arXiv:2302.03668v2 [cs.LG] 1 Jun 2023

让困难提示词变得简单：基于梯度的离散优化：提示调优与发现

Yuxin Wen *1 **Neel Jain** *1 **John Kirchenbauer** [1] **Micah Goldblum** [2] **Jonas Geiping** [1] **Tom Goldstein** [1]1马里兰大学， [2]纽约大学{ywen, njain17, jkirchen, jgeiping,tomg}@umd.edu, goldblum@nyu.edu

摘要

现代生成模型的强大之处在于能够通过基于文本的提示进行控制。典型的“硬”提示由可解释的单词和词元构成，必须由人类手工编写。此外，还有“软”提示，它们由连续的特征向量组成。虽然可以使用强大的优化方法来发现这些提示，但它们难以解释、难以在不同模型间重复使用，也无法轻松接入基于文本的界面。

我们介绍一种方法，通过高效的基于梯度的优化，稳健地优化硬文本提示。该方法可自动为文本到图像和文本到文本应用生成硬文本提示。在文本到图像场景中，该方法为扩散模型创建硬提示，使 API 用户无需预先了解如何提示模型，就能轻松生成、发现并组合图像概念。在文本到文本场景中，我们展示了可以自动发现有效的硬提示，用于调优语言模型的分类能力。

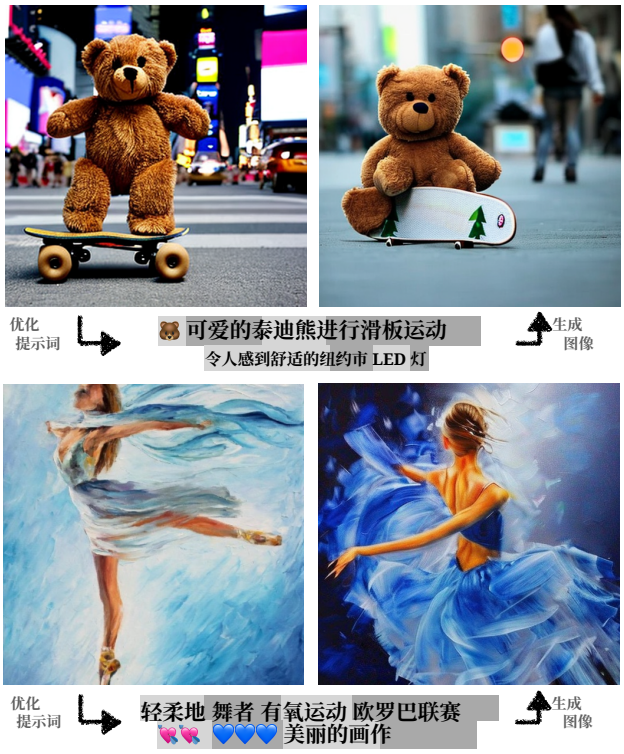

图 *1*。通过优化进行困难提示词发现的两个示例。给定一幅图像（左），使用 CLIP 发现离散文本提示词，并将其用于提示稳定扩散，从而生成新的图像（右）。使用两种深浅不同的灰色来显示所恢复提示词中的词元边界。

1. 引言

提示工程是创建用于引导生成模型的指令的艺术。它是释放大型模型在图像生成和语言任务方面能力的关键。如今，提示工程方法大致可以分为两类。首先是硬提示方法，这类方法使用人工设计的、由可解释词元组成的序列来诱导模型行为。困难提示词发现是一种专门的炼金术，许多优质提示词都是通过反复试验或纯粹的直觉发现的。其次是软提示，它们由连续值语言嵌入组成，并不对应任何人类可读的词元。软提示发现是一门数学科学；利用基于梯度的优化器和大型精选数据集，可以为专门任务生成性能极高的提示词。

尽管设计硬提示颇具挑战，但它们也有自己的优势。硬提示及其利用的技巧可以混合、搭配和变换，以执行各种不同的任务，而软提示则高度专用。硬提示具有可移植性；它们可以通过一个模型发现，随后部署到另一个模型上。

*贡献相同。代码可从 https://github.com/YuxinWenRick/hard-prompts-made-easy获取。

这种可移植性对于软提示而言是不可能的，因为不同模型的嵌入维度和表示空间存在差异。最后，当只能通过 API 访问模型，且无法控制输入的嵌入时，也可以使用硬提示。

本研究探讨了如何运用高效的梯度方法来优化和学习离散文本，重点关注其在提示工程中的应用。借此，我们实现了通过优化学习硬提示。学习得到的硬提示兼具软提示的简便性和自动化能力，以及硬提示的可移植性、灵活性和简洁性。我们的主要贡献总结如下：

- 我们提出了一种利用连续优化学习硬提示的简单方案。该方案以现有的文本优化梯度重投影方案为基础，并借鉴了大规模离散优化领域中量化网络的研究经验。
- 我们展示了这种优化方法可用于学习图像生成的硬提示，从而为我们提供了一种通用工具，用于创建能够引出特定图像风格、物体和外观的提示。尽管使用的词元少得多，且不包含任何手工设计的组件，这些学习到的提示仍能与高度专业化的提示生成工具相媲美。
- 我们还展示了学习到的硬提示在语言分类任务上的良好表现，优于其他文本优化方案。这些学习到的提示能够很好地迁移到不同网络；通过加入流畅度约束进行正则化，以提高可解释性后，其迁移效果会进一步提升。

除了捕捉学习得到的提示词所带来的可量化收益外，所提出的方案还可用于促进提示探索与发现，因为优化往往能找出既高度可解释、又出人意料的单词和词元。

2. 相关工作

语言模型中的提示。 Brown 等人 (2020年) 首批展示了提示在预训练语言模型任务适配中的强大能力。这种“指令微调”范式自此成为提升大型模型遵循复杂、特定任务指令能力的标准方法（Sanh 等人，2022年；Chung 等人，2022年）。然而，要自动找到适用于这些目的的文本提示集（即硬提示），仍然是一项尚未解决的挑战。 Lester等人(2021b) 简化了 Li 和 Liang (2021年) 提出的“前缀微调”技术，从而确立了这一流程被称为标准的 软 “提示词微调”，即优化一系列连续值嵌入，并将其置于输入词元的真实嵌入之前。然而， Khashabi 等人(2022年) 后续的研究表明，这种技术生成的嵌入序列可能映射为语义可解释性有限的词元序列。为解决这些局限，本文提出一种方法，将连续软提示优化与硬词汇表约束结合起来，从而生成特定任务且可解释的词元。

语言离散优化。 AutoPrompt（Shin 等人，2020年）是最早面向Transformer语言模型的离散提示词优化框架之一。此后的方法包括无梯度短语编辑方法（Prasad 等人，2022年）、基于朗之万动力学的嵌入优化方法（Shi 等人，2022年）以及一种强化学习方法（Deng 等人，2022年）。

