面向多属性可控对话生成的语义空间落地加权解码

张志凌 MengyueWu*上海交通大学，中国上海
{blmoistawinde,mengyuewu}@sjtu.edu.cn

Kenny Q. Zhu*德克萨斯大学阿灵顿分校，美国得克萨斯州阿灵顿kenny.zhu@uta.edu

arXiv:2305.02820v2 [cs.CL] 5 Nov 2023

摘要

通过个性、情绪和对话行为等多种属性控制聊天机器人的话语生成，在实践中很有用，但尚未得到充分研究。我们提出了一种名为 DASC 的新型框架，该框架采用加权解码范式，具备很强的可控性，同时通过在属性语义空间中进行落地来提升生成质量。随后，通过对多个属性嵌入进行插值，直观地实现多属性生成，从而大幅缩减模型规模。实验表明，DASC 在生成任务中能够同时控制三个方面，并实现较高的控制准确率；此外，即使在分布外鲁棒性测试中，它也能生成有趣且较为合理的回复。[1]

1 引言

个性化对话系统是在人机交互和情感陪伴领域颇具前景的自然语言处理应用。我们期望这类系统具有个性、能够表达情绪、采取对话行为，甚至运用复杂的策略（Liu 等人，2021b），因此需要开展可控文本生成方面的研究。尽管这一领域近期取得了进展（Dathathri 等人，2020年；Keskar 等人，2019年；Krause 等人，2021年），现有研究主要针对单属性控制，忽略了人类互动通常会同时传达多种属性这一事实。因此，我们探索一项全新的任务——多属性可控对话生成，这项任务能够显著提升聊天机器人的表现力、拟人性和可解释性。然而，属性的组合众多，可能导致每种设定下可用的数据十分稀缺，这给该任务带来了巨大挑战。

在以往的研究中，加权解码方法在单属性控制任务中取得了巨大成功（Arora 等人，2022；Liu 等人，2022）。加权解码方法会训练一个词元级别的属性分类器，该分类器根据生成词汇表中每个词元的情况，预测文本传达目标属性的概率。随后，在解码过程中利用这些预测概率重新调整词元生成的权重，从而引导模型表达该属性。尽管这类方法在单属性控制中表现出色，但将多个分类器预测的多种属性概率相乘，以扩展到多属性情形时，它们存在一定局限。由此会引入与庞大的词汇表规模成正比的额外参数 $|V|$，而且参数量还会随着属性数量增加而成倍增长。随之产生的大量参数不仅会降低模型效率，还会损害生成质量。由于每种属性组合的数据通常都很少，模型容易过拟合，从而增加退化的风险（Holtzman 等人，2020年）。

为克服这些局限性，我们提出**对**话**属**性**空**间**控**制器（**DASC**）。我们建立了一个属性语义空间，通过跨属性共享的属性词元嵌入将词汇表中的每个词元投影到该空间中。语言模型的隐藏状态也通过属性专用层转换为空间中的属性上下文嵌入。我们将训练属性空间，使适合传达所需属性的词元接近当前上下文嵌入。随后，我们可以在解码过程中为这些词元分配更高的权重。我们将展示，DASC 不仅能够继承加权解码的强大可控性，还能通过在该空间中对多个属性嵌入进行插值，实现多属性控制的自然解决方案。此外，共享的属性词元嵌入还能缓解过度参数化问题，并提升模型的鲁棒性。

*通讯作者。[1]代码和数据可在以下网址获取：https://github.com/blmoistawinde/DASC。

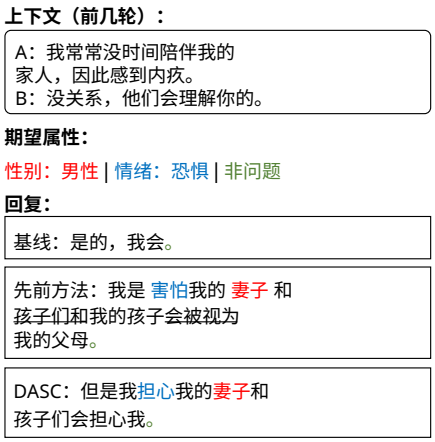


图1：多属性可控对话生成示例。基线系统未尝试进行任何控制，因此生成的回复枯燥乏味；而一种先前的属性控制方法则生成了重复且不合逻辑的文本。DASC 成功生成了既流畅又符合属性要求的回复。

我们在一个属性丰富的开放域对话数据集（Xu 等，2022）上进行了实验，以同时控制 3 个属性方面：性别风格（男性、女性、中性）、情绪（8 个类别）以及对话行为的简单划分（问题与非问题）。如图 1所示，与以往方法相比，DASC 在组合式控制任务中实现了很强的可控性，同时避免生成低质量文本。属性词元嵌入的可视化结果（见图 3）呈现出有利于控制的特定模式，这些模式不同于通用语言模型词元嵌入所呈现的模式。我们进一步在分布外设置下进行了鲁棒性测试，验证了 DASC 的可控性具有良好的泛化能力。我们的贡献如下：1）我们提出了用于可控对话生成的语义空间落地加权解码，该方法通过在空间中对嵌入进行插值，能够直观地解决多属性控制任务；2）与其他加权解码替代方案相比，DASC 使用更少的参数，同时通过共享属性嵌入的设计实现更好的性能；3）DASC 能够在同时控制 3 个方面的情况下实现高准确率，同时在传统测试设置和分布外鲁棒性测试中保持具有竞争力的生成质量。

2 方法

在本节中，我们将首先定义任务和用于可控生成的加权解码方法。作为背景介绍。随后，我们将介绍所提出的**DASC**框架。

2.1 任务定义

给定对话**上下文** C 和**属性** A =(a_1、 a_2.....、 a_K)，可控对话生成旨在生成与上下文一致且具有这些属性的**回复** R =(r_1、 r_2.....、 r_N)。[2] 可以从多个**方面** 对属性进行分组，本文将重点关注性别风格、情绪和对话行为。一个方面涵盖多个相关属性，例如快乐、悲伤 属于情绪方面。每个属性可以取三个值：1表示使用该属性，0表示不使用， ϕ 表示该属性不适用于该回复。

2.2 面向可控生成的加权解码

标准的、不可控的对话生成可以用标准的条件语言建模目标来表述： L_{CLM} = $-\sum_{n=1}^{N} logP(r_n|r_{1:n-1},\ C)$

我们可以使用类似 BART 的基于 Transformer 的编码器-解码器架构（Lewis 等人，2020年）来对其建模，其中编码器将上下文编码为隐藏状态，作为解码器生成回复的条件。为简洁起见，下面我们将省略 C。

在可控对话生成中，我们还在生成条件中引入属性。假设我们使用单个属性a进行生成，那么目标是对 $P(r_n|r_{1:n-1}, a)$ 建模。使用贝叶斯规则，可将其转换为：

$$P(r_n|r_{1:n-1}, a) \propto P(r_n|r_{1:n-1})P(a|r_{1:n-1}, r_n)^{\alpha} \quad (1)$$

其中， α 是一个可用于调整控制强度的超参数。这意味着，我们可以将生成概率分解为标准的因果语言模型（CLM）概率与另一个逐词属性分类器在解码过程中给出的预测值的加权组合。因此，基于这种分解建立的方法被称为**加权解码** 模型。

