于监督微调需要高质量的平行文本对数据集作为强监督，其中每一对文本由输入文本及其经过风格转换的对应文本组成。由于缺少适用于 UTST 任务的平行文本对，我们提出通过以下流程合成用于微调的数据。

• 对于从语料库中 X 采样的每个源文本，我们将其与 $\mathcal{S}$ 中的多种强度进行配对，形成一组输入（X、s_y），其中 $s_y \in \mathcal{S}$ 是我们希望将文本迁移 X 到的目标风格，而迁移后的文本风格强度为 $s_y \neq s_x$。

• 给定一个输入，我们利用 GPT-4o-mini（OpenAI 等人，2024年）将其迁移 X 为具有目标风格强度的文本 $\tilde{Y}$[1]。

• 我们使用非平行语料库训练一个分类器 $\mathcal{C}(\cdot)$ 作为评判器，以确定合成的目标文本是否满足所要求的强度。˜如果合成文本 Y 被预测为目标˜强度，我们就将平行数据（X、s_y、Y）加入数据集。否则，我们会重复上述步骤，直到大语言模型生成所需的文本。如果经过 10 次迭代仍未生成合适的输出，我们将其丢弃 X。

因此，我们获得了一组用于风格迁移的平行数据 $\mathcal{D} = \{X_i, s_y, \tilde{Y}_i\}_{i=1}^{N}$，其中包含对齐文本对。我们将一个大语言模型视为 $\mathcal{G}(\cdot)$，并使用以下目标在 $\mathcal{D}$ 上对其进行微调：

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{N}\sum_{j=1}^{k} \log P(\tilde{Y}_i | X_i, s_j),$$

其中，概率由大语言模型建模。微调后，模型 $\mathcal{G}(\cdot)$能够生成符合目标风格的ˆtext Y。

4.2 近端策略优化的奖励函数

虽然 SFT 通过强监督有效地引导模型捕捉风格特征，但在不同输入中确保模型始终遵循细微或复杂的风格属性方面，往往仍存在不足。为此，我们借鉴先前的 RL 方法（Gong et al.，2019年；Luo 等，2019a），提出了一种基于强化学习的方法，以优化模型的文本生成。不同于以往使用单尺度奖励判断极性符合度（例如正面与负面）的基于强化学习的方法，我们方法的贡献在于采用层次化奖励优化来量化强度符合度（例如非常正面与正面），从而显式评估风格化文本的全局和局部属性。我们引入了两级奖励：句子级奖励，以两种不同形式衡量生成文本的风格强度；以及词典级奖励，通过分析词汇分布，使生成文本与细粒度语言特征保持一致。我们将上述奖励整合到最终的奖励函数中。

4.2.1 层次化风格奖励

句子级奖励。句子级奖励源自基于分类器的评估，能够捕捉整体风格属性。由于风格强度既可以通过细粒度分类（例如，情感分析（Zhang 等人，2015年））， 也可以通过回归（例如，可读性评估（Feng 等人，2010年））来表示，因此我们针对回归和分类两种方式定义句子级奖励。

• 基于回归的：当使用回归模型 $\mathcal{C}(\cdot)$来衡量生成文本的强度时，ˆ将采用此类奖励 Y。当生成文本与目标强度的偏差较小时，奖励值较高；否则，奖励值较低。受现有研究（Ribeiro 等人，2023年）启发，我们采用以目标强度为中心的归一化高斯分布来建模基于回归的分类器的奖励 $f \sim \mathcal{N}(s_y, \sigma^2)$：

$$\hat{s} = \mathcal{C}(\hat{Y});$$
$$\mathcal{R}_{\text{sent}}(\hat{Y}) = f(\hat{s})/f(s_y).$$

如果观测到的强度等于目标强度，奖励将取最大值 1；随着 ˆs 偏离s_y，奖励将呈指数级下降。

• 基于分类的：当使用分类模型 $\mathcal{C}(\cdot)$ 来衡量生成文本的强度时，将采用此类奖励。奖励的定义如下：

$$\mathcal{R}_{\text{sent}}(\hat{Y}) = \mathcal{C}(\hat{Y}),$$

ˆ其中，$\mathcal{C}(Y)$输出给定生成文本时目标强度的概率。奖励

[1]我们在附录A中展示提示词。

如果预测生成文本具有目标风格，奖励将为其赋予最大值 1。

如我们所见，这两种奖励都鼓励模型生成符合目标风格强度的文本。

词典级奖励。句子级奖励关注整体特征，而词典级奖励则从风格特定词汇和词汇分布中捕捉词汇强度。由于风格差异往往体现在不同的用词模式上（Mir 等人，2019年），识别并奖励风格特定词汇能够增强大语言模型生成符合目标风格语言特征文本的能力。

为识别风格特定词汇，我们采用词频-逆文档频率（TF-IDF）来衡量某个词在集合或语料库文本中的重要性。因此，对于 UTST 任务，风格特定词语通常是 TF-IDF 值较高的词。我们选取每个强度等级中 TF-IDF 值排名前1000 的词，构成风格特定词汇。附录E展示了语料库中体现不同风格强度的代表性词语。随后，对于每个风格强度等级，我们将语料库中的所有文本串联起来，并构建一个 TF-IDF 向量，以表示词汇分布的中心。对于每段生成文本，我们也构建一个 TF-IDF 向量来表示其词汇分布：

$$\mathbf{v}_s = \text{TF-IDF}(\mathcal{D}^s);$$
$$\hat{\mathbf{v}} = \text{TF-IDF}(\hat{Y}),$$