我们将两种基于梯度的方法作为基线：FluentPrompt 和 AutoPrompt（Shi 等人，2022；Shin 等人，2020）。AutoPrompt 使用 Ebrahimi 等人（2018）提出的 HotFlip，利用梯度为提示词中的每个位置贪心地选择最优词元，从而筛选出一组优质候选项。然而，AutoPrompt 的计算成本很快就会变得很高。每次梯度更新时，该方法都需要评估提示词中每个位置上的每个候选项，因此会增加大量额外的前向传播。为避免额外的前向传播，我们最初考虑了添加和不添加流畅度约束的 $\text{AutoPrompt}_{k=1}$，但发现带有流畅度约束的 $\text{AutoPrompt}_{\text{SGD}}$ 表现优于其他版本，如图12所示。因此，与 Shi 等人（2022）类似，我们采用 AutoPrompt 的随机梯度下降版本作为另一种基线。FluentPrompt 与 AutoPrompt 的不同之处在于，它使用朗之万动力学（Kumar 等人，2022）优化提示嵌入，并加入了流畅度惩罚。

对于上文讨论的基线，在每个更新步骤结束时，都会将优化后的提示嵌入投影到其最近邻嵌入上，以确保优化是在自然语言词元构成的离散集合上进行的。然而，如果最近邻与嵌入相距甚远，且学习率未经过适当调优，嵌入可能会停滞不前，因而需要大量超参数调优，如图所示 8。这种约束的代价是，优化所能找到的解会失去灵活性。另一方面，虽然软提示不受这种限制，但正如 Khashabi 等人 (2022年) 所观察到的，将训练良好的软提示直接截断为最近的离散提示会严重降低性能。

从图像中发现提示。 从图像中提取丰富信息并通过自然语言文本传达的过程称为 图像描述。 Zhang等人(2021年)、 Hu等人(2022年)和 Li等人(2022年)通过在图文对上训练大型图像描述模型来实现这一目标

。然而，这些图像说明文字通常较为笼统，可能无法准确反映新出现或未见过的物体。在 Gal等人 (2022年)中，作者提出了一种利用软提示优化文本引导的扩散模型的方法，从而生成与原始图像中的视觉概念相似的内容。不过，在这种情况下，尽管最终的软提示效果不错，但通过扩散模型进行优化的成本非常高，而且提示既不可解释，也不可移植。

离散优化。离散优化器长期以来一直用于训练具有量化（例如二值）权重的神经网络。在这一背景下，在梯度更新步骤之间重新投影的方法被称为 随机舍入。然而，已知这种方法不具备连续优化所具有的融合保证（Li 等人， 2017年）。在过去十年中，随机舍入已被更新的优化器所取代，这些优化器保留权重的连续表示，而非离散表示（Courbariaux 等人，2015年）。这些优化器始终能够带来更高的准确率（Rastegari 等人， 2016; Courbariaux 等人，2016），并避免局部极小值（Li 等人， 2017年）。

我们从二值网络领域的这些经验中汲取灵感，并将其应用于改进和简化语言任务中的离散优化器。

3. 方法

硬提示学习。现在，我们介绍一种有效且易于使用的离散提示优化技术。该过程需要以下输入：一个冻结模型、θ、一系列可学习嵌入、$\mathbf{P} = [\mathbf{e_i}, ...\mathbf{e_M}], \mathbf{e_i} \in \mathbb{R}^d$，其中 M 是要优化的向量所对应的“词元”数量， d 是嵌入的维度。此外，我们使用一个目标函数 $\mathcal{L}$。词元空间的离散性通过投影函数 $\text{Proj}_\mathbf{E}$ 实现，该函数接收提示中的各个嵌入向量$\mathbf{e_i}$，并将其投影到嵌入矩阵 $E^{|V| \times d}$ 中的最近邻，其中 $|V|$ 是模型的词汇表大小。我们将此操作的结果记为 $\mathbf{P}' = \text{Proj}_\mathbf{E}(\mathbf{P}) := [\text{Proj}_\mathbf{E}(\mathbf{e_i}) ... \text{Proj}_\mathbf{E}(\mathbf{e_M})]$。此外，我们定义一个广播函数，$\mathcal{B}$ ： $\mathbb{R}^{(M \times d)} \rightarrow \mathbb{R}^{(M \times d \times b)}$，该函数在批次维度上将当前提示嵌入（$\mathbf{P}$）重复 b 次。

形式化地说，为了学习硬提示词，我们通过衡量$\mathbf{P}$ 在任务数据上的表现，最小化以下风险：$R(\mathbf{P}') = \mathbb{E}_D(\mathcal{L}(\theta(\mathcal{B}(\mathbf{P}, \mathbf{X})), \mathbf{Y}))$。

我们的方法。我们提出了一种简单而高效的基于梯度的离散优化算法，结合了基线离散优化方法与软提示优化的优点。我们将该方案称为 PEZ，其具体步骤定义于算法 1中。该方法维护连续迭代值，在我们的

算法1轻松**P**提示词**EZ**：PEZ 算法

输入： 模型 θ，词汇嵌入 $\mathbf{E}^{|V|}$，投影函数 Proj，广播函数 $\mathcal{B}$，优化步数 T，学习率 γ，数据集 D
从真实嵌入中采样：
$\mathbf{P} = [\mathbf{e_i}, ...\mathbf{e_M}] \sim \mathbf{E}^{|V|}$
对于 $1, ..., T$ **do**
  获取当前小批次（X， Y） $\subseteq D$。
  前向投影：
  投影$'$ =投影$_\mathbf{E}$(**投影**)
  计算关于投影后 嵌入的梯度：
  $g = \nabla_{\mathbf{P}'} \mathcal{L}_{\text{task}}(\mathcal{B}(\mathbf{P}', X_i) Y_i, \theta)$
  对连续 嵌入应用梯度：
  $\mathbf{P} = \mathbf{P} - \gamma g$
结束循环
最终投影：
$\mathbf{P}$ =Proj$_\mathbf{E}$[$\mathbf{P}$]
返回 P

应用中对应于一个软提示。在每次前向传播期间，我们首先将当前嵌入$\mathbf{P}$投影到最近邻$\mathbf{P}'$，然后计算梯度。接着，利用离散向量$\mathbf{P}'$的梯度，更新连续/软迭代值$\mathbf{P}$。

4. 使用 CLIP 进行提示反演

我们用于学习硬提示的方法非常适合多模态视觉语言模型。对于这类模型，例如 CLIP（Radford 等人，2021年），我们可以使用 PEZ 找出描述一个或多个目标图像的图像说明文字。进而，这些发现的图像说明文字可以作为提示词，用于图像生成应用。由于大多数文本引导扩散模型都会使用预训练文本编码器（例如 CLIP 文本编码器），并在训练期间将其冻结，因此我们可以利用这些预训练文本编码器来发现与下游扩散模型直接相关的提示词。例如，我们可以优化一段描述图像的文字，并将其用作扩散模型的提示词，以生成内容相同的其他图像。

由于 CLIP 模型有自己的图像编码器，我们可以将其用作损失函数来驱动 PEZ 方法。这样，我们优化提示词时只需考虑它们与 CLIP 图像编码器之间的余弦相似度，而完全避免对整个扩散模型进行梯度计算。

形式上，给定文本编码器函数 f 和图像编码器函数 g，我们通过最小化以下目标函数来优化与目标图像 x 对应的硬提示嵌入$\mathbf{P}$： $\mathcal{L}(\mathbf{P}\ x) = 1 - \mathcal{S}(f(\mathbf{P})\ g(x))$, 其中$\mathcal{S}$ 是两个向量之间的余弦相似度。