Director（Arora 等人，2022），这是一种具有代表性的加权解码方法，它将属性分类器实现为位于解码器隐藏状态之上的线性层。该方法在每一步都会进行二元分类，以判断生成的句子是否体现了期望的属性（例如，开心或

[2]在我们的工作中，我们预先假设属性由对话策略提供，且不涵盖端到端场景。

不开心）。对于训练集中的句子，其各个词元都可以使用整句的属性进行训练，采用二元交叉熵（BCE）损失。我们将这种词元级损失记为 L_t。

$$\begin{aligned} L_t &= BCE(P(a|r_{1:n-1}, r_n)) \\ &= BCE(\sigma([W_a h_n]_{r_n})) \end{aligned} \quad (2)$$

其中，$h_n \in \mathbb{R}^d$ 表示第 n个 token 的隐藏状态，$W_a \in \mathbb{R}^{|V| \times d}$ 表示在词汇表中给定每个 token 的生成结果时，用于属性预测的可学习权重矩阵，而 $[\cdot]_{r_n}$ 表示对下一个 token r_n进行索引选择。需要注意的是，它仅收集真实响应中 token r_n 对应的属性 logits。对于词汇表 V中的其他 $|V|-1$ 个 token，由于没有相应的标签，因此无法通过训练获得。因此，它使用一个额外的正则化项，通过均方误差损失（MSE loss）训练这些 token 的预测值，使其尽可能接近 0.5。

在处理多属性控制时，我们可以通过引入多个属性分类器的乘积来扩展式（1），并假设各属性之间条件独立：

$$P(r_n|r_{1:n-1}, a) \propto P(r_n|r_{1:n-1}) \prod_{\substack{k=1 \\ a_k \neq \phi}}^{K} P(a_k|r_{1:n})^{\alpha} \quad (3)$$

概率的乘积通常通过 logits 的求和来实现：

$$\delta(r_n|r_{1:n-1}, a) = \delta(r_n|r_{1:n-1}) + \alpha \sum_{\substack{k=1 \\ a_k \neq \phi}}^{K} \delta(a_k|r_{1:n}) \quad (4)$$

现有工作通过多次运行属性条件语言模型（Lin 和 Riedl，2021年），或单次运行多个模型（Liu 等人，2021a），实现了此类扩展；随着属性数量增加，这两种方法的成本都可能非常高。在此，我们介绍 Director 的一种相对简单的扩展：只需添加 K 线性分类器头，以预测多个属性。我们将这一简单扩展称为 M-Director，或为简便起见，简称 Director。请注意，尽管 M-Director 比以往方法更高效，但仍会引入$d \times |V| \times K$ 额外参数。鉴于 $|V|$ 通常多达数万个，该模型将包含海量参数，导致训练或推理效率低下，也容易过拟合。

2.3 对话属性空间控制器

我们假设，上述加权解码的典型方法可能不是学习词元级属性语义的最高效方式，尤其是在多属性情形下。学习目标仅针对目标句子中的单个词元，而所有其他词元受到同等程度的正则化。这通常并不合理，因为给定属性后，一些与目标词元相似的词元也应具有较高的概率，而与其不同的其他词元则不太可能被生成。例如，对于一条 开心回复“很高兴认识你”的第一个词元，“很高兴”也是一个合理的替代选项，而“悲伤”则不是，但它们在训练中的属性标签都为 0.5。

我们可以在高维空间中纠正这种反直觉现象。一方面，每个令牌都有一个 p 维嵌入，用于编码其属性语义（属性令牌嵌入，$ATEMB$）。另一方面，语言模型（LM）的隐藏状态（h_n）也会通过特定于属性的线性层（$W^k \in \mathbb{R}^{p \times d}$）投影到同一空间中，以获得属性上下文嵌入，$\hat{}h_n^k = \hat{}W^k h_n$。因此，该空间中的不同向量传达不同的语义，我们将这个空间称为属性语义空间。

为了利用这一潜在空间进行加权解码，对于每个 $\hat{}h_n^k$，我们根据空间中的嵌入相似度找出与其属性相关的词元，并在解码过程中为这些词元赋予更高的权重。具体而言，这是通过基于点积的词元级属性分类器实现的。

$$\delta(a_k|r_{1:n}) = \hat{h}_n^k \cdot ATEMB(r_n) \quad (5)$$

在这种情况下，当某个词元针对特定属性以较高概率进行训练时，它在属性空间中的邻居也会具有较高的概率。这缓解了以往加权解码方法的局限，并消除了对其他词元使用正则化器的必要性。此外，将其应用于多属性加权解码时，我们得到：

$$\begin{aligned} \delta(r_n|r_{1:n-1}, a) &= \delta(r_n|r_{1:n-1}) \\ &+ \alpha(\frac{1}{K} \sum_{\substack{k=1 \\ a_k \neq \phi}}^{K} \hat{h}_n^k) \cdot ATEMB(r_n) \end{aligned} \quad (6)$$

其中，第二项中括号内的部分可解释为多个属性上下文嵌入的平均／等权插值。[3]

[3]在插值过程中，可以为每个嵌入分配不同的权重，这将留待未来的工作。

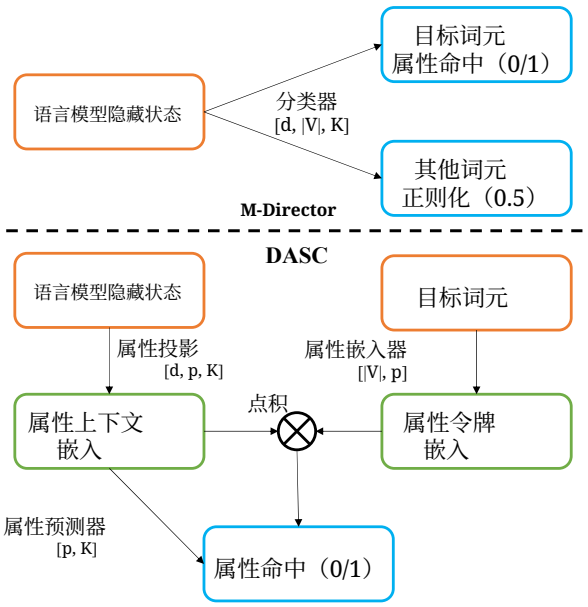


图 2：M-Director 与 DASC 的框架对比。M-Director 使用分类器头，对目标句子中的每个词元进行二元属性命中分类，并对其他词元进行正则化。DASC 将语言模型隐藏状态和目标词元都投影到属性空间，并通过二者的点积对属性命中进行分类。对于每个参数化模型组件，我们在方括号中标出其形状。

这一表述表明，如果属性空间得到恰当的学习和表示，嵌入插值就能准确反映所需属性的语义，进而使 DASC 能够实现合理的属性组合。

为了帮助属性嵌入的学习，我们在每个步骤的属性上下文嵌入之上引入另一个线性层，以直接预测完整回复的属性。这有助于使属性上下文嵌入与其属性对应的区域更好地对齐。我们将新的句子级分类损失记为 L_s。为清晰起见，我们给出单属性情况下的公式，该公式只需对所有非空属性求和即可简单扩展到多属性场景。

$$\begin{aligned} L_s &= BCE(P(a|r_{1:n-1})) \\ &= BCE(\sigma(v_a \cdot \hat{h}_n)) \end{aligned} \tag{7}$$

