arXiv:2304.00185v1 [cs.CV] 1 Apr 2023

PrefGen：基于偏好的相对属性图像生成

Alec Helbling Christopher J. Rozell Matthew O'Shaughnessy Kion Fallah 佐治亚理工学院 亚特兰大，GA 30332，美国alechelbling@gatech.edu

摘要

深度生成模型能够生成具有人脸等内容的高保真图像。近年来，基于特定定量属性（例如人脸所传达的情绪）有条件地生成图像的研究取得了显著进展。这些方法通常要求用户明确量化视觉属性的期望强度。这种方法的局限在于，许多属性（例如人脸看起来有多"愤怒"）很难由用户进行精确量化。然而，用户能够可靠地判断两张人脸中哪一张看起来"更愤怒"。基于这一前提，我们开发了 *PrefGen* 系统，通过向用户呈现形式为"您更喜欢图像 a 还是图像 b？"的简单成对比较查询，使用户能够控制生成图像的相对属性。利用一系列查询响应中的信息，我们可以估计用户对一组图像属性的偏好，并执行偏好引导的图像编辑和生成。此外，为了使偏好定位切实可行且高效，我们采用了主动查询选择策略。我们在使用 *StyleGAN2* 生成器执行人脸编辑的任务中展示了该方法的有效性。此外，我们还展示了如何将该方法与 *CLIP* 结合，使用户能够编辑由文本提示指定的属性相对强度。 [1]



图1. **相对文本属性的偏好估计**。我们使用 *PrefGen* 估计预言机对"愤怒"和"红发"这两种文本属性的偏好。给定一张起始图像，PrefGen 会提出一系列成对比较，以确定预言机在目标图像中偏好的属性。顶行展示了预言机每回答一次查询后生成的偏好估计结果。第二行和第三行展示了向预言机提出的查询，其中第二行的图像比第三行的图像更受偏好。随着回答的查询越来越多，根据估计偏好生成的图像会逐渐收敛到最后一列中的真实目标。

1. 引言

深度生成模型已展现出生成高质量人脸等内容图像的能力[23]。为了拓展其能力，人们正投入大量精力来提升这些模型的可控性，使最终用户能够修改输出图像的连续属性 [13, 33, 1, 22, 34]。其中一些技术允许用户编辑通过无监督学习发现的一组视觉属性 [13, 33], ，而另一些技术则利用监督学习或预训练分类器来建立从属性到生成模型潜空间的映射 [1, 22, 34]。然而，对于所有这些方法，如果用户希望精确调整某一视觉属性的强度，就必须显式调节一组定量属性。对于修改图像的相对属性 [24] 而言，这种用户界面并不适用，例如调整一个人看起来有多"悲伤"或"疲惫"。不同于"温度"这类属性，人们对于"愤怒" [37], 等属性没有共同的尺度概念，因此很难显式量化一张人脸看起来有多"愤怒"。这限制了属性条件生成模型的适用范围，使其只能用于那些易于量化的属性。