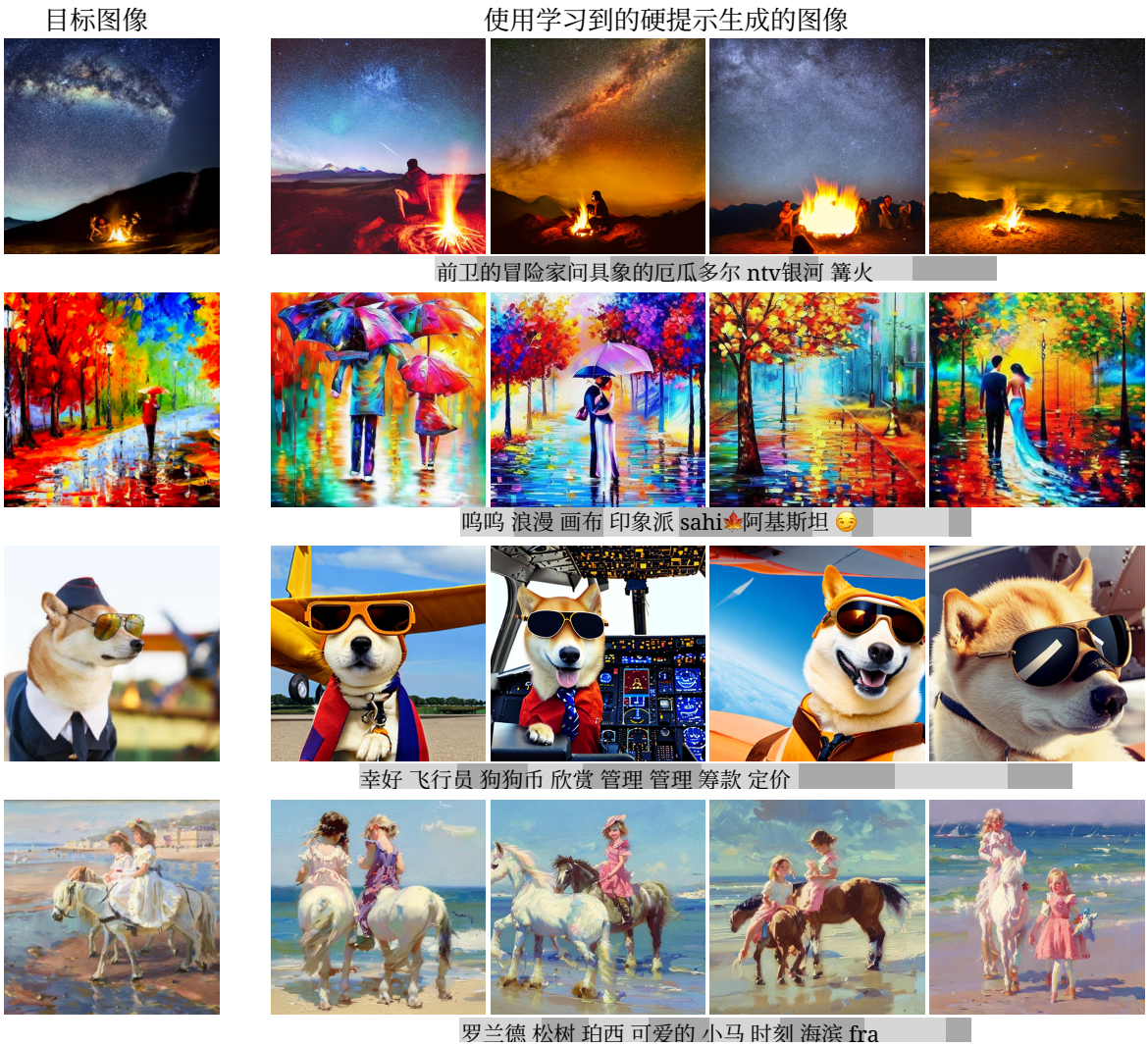

图 2。使用学习到的硬提示生成的图像，展示了针对四种不同目标图像的生成结果。对于给定的目标图像（左），使用 CLIP 找到离散文本提示词，并将其用于提示稳定扩散，从而生成新图像（右）。使用两种深浅不同的灰色显示恢复出的提示词中的词元边界。

4.1. 实验设置

我们在四个分布各异的数据集上进行实验：LAION（Schuhmann 等人，2022年）、MS COCO（Lin 等人，2014年）、Celeb-A（Liu 等人，2015年）和 Lexica.art（San-tana，2022年）。LAION 包含从互联网上抓取的超过50亿张各类图像，包括照片和画作。MS COCO 主要包含包含多个常见物体的真实生活照片，而 Celeb-A 由名人肖像组成。Lexica.art 是一个包含 AI 生成画作及其提示词的数据集。

我们通过原始（目标）图像与使用学习到的硬提示生成的图像之间的图像相似度来衡量提示词质量。为此，我们使用更大型的参考 CLIP 模型 OpenCLIP-ViT/G。该模型未参与优化，可作为中立的评估指标，用于衡量 se-图像之间的语义相似度。

我们选择稳定扩散-v2（Rombach 等人，2022年）作为生成模型，并使用开源 CLIP 模型 OpenCLIP-ViT/H（Cherti 等人，2022年）来构造提示词，因为两者使用相同的文本编码器。在提示词优化过程中，我们采用通用学习率 0.1，并使用 AdamW 优化器运行 3000 个优化步数（Loshchilov 和 Hutter，2017年）。对于稳定扩散-v2，我们将引导尺度设为 9，推理步数设为25。对于每个数据集，我们随机抽取100 个数据点，并对使用不同随机种子进行的5 次运行所得的 CLIP 分数取平均值。

一种自然的 CLIP 困难提示词发现基线

表1。对学习得到的硬提示进行定量评估。我们报告原始图像与硬提示生成的图像之间的 CLIP 分数。分数越高，表示生成图像与目标图像包含的语义内容越相似。

	令牌数	要求	LAION	MS COCO	Celeb-A	Lexica.art
PEZ（我们的方法）	8	CLIP	0.697	0.674	0.602	0.711
CLIP Interrogator	∼ 77	CLIP +Bank +BLIP	0.707	0.690	0.558	0.762
不含 BLIP 的 CLIP Interrogator	∼ 77	CLIP + Bank	0.677	0.674	0.572	0.737
PEZ（我们的方法） + Bank	8	CLIP + Bank	0.702	0.689	0.629	0.740
CLIP Interrogator	8	CLIP + Bank +BLIP	0.539	0.575	0.360	0.532
CLIP Interrogator	16	CLIP + Bank +BLIP	0.650	0.650	0.491	0.671
CLIP Interrogator	32	CLIP + Bank +BLIP	0.694	0.663	0.540	0.730
软提示	8	CLIP	0.408	0.420	0.451	0.554



图3。 用于风格迁移的学习到的硬提示。给定几张风格相同的样例图像，我们可以用硬提示提取其风格，并将其迁移到其他物体或场景中。详细的模板和硬提示见附录 A.1。样例图像来源： Qinni 和 facundo-lopez。