其中 $v_a \in \mathbb{R}^p$ 是属性预测的可学习权重。与 L_t (公式（2）)相比，这是一项独立于r_n的句子级分类任务，也可以解释为在决定属性之前预测其先验概率

下一个要生成的词元，因此参数不会随 $|V|$ 扩展。最终损失为： $L_{train} = L_{CLM} + \beta(L_s + L_t)$，其中 β 是一个超参数，用于控制属性相关损失的权重。

我们将所提出的框架命名为对话属性空间控制器（**DASC**）。图 2展示了 DASC 的示意图及其与 M-Director 的对比。DASC 引入的参数较少（$d \times p \times K + |V| \times p$），而 M-Director 引入的参数较多（$d \times |V| \times K$）。由于我们设定了 $p << |V|$，属性投影的参数会少得多。此外，在处理 $K > 1$时，各属性共享的词元嵌入也能节省参数，而属性预测器的参数则几乎可以忽略不计。

3 实验

本节将开展实验，以检验以下假设：（1）在多属性可控生成中，DASC 能够实现较强的可控性，同时保持良好的生成质量；（2）DASC 的性能得益于属性语义空间中具有意义的表示以及参数数量的减少；（3）DASC 还可以灵活扩展到其他控制任务，例如组合多种情绪，或采用特定策略提供情感支持。

3.1 实验设置

我们在 *self* 划分的 DuLemon 数据集（Xu et al., 2022）上进行实验。该数据集是一个中文开放域对话数据集，包含丰富的个性化内容，因此我们可以找到希望控制的各种属性。我们将数据划分为训练集、开发集和测试集，分别包含 352,999、2439 和 2412 条话语。由于原始数据集不包含控制属性的标注，我们开发了若干分类器，每种属性类型对应一个，用于标注数据集。对于性别风格（男性、女性、中性），我们使用 Suet al.（2020 年）发布的数据集训练 MacBERT 分类器（Cuietal.，2020年），其准确率达到=9498%。对于情绪，我们遵循 Zhouetal.（2018），使用 NLPCC2013 和 NLPCC2014 数据集（8 种情绪类别）训练另一个 MacBERT 分类器，其准确率为 93.96%。对于问题对话行为（问题与非问题），我们使用简单的启发式规则进行标注：如果句子包含问号（?），就将其视为问题，否则视为非问题。随后，我们使用这 3 个分类器为

数据集中的每条回复标注这 3 类属性（共 13 种）。

3.1.1 竞争方法

我们将所提出的 DASC 框架与不同类型可控生成方法中的代表性方法进行比较。所有竞争方法均采用`fnlp/bart-base-chinese`（Shao 等人，2021年) 模型作为骨干网络 [4]：**基线** 仅在数据集上对骨干网络进行微调，不使用控制属性。**Rerank**使用基线模型进行 top-p 采样（Holtzman 等人，2020年），为每个上下文生成 5 个回复候选，然后使用属性分类器（即我们用于自动标注的同一组独立模型）对候选进行重排序。遵循Thoppilan 等人 (2022) 的做法，我们使用各方面预测概率之和进行排序。**CTRL** 我们重新实现了Keskar 等人(2019年) 的对话生成方法：为每个方面定义 3 组特殊控制代码，并在微调期间将相应的 3 个属性标记追加到每个对话上下文中。**Director**这是 Director 的多属性扩展（Arora 等人，2022），详见第2.2节。更多实验细节见附录。

3.1.2 评估

自动评估为评估可控性，我们使用与数据集标注时相同的属性分类器，计算生成内容中各属性的准确率（Acc_G、Acc_E、Acc_Q 分别对应性别、情绪和问题）。对于生成质量，我们使用 BertScore（BScore）（Zhang 等人，2020年）评估生成内容与参考回复的相似度，并使用 Distinct-2（Li 等人，2016年）衡量多样性。

人工评估我们从测试集中抽取了 100 个上下文进行人工评估。由于原始测试集的分布极不均衡，我们在抽样过程中设定了约束条件，以确保各类情绪的分布更加均衡，从而保证评估结果具有代表性（21 个无情绪、16 个悲伤、16 个厌恶、16 个快乐、16 个喜爱、5 个愤怒、5 个惊讶、5 个恐惧）。我们

[4]我们还针对利用人物描述文本的方法进行了实验，包括 BoB（Song 等人，2021年）以及使用 ChatGPT 提示（OpenAI，2022），尽管这些方法可能并不特别适合控制此处的稀疏属性。我们将在附录中提供实验细节和自动评估结果。

	BScore	Dist-2	Acc_G	Acc_E	Acc_Q
基线	68.18	19.25	68.49	46.31	69.61
Rerank	69.23	19.28	75.46	54.93	82.42
CTRL	**71.09**	18.91	85.32	77.49	**100.00**
Director	69.54	<u>21.40</u>	<u>95.81</u>	**86.73**	**100.00**
DASC	<u>70.42</u>	**21.94**	**95.85**	<u>86.07</u>	**100.00**

表1：在DuLemon测试集上的自动评估结果。最佳结果以粗体显示，次佳结果加下划线。

邀请了 2 名以中文为母语的志愿者，从 3 个角度评估每条生成结果。**属性准确性**：回复是否传达了给定属性。**合理性**(1−4)：回复是否流畅、与上下文连贯且符合常识。**趣味性**(1−4)：回复是否具体、新颖，并能促使对话以更有趣的方式继续。

3.2 结果

自动评估结果如表1所示。我们可以看到，*Rerank* 未能展现出较强的可控性，因为基础模型在未使用属性进行微调的情况下，难以生成具有属性排序的候选结果。*CTRL* 在微调过程中利用了这些属性，并取得了更高的控制准确率和BertScore，但总体上并未生成更多样化的响应。*Director* 和*DASC* 展现出最佳的可控性，并且*DASC*根据Distinct-2和BertScore，生成了更多样且合理的响应。

	Acc_G	Acc_E	Acc_Q	兴趣	合理性
基线	0.80	0.55	0.64	2.04	**3.46**
Rerank	0.81	0.62	0.82	2.13	3.44
CTRL	0.85	0.82	**0.97**	2.24	**3.46**
Director	0.87	0.87	0.96	2.25	3.26
DASC	**0.88**	**0.88**	**0.97**	**2.37**	3.28

表 2：DuLemon测试集上的人工评判 .