虽然人们可能无法明确描述一张人脸看起来有多"愤怒"，但他们更容易说出两张人脸中哪一张看起来"更愤怒" [17]。受这一观察启发，

[1]代码：https://github.com/helblazer811/PrefGen

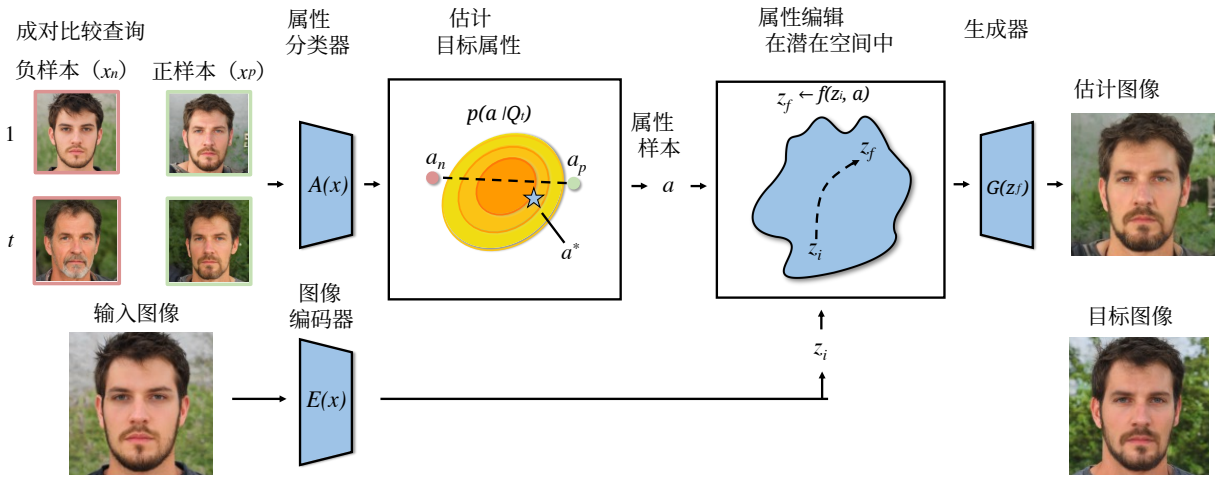


图2. **所提出的偏好引导图像编辑系统**。将图像映射到属性空间 $A(x)$ 后，我们向预言机呈现一系列成对比较查询 $Q_t = \{q_1 \dots q_t\}$。利用对成对查询的响应，计算属性空间上的后验分布 $p(a \mid Q_t)$。我们从该后验分布中采样，以估计用户的理想点 a^*。我们将理想点估计值与输入图像的潜向量 $E(x_i) = z_i$ 结合，生成生成模型潜在空间中的最终向量 $z_f \leftarrow f(z_i, a)$。最后，使用生成器将修改后的潜向量 z_f 映射到图像空间 $G(z_f)$。

我们开发了*PrefGen* 系统，利用以下形式的成对比较查询所提供的信息来控制 [24] 图像的相对属性：“您更喜欢图像 x_a还是图像 x_b？”通过向用户呈现一系列此类查询，我们可以推断出他们对一组相对属性的偏好，而无需对其进行明确量化。我们利用一系列查询所提供的信息 Q 来推断用户在相对属性空间中的偏好分布 $\mathcal{A}$。对于用户偏好的属性，我们采用贝叶斯框架，从而能够主动选择查询 [32, 5, 14]并量化不确定性。这样，我们既能高效搜索属性空间，也能分别应对用户回答中的不一致性。

长期以来，成对比较一直被视为一种丰富的信息形式，可用于推断用户对一组项目的偏好 [17]，也可用于推理难以量化的心理量表 [37]。此外，还可以从搜索引擎鼠标点击数据等广泛可用的来源被动提取成对比较查询 [27]。这使得所提出的框架有望利用来自用户的隐式反馈来控制生成模型的输出 [13, 33, 1, 22, 34]，而无需用户明确量化自己的偏好。

在我们的实验中，我们展示了本方法在 StyleGAN 2 [16] 架构上的有效性。StyleGAN 是一种广泛使用的生成对抗网络（GAN） [11] 架构。不过，原则上，本方法具有足够的通用性，可与任何能够根据不同定量属性进行条件生成的生成模型结合。我们将此方法应用于人脸属性编辑任务。我们展示了 PrefGen 可以轻松结合现有的属性条件图像生成方法使用。具体而言，我们训练从属性（如面部年龄和姿态）到潜向量的监督映射，正如 [34], 中所做的那样，并通过成对比较推断用户对这些属性的偏好。近期研究表明，可以利用文本信息修改图像的视觉属性 [25]。在此基础上，我们展示了如何借助 CLIP [26] 等联合视觉-语言模型，利用成对比较的信息编辑视觉属性的相对强度。