其中，$\mathbf{v}_s$ 表示请求强度 s 对应的枢轴词频-逆文档频率向量，$\hat{\mathbf{v}}$ 表示生成文本的词频-逆文档频率向量。通过计算生成文本向量与枢轴向量之间的相似度，可以如下衡量词典级奖励：

$$\mathcal{R}_{\text{lex}} = \frac{\exp(\text{sim}(\mathbf{v}, \mathbf{v}_{s_y})/T)}{\sum_{j=1}^{k} \exp(\text{sim}(\hat{\mathbf{v}}, \mathbf{v}_{s_j})/T)},$$

其中，$\mathbf{v}^{s_y}$ 是目标强度对应的枢轴词频-逆文档频率向量，$\text{sim}(\cdot,\cdot)$ 表示余弦相似度函数，T 是一个温度参数，用于控制分布的尖锐程度。

这一奖励鼓励模型突出目标风格的典型词汇模式，从而提高风格保真度。

4.2.2 一致性奖励

为了在风格迁移过程中保留输入的语义信息，我们遵循现有研究（Zhang* 等人，2020年），提出一致性奖励：

$$\mathcal{R}_{\text{cons}} = \frac{1}{|\hat{Y}|} \sum_{y_j \in \hat{Y}} \max_{x_i \in X} (\text{Emb}(y_j) \cdot \text{Emb}(x_i)),$$

其中，$|\hat{Y}|$ 是 $\hat{Y}$ 中的词元数量，$\text{Emb}(\cdot)$ 表示词元的 RoBERTa （Liu 等人，2019年）嵌入。该分数经过归一化至 $[0, 1]$，表示源文本与目标文本之间的语义对齐程度。

4.2.3 总奖励

总奖励表示为上述奖励的加权和：

$$\mathcal{R} = \lambda_1 \cdot \mathcal{R}_{\text{sent}} + \lambda_2 \cdot \mathcal{R}_{\text{lex}} + \lambda_3 \cdot \mathcal{R}_{\text{cons}},$$

其中 λ_1、λ_2 和 λ_3 是用于平衡奖励的三个超参数。我们的最终模型通过第 4.1 节中的微调和近端策略优化获得。因此，我们的模型 $\mathcal{G}(\cdot)$ 无需任何平行数据，即可实现具有可控强度的文本风格迁移。

5 实验设置

5.1 数据集

与大多数以往工作只关注不同风格极性之间的文本风格迁移（例如，正面 → 负面）不同，我们在两个用于实现可控强度的 UTST 数据集[2] 上进行实验：

- **CNN/DM**：这是一个用于可读性迁移的数据集，收集自 CNN Daily 邮件（Hermann 等人，2015年），其中包含原始文本以及可读性等级各异的摘要。为构建 UTST 任务的数据集，我们计算摘要的弗莱施阅读易读性（FRE）分数（Flesch，1948年），并将其划分为四个可读性等级，即 {小学（*1*） → 初中 （2） →高中 （3） → 大学（4)}。这些集合中的文本数量分别为42824，176262，61099和 6459。
- **YELP**：这是一个最初用于情感分析的数据集，收集自 Yelp 评论

[2]数据集的详细信息见附录D

(Zhang 等人，2015年）。其中包含从 1 到 5 星级的评分，分别对应 {非常负面（1） → 负面（2） → 中性（3） → 正面（4） → 非常正面（5）。我们将其视为 UTST 数据集，并将其划分为五个非平行语料库集合用于训练。每个集合中的文本数量为130k。

5.2 评估指标

按照 TST 任务的标准评估方式，我们采用评估风格对齐和内容保留的指标。对于风格对齐，我们使用**FRE**衡量预测文本的可读性强度，并用$\mathbf{FRE}_{\Delta}$ 表示预测文本与目标文本之间的强度偏差。我们使用星级评分（**STAR**）衡量预测文本的情感强度，并用$\mathbf{STAR}_{\Delta}$ 表示预测文本与目标文本之间的强度偏差。我们还采用第**H-Re**节所示的层次化奖励之和4.2来衡量测试数据的风格对齐情况。对于内容保留，我们使用 ROUGE-L（**RG-L**）衡量预测文本与输入文本之间的重叠度。详细说明见附录C。

5.3 可比方法

为了评估我们提出的具有可控强度的 UTST 方法的性能，我们全面比较了多种 SFT 后接 PPO 范式：**SFT**：我们采用 flan-t5-large 模型作为骨干大语言模型，并使用合成的平行文本对其进行监督微调。**SFT + PPO** ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{cons}}$)：这是我们采用 SFT 后接 PPO 范式的方法，并使用句子级奖励作为风格对齐的判定依据。**SFT + PPO** ($\mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$)：这是我们采用 SFT 后接 PPO 范式的方法，并使用词典级奖励作为风格对齐的判定依据。**SFT + PPO** ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$)：这是我们采用 SFT 后接 PPO 范式的方法，并使用层次化奖励作为风格对齐的判定依据。