随后我们展示表中的人工判断结果2。对于 Acc_G、Acc_E 和 Acc_Q，标注者间的一致性在 Cohen's κ中分别为 0.65、0.55 和 0.64，表明一致性达到中等至较高水平。对于 $Interestingness$ 和$Sensibleness$，一致性在 Pearson's r中分别为 0.48 和 0.44。这并不令人意外，因为后两种视角具有很强的主观性。属性准确率的评估结果与自动评估结果相似，不同之处在于性别准确率略有下降。我们发现，人工评估者能够发现与性别刻板印象相关的错误（Bolukbasi 等人，2016年），例如男性风格生成士兵、女性风格生成育婴者，而这些职业本应

性别中立。标注人员还会检查没有问号的问题，这也解释了 Acc_Q 略有差异。

总体而言，根据人工评估，可控性的排名仍然成立，其中 DASC 表现最佳。基线、Rerank 和 CTRL 的合理性略高于加权解码方法，这与先前文献中普遍观察到的可控性与质量之间的权衡一致（Dathathri 等人，2020年；Yang 和 Klein，2021年；Qian 等人，2022）。所有可控生成方法的趣味性得分都高于基线，这证明了可控生成的优势。在属性准确率与 Director 相近的情况下，DASC 的趣味性得分最高，这表明属性语义空间是有效的：它能够更好地表示属性语义，并在加权解码中以更合理的方式组合控制属性。

3.3 鲁棒性测试

在先前的实验中，提供给模型的控制属性来自参考回复。因此，模型在为上下文生成最可能的回复时，可能会碰巧命中预期属性，却无法对任意给定的属性实现真正可靠的可控性。因此，我们进一步开展实验，测试可控生成方法在分布外场景中的鲁棒性。

具体而言，我们从测试集中抽取了 100 个上下文，并将 8 种情绪分别作为生成条件，与原始性别和问题行为配对后提供给模型[5]。随后，我们使用贪心解码为每个（上下文，属性）组合生成回复，并对这 800 条生成结果进行类似的自动评估和人工评估。

	Dist-2	Acc_E	兴趣	明智
Rerank	17.55	17.00	-	-
CTRL	21.07	43.38	1.91	**3.00**
Director	*34.73*	61.88	1.62	2.27
DASC	**26.71**	**65.38**	**2.08**	2.82

表3：鲁棒性测试结果。

表3展示了鲁棒性测试结果。[6] 与表1相比，我们可以看到情绪准确率方面，Rerank 和 CTRL 均显著下降，这表明它们的可控性无法泛化。另一个值得注意的现象是 Director 达到了异常的*Distinct-2*。随后，我们进一步通过人工评估分析它们的表现（不包括无法控制属性的 Rerank）。我们发现，Director 经常生成语法不通、逻辑混乱且重复冗长的回复（如图1中的第二个回复）。Director 在情绪准确率方面的损失也高于 DASC，这表明由于其参数量较大，它可能过拟合训练分布，因此在这种分布外设置下表现更差。与 CTRL 相比，DASC 的合理性较低，但趣味性较高，同时在多样性和可控性方面也有显著优势。

3.4 语义空间可视化

为了清楚了解所提出的属性语义空间如何帮助实现可控生成，我们使用 t-SNE（Van der Maaten 和 Hinton，2008）将其可视化于二维空间中。首先，我们可视化一些有代表性的属性相关词元的属性词元嵌入，并将其与原始语言模型中对应的嵌入进行比较（图3）。比较这两幅图可以发现：（1）不同方面的词元嵌入在属性空间中更易区分（见不同颜色的点），而同一方面的词元彼此更接近，尽管它们在词性等其他语言特征上存在差异（如“handsome”和“male”）。（2）同一方面中不同属性的词元嵌入在属性空间中也有所区分（如“male”-“female”、“love”-“miserable”）。这些特征使 DASC 能够成功控制生成具有鲜明差异的属性，并组合来自不同方面的属性。

接下来，我们还对属性上下文嵌入进行了可视化。具体而言，我们从数据集的开发集中选取具有特定属性的回复，将其输入模型，并对每个解码器词元处的属性上下文嵌入求平均，作为句子级表示，再将其与句子级属性标注配对进行分析。为简洁起见，我们仅在图4中展示情绪标签的可视化结果，并在附录中提供性别和疑问行为标签的可视化结果。可以看到，来自不同情绪句子的上下文嵌入在空间中明显分离，这表明 DASC 具有很强的多属性可控性

[5]我们不改变这两个属性，因为有时需要根据上下文来确定它们。[6]此处未报告 BertScore，因为模型可以被引导生成不同于真实值的属性，因此基于相似度的指标无法作为生成质量的替代指标。

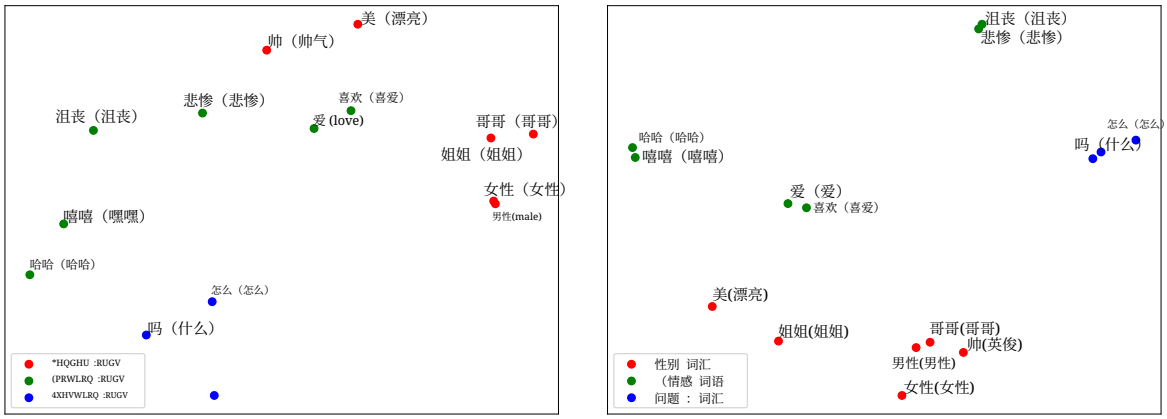


图3：通过 t-SNE 可视化对比两组词元嵌入：一组来自语言模型（左），另一组来自属性语义空间（右）。

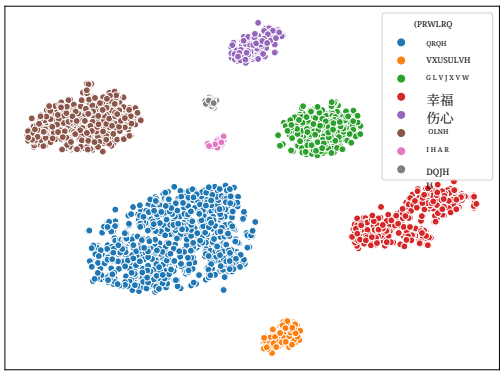


图4：不同情绪的回复的属性上下文嵌入的 t-SNE 可视化。

具有多个属性的 DASC。

3.5 参数分析

如前所述，DASC 只需相对较少的参数即可实现多属性可控生成的加权解码。本文通过调整属性空间的维度 p来研究参数数量的影响，并与不使用参数或使用大量参数进行属性控制的基线和 M-Director 进行比较。我们使用 BertScore 评估生成质量，并计算三个方面的平均控制准确率，以反映可控性。

结果如表 4 所示。比较不同的 DASC p，可以发现，参数量增大通常能够提升模型的可控性，但即使相对较小的 p (p=512) 也已能够实现较高的控制准确率。在生成质量方面，Director 使用的参数量几乎是基础模型的两倍，