本文的贡献如下：

- PrefGen 系统通过允许用户回答一系列成对查询来编辑生成图像的相对属性，从而为预训练生成模型增添可控性。

- 通过两种构建属性编码器的方法进行实证验证，并扩展了先前的主动查询选择方法，以利用连续的生成属性空间。

2. 背景与相关工作

相对属性 许多属性本身具有一定的排序关系，但人们很难明确地量化它们。例如，人们可以根据服装看起来有多“时髦”来排序，也能理解一顿晚餐有多“精致”，但要始终如一地量化这些属性，很可能会感到困难。这些属性称为相对属性 [24]，已有大量研究致力于对其进行建模

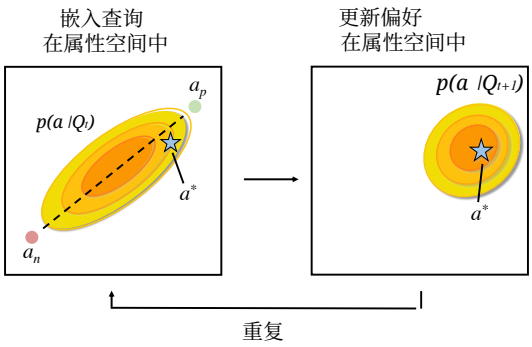


图3。**成对查询后的偏好估计更新**。向用户呈现成对查询 q 使我们能够估计用户偏好 $p(a \mid q)$ 的分布。在提出查询 q后，我们可以将该信息与先前查询的集合 Q_t结合起来，使得 $Q_{t+1} = Q_t \cup q$，并缩小属性空间中可能与用户偏好相一致的区域。

。许多研究侧重于学习对视觉属性的相对强度进行排序 [19, 41, 35]。另一个相关研究方向则侧重于学习相似度度量或 相似度嵌入，使嵌入空间中各项之间的距离对应于相似性的概念 [36, 2, 30]。WhittleSearch 的作者 [18]与我们的研究动机相似：他们希望通过相对属性反馈（成对比较查询）在有限的图像集合中进行搜索。然而，我们的方法允许借助生成模型，从属性的连续空间中采样，而不局限于有限的图像集合。

可控生成模型 生成建模中一个困难且尚未解决的问题，是如何控制生成内容的属性。许多生成模型能够生成高保真示例[11, 29], , 但需要额外的方法来进行控制。大量研究致力于在不修改底层生成对抗网络权重的情况下，使用生成对抗网络[22, 1, 33, 25], 对具有各种连续属性的图像进行条件采样。一些方法采用基于不同底层架构的监督映射，例如归一化流 [1]和基于能量的模型 [22]。另一些方法则尝试推断 StyleGAN 潜在空间中的可解释方向 [33]。 [34]训练从连续属性映射到 StyleGAN 潜在空间的神经网络映射，并微调其表征，使其实现解耦（即不同的潜在子空间对应不同的可解释属性）。所有这些方法都要求用户明确量化连续属性。然而，许多属性具有相对性，用户难以明确量化。受此问题启发，我们开发了一种方法，让用户通过回答成对比较查询来控制生成图像的属性。在实验中，我们展示了如何轻松地将我们的方法应用于现有条件生成方法之上，例如 [34]利用成对查询中的信息，有条件地生成具有不同连续属性的图像。

文本驱动的图像生成近年来，使用文本引导生成模型输出的研究取得了实质性进展 [21, 28]。在 StyleGAN 相关文献中，大量工作致力于寻找 StyleGAN 潜在空间中对应不同文本属性 [25, 1]的方向。尽管用于生成具有不同属性图像的有监督方法展现出令人印象深刻的能力，但它们需要使用大量标注数据或预训练分类器。这将用户的编辑能力限制在一组狭窄的预定义属性上。在 StyleCLIP [25], 中，作者展示了可以编辑 StyleGAN 生成图像，使其符合特定的文本提示，例如“愤怒的人”或“可爱的猫”。他们使用 CLIP [26] 联合文本-图像嵌入架构，推断 StyleGAN 潜在空间中对应不同文本属性的方向。语言提供了一套固定的形容词和修饰语，用于描述视觉属性的强度。这使得仅凭文本精确控制属性的相对强度变得具有挑战性。在我们的实验中，我们采用类似于 StyleCLIP 的方法推导 CLIP 空间中的方向，并证明可以使用 *PrefGen* 修改相对文本属性的强度。