是 *CLIP Interrogator*[1]。为了生成描述性硬提示，该工具首先使用预训练图像描述模型 BLIP（Li 等人，2022年）为目标图像生成说明文字。然后，根据关键词与目标图像之间的 CLIP 分数，将排名靠前的k 关键词从预先收集的关键词库中添加到说明文字后。这些关键词来自多个渠道，包括 5,265 个艺术家姓名（如“梵高”）以及 100,970 条提示工程短语，因此形成了丰富多样的集合。我们发现，该关键词库涵盖了 Lexica.art 数据集中的大多数短语。 *CLIPInterrogator* 随后贪心地采样关键词，直到提示词达到 CLIP 的词元长度限制77。

4.2. 结果

我们在图2中展示了使用我们的方法学习到的硬提示及其对应的生成结果。生成的图像清楚地表明，这些提示有效地捕捉了目标图像的语义特征。此外，无论是根据 CLIP 分数衡量，还是通过视觉检查，生成结果都与原始图像高度相似。另外，硬提示不会对原始目标图像过拟合，并且在使用不同随机种子时能够生成多样化的图像。

提示词是人类可读的，由真实单词和乱码（非词语词元序列）混合组成。然而，提示词中包含的有效单词为图像提供了大量信息。例如，在第一行中，我们可以看到“milkyway”和“campfire”这两个单词，它们正是目标图像中的两个主要元素。有趣的是，优化后的提示词也可能包含表情符号，如第二行所示。它代表了画面一侧的树木，也体现了图像的色彩主题。优化过程似乎会选择这些表情符号，在保持提示词简洁的同时纳入有用信息。

[1]https://github.com/pharmapsychotic/clip-interrogator



图 *4.* 拼接后的学习得到的硬提示。我们展示了在两张互不相关的图像上学习得到的硬提示可以拼接起来，从而融合其中的语义概念。

此外，我们在表1中给出了定量评估结果。我们的方法在全部四个数据集上表现稳定，优于其他基于梯度的优化基线（完整表格见表7）。值得注意的是，我们取得了与 *CLIP Interrogator*相近的性能；它在 LAION、MS COCO 和 Lexica.art 上的 CLIP 分数最高，但在 Celeb-A 上并非如此（*CLIPInterrogator* 的关键词库未包含许多与真实人脸相关的单词）。不过，*CLIPInterrogator* 使用了大型精选提示数据集、图像描述模型 BLIP，以及大量词元（最多77个）；而我们提出的方法仅使用 CLIP 模型进行提示发现，总共只用8 个词元，充分展现了其兼具简洁性与强大性能。

我们对上述每项差异分别进行了消融实验。为此，我们将关键词库纳入优化方法，并仅允许投影到关键词库中的词元上。总体而言，我们发现，在模型中加入这一约束并禁用 BLIP，以便在公平的条件下比较两种方法后，便能弥补这两种方法在 LAION 和 Lexica.art 上的大部分量化差距。此外，缩短 *CLIP Interrogator* 的词元长度会导致性能大幅下降；再次强调，在将两种方法的词元长度统一为 8 进行比较时，这一现象尤为明显。我们注意到，尽管稳定扩散和 CLIP 使用相同的文本编码器，但在我们的评估中，与所有硬提示方法相比，软提示的迁移效果并不好。

提示长度。 我们进一步消融研究了最佳词元数量。在图 5中，我们发现，使用稳定扩散生成图像时，提示越长并不一定效果越好，尽管更长的提示会严格降低 CLIP 图像编码器上的损失。因此，长提示会过拟合，迁移能力也较弱；我们通过实验证明，长度为 16 时能够获得泛化性最强的表现。

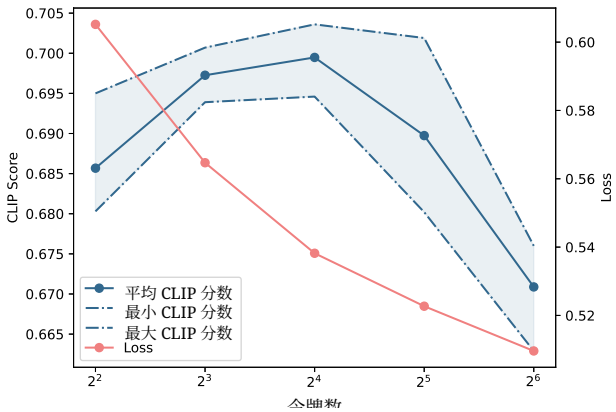

图*5*。提示长度消融实验，展示了随着提示长度增加，CLIP 图像编码器上的训练损失，以及生成的稳定扩散图像的验证集 CLIP 分数。

4.3. 风格迁移

所提出的方法也可以轻松适用于风格迁移。我们沿用 Gal 等人（2022年）研究软提示时采用的设置，但使用我们的硬提示。给定若干具有相同风格的示例，我们将它们共有的风格特征提取到一个硬提示中，并用该提示将这种风格应用于新的物体或场景。图 3 展示了两个风格迁移示例，表明我们的方法能够轻松地将共有的风格元素嵌入提示中，并将其应用于新概念。模板和学习到的提示可见附录 A.1。



图6。提示词蒸馏 。尽管词元更少，硬提示仍能生成与原始图像概念上非常相似的图像 al.

4.4. 提示拼接

学习得到的硬提示也非常适合作为可组合构建块，用于构建复杂场景。我们在图 4中对此进行了测试：我们分别为两张无关的图像生成提示，然后通过拼接它们的提示来融合两张图像。我们发现，即使是不同的概念，例如海滩上的彩绘马和森林中逼真的日落，也可以通过生成的提示组合在一起。

4.5. 提示蒸馏

我们还可以将提示优化方法用于提示蒸馏，在保留提示能力的同时缩短其长度。在扩散模型的文本编码器最大输入长度有限时，蒸馏尤其有用，例如 CLIP 模型的最大输入长度为 77 个词元。此外，较长的提示可能包含冗余且不重要的信息，尤其是手工编写的提示，因此我们的目标是提炼其精髓，只保留提示中的重要信息。我们仅依据文本编码器 f 优化较短的提示，使其特征与较长提示的特征相匹配。给定目标提示的嵌入 $\mathbf{P}_{\text{target}}$和可学习嵌入 $\mathbf{e}$，我们只需将损失修改为：$\mathcal{L} = 1 - Sim(f(\mathbf{P}_{\text{target}}) \; f(\mathbf{P}))$。我们将蒸馏比率定义为 $|\mathbf{P}|/|\mathbf{P}_{\text{target}}|$。

在图 6中，我们展示了原始提示词和经过四种不同蒸馏比率蒸馏的提示词生成的图像：0。7、0。5、0。3、0。1。我们可以看到，即使只有 3 或 4 个词元，硬提示仍能生成概念上与原始图像非常相似的图像，成功蒸馏了较长的人工编写指令。

5. 使用语言模型进行离散提示调优

在文本到文本设置中，算法1 的目标是找出一个离散词元序列，即硬提示，以引导语言模型预测分类任务的结果。由于流畅度是文本的一项重要属性，Shi 等人 (2022年)发现，流畅度能够提升提示的可读性和性能。因此，我们在本节中将优化目标定义为任务损失与流畅度损失的加权函数，

$$\mathcal{L} = (1 - \lambda_{\text{fluency}})\mathcal{L}_{\text{task}} + \lambda_{\text{fluency}}\mathcal{L}_{\text{fluency}}.$$