由于先前的 TST 方法不适用于极性迁移，我们采用 **GPT-4o-mini** (OpenAI 等人，2024年），将这一先进的大语言模型作为 UTST 任务的强基线。在零样本设置下，我们只需通过提示指示它生成风格强度各异的文本，详情见附录 B。

5.4 实现细节

对于 YELP，我们采用在非平行语料库上微调的 DistilBERT[3] 作为基于分类的方法 $\mathcal{C}(\cdot)$。对于 CNN/DM，我们采用 FRE 函数（一种回归模型）作为 $\mathcal{C}(\cdot)$。

在强化学习阶段，我们使用 TRLX 库（Havrilla 等人，2023年），并采用近端策略优化。可读性和情感的奖励权重设为 $\lambda_1 = 0.5$、$\lambda_2 = 0.3$ 和 $\lambda_3 = 0.2$，具体说明见附录 G.1；词汇层面奖励的 Softmax 温度为 $T = 0.01$ 。详细参数设置见表7。

6 结果

6.1 主要结果

可读性迁移。表 1比较了 UTST 范式与基线方法在 CNN/DM 数据集上针对四种目标可读性等级的表现。即使不进行针对特定任务的训练，GPT-4o-mini 在零样本设置下的表现仍不理想，这凸显了基于提示的方法的局限性——即便使用了强大的大语言模型。相比之下，我们的 UTST 范式取得了显著提升，平均 FRE 提高至多 64.05%$_{\Delta}$，H-Re 提高至多 40.74%，证明了先进行监督微调、再进行近端策略优化这一流程的有效性。

情感迁移。我们进一步在用于情感迁移的 YELP 数据集上评估了 UTST 方法，结果见表2。与可读性迁移任务的结果一致，UTST 显著优于 GPT-4o-mini，在 STAR_{Δ} 中提升了26.85%，在 H-Re 中提升了 36.36%，进一步证明了其在文本风格迁移中的有效性。

总体分析。**H-Re**是 UTST 任务的主要指标，因为它全面评估了句子和词汇层面的表现。总体而言，SFT 后接 PPO 范式优于单独使用 SFT（表1中 H-Re 为 0.38 vs. 0.34；表2中为 0.45 vs. 0.43）。SFT 只能学会模仿数据集表面的风格特征，无法主动适应复杂或连续的风格控制目标。相比之下，PPO 借助奖励信号直接优化不可微的风格指标，从而提升了对细粒度风格的控制能力。

最佳奖励组合会因不同的风格类型而异：对于可读性，$\mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$ 被视为最佳选择，因为可读性取决于词汇选择（由 $\mathcal{R}_{\text{lex}}$捕捉）

[3] https://huggingface.co/distilbert

可读性 (s_y)	FRE	FRE$_\Delta$ ↓	RG-L ↑	H-Re ↑
GPT-4o-mini				
小学	70.96		19.04	54.12	0.24
初中	65.41	4.59	59.64	0.40
高中	60.73	10.73	62.54	0.30
大学	47.42	27.42	60.92	0.14
平均值	-	15.44	59.31	0.27
SFT				
小学	73.97		16.03	61.63	0.29
初中	64.96	5.04	63.38	0.40
高中	51.89	1.89	61.91	0.38
大学	32.27	12.27	55.22	0.29
平均值	-	8.81	**60.54**	0.34
监督微调 + 近端策略优化 ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{cons}}$)				
小学	78.77		11.23	58.62	0.38
初中	65.25		4.75	63.11	0.40
高中	50.70	0.70	61.27	0.38
大学	28.87	8.87	53.06	0.32
平均值	-	6.39	59.02	0.37
监督微调 + 近端策略优化 ($\mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$)				
小学	79.96		10.04	56.73	0.40
初中	64.65	5.35	61.65	0.40
高中	48.92	1.08	59.50	0.38
大学	25.71	5.71	49.99	0.34
平均值	-	**5.55**	56.97	**0.38**
监督微调 + 近端策略优化 ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$)				
小学	78.87		11.13	58.12	0.38
初中	65.25	4.75	62.69	0.40
高中	50.89	0.89	60.90	0.38
大学	26.67	6.67	51.49	0.33
平均值	-	5.86	58.30	0.37