方法	#参数	BScore	平均准确率
基线	-	68.18	61.47
DASC (p=512)	15.94M	70.18	92.56
DASC (p=1024)	31.88M	70.12	92.72
DASC (p=2048)	63.75M	**70.42**	93.97
DASC (p=4096)	127.50M	70.26	**94.42**
Director	210.98M	69.54	94.14

表 4：用于可控性和生成质量的额外参数数量的影响。

（210.98M 对比 116.26M），可能存在过度参数化的问题，因此损害了其生成质量。中等规模的 DASC 能够取得最佳 BertScore，而更小规模的模型并不会显著降低性能。这表明 DASC 在应用中前景可期，因为其参数量少于其他方案，并且比通常拥有超过 6B 参数的大语言模型少几个数量级。

3.6 案例研究

除了图1所示的多方面控制外，我们还展示了一个概念验证应用：DASC 可以在生成的单条回复中自然融合两种情绪。只需将这两个属性的值都设为 1，而不是 ϕ 即可实现。结果如图 5和图 8所示。可以看到，DASC 能够成功生成带有单一情绪或融合两种情绪的回复；后者可能生成更加生动的回复。

3.7 ESConv 实验

为进一步探索 DASC 的潜力，我们还在另一个数据集 ESConv（Liu 等人，2021b）上进行了实验。该英文数据集旨在提供



图 5：给定不同的情绪及其组合作为控制属性，DASC 会针对相同的上下文生成不同的回复。

情感支持，帮助求助者应对 8 种预定义策略。我们将人工标注的策略标签用作控制属性，并测试了 3 种方法：**基线**、**CTRL**和**DASC**。由于效率低下，我们在此排除了 Director。我们报告自动评估指标**Distinct-2** 以及人工评估的**策略准确率**、**有用性**(1−4) 和**合理性**(1−4)。从表5可以看出，对相对复杂的策略进行控制更加困难，因此准确率低于此前的实验（表2）。尽管如此，DASC 仍实现了合理的控制准确率，并且在所有指标上均优于其他方法。这些结果表明，DASC 不受语言限制，可有效应用于多种属性控制。更多细节和生成示例见附录。

	Dist-2	Acc	有用的	合理性
基线	19.28	0.27	1.92	3.30
CTRL	21.20	0.52	2.04	3.31
DASC	**25.86**	**0.70**	**2.24**	**3.48**

表 5：ESConv 上的测试结果。

4 相关工作

近年来，可控生成受到广泛关注。PPLM（Dathathri 等人，2020年）提出了一种即插即用框架，通过额外的属性分类器来控制生成。此后的研究进展大致可分为三类。 重排序 方法利用属性分类器，对完整的生成候选结果进行简单排序（Thoppilan 等人，2022），或对部分生成结果进行排序，以引导后续输出（Yang 和 Klein，2021年）。集成式 方法则将.....

将与属性相关的可训练参数集成到生成模型中进行微调，例如离散控制代码（Keskar 等人，2019年）或连续提示前缀（Qian 等人，2022）。加权解码 方法利用令牌级属性分类器来引导每一步解码。例如，Krause 等人（2021年）和 Liu 等人（2021a）利用一个或两个额外的类别条件语言模型来提供属性判别。Director（Arora 等人，2022）将属性分类器作为简单的线性层集成在语言模型隐藏状态之上。

目前，多属性可控生成尚未得到充分研究。Lin 和 Riedl (2021年) 提出将加权解码扩展到多属性情形，方法是简单地将多个属性条件语言模型相乘。Gu 等人 (2022) 提出了一种结合交集搜索算法的基于变分自编码器的方法，用于多方面可控生成，但该方法无法直接用于对话生成等条件生成任务。Mireshghallah 等人 (2022) 提出了一种基于能量的可控生成方法，可以结合多种控制，但主要适用于固定长度生成。

可控生成技术在对话系统中尤为重要，研究人员也探索了多种控制方面的应用。例如，我们可以通过对话行为来调节生成内容，从而真实地体现期望的行为（Wen 等人，2015）；在回复中加入情绪，以增强聊天机器人的表达力（Zhou 等人，2018）；还可以加入性别（Su 等人，2020年）和角色设定（Zhang 等人，2018）等个人资料，打造更像人类的陪伴者。近年来，大语言模型的进步推动了对话生成的发展，使具有复杂个性和记忆的角色扮演等应用成为可能（Park 等人，2023年）。然而，其高昂的成本和隐私顾虑使其在某些部署场景中的适用性较低。

另一种可控生成方法采用稠密的角色描述（Zhang 等人，2018）。这种范式能够通过自由文本表达丰富的角色信息，例如个人状况、爱好和职业。自然语言形式允许整合其他语言资源，从而提升生成质量。例如，Song 等人 (2021年)将角色一致性学习任务与回复生成任务解耦，并利用非对话自然语言推理数据集

以帮助前者并进而提升后者。然而，尽管属性也可以通过自由文本描述来表达，但其中可能包含噪声，因此效果不如特定于属性的方法，正如我们的实验所示（附录C）。进一步结合这两种范式以实现更通用的可控生成，值得探索（Tang et al., 2023年）。

5 结论

本文提出了 DASC，一种用于多属性可控对话生成的新型框架。该框架建立在加权解码范式之上，以实现强可控性，并进一步将其置于属性语义空间中，从而能够通过对多个属性嵌入进行插值，同时控制多个属性。实验表明，DASC 能够在实现多属性生成强可控性的同时，在分布外场景中保持较高质量。与替代方案和大语言模型相比，DASC 所需的参数数量少得多，因此效率很高，是一种很有前景的部署选择。

局限性

所提出方法的一些局限性仍有待未来研究解决。

首先，我们的实验设置假设生成时所需的属性已知，这就需要单独的对话策略来决定提供给模型的属性标签。因此，我们的模型无法直接应用于端到端对话模型，也可能受到对话策略模型潜在误差传播的影响。由于 DASC 的预期用途是作为流水线式对话系统的一个组件，此类系统中这些常见问题不在本研究的讨论范围之内。

此外，我们需要使用带有多个属性的标注数据集进行评估，但这类数据集十分稀少。因此，我们主要在一个数据集上评估多属性控制能力。或许需要在更多数据集上进行实验，尤其是包含额外属性的数据集，以进一步验证所提方法的有效性。

最后但同样重要的是，DASC 并不直接适用于以自由文本作为控制信号的可控生成，例如角色设定描述（Zhanget al., 2018），这可能会限制其应用范围，尽管我们可以简单地将 DASC 与其他技术，例如拼接描述，也能实现这一目标，但还需要进一步探索。

伦理声明

所提出的方法用于控制性别风格。正如我们在第 3.2 节中注意到并讨论的那样，模型可能会依赖性别刻板印象来生成符合该性别的回复。潜在原因是，用于训练分类器的数据集本身就包含性别偏见标签，而分类器利用了这些偏见，并通过自动标注的标签将其传递给生成模型。为避免这种影响，我们可以仔细清理数据集，去除这类偏见标签（Gehman 等人，2020年），或者挖掘这类偏见词元，并在加权解码过程中对其进行惩罚。我们还可以应用基于人类反馈的强化学习，进一步减轻偏见（Ouyang 等人，2022）。