我们设定 $\lambda = 0.003$ 与Shi 等人（2022年）相似，并消融去除流畅度因素后的方法（$\lambda = 0$），记为无流畅度。我们旨在证明，这种方法生成的硬提示在不同基于Transformer的语言模型之间迁移时效果良好，也能用于在少样本设置下发现提示词。这些提示词的一个突出优点，尤其是在语言应用中，是可以先在较小的语言模型上进行优化，再迁移到其他规模大得多的模型上。

表 2。在 GPT-2 Large 上学习的各方法的五个提示，在 SST-2 验证集上的准确率（%）和标准误差，以及将这些提示迁移到参数量为 1.3B 至 6.7B 的更大型模型后的结果。软提示 的基线准确率为 $\mathbf{93.35}_{\pm 0.01}$（针对 GPT-2 Large 优化），但无法在不同模型之间迁移。Empty^{Template} 表示前面没有提示，但包含预先设定的模板。

方法	GPT-2 Large（755M，**来源**）	GPT-2 XL（1.3B）	T5-LM-XL（3B）	OPT（2.7B）	OPT（6.7B）
空^{Template}	80.84	73.85	52.75	72.48	58.72
AutoPrompt$_{\text{SGD}}$	$87.56_{\pm 0.35}$	$78.19_{\pm 2.68}$	$56.01_{\pm 1.67}$	$73.69_{\pm 1.63}$	$65.28_{\pm 1.75}$
FluentPrompt	$\mathbf{88.33}_{\pm 0.35}$	$78.53_{\pm 2.82}$	$55.64_{\pm 0.59}$	$70.39_{\pm 2.08}$	$61.74_{\pm 1.25}$
PEZ$^{\text{No Fluency}}$（我们的）	$88.12_{\pm 0.15}$	$77.8_{\pm 3.45}$	$61.12_{\pm 2.94}$	$76.93_{\pm 1.29}$	$71.72_{\pm 3.16}$
PEZ^{Fluency}（我们的）	$88.05_{\pm 0.55}$	$\mathbf{79.72}_{\pm 3.26}$	$\mathbf{63.30}_{\pm 2.30}$	$\mathbf{77.18}_{\pm 3.82}$	$\mathbf{72.39}_{\pm 1.82}$

5.1. 数据集与设置

我们将算法1 与相关算法在三个分类任务上进行比较，其中包括两个情感分析任务：SST-2（Socher等人，2013年）和亚马逊情感极性数据集（McAuley和Leskovec，2013年），以及一个四分类任务 AGNEWS（Zhang等人，2015年）。我们基于Ding等人（2022年）探索的设置，使用 GPT-2 Large（7.74亿参数）（Radford等人，2019）和 Adafactor优化器（Shazeer和Stern，2018年），以32为批量大小（Lester等人，2021a），对硬提示进行优化。提示模板和标签映射词的详情见表4。

可迁移性设置。 为了测试可迁移性，我们使用 GPT-2 Large 生成提示，训练 5000 步。然后，我们为每种技术选出平均验证准确率最高的五个提示，并在更大的模型上进行测试。我们在以下模型上测试迁移后的文本：GPT-2 XL、T5-LM-XL、OPT-2.7B 和 OPT-6B（Radford等人，2019；Lester等人，2021b；Zhang等人，2022年），以验证所提出算法相较于相关技术的可靠性，并测试硬提示能否稳定提升性能。因此，我们还将仅包含模板的空提示作为基线。

少样本设置。 在少样本设置下，我们在 AGNEWS 数据集上使用 GPT-2 Large 对每个提示进行 100 个 epoch 的优化，从每个类别中抽取两个示例（$k = 2$）和四个示例（$k = 4$）来构建训练集。此外，我们还创建一个规模相同的留出集，最后在整个验证集上验证这些提示。

5.2. 结果

我们验证了，在情感分析任务中，我们的方法与其他方法表现相当；在 AGNEWS 上，我们的方法则优于其他方法约 2 %。详见表5 。

提示迁移能力。 表 2 展示了每种方法中在 GPT-2 Large 上训练的五个提示迁移到其他大语言模型后的表现。有趣的是，仅仅扩大模型规模——而不进行额外训练——并不能保证性能相应提升。[2] 我们看到，所有基于梯度的方法

[2]在 GPT-2 Large 和 XL 上进行的一项快速实验显示，使用模板与不使用模板相比，模板能提升性能

表3。在 AGNEWS 上，每个类别使用 k 个示例/样本时，各方法在 100 个随机种子上的平均验证准确率及标准误；所有方法均采用提前停止（包括软提示），并使用三个词元**附加到文本末尾**，与原始模板（“It was about”）类似。我们在这些实验中设置 $\lambda = 0.03$ 。“Empty”表示不含额外提示词的模板。

方法	k=2	k=4
空^{Template}	58.34	58.34
PEZ$^{\text{No Fluency}}$（我们的方法）	$70.07_{\pm 0.81}$	$73.99_{\pm 0.45}$
PEZ^{Fluency}（我们的方法）	$70.93_{\pm 0.60}$	$74.15_{\pm 0.48}$
软提示	$74.92_{\pm 0.58}$	$79.93_{\pm 0.36}$

与仅评估模板相比，都能够实现迁移；我们发现，采用流畅度约束训练的提示比其他提示迁移效果更好。此外，我们可以看到，使用我们的流畅方法时，OPT-6.7B 带来的提升最大，比仅使用模板的基线高出约 14% 。此外，从附录的表 6 中可以看到，我们的 AGNEWS 提示能够从 GPT-2 Large 迁移到 GPT-2 XL。

提示发现。 表 3 表明，即使只有少量样本，与前置提示相比，我们也能获得较高的验证准确率。值得注意的是，每次少样本运行大约需要5 分钟。

我们运行了 100 个随机种子，每个训练集都包含来自每个类别的 k 个样本，并对排名靠前的提示进行定性考察。尽管许多提示难以解读，但其中也有许多连贯易懂。例如，即使对于 $k = 2$，一些提示也包含了“*BBC*”等新闻来源；另一些提示则针对来自博客的文本，提出了处理新闻分类任务的新思路：“*Brian* 博客，”或“博客革命分析。”得益于这些基于梯度的方法的高效性，提示工程师可以借助它们探索发现新颖提示的新途径。

6. 安全问题

文本到图像扩散模型 API 中常使用基于词元或词语级别的内容过滤器，以防止生成

针对不同模型的表现也有所不同。



图 7。通过 Midjourney 生成的受版权保护图像。此处仅为研究目的通过 API 请求生成。

不适宜工作场所内容或受版权保护的内容。例如，图像生成 API `Midjourney` 因著名的阿富汗女孩照片涉及版权问题，禁止提示中包含子字符串“阿富汗人”[3]。

然而，提示词优化可以被用作绕过简单的基于规则的内容过滤器的手段。PEZ 可以生成一个提示词，避开禁用词元，同时仍与原始目标提示“阿富汗女孩”的文本特征相匹配。图 7 展示了`Midjourney` 使用优化后的提示词生成的结果，该提示词在不包含禁用词“阿富汗人”的情况下，成功复现了被禁图像。请注意，提示词似乎错误地将图像中的人物沙尔巴特·古拉与塔利班联系起来。

即使防御者此时反复更新屏蔽列表，将对抗性提示词中的其他单词也列为禁用词，攻击者仍能持续绕过新增的内容限制，正如补充材料图 10所示。总的来说，我们认为，只有基于完整特征的内容检测器才有望缓解模型所有者面临的这些问题（Rando 等人，2022年）。