表1：CNN/DM 的结果。展示了各项细分结果和平均值。

情感(s_y)	STAR	STAR$_\Delta$ ↓	RG-L ↑	H-Re ↑
GPT-4o-mini				
非常负面	1.5190	0.5190	45.39	0.35
负面	1.8018	0.1982	47.11	0.38
中性	2.9476	0.0524	46.58	0.23
正面	4.3652	0.3652	49.78	0.30
非常正面	4.6032	0.3968	48.98	0.41
平均值	-	0.3063	47.57	0.33
SFT				
非常负面	1.2012	0.2012	43.68	0.50
负面	2.3338	0.3338	45.88	0.38
中性	3.3854	0.3854	46.24	0.37
正面	4.1382	0.1382	46.20	0.32
非常正面	4.9416	0.0584	43.92	0.60
平均值	-	0.2234	45.18	0.43
监督微调 + 近端策略优化 ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{cons}}$)				
非常负面	1.1954	0.1954	46.16	0.50
负面	2.3254	0.3254	48.55	0.38
中性	3.3760	0.3760	48.95	0.38
正面	4.1330	0.1330	49.21	0.32
非常正面	4.9374	0.0626	47.88	0.60
平均值	-	**0.2185**	**48.15**	0.44
监督微调 + 近端策略优化 ($\mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$)				
非常负面	1.2004	0.2004	46.19	0.50
负面	2.3146	0.3146	48.80	0.38
中性	3.3904	0.3904	49.14	0.38
正面	4.1370	0.1370	49.21	0.32
非常正面	4.9410	0.0590	46.53	0.60
平均值	-	0.2203	47.97	0.44
监督微调 + 近端策略优化 ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$)				
非常负面	1.2114	0.2114	46.18	0.51
负面	2.3234	0.3234	48.49	0.40
中性	3.3748	0.3748	48.89	0.39
正面	4.1462	0.1462	49.04	0.33
非常正面	4.9374	0.0374	46.39	0.62
平均值	-	0.2186	47.80	**0.45**

表2：YELP 的结果。展示了细分结果和平均值结果。

以及内容保留（由 $\mathcal{R}_{\text{cons}}$捕捉）——鉴于词语难度与句子结构高度相关，句子特征在此处是多余的。对于情感， $\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{cons}}$ 被视为最佳选择，因为情感含义依赖上下文（例如，“忙碌”可以传达正面或负面的内涵），因此需要 $\mathcal{R}_{\text{sent}}$来确定词语在其上下文框架中的情感极性。这种适应性凸显了所提出模型在不同风格维度上的泛化能力。

然而，**RG-L** 略有下降，这归因于其内在的计算机制。详细说明见附录 F。

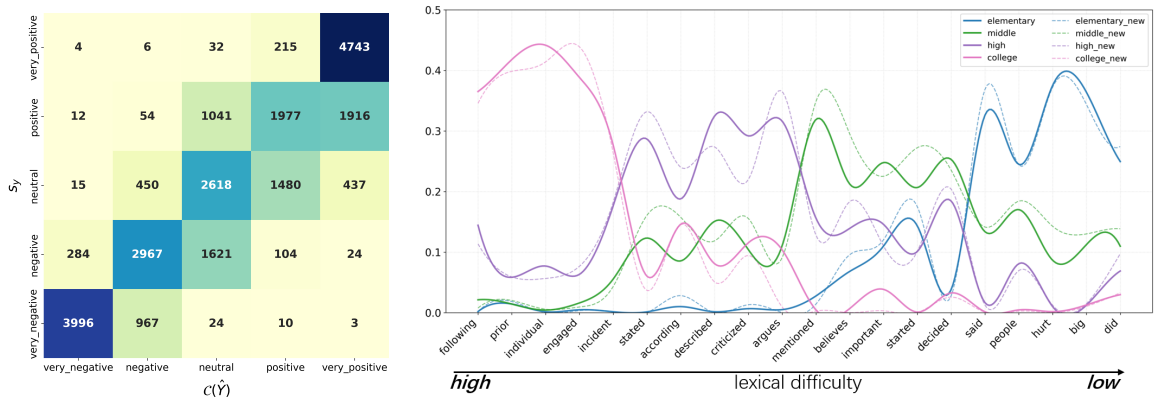
6.2 进一步分析

6.2.1 事实一致性分析

图中的热力图 3a显示预测结果呈现明显的对角线分布，尤其是在非常正面（4743 个案例）和非常负面（3996 个案例）类别中，这表明模型能够准确识别情感强烈的表达。大多数误分类发生在相邻类别之间。例如，正面被判为非常正面（1916 个案例），凸显了区分细微情感变化的难度。中性样本表现出较高的模糊性，常与相邻类别混淆（共 4985 个实例），这可能是由于词汇模式存在重叠。尽管面临这些挑战，模型仍表现出色，能够有效捕捉明显且细腻的情感信号。

6.2.2 词典分布的可视化

图3b 比较了参考文本与生成文本在不同可读性强度下的词典分布。结果显示，词汇一致性较高，尤其是在风格极性——词汇方面



(a) YELP 数据集上的事实一致性热力图。每个单元格表示实例数量。迁移文本的情感由经过微调的 DistilBERT 分类器预测。

(b) 词典分布可视化。我们展示了枢轴向量和生成文本中的一些风格特定词语及其 TF-IDF 值。实线表示标注语料库的分布，虚线表示生成文本的分布。从左到右排列的词语，其词汇难度逐渐降低。

图 3：SFT + 近端策略优化 ($\mathcal{R}_{\text{sent}} + \mathcal{R}_{\text{lex}} + \mathcal{R}_{\text{con}}$ 的事实一致性与词典分布分析 $_{\text{s}}$).