尽管所提出的方法主要旨在提升聊天机器人的趣味性，并赋予模型情感支持等能力，但此类方法也可能被用于恶意用途。例如，人们可能将毒性作为一项属性，鼓励模型生成更多有毒回复。因此，应谨慎限制此类技术的应用范围。

我们遵守所用数据集的许可协议。

参考文献

Kushal Arora、Kurt Shuster、Sainbayar Sukhbaatar 和 Jason Weston。2022年。Director：用于监督式语言建模的生成器-分类器。载于 计算语言学协会亚太分会第二届会议暨第十二届国际自然语言处理联合会议，第512–526页。

Tolga Bolukbasi、Kai-Wei Chang、James Y. Zou、Venkatesh Saligrama 和 Adam Tauman Kalai。2016年。男人之于计算机程序员，是否如同女人之于家庭主妇？消除词嵌入中的偏见。载于 《神经信息处理系统进展》第*29*卷：*2016*年神经信息处理系统年度会议，*2016*年*12*月*5*日至*10*日，西班牙巴塞罗那，第4349–4357页。

Yiming Cui、Wanxiang Che、Ting Liu、Bing Qin、Shijin Wang 和 Guoping Hu。2020年。重新审视中文自然语言处理的预训练模型。收录于计算语言学协会论文集：*EMNLP 2020*，第 657–668 页，线上。计算语言学协会。

Sumanth Dathathri、Andrea Madotto、Janice Lan、Jane Hung、Eric Frank、Piero Molino、Jason Yosinski 和 Rosanne Liu。2020年。即插即用语言模型：一种实现可控文本生成的简单方法。载于第八届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。

Samuel Gehman、Suchin Gururangan、Maarten Sap、Yejin Choi 和 Noah A. Smith。2020年。RealToxi-cityPrompts：评估语言模型中的神经毒性退化。载于计算语言学协会年会论文集：*EMNLP 2020*，第 3356–3369 页，线上。计算语言学协会。

Yuxuan Gu、Xiaocheng Feng、Sicheng Ma、Lingyuan Zhang、Heng Gong 和 Bing Qin。2022。用于多方面可控文本生成的分布视角。*arXiv* 预印本，abs/2210.02889。

Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。2020年。神经文本退化的奇特案例。载于第*8*届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。

Nitish Shirish Keskar、Bryan McCann、Lav R Varshney、Caiming Xiong 和RichardSocher。2019年。CTRL：一种用于可控生成的条件变换器语言模型。*arXiv*预印本，abs/1909.05858。

Ben Krause、Akhilesh Deepak Gotmare、Bryan McCann、Nitish Shirish Keskar、Shafiq Joty、Richard Socher 和 Nazneen Fatema Rajani。2021年。GeDi：生成式判别器引导的序列生成。载于计算语言学协会论文集：*EMNLP 2021*年研究成果，第 4929–4952 页，多米尼加共和国蓬塔卡纳。计算语言学协会。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于 第 *58* 届计算语言学协会年会论文集，第 7871–7880 页，线上。计算语言学协会。

Jiwei Li、Michel Galley、Chris Brockett、Jianfeng Gao 和 Bill Dolan。2016年。一种促进多样性的神经对话模型目标函数。载于《计算语言学协会北美分会*2016*年会议：人类语言技术论文集》，第110–119页，圣迭戈，加利福尼亚。计算语言学协会。

Zhiyu Lin 和 Mark O Riedl。2021年。Plug-and-blend：一种结合草图进行即插即用可控故事生成的框架。载于*AAAI*人工智能与互动数字娱乐会议论文集，第17卷，第58–65页。

Alisa Liu、Maarten Sap、Ximing Lu、Swabha Swayamdipta、Chandra Bhagavatula、Noah A. Smith 和 Yejin Choi。2021a。DExperts：借助专家与反专家实现解码时可控文本生成。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第6691–6706页，线上。计算语言学协会。

Siyang Liu、Chujie Zheng、Orianna Demasi、Sahand Sabour、Yu Li、Zhou Yu、Yong Jiang 和 Minlie Huang。2021b。迈向情感支持对话系统。载于《计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集》（第*1*卷：长篇论文），第3469–3483页，线上。计算语言学协会。

Yizhu Liu、Qi Jia 和 Kenny Q. Zhu。2022。通过预训练信息选择实现抽象式摘要的长度控制。载于《计算语言学协会第*60*届年会论文集》（第*1*卷：长篇论文），第6885–6895页，都柏林，爱尔兰。计算语言学协会。

Fatemehsadat Mireshghallah、Kartik Goyal 和 Taylor Berg-Kirkpatrick。2022。Mix and match：基于能量语言模型的免学习可控文本生成。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第401–415页，爱尔兰都柏林。计算语言学协会。

OpenAI。2022。Introducing ChatGPT。https://openai.com/blog/chatgpt。

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

Joon Sung Park、Joseph CO'Brien、Carrie J Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023年。生成式智能体：人类行为的互动模拟体。*arXiv*预印本 *arXiv:2304.03442*。

Jing Qian、Li Dong、Yelong Shen、Furu Wei 和 Weizhu Chen。2022。利用对比前缀实现可控自然语言生成。载于计算语言学协会年会 *ACL 2022* 论文集，第 2912–2924 页，爱尔兰都柏林。计算语言学协会。

Yunfan Shao、Zhichao Geng、Yitao Liu、Junqi Dai、Fei Yang、Li Zhe、Hujun Bao 和 Xipeng Qiu。2021年。CPT：一种用于中文语言理解和生成的预训练非平衡Transformer。*arXiv* 预印本 *arXiv:2109.05729*。

Haoyu Song、Yan Wang、Kaiyan Zhang、Wei-Nan Zhang 和 Ting Liu。2021年。BoB：用于利用有限个性化数据训练基于人物角色的对话模型的 BERT over BERT。*arXiv* 预印本 *arXiv:2106.06169*。

Yixuan Su、Deng Cai、Yan Wang、Simon Baker、Anna Korhonen、Nigel Collier 和 Xiaojiang Liu。2020年。通过信息引导的强化学习策略进行风格化对话生成。*arXiv* 预印本，abs/2004.02202。

Yihong Tang、Bo Wang、Miao Fang、Dongming Zhao、Kun Huang、Ruifang He 和 Yuexian Hou。2023年。通过对比潜变量增强个性化对话生成：结合稀疏和稠密人物角色。*arXiv* 预印本 *arXiv:2305.11482*。

Romal Thoppilan、Daniel DeFreitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du 等。2022。LaMDA：面向对话应用的语言模型。*arXiv* 预印本，abs/2201.08239。

Laurens Van der Maaten 和 Geoffrey Hinton。2008年。使用 t-SNE 可视化数据。机器学习研究杂志，9(11)。

Tsung-Hsien Wen、Milica Gaši´c、Nikola Mrkši´c、Pei- Hao Su、David Vandyke 和 Steve Young。2015年。面向口语对话系统的语义条件 LSTM 自然语言生成。载于*2015*年自然语言处理经验方法会议论文集，第1711–1721页，葡萄牙里斯本。计算语言学协会。