通过成对比较估计偏好形式为“您更喜欢图像 x_a 还是图像 x_b” 的成对比较查询，为建模用户对相对属性的偏好 [10, 3]提供了一种精妙的方法。它们使我们能够对那些难以明确描述或量化的直观概念进行建模 [9, 37]。此外，还可以从搜索引擎点击等广泛可用的数据中被动收集成对比较查询 [27, 4]。这意味着我们可以从丰富可用的数据源中合成用户偏好信息，并利用生成模型生成符合用户偏好的内容。

我们的工作密切借鉴了利用成对比较中的信息来高效搜索现有相似度嵌入的方法 [5, 14, 8, 40]。该领域一个常见的简化假设是：用一个称为理想点的单一点来表示用户偏好的特征或属性集合 $a^* \in \mathcal{A}$ [7]。我们假设，对于嵌入空间中的两个点，用户会偏好距离 a^* 更近的那个点。我们的目标是利用成对查询中的信息来估计这个理想点。可能的成对查询数量随图像数量 N 按 $O(N^2)$ 的数量级增长，因此向用户展示所有可能的查询通常既不可行又存在冗余。 [14] 的作者提出了一种方法，通过主动选择成对查询来更高效地对固定的一组项目进行排序。

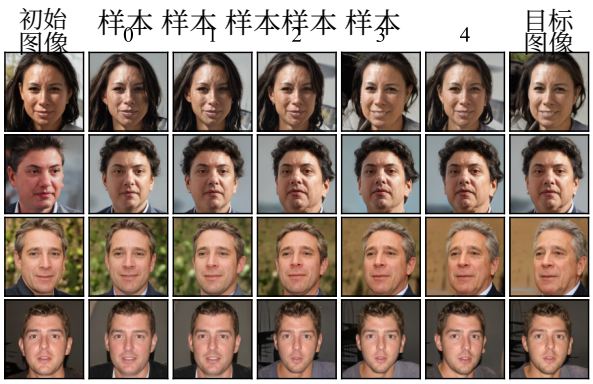


图 4. **随收集的查询而编辑的图像**。经过 N 次查询后生成的图像，以匹配预言机的偏好。随着时间推移，生成的样本会从输入图像逐渐变得更像最后一列中的目标图像。

其他方法以这一方法论为基础，尝试在连续嵌入空间中搜索理想点[8, 40, 5]。

[6] 提出一种概率方法来估计偏好，并开发基于信息论主动选择查询的方法。与我们最密切相关的是 [5]，的工作，他们设计了指标，以便从固定集合中主动选择查询，并在选择时考虑响应中的噪声。我们将他们的方法应用于估计用户对相对视觉属性空间的偏好，并进一步将其扩展为从连续空间而非固定集合中选择最优查询。据我们所知，我们的工作首次将成对比较搜索与主动查询结合起来，以控制生成模型的属性。

与我们的方法最为相似的可能是以下工作：[12]，作者试图利用一组成对比较约束来控制生成对抗网络（GAN）。然而，他们借助长短期记忆网络（LSTM），直接预测一个潜向量，以生成满足成对约束的图像。我们通过实验表明，与解析的贝叶斯框架相比，这种方法的性能较差 [6, 5]，这可能是因为从训练查询推广到特定用户查询响应这一问题本身存在欠定性。我们在第4节将此方法用作基线。