类别 文本 原文 这里的食物实在太棒了。这里总是很忙，这足以说明它有多受欢迎，但员工效率很高，确保你能迅速入座并得到服务。记得点一扎啤酒——非常划算，还能省下不少钱。**小学** 这里的食物真的 很好吃。这里总是很忙，这很好，因为有很多人来这里。员工非常热心。他们确保你能很快入座 并且 拿到食物。别忘了点一扎啤酒。非常划算，还能省下很多钱。**中等** 这家餐厅的食物一直很棒，很容易看出它为何如此受欢迎。员工总是很友善，也乐于提供帮助，确保你的餐点很快送上。点一扎啤酒也是个 不错的主意– 常划算，还能省下不少钱。 家店的食物 始 上乘，也显而易见它为何如此 受欢迎。服 到，确保餐点及时送上。此外 扎啤 是个 不错的主意——非常 划算，还能省下相当多的钱。**大学** 这家 餐体验 充分体现了卓越的品质与服务水平。尽管店内氛围始终热闹非凡，足见这家 餐厅 极具吸引力，员工仍展现出非凡的效率，确 供服务，并 照顾 顾客。此外，强烈建议点 酒——这非常划算，幅减少总体开支。

表 3：不同可读性强度下迁移的餐厅评论文本。高亮词语为风格特定词汇，颜色越深表示强度越高。

生成文本中的词汇分布与参考文本中的分布十分接近。即使在中等强度下，差异仍然很小，这凸显了模型再现可读性特定词汇分布的能力。在层次化奖励设计的支持下，这种高度的词汇对齐进一步证实了我们的方法在捕捉细粒度风格属性方面的有效性。

6.2.3 案例研究

如表 3所示，我们将 YELP 数据集中的一条餐厅评论转换为四个版本，其可读性等级各不相同。随着阅读难度提高，文本长度逐渐增加。此外，高阅读难度版本中突出显示的词语数量表明，这些版本往往会使用更高级的词汇，例如“体现”和“必要的”。该示例展示了我们的方法在生成具有不同风格强度的文本时的有效性和灵敏度。

7 结论

我们提出一种用于可控强度 UTST 的 SFT 后接 PPO 范式，使用经过筛选的伪平行语料库进行 SFT，并通过分层近端策略优化实现细粒度风格控制。在 CN-N/DM 和 Yelp 上的实验表明，可读性和

情感迁移均有所提升，与 GPT-4o-mini 相比，风格对齐更准确、内容保留效果更好，词汇精确度也更高。这种方法推动了 UTST 的发展，并拓展了文本转换的应用范围。未来工作将探索把强度控制扩展至更多风格维度，以增强其在现实场景中的适用性。

局限性

我们的 UTST 范式在可控强度文本风格迁移方面展现出良好的性能，但评估的语言范围仍存在潜在局限。目前的实验均基于英语数据集（CNN/DM 和 Yelp）开展。尽管这些数据集具有代表性，能够用于常见的风格迁移基准，但可能无法充分体现该方法在多样化语言和文化语境中的有效性。鉴于所提出的方法本质上与语言无关，未来的工作可以将评估扩展至多语言或非英语数据集，从而在更广泛的现实场景中，对其泛化能力和实际效用进行更全面的评估。

参考文献

Eleftheria Briakou、Sweta Agrawal、Joel Tetreault 和 Marine Carpuat。2021年。评估风格迁移的评估指标：多语言正式度迁移案例研究。在 *2021*年自然语言处理实证方法会议论文集 中，第 1321–1336 页。

Mingkai Deng、Jianyu Wang、Cheng-Ping Hsieh、Yi- han Wang、Han Guo、Tianmin Shu、Meng Song、Eric Xing 和 Zhiting Hu。2022年。Rlprompt：使用强化学习优化离散文本提示。在*2022*年自然语言处理经验方法会议论文集中，第3369–3391页。

Lijun Feng、Martin Jansche、Matt Huenerfauth 和 Noémie Elhadad。2010 年。自动可读性评估特征比较。在*Coling 2010*：海报中，第276–284页。

Rudolf Franz Flesch。1948年。一种新的可读性标尺。应用心理学杂志，32(3)：221–233。

Zhenxin Fu、Xiaoye Tan、Nanyun Peng、Dongyan Zhao 和 Rui Yan。2018年。文本中的风格迁移：探索与评估。在*AAAI*人工智能会议论文集中，第32卷。

Hongyu Gong、Suma Bhat、LingfeiWu、JinJunXiong 和 Wen-mei Hwu。2019年。强化学习基于强化学习的文本风格迁移，无需平行训练语料。载于 *2019*年北美计算语言学协会年会论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 3168–3180 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Jingxuan Han、Quan Wang、Zikang Guo、Benfeng Xu、Licheng Zhang 和 Zhendong Mao。2024年。利用合成平行数据进行文本风格迁移的解耦学习。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第15187–15201页，泰国曼谷。计算语言学协会。

Alexander Havrilla、Maksym Zhuravinskyi、Duy Phung、Aman Tiwari、Jonathan Tow、Stella Biderman、Quentin Anthony 和 Louis Castricato。2023年。trlX：一个用于大规模基于人类反馈的强化学习框架。载于*2023*年自然语言处理实证方法会议论文集，第8578–8595页，新加坡。计算语言学协会。

Karl Moritz Hermann、Tomás Kociský、Edward Grefenstette、Lasse Espeholt、Will Kay、Mustafa Suleyman 和 Phil Blunsom。2015年。教机器阅读与理解。载于神经信息处理系统大会，第1693–1701页。