Liang Xu、Hai Hu、XuanweiZhang、Lu Li、Chenjie Cao、Yudong Li、Yechen Xu、Kai Sun、Dian Yu、Cong Yu、Yin Tian、Qianqian Dong、Weitang Liu、Bo Shi、Yiming Cui、Junyi Li、Jun Zeng、Rongzhao Wang、Weijian Xie、Yanting Li、Yina Patterson、Zuoyu Tian、Yiwen Zhang、He Zhou、Shaoweihua Liu、Zhe Zhao、Qipeng Zhao、Cong Yue、Xinrui Zhang、Zhengliang Yang、Kyle Richardson 和 Zhenzhong Lan。2020年。CLUE：中文语言理解评测基准。载于 第*28*届国际计算语言学会议论文集，第4762–4772页，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Xinchao Xu、Zhibin Gou、Wenquan Wu、Zheng-Yu Niu、Hua Wu、Haifeng Wang 和 Shihang Wang。2022。好久不见！具有长期人物角色记忆的开放域对话。载于 计算语言学协会：*ACL 2022* 论文集，第 2639–2650 页，爱尔兰都柏林。计算语言学协会。

Kevin Yang 和 Dan Klein。2021年。FUDGE：受控使用未来判别器进行文本生成。载于《*2021*年北美计算语言学协会分会会议论文集：人类语言技术》，第3511–3535页，线上。计算语言学协会。

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？载于《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，澳大利亚墨尔本。计算语言学协会。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 Yoav Artzi。2020年。BertScore：使用 BERT 评估文本生成。发表于第八届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。

Hao Zhou、Minlie Huang、Tianyang Zhang、Xiaoyan Zhu 和 Bing Liu。2018年。情感聊天机器：利用内外部记忆进行情感对话生成。发表于第三十二届 *AAAI* 人工智能会议（*AAAI-18*）、第 *30* 届人工智能创新应用会议（*IAAI-18*）和第八届 *AAAI* 教育人工智能进展研讨会（*EAAI-18*）论文集，路易斯安那州新奥尔良，美国，*2018*年*2*月*2*日至*7*日，第730–739 页。AAAI出版社。

A 控制强度的影响

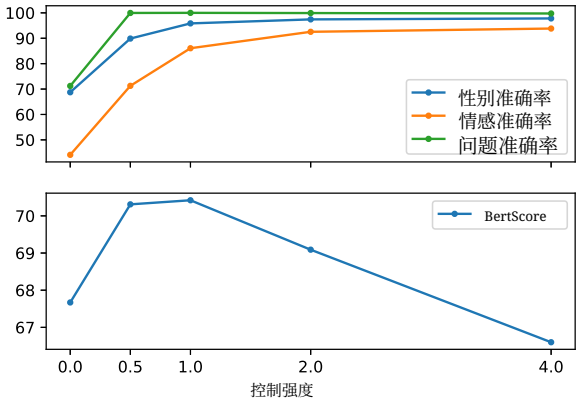


图 6：控制强度对可控性和生成质量的影响

我们展示了控制强度 α （公式（6））对 DASC 可控性和生成质量的影响，如图6所示。从图中的趋势可以看出，对于问题这类易于控制的属性，即使控制强度较低，也能实现完美控制；而对于 α等较难控制的属性，如情绪，则需要更高的 α 才能获得较高的成功率。因此，我们还可以进一步推测，为每个方面设置不同的控制强度，例如为情绪设置较高的 α，为问题设置较低的 α，或许能够更好地平衡各属性的控制准确率与生成质量。仔细调节参数或采用特定的搜索算法

（Gu 等人，2022）或许有助于实现这一目标，我们将此留待未来工作。[7]

B 实验细节

对于所有实验方法，均使用bart-base[8]作为基础模型。所有模型均在数据集上微调 6 个 epoch。在加权解码范式（公式（4））下对 Director 和 DASC 进行多方面控制时，我们将目标属性的变量设为 1，将其他变量设为 ϕ。解码方法为 top-p 采样，采样参数为 $p = 0.5$。DASC 使用控制权重$\alpha = 1$ 和分类器损失权重 $\beta = 0.1$，与之前的实验类似。本文的所有实验均在 Linux 服务器上进行，每个实验均在单个 NVIDIA A100 GPU 上运行。为避免过拟合，我们选择开发集上 BertScore 最佳的检查点进行最终测试。我们固定实验中的随机种子，并报告单次运行得到的结果。下面给出 ESConv 实验的具体细节。

对于 ESConv（Liu 等人，2021b）上的实验，我们使用最新发布的版本[9]，，其中包含 1,300 段对话，并将其划分为 1,100/100/100 的训练集/开发集/测试集，分别包含 15,605/1,403/1,369 条话语。在人工评估中，我们进一步从数据集定义的 7 种情感支持策略中，每种抽取 15 条话语（不包括含义模糊的 *Other* 类别），总计获得 105 条话语。 **合理性**(1–4) 的含义与前一个数据集中的实验类似：即响应是否流畅、与上下文连贯，并且符合常识。对于 **有用性**(1–4)，我们考察响应是否深入探讨了求助者面临的问题，是否具有安慰作用，是否包含有用的建议，或是否鼓励进一步深入讨论。

C 描述控制实验

稠密人物角色描述是对话生成中另一种常见的控制信号形式（Zhang 等人，2018年），我们也可以将稀疏属性转换为描述性文本，以兼容这些方法。因此，我们现在补充了两种利用角色描述进行控制的代表性方法的新实验结果。

BoB （Song 等人，2021年）将角色一致性学习与响应生成任务解耦，并利用非对话类自然语言推理数据集来提升一致性学习的性能，最终改善个性化生成效果。

为了在我们实验所用的数据集上应用 BoB，我们依据规则将离散的属性标注转换为文本描述。例如，男性/女性性别将转换为“我是一个女孩/男孩。”，问题则描述为“我想问一个问题”。对于情绪，我们将其填入模板“我感觉{emotion}。”中。我们将这三段描述文本拼接起来，作为 BoB 使用的角色设定文本。根据官方 GitHub 仓库的建议，我们采用中文自然语言推理数据集 CMNLI（Xu 等人，2020年）作为辅助推理数据集。

ChatGPT （OpenAI，2022）是一种具有代表性的大型语言模型（LLM），能够遵循人类指令并给出回应。因此，我们可以将属性值编码到 ChatGPT 系统消息中，以实现对这些属性的控制。我们使用以下模板：

> 下面给出一段对话历史。请以 {current speaker} 的身份，用 {gender and emotion} 的语气，通过 {dialog act} 回应下一句话。

然后，我们将对话历史用作用户消息，并让 ChatGPT （ gpt-3.5-turbo ）生成下一句话。我们使用温度=0，并在=“\n”处停止，即采用贪心搜索生成一行文本，这与原始数据集的设置类似。

	BScore	Dist-2	Acc_G	Acc_E	Acc_Q
基线	68.18	19.25	68.49	46.31	69.61
CTRL	**71.09**	18.91	85.32	77.49	**100.00**
DASC	70.42	21.94	**95.85**	**86.07**	**100.00**
BoB	65.47	23.44	74.59	64.76	98.51
ChatGPT	66.21	**30.98**	69.49	56.88	98.22