3. 方法

我们的目标是让预言机通过提出简单的成对比较查询来控制生成图像的相对属性，查询形式为“你更喜欢图像 x_a 还是图像 x_b？”利用一系列此类查询中的信息 $Q_t = \{q_1, \dots, q_t\}$，我们希望估计预言机对一组相对属性的偏好 $\mathcal{A}$。我们可以

算法1基于成对比较的图像生成**输入：** 输入图像 x_i将输入图像编码到潜在空间 $z_i = E(x_i)$ 初始化查询集 $Q_0 = \{\}$**从** $t = 1$**到** T **执行**主动采样一个查询 $q = S(Q_t)$ 提示预言机对 q作出响应将已回答的查询添加到 $Q_t = Q_{t-1} \cup q$**循环结束**使用来自马尔可夫链蒙特卡罗的样本计算 $\hat{a}^* = \mathbb{E}[a \mid Q_T]$。使用我们的估计 $z_f = f(z_i,\ \hat{a}^*)$ 修改 z_i生成最终编辑图像 $x_f = G(z_f)$ **输出：** x_f

向预言机$O(q_i)$提交一个尚未回答的查询 $q_i = (x_a, x_b)$，预言机返回一个已回答的查询 $O(q) = (x_p, x_n)$。因此，

$$q_i = (x_p, x_n) \iff x_p \prec x_n, \tag{1}$$

其中，$x_p \prec x_n$ 表示相较于图像 x_n，预言机更“偏好”图像 x_p。我们假设存在某个向量$a^* \in \mathcal{A}$，称为目标属性，它最能代表预言机偏好。我们的目标是根据查询 Q_t中的信息推断 a^*。

一旦得到预言机理想点的估计值 $\hat{a}^*$，我们就需要一种方法来生成符合这些偏好的图像。为此，我们将属性映射到生成模型潜空间 $\mathcal{Z}$。我们用属性映射函数来表示这一过程，该函数将潜向量 z_i 和属性 a 映射为另一个潜向量 $z_f \leftarrow f(z_i, a)$。我们既可以修改初始潜向量 z_i（该向量或许通过GAN反演 $z_i = E(x_i)$ [38], 得到），也可以仅利用属性 $z_i \sim p(z)$中的信息来生成图像。我们的目标是对 $f(z_i, a)$进行建模，使得 $A(G(z_f)) \approx a$，且 $G(z_f)$ 与 $G(z_i)$的身份相同，其中 $A(x)$是预测图像 $x \in \mathcal{X}$属性的函数。我们将在第3.3节探讨对 $A(\cdot)$和 $f(\cdot)$进行建模的两种主要方法。

3.1. 用户偏好的贝叶斯模型

我们使用贝叶斯框架，根据查询 Q_t 中的信息，估计属性空间 $\mathcal{A}$ 上用户偏好的分布 $p(a \mid Q_t)$。我们采用一个称为理想点模型 [7], 的简化假设：用户对一组属性 $a \in \mathcal{A}$ 的偏好与该向量 a 和目标属性 a^* 之间的距离相对应。设 $A(x)$ 是一个从图像空间 $\mathcal{X}$ 到属性 $\mathcal{A}$ 的映射（可能未知）。当正图像 x_p 的属性 $a_p = A(x_p)$ 比负图像 x_n 的属性 $a_n = A(x_n)$ 更接近目标属性 a^* 时，我们称用户“更偏好”正图像 x_p 而非负图像 x_n：