Zhiqiang Hu、Roy Ka-Wei Lee 和 Charu C.Aggarwal。2020年。文本风格迁移：综述与实验评估。*CoRR*，abs/2010.12742。

Joosung Lee。2020年。稳定风格转换器：一种用于文本风格迁移的编码器-解码器删除与生成方法。发表于第*13*届国际自然语言生成会议论文集，第195–204页。

Juncen Li、Robin Jia、He He 和 Percy Liang。2018年。删除、检索、生成：一种简单的情感与风格迁移方法。发表于计算语言学协会北美分会*2018*年会议论文集：人类语言技术，第*1*卷（长篇论文），第1865–1874页。

Xiaolin Li、Lei Huang、Yifan Zhou 和 Changcheng Shao。2021年。Tst-gan：一种基于文本风格迁移的法律文档生成模型。发表于*2021*年第*4*届机器人、控制与自动化工程国际会议（*RCAE*），第90–93页。电气电子工程学会。

Pusheng Liu、Lianwei Wu、Linyong Wang、Sensen Guo 和 Yang Liu。2024年。循序渐进：利用大语言模型控制文本中的任意风格。载于*2024*年计算语言学、语言资源与评估国际联合会议论文集（*LREC-COLING 2024*），第 15285–15295 页。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*CoRR*，abs/1907.11692。

Fuli Luo、Peng Li、Jie Zhou、Pengcheng Yang、Baobao Chang、Zhifang Sui 和 Xu Sun。2019a。用于无监督文本风格迁移的双重强化学习框架。载于第 *28* 届国际人工智能联合会议论文集，*IJCAI 2019*。

Fuli Luo、Peng Li、Jie Zhou、Pengcheng Yang、Baobao Chang、Xu Sun 和 Zhifang Sui。2019b年。用于无监督文本风格迁移的双重强化学习框架。载于第二十八届国际人工智能联合会议论文集，*IJCAI-19*，第 5116–5122 页。国际人工智能联合会议组织。

Remi Mir、Bjarke Felbo、Nick Obradovich 和 Iyad Rahwan。2019年。评估文本风格迁移。载于*2019*年北美计算语言学协会年会论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 495–504 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Sourabrata Mukherjee 和 Ondˇrej Dušek。2023年。利用低资源平行数据进行文本风格迁移。载于第十六届国际自然语言生成会议论文集，第 388–395 页。

Sourabrata Mukherjee、Atul Kr. Ojha、John Philip Mc- Crae 和 Ondrej Dusek。2025年。文本风格迁移评估：是否存在任何可靠指标？载于*2025*年计算语言学协会美洲分会会议论文集：人类语言技术（第*4*卷：学生研究研讨会），第418–434页，美国阿尔伯克基。计算语言学协会。

Kai North、Marcos Zampieri 和 Matthew Shardlow。2023年。词汇复杂度预测：综述。*ACM*计算调查，55(9):1–42。

OpenAI、：、AaronHurst、Adam Lerer、AdamP.Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford、Aleksander M.adry、Alex Baker-Whitcomb、Alex Beutel、Alex Borzunov、Alex Carney、Alex Chow、Alex Kirillov，以及另外401人。2024年。GPT-4o系统卡。预印本，arXiv:2410.21276。

Shrimai Prabhumoye、Yulia Tsvetkov、Ruslan Salakhut- dinov 和 Alan W Black。2018年。通过回译进行风格迁移。载于第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第866–876页，澳大利亚墨尔本。计算语言学协会。

Emily Reif、Daphne Ippolito、Ann Yuan、Andy Coenen、Chris Callison-Burch 和 Jason Wei。2022年。利用大语言模型实现任意文本风格迁移的方法。在计算语言学协会第*60*届年会论文集（第*2*卷：短篇论文），第837–848页。

Leonardo F. R. Ribeiro、Mohit Bansal 和 Markus Dreyer。2023年。生成可控制可读性等级的摘要。载于*2023*年自然语言处理实证方法会议论文集，第11669–11687页，新加坡。计算语言学协会。

Akhilesh Sudhakar、Bhargav Upadhyay 和 Arjun Ma- heswaran。2019年。“转换式”删除、检索、生成方法，用于可控文本风格迁移。载于*2019*年自然语言处理实证方法会议和第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第3269–3279页。

Mirac Suzgun、Luke Melas-Kyriazi 和 Dan Juraf- sky。2022年。提示与重排序：一种利用小型语言模型进行零样本和少样本任意文本风格迁移的方法。载于*2022*年自然语言处理实证方法会议论文集，第2195–2222页。

Keren Tan、Kangyang Luo、Yunshi Lan、Zheng Yuan 和 Jinlong Shu。2024年。一种用于词汇简化的 LLM 增强型对抗编辑系统。载于计算语言学、语言资源与评估国际联合会议（*LREC-COLING 2024*）论文集，第 1136–1146 页。

Shengyu Zhang、Linfeng Dong、Xiaoya Li、Sen Zhang、Xiaofei Sun、Shuhe Wang、Jiwei Li、Runyi Hu、Tian- wei Zhang、Fei Wu 等。2023年。大语言模型的指令微调：综述。*arXiv* 预印本 *arXiv:2308.10792*。