表6：采用基于人物描述的控制方法在 DuLemon 测试集上进行的自动评估。

	Dist-2	Acc_E
CTRL	21.07	43.38
DASC	26.71	**65.38**
BoB	24.25	30.13
ChatGPT	**37.02**	30.00

表7：采用基于人物描述的控制方法，在 DuLemon 鲁棒性测试集上的自动评估。

[7]在实践中，我们进一步将 α 乘以属性 K 的数量，以适应属性数量的变化。为清晰起见，公式（6）和图6 中未计入此项。[8] https://huggingface.co/facebook/bart-base[9] https://github.com/thu-coai/Emotional-Support-Conversation

首先，我们可以看到，ChatGPT 能够生成高度多样化的文本（测试集上的 Dist-2 与人工真实值相近，分别为 30.98 和 30.63）。然而，ChatGPT 无法达到与 CTRL 和 DASC 等微调方法相当的可控性，尤其是在提示中难以轻易描述其表现形式的属性上（例如性别和情绪）。

此外，我们可以看到，与基线和零样本 ChatGPT 相比，BoB 也表现出很强的可控性。除了 ChatGPT 之外，它的生成多样性也高于其他方法，这可能是引入辅助推理数据集所致。然而，它的可控性相对不如 DASC。它的生成质量也较差，这既体现在较低的 BertScore 上，也体现在我们的人工检查中；我们发现了许多不通顺的案例，例如“是的。毕竟，我是律师，否则很难找到一具尸体。”在鲁棒性测试中，BoB 对控制信号变化的敏感度也较低，这体现在较低的 Dist-2 和情绪准确率上。

总之，我们认为，像 ChatGPT 和 BoB 这样的基于描述的可控生成方法有可能更好地控制离散属性，但这需要在提示工程上付出大量努力（例如，为 ChatGPT 提供更详细的描述，或让 BoB 的描述更贴近辅助推理数据集）。不过，当我们拥有相对充足的离散控制属性标注数据时，在“小型”语言模型上使用 DASC 无疑是一个简单且有竞争力的选择。

D 示例与可视化

本节中，我们提供补充示例和图示。

图7 展示了图5中情绪组合示例的中文原文，我们还在图8中提供了另一个示例，展示了 DASC 甚至可以在同一回复中组合积极情绪喜爱和消极情绪恐惧，以表达复杂含义。

我们在图*9*和图*10*中展示了不同性别风格和问题行为句子的属性上下文嵌入的 t-SNE 可视化结果。我们发现了与情绪分析相似的结果：不同属性的嵌入明显分离。

我们还展示了 ESConv 上生成结果的 2 个示例。在图 11中，两个基线

上下文（最近几轮）：

A：你好，很高兴认识你，你没上班吗？B：我是个小演员，最近不用拍戏。

期望属性：

性别：中性 | 非问题

不同情绪下的 DASC：

惊讶： 原来是这样啊，我还以为你是大明星呢。

喜爱： 真羡慕你，我还在上学呢。

惊讶+喜爱: 哇塞,太厉害了,我以前也是个演员，不过现在只是一名普通白领。

图 7：图 5的原始中文文本

上下文（最近几轮）：

A：你喜欢看什么电影呢？（你喜欢看什么电影？）B：我比较喜欢看恐怖电影，你喜欢吗？（我比较喜欢看恐怖电影。你喜欢吗？）

期望属性：

性别：中性 | 非问题

不同情绪下的 DASC：

恐惧： 我不喜欢，我觉得恐怖电影太吓人了。（我不喜欢它们。我觉得它们太 吓人。）

喜爱： 我也喜欢看恐怖电影，我觉得看着特别刺激。(我也喜欢看恐怖电影，我觉得特别刺激。)

恐惧+喜爱： 我也挺喜欢的，但是我不敢看。 (我也喜欢 它们，但我太 害怕看它们了。)

图 8：DASC 在生成的回复中组合喜爱和恐惧情绪的示例。

并且 DASC 成功采用了预期的策略，而 CTRL 未能做到这一点。然而，基线在结尾处也包含了一个重复的问题，而 DASC 则给出了更全面的重新表述，体现出对情境的深入理解，因此对于寻求帮助者而言，这种回应会被认为更有帮助。在图12中，只有 DASC 在生成的回复中采用了正确的策略，而这种对焦虑情绪的精准反映使回应更具同理心。

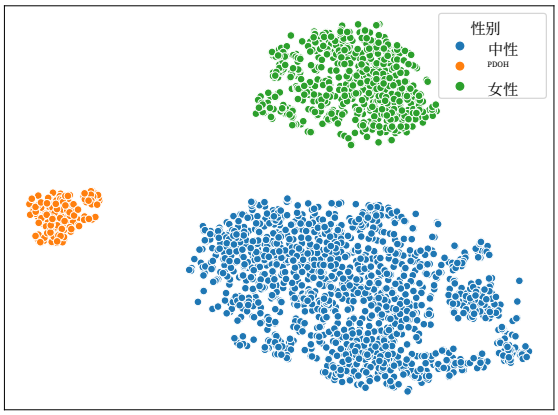


图 9：不同性别风格的回复的属性上下文嵌入可视化。

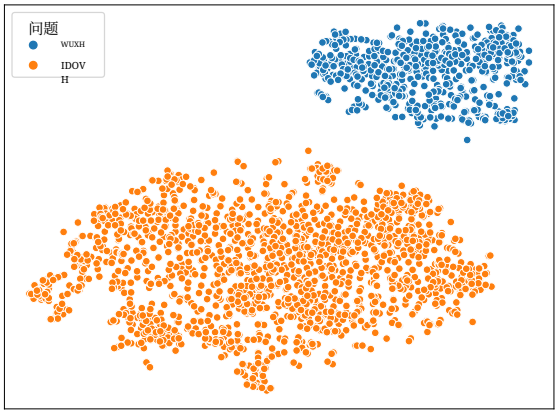


图 10：不同问题行为下回复的属性上下文嵌入可视化。

上下文（最近几轮）：

A：很高兴得知你一切都好。今天我能为你提供什么帮助？B：我曾参加过一场饮酒挑战，这件事已经影响我的婚姻有一段时间了。

策略： [重述或释义]

回复：

基线：听说 您正在面临饮酒挑战，我很遗憾。 我可以如何帮助您你今天怎么样？

CTRL：听到这个消息我很难过。我以前也经历过类似的情况。

DASC：我明白了。所以你正面临酗酒问题而且你正在努力维持你们的婚姻？

图 11：采用“复述或释义”策略的 ES-Conv 示例中的系统生成结果。

上下文（上一轮）：

A：你在想什么？B：由于新冠肺炎病例上升，我对可能发生的第二次停业感到非常焦虑。B：第一次停业期间，很多人失去了工作

策略： [情感反映]

回复：

基线：听到这些我很难过。你有没有试着和你的老板谈过这件事？

CTRL：听到这个消息我很遗憾。我相信你一定会渡过难关的。

DASC：我能理解你的感受，我自己也曾处于这种境地，我能看出你非常焦虑。

图12：采用“情感反映”策略时，ES-Conv 一个示例中的系统生成内容。