7. 结论

我们提出了一种新方法，利用连续嵌入可靠地优化硬提示。我们的方法的关键优势在于，在优化硬提示词元时，将连续的（即软）提示作为中间变量，并利用基于梯度的优化。这样一来，算法便能在嵌入空间中选出离散嵌入有效的位置，而不是仅仅优化软提示，再将其投影到附近的词元嵌入上，寄希望于这些邻近的硬提示也能表现良好。此外，由于我们的方法会在所有步骤中利用梯度，并将其累积到软提示中，因此这一过程能让优化对学习率和数据中的潜在噪声更加稳健。

尽管我们的工作推动了提示词优化的发展，但社区对语言模型嵌入空间的理解仍处于起步阶段。未来，若能更深入地理解嵌入空间的几何结构，很可能就能实现更强大的提示词优化。

总的来说，我们通过实验表明，硬提示可以轻松生成，并能灵活应用于实际场景。然而，硬提示的一个局限是，尽管它们可供人类阅读，却仍可能包含若干难以解读的词元。此外，硬提示还可能从语言模型的训练数据中提取有害短语或敏感内容。尽管我们没有观察到这类行为的具体实例，但这是未来应用中应当考虑的一个问题。

8. 致谢

本研究得以完成，离不开美国海军研究办公室（N000142112557）、美国海军研究办公室多学科大学研究计划、美国国家科学基金会（IIS-2212182）以及第一资本银行的支持。

参考文献

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. 语言模型是少样本学习者。发表于 第*34*届神经信息处理系统大会（*NeurIPS 2020*），2020年12月。网址：http://arxiv.org/abs/2005.14165。

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C.,Schmidt, L. 和 Jitsev, J. 对比式语言-图像学习的可复现缩放定律。*arXiv* 预印本 *arXiv:2212.07143*，2022年。

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S. 等。指令微调语言模型的缩放。*arXiv* 预印本 *arXiv:2210.11416*，2022年。

Courbariaux, M., Bengio, Y., and David, J.-P. Binarycon- nect：在传播过程中使用二进制权重训练深度神经网络。神经信息处理系统进展，28，2015年。

[3] https://en.wikipedia.org/wiki/Afghan_女孩

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. 二值化神经网络：将权重和激活值限制为+ 1或-1，以此训练深度神经网络。*arXiv* 预印本 *arXiv:1602.02830*，2016年。

Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E. P., and Hu, Z. Rlprompt：通过强化学习优化离散文本提示。*arXiv*, abs/2205.12548, 2022年。

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H., and Sun, M. OpenPrompt：一个用于提示学习的开源框架。载于计算语言学协会第*60*届年会论文集：系统演示，第105–113页，爱尔兰都柏林，2022年5月。计算语言学协会。doi: 10.18653/v1/2022.acl-demo.10.
URLhttps://aclanthology.org/2022.acl-demo.10。

Ebrahimi, J., Rao, A., Lowd, D. 和 Dou, D. Hot-Flip：用于文本类别分类的白盒对抗样本。载于第*56*届计算语言学协会年会论文集（第*2*卷：短论文），第31–36页，澳大利亚墨尔本，2018年7月。计算语言学协会。doi：10.18653/v1/P18-2006。网址：https://aclanthology.org/P18-2006。

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G. 和 Cohen-Or, D. 一张图像胜过一个词：使用文本反演实现文本到图像生成的个性化。*arXiv*预印本 *arXiv:2208.01618*，2022年。

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y. 和 Wang, L. 扩大用于图像描述的视觉语言预训练规模。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第17980–17989页，2022年。

Khashabi, D., Lyu, X., Min, S., Qin, L., Richardson, K., Welleck, S., Hajishirzi, H., Khot, T., Sabharwal, A., Singh, S., and Choi, Y. 提示偏离：离散化解释连续提示的奇特案例。载于 计算语言学协会北美分会*2022*年会议论文集：人类语言技术，第 3631–3643 页，美国西雅图，2022年7月。计算语言学协会。doi: 10.18653/v1/2022.naacl-main.266。网址：
https://aclanthology.org/2022.naacl-main.266。

Kumar, S., Paria, B., and Tsvetkov, Y. 基于梯度的语言模型约束采样。*arXiv* 预印本 *arXiv:2205.12558*，2022年。

Lester, B.、Al-Rfou, R. 和 Constant, N. 参数高效提示调优的规模效应。载于神经信息处理系统进展 *2021*年会议论文集

自然语言处理，第 3045–3059 页，线上会议及多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。doi: 10.18653/v1/2021.emnlp-main.243。网址：
https://aclanthology.org/2021.emnlp-main.243。
Lester, B.、Al-Rfou, R. 和 Constant, N. 参数高效提示调优的规模效应。*arXiv:2104.08691 [cs]*, 2021年9月。网址：http://arxiv.org/abs/2104.08691。Li, H.、De, S.、Xu, Z.、Studer, C.、Samet, H. 和 Goldstein, T. 训练量化网络：更深入的理解。神经信息处理系统进展，30，2017年。

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP：引导语言-图像预训练，以统一视觉-语言理解与生成。*arXiv* 预印本 *arXiv:2201.12086*，2022年。

Li, X. L. and Liang, P. 前缀调优：优化用于生成的连续提示。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4582–4597页，2021年。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., and Zitnick, C. L. Microsoft COCO：情境中的常见物体。载于《欧洲计算机视觉会议》，第740–755页。施普林格，2014年。

Liu, Z., Luo, P., Wang, X., 和 Tang, X. 自然环境中的深度学习人脸属性。在*IEEE*国际计算机视觉会议论文集，第3730–3738页，2015年。

Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*，2017年。

McAuley, J. 和 Leskovec, J. 隐因子与隐藏主题：利用评论文本理解评分维度。在第*7*届*ACM*推荐系统会议，RecSys ’13，第165–172页，纽约，纽约州，美国，2013年。计算机协会。ISBN 9781450324090。doi: 10.1145/2507157.2507163。
URLhttps://doi.org/10.1145/2507157.2507163。

Prasad, A., Hase, P., Zhou, X., 和 Bansal, M. Grips：用于提示大语言模型的无梯度、基于编辑的指令搜索。*arXiv* 预印本 *arXiv:2203.07281*，2022年。

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., 和 Sutskever, I.《语言模型是无监督多任务学习者》。*OpenAI*，第 24 页，2019。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等。从自然语言监督中学习可迁移视觉模型。载于 国际机器学习大会，第 8748–8763 页。PMLR，2021。Rando, J., Paleka, D., Lindner, D., Heim, L., 和 Tram`er, F. 对稳定扩散安全过滤器进行红队测试。*arXiv*，abs/2210.04610，2022。Rastegari, M., Ordonez, V., Redmon, J., 和 Farhadi, A. Xnor-net：使用二值卷积神经网络进行 ImageNet 分类。载于 欧洲计算机视觉会议，第 525–542 页。施普林格，2016。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 和 Ommer, B. 使用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 10684–10695 页，2022。Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., 和 Rush, A. M. 多任务提示训练实现零样本任务泛化。载于国际学习表征会议，2022。网址 https://openreview.net/forum?id=9Vrb9D0WI4。