Tianyi Zhang*、VarshaKishore*、FelixWu*、Kilian Q. Weinberger 和 Yoav Artzi。2020年。BERTScore：使用 BERT 评估文本生成。载于国际学习表征会议。

Xiang Zhang、Junbo Jake Zhao 和 Yann LeCun。2015年。用于文本分类的字符级卷积神经网络。*CoRR*，abs/1509.01626。

Chang Zong、Yuyan Chen、Weiming Lu、Jian Shao、Yongfeng Huang、Heng Chang 和 Yuet- ing Zhuang。2024年。Proswitch：知识引导的指令微调，用于在专业与非专业回复之间切换。*arXiv* 预印本 *arXiv:2403.09131*。

A 平行数据生成提示词

我们介绍了用于生成可读性和情感平行数据的提示词。

A.1 可读性

将以下文本改写为符合 {outputDesc}可读性级别（弗莱施阅读易读性分数{FREScore}）。调整词汇、句子长度和复杂度，使其充分体现目标可读性级别。对于小学水平，请使用非常简单的词语和短句。对于初中水平，请使用较简单的词语和稍长的句子。对于高中水平，请使用更复杂的词语和多样的句子结构。对于大学水平，请使用高级词汇和复杂的句子结构。保持核心含义和内容一致，同时调整风格，使其自然简洁。输入： {text}输出：

A.2 情感

将以下 {inputRating}Yelp 评论改写为一条 {outputRating} 评价。转换语气和措辞，使其充分体现目标评分所表达的情感。对于五星评价，请使用“惊艳”或“出色”等非常正面的词语；对于一星评价，请使用“糟糕”或“恶劣”等强烈负面的词语。保持核心方面（食物、服务、氛围）一致，但调整对它们的描述，使其符合目标评分。确保输出自然、真实、简洁（长度与输入相近），并避免使用“还行”或“不错”等含糊的词语。输入： {text}输出：

B 可控文本生成提示词

我们给出了用于生成具有可控强度的可读性和情感文本的提示词 ty.

B.1 可读性

将以下文本改写为 {可读性级别}输入： {文本}输出：

B.2 情感

将以下文本改写为 {情感级别}输入： {文本}输出：

C 评估指标

- **FRE**：使用弗莱施阅读易读性分数衡量可读性，评估文本复杂度，以实现可读性迁移：

$$\text{FRE} = 206.835 - 1.015 \cdot \frac{\text{total words}}{\text{total sentences}} - 84.6 \cdot \frac{\text{total syllables}}{\text{total words}}.$$

FRE$_\Delta$ 量化与目标范围中点的偏差：

$$\text{FRE}_\Delta = |\text{FRE}_{\hat{y}} - \text{FRE}_{s_y}|,$$

其中 FRE$_{\hat{y}}$ 是生成文本的 FRE，而 FRE$_{s_y}$ 是目标范围的中点（表4）。

• **STAR**：表示使用分类器预测的风格类别（可读性为 1–4，情感为 1–5），例如，用于情感的 DistilBERT、用于可读性的 FRE 到等级映射。**STAR**$_\Delta$ 用于衡量与目标类别的偏差：

$$\text{STAR}_\Delta = |\text{STAR}_{\hat{y}} - \text{STAR}_{s_y}|,$$

其中 STAR$_{\hat{y}}$ 是预测类别，STAR$_{s_y}$ 是目标类别。

• **RG-L**：通过 ROUGE-L 评估内容保留，测量输入文本与生成文本之间的最长公共子序列：

$$\text{RG-L} = \frac{\text{LCS}(X, \hat{Y})}{|\hat{Y}|},$$

ˆ其中 LCS(X, Y) 是输入文本 X 与生成文本 ˆ ˆ ˆ Y之间最长公共子序列的长度，而 $|Y|$ 是 Y 的长度。

• **H-Re**：评估整体迁移质量，相当于第 4.2.3 节中用于可读性和情感的奖励函数，但将 BERTScore 系数设为 0：

$$\text{H-Re} = 0.5 \cdot \mathcal{R}_{\text{sent}} + 0.5 \cdot \mathcal{R}_{\text{lex}},$$

其中 $\mathcal{R}_{\text{sent}}$ 和 $\mathcal{R}_{\text{lex}}$ 定义见第 4.2.1 节。

D 数据集概览

表 4 展示了可读性和情感数据集的详细信息，包括等级分数、强度属性和数据分布。

风格	数据集	分数	属性	训练	Test
可读性	CNN/DM	$80 \leq \text{FRE} \leq 100$	小学生	4,000	4,000
		$60 \leq \text{FRE} < 80$	中学生	4,000	
		$40 \leq \text{FRE} < 60$	高中生	4,000	
		$0 \leq \text{FRE} < 40$	大学生	4,000	
情感	Yelp	五星	非常积极的语气	4,000	5,000
		四星	积极的语气	4,000	
		三星	中性语气	4,000	
		2 星级	负面语气	4,000	
		1 星级	非常负面语气	4,000	

表 4：按可读性和情感分类的风格属性。此处的训练集数量指构建的伪平行数据集的数量。

E 风格强度的代表性词语

如表 5所示，列出了情感语料库中代表不同风格强度的一些可读性和情感词语，构成词典级奖励的词汇分布。

F RG-L 下降原因说明

在可读性数据集上，RG-L指标仅呈现出轻微下降。这一现象可归因于该指标的核心计算逻辑：RG-L将风格转换前后的文本的最长公共子序列（LCS）纳入分析。该指标的设计重点在于衡量转换前后文本之间的文本内容保留程度。然而，对此类内容保留的要求不可避免地会导致词汇层级发生变化，

分数	特征词
	可读性
小学	说、人们、大、伤害、做了、帮助、得到、年、打电话、坏
初中	提及，只是，相信，喜欢，旧的，决定，分享，知道，正确的，重要的
高中	陈述，根据，表达，此外，目前，描述，最近，声称
大学	关于，显著的，个人的，接下来的，随后，值得注意的是，此外
	情感
非常正面	出色、令人惊叹、绝对、不可思议、真正、极好、珍品、极其
正面	扎实、很棒、总体而言、确实、令人愉快、令人满意、令人欣喜、美味、非常
中性	一点、不错、相当、改进、潜力、总体而言、好、考虑、欣赏
负面	令人失望、遗憾、留下、期望、令人沮丧、更好、预期、缺乏
非常负面	糟糕、可怕、平淡、灾难、失望、避免、最差、完全、几乎不

表5：可读性与情感语料库中不同风格强度的代表性词语。

使RG-L指标偏离了“评估风格转换强度”这一核心目标。因此，仅依据该指标无法全面评估风格转换方法的有效性。

F.1 表格案例的证据

表6 呈现了一个具体的实证案例，支持上述结论。

文本类别	FRE	RG-L	文本内容
原文	66.10	—	这位科学家进行了仔细的测试，以获得准确的结果。
重写文本1	33.07	66.67	这位科学家进行了细致的实验，以获得精确的结果。
重写文本2	15.64	52.63	这位科学家认真地实施了实验，以确保结果精确。

表6：不同强度等级下的文本风格转换示例

F.2 指标—目标适应性分析 ty

本研究最初的目标是将原文重写为符合“大学及以上阅读标准”的版本，对应的 FRE 分数范围为0–40。如表中所示，重写文本1和重写文本2均满足这一要求。

对风格转换强度的进一步分析表明，重写文本2的 FRE 得分（15.64）低于重写文本1（33.07）。这一数值差异说明，重写文本2表现出更显著的风格变化。尽管重写文本2的 RG-L 值（52.63）低于重写文本1（66.67），但这一 RG-L 值差异并未削弱其在风格转换效果方面的优势，反而进一步证实了 RG-L 指标无法有效反映风格转换强度这一结论。

G 超参数设置

表 7 展示了采用先监督微调后近端策略优化范式的 UTST 的部分设置，包括计算基础设施、GPT-4o-mini 设置、DeepSpeed 设置、监督微调设置和近端策略优化设置。

G.1 组合奖励函数的超参数设置策略

对于组合奖励函数， λ_1、 λ_2和 λ_3 采用固定权重策略进行配置，不进行耗费资源的网格搜索。权重遵循按重要性排序的顺序保持原则：句子级奖励（R_{sent}）的权重最高，旨在优先确保生成文本的整体风格与目标强度相匹配；其次是词级奖励（R_{lex}），用于增强细粒度词汇特征与目标风格之间的匹配；一致性奖励（R_{cons}）的优先级最低，因为它并非核心关注点

本研究。此奖励仅要求在风格迁移过程中保留核心语义，且未针对该组件进行消融实验。

参数	值
计算设置	
GPU	NVIDIA A6000 48GB（2 台）
flan-t5-large	7.83 亿个参数
RoBERTa-large	3.55 亿个参数
DistilBERT 基础模型	6700 万个参数
GPT-4o-mini 设置	
温度	0.7
DeepSpeed 设置	
ZeRO 优化阶段	2
混合精度	bf16
梯度累积步数	8
梯度裁剪	1.0
通信窗口大小	512
监督微调设置	
优化器	AdamW
优化器参数	$\beta = (0.9, 0.999),\ \epsilon = 10^{-8}$
学习率	10^{-4}
权重衰减	0.0
丢弃率	0.1
批次大小	16
训练轮数	20
*PPO*设置	
优化器	AdamW
优化器参数	$\beta = (0.9, 0.999),\ \epsilon = 10^{-8}$
学习率	10^{-5}
权重衰减	10^{-6}
调度器参数	$T_{\max} = 10^4,\ \eta_{\min} = 10^{-6}$
批次大小	2
训练轮数	10
总步数	100000
采样轨迹数	512
块大小	4
初始 KL 系数	0.05
目标 KL 散度	6
时域	10000
折扣因子（γ）	0.99
广义优势估计（GAE）λ（λ）	0.95
策略裁剪范围	0.2
价值裁剪范围	0.2
价值函数系数	1.0
奖励裁剪范围	10
最大生成新词元数	256
采样温度	1.0
Top-k 采样	50
Top-p 采样	0.95
常规	
随机种子	42

表 7：超参数设置