Santana, G. Gustavosta/Stable-Diffusion-Prompts ·Hugging Face 上的数据集，2022 年 12 月。网址 https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts。

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 等。Laion-5b：用于训练下一代图文模型的开放大规模数据集。*arXiv* 预印本 *arXiv:2210.08402*，2022。Shazeer, N. 和 Stern, M. Adafactor：具有次线性内存开销的自适应学习率。载于 国际机器学习大会，第 4596–4604 页。PMLR，2018。Shi, W., Han, X., Gonen, H., Holtzman, A., Tsvetkov, Y., 和 Zettlemoyer, L. 迈向可读提示调优：库布里克的《闪灵》是一部好电影，也是一个好提示吗？*arXiv* 预印本 *arXiv:2212.10539*，2022。

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. AutoPrompt：利用自动生成的提示从语言模型中 eliciting 知识。发表于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 4222–4235 页，线上，2020年11月。计算语言学协会。doi: 10.18653/v1/2020.emnlp-main. 346。网址：https://aclanthology.org/2020.emnlp-main.346。

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. 基于递归深度模型的情感树库语义组合性研究。发表于*2013*年自然语言处理实证方法会议论文集，第 1631–1642 页，美国华盛顿州西雅图，2013年10月。计算语言学协会。网址：https://www.aclweb.org/anthology/D13-1170。

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y. 和 Gao, J. Vinvl：重新审视视觉语言模型中的视觉表示。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 5579–5588 页，2021年。

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V. 等。Opt：开放式预训练Transformer语言模型。*arXiv* 预印本 *arXiv:2205.01068*，2022年。

Zhang, X.、Zhao, J. 和 LeCun, Y. 字符级卷积网络用于文本分类。神经信息处理系统进展，28，2015年。

A. 附录

附录 A.1：使用 CLIP 进行提示反演的补充结果

我们在图中提供了更多定性结果9。

对于图3中的每个示例，我们分别使用以下模板：“一只采用 {}风格的老虎”，“一条采用 {}风格的巴黎街道”，“一个采用 {}风格的计算器”，“一枚采用 {}风格的火箭”，其中 {} 替换为以下硬提示词：

resonvillains 星空观测 插画 教程 sma 国际妇女节 水彩 fiberlilycamila 横滨 -悲伤 fluidslatest

npr 动漫 小说 pureibanganesha irvin paints encapsulmondo illustrillustroversized sultanconan ¢

分别对应实验 1 和实验 2。

A.2. 文本到文本硬提示的补充实验与详细信息

基线目标函数的形式化定义 形式上，我们将 *AutoPrompt*$_{SGD}$ 步骤定义为：

$$\mathbf{P}_{i+1} = \text{Proj}_{\mathbf{E}}[\mathbf{P}_i - \eta \nabla_{\mathbf{P}_i} \mathcal{L}(\mathcal{B}(\mathbf{P}_i, X_i), Y_i, \theta)]$$

此外，*FluentPrompt* 的更新定义如下：

$$\mathbf{P}_{i+1} = \text{Proj}_{\mathbf{E}}[\mathbf{P}_i - \eta \nabla_{\mathbf{P}_i} \mathcal{L}(\mathcal{B}(\mathbf{P}_i, X_i), Y_i, \theta) + \sqrt{2\eta\beta_i} z]$$

第 5 节的详细信息 对于表5，我们报告三个学习率（0.1、0.3 和 0.5）下的最佳验证准确率；对于 *FluentPrompt* 和 *AutoPrompt*$_{\text{SGD}}$，我们使用文中报告的学习率（1、3 和 10），其余超参数则遵循 Shi 等人 (2022年) 对 *FluentPrompt*的设置。在这些实验中，我们会将长度为 10 个词元的提示词前置于每条输入文本。所有方法均采用提前停止：每个数据集使用 5000 个样本作为留出集，并每隔 100 步进行一次评估。

表 5 显示，我们在情感分析方面与其他方法表现相当，而在 AGNEWS 上则领先其他方法约 2%。检查提示后，我们发现所有方法生成的提示都不是连贯的英语。不过，它们确实能生成相关的词元和短语。例如，我们的方法在 SST-2 上加入流畅度约束后，生成了“负面氛围这部不成熟的好莱坞普通话影片——好莱坞这部充满活力的剧本。”[4] 这表明优化过程能够找到与任务相关的单词，但缺乏生成完整句子的能力。

[4]尽管我们使用类别标签词元初始化这些词元，但在考察提示词在优化过程中的变化时，发现所有词元都偏离了初始词元。这表明该过程能够重新学习类别标签。

表*4*。每个数据集所使用的模板和言语化器。

数据集	模板	言语化器
SST-2	<s>它 <mask>	正面，负面
Amazon	<s>那是 <mask>	正面，负面
AGNEWS	<s>那是关于 <mask>	政治、体育、商业、科技

表*5*。在输入文本开头**添加**10 个离散标记进行训练的验证准确率。报告三次学习中的最佳准确率，并给出在 5 次运行中的标准误。

方法	SST-2	AGNEWS	Amazon
AutoPrompt$_{\text{SGD}}$	$87.56_{\pm 0.35}$	$74.36_{\pm 0.47}$	$87.75_{\pm 0.17}$
FluentPrompt	$88.33_{\pm 0.35}$	$74.62_{\pm 0.24}$	$87.42_{\pm 0.18}$
PEZ$_{\text{No 流畅度}}$（我们的方法）	$88.12_{\pm 0.15}$	$77.06_{\pm 0.20}$	$87.70_{\pm 0.21}$
PEZ$_{\text{Fluency}}$（我们的方法）	$88.05_{\pm 0.55}$	$76.94_{\pm 0.48}$	$87.78_{\pm 0.19}$
软提示	$93.35_{\pm 0.01}$	$92.76_{\pm 0.01}$	$94.65_{\pm 0.01}$

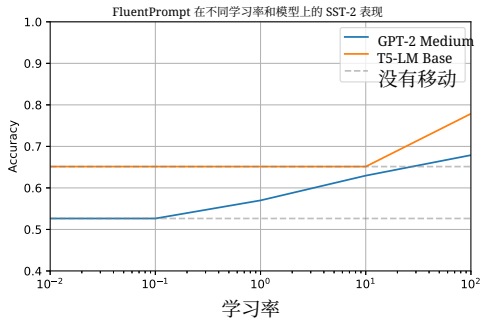

图*8*。结果表明，通过对每一步进行投影，*FluentPrompt*以及进一步扩展的*AutoPrompt*$_{SGD}$，可能会受到一些有趣的、强烈依赖于模型的学习率的影响。

表*6*。展示了将 GPT-2 Large 上学习到的硬提示迁移到 GPT-2 XL 后的验证准确率及标准差。

方法	GPT-2 Large（755M）	GPT-2 XL（1 .3B）
空$_{\text{template}}$	58.34	52.42
AutoPrompt$_{\text{SGD}}$	$74.36_{\pm 0.47}$	$63.79_{\pm 3.61}$
FluentPrompt	$74.62_{\pm 0.24}$	$61.57_{\pm 5.1}$
PEZ$_{\text{No 流畅度}}$（我们的）	$77.06_{\pm 0.20}$	$59.45_{\pm 8.63}$
PEZ$_{\text{Fluency}}$（我们的）	$76.94_{\pm 0.48}$	$67.59_{\pm 2.67}$

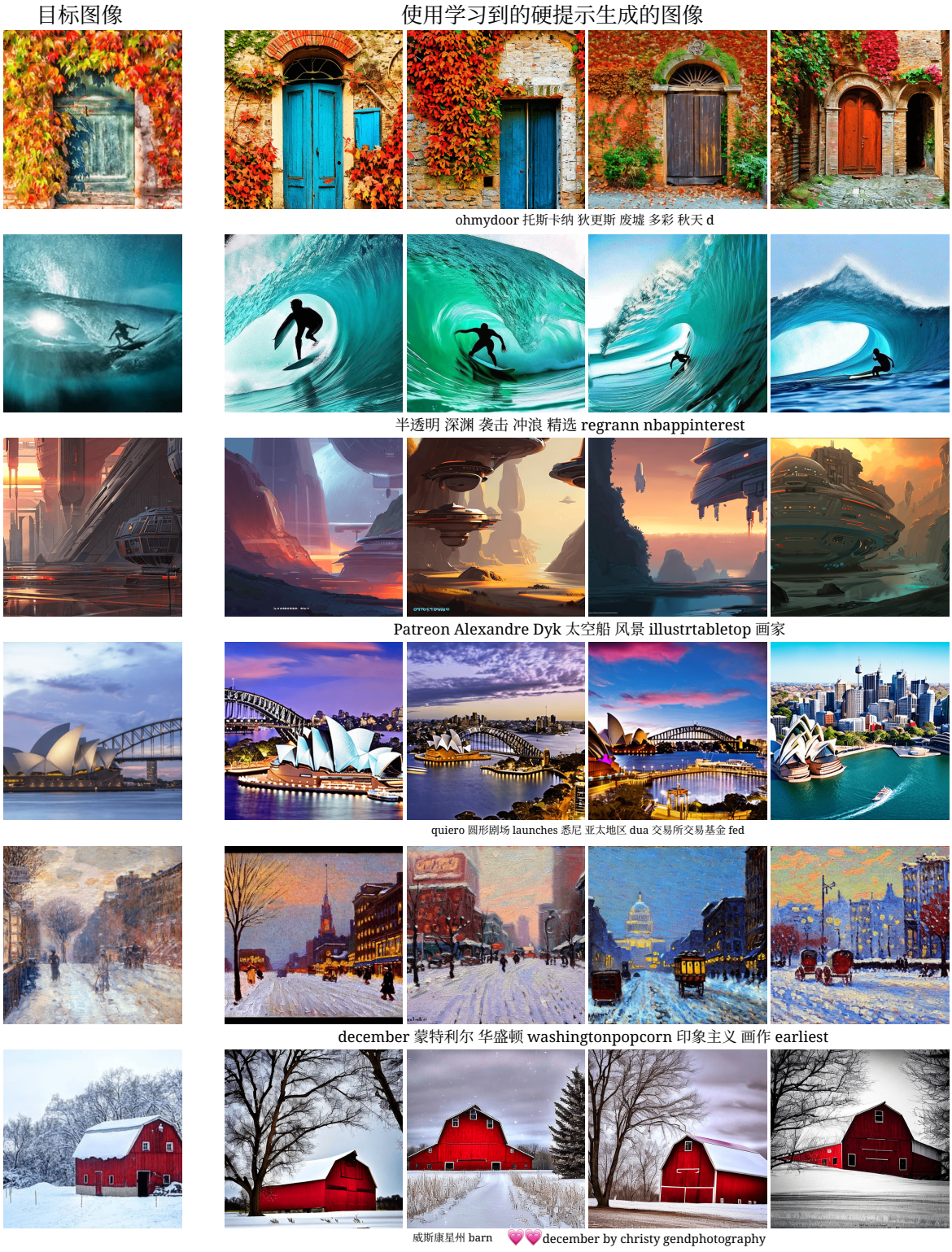

图*9*。 使用学习得到的硬提示获得的更多定性结果。



图*10.* 迭代规避 Midjourney 内容过滤器，并移除敏感词/词元。

表 7。 学习得到的硬提示的定量结果。我们报告原始图像与由硬提示生成的图像之间的 CLIP 分数

	令牌数	要求	LAION	MS COCO	Celeb-A	Lexica.art
$\text{AutoPrompt}_{\text{SGD}}$	8	CLIP	0.689	0.669	0.595	0.702
FluentPrompt	8	CLIP	0.688	0.671	0.583	0.702
PEZ（我们的方法）	8	CLIP	0.697	0.674	0.602	0.711
CLIP Interrogator	~ 77	CLIP + Bank + BLIP	0.707	0.690	0.558	0.762
不使用 BLIP 的 CLIP Interrogator	~ 77	CLIP +Bank	0.677	0.674	0.572	0.737
PEZ（我们的方法） + Bank	8	CLIP +Bank	0.702	0.689	0.629	0.740
CLIP Interrogator	8	CLIP + Bank + BLIP	0.539	0.575	0.360	0.532
CLIP Interrogator	16	CLIP +Bank +BLIP	0.650	0.650	0.491	0.671
CLIP Interrogator	32	CLIP +Bank +BLIP	0.694	0.663	0.540	0.730
软提示	8	CLIP	0.408	0.420	0.451	0.554

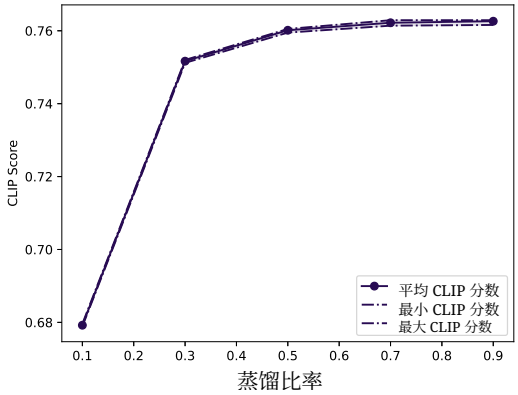

图 *11.* 不同蒸馏比例下提示蒸馏的定量结果。CLIP 分数是通过比较原始提示词生成的图像与蒸馏后提示词生成的图像计算得出的。

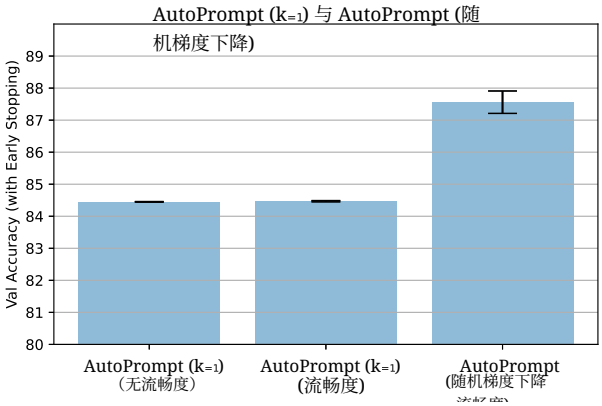

图 *12*。SST-2 验证准确率：比较 AutoPrompt (k=1) 与采用随机梯度下降的 AutoPrompt。从图中可以看出，采用随机梯度下降的 AutoPrompt 优于 AutoPrompt (k=1)，其中 k 是贪心过程评估的候选项数量。