$$x_p \prec x_n \iff \|a^* - a_p\| < \|a^* - a_n\|. \tag{2}$$

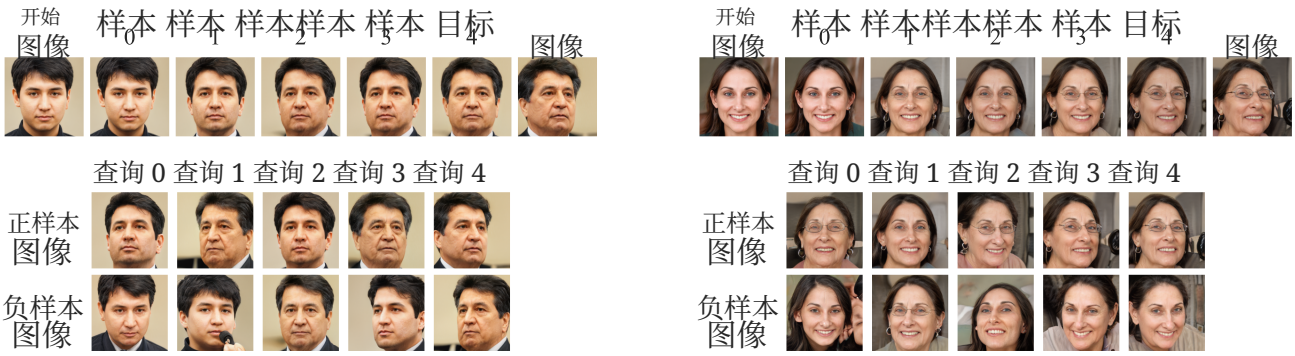


图 5. **PrefGen 的定性试验运行。** 我们使用监督属性映射，在由人脸的偏航角、俯仰角和年龄构成的四维空间中展示 PrefGen 的两次试验运行。图示详情请参见 图 1 。我们展示了提交给预言机的每个查询，以及每次查询后从后验分布中采样生成的图像。随着查询响应不断收集，样本图像在定性上逐渐收敛至目标图像。

遵循公式 2中的假设以及[5] 的工作，我们将 $x_p \prec x_n$ 的似然表示为

$$p(q \mid a) = \sigma\big(k_q(\|a - a_n\|^2 - \|a - a_p\|^2)\big), \qquad (3)$$

其中， σ 是逻辑函数， k_q 是一个标量值，用于衡量预言机响应中的噪声程度。直观而言，当用户更偏好属性为 a_p 的图像，而非属性为 a_n的图像，且前者比后者更接近理想点 a^* 时，该似然会赋予很高的概率。噪声常数 k_q 可以视为信噪比，用于确定每个查询所包含的信息量。预言机可以通过在不同查询中调整 k_q ，来适应不同查询中噪声量预期不同的情况 [5]。

通过以概率方式表示这个问题，我们可以纳入各种误差来源，例如用户对成对查询的回答中的“噪声”。我们可以通过递归应用贝叶斯规则，整合多个成对查询所提供的信息，

$$p(a \mid q_1, ..., q_t) = \frac{p(a)\prod_{j=1}^{t} p(q_j \mid a)}{p(q_t \mid q_1, ..., q_{t-1})}, \qquad (4)$$

其中，我们对用户偏好的先验分布 $p(a)$ 是均匀分布。最后，我们使用马尔可夫链蒙特卡罗从后验分布 $p($ $a \mid Q_t)$ 中进行样本采样，其中 $Q_t = \{q_1, \ldots, q_t\}$并取其期望值 $\mathbb{E}[a \mid Q_t]$ ，以形成对目标属性 a^*的估计。直观地说，这一过程可以看作二分搜索算法的多维推广。该过程如算法 1和图 3所示。

3.2. 用于高效偏好估计的主动查询选择

通常无法向用户展示所有可能的成对比较查询。给定一组固定的 N 图像，可以据此构造 $O(N^2)$ 个成对查询。

此外，在我们的设定中，我们拥有一个能够从属性空间连续采样的生成模型，因此可供选择的查询有无限多个。问题依然存在：我们该如何选择一组查询，以便最好地估计用户的偏好？一种简单的解决方案，我们称之为随机查询，就是从众多可能的查询中随机选取一个较小的子集。然而，鉴于此前向用户展示的查询所提供的信息，许多查询都是冗余的。

我们基于主动学习的方法 [32] 来选择能够最小化后验分布 $p(a \mid Q_t)$中不确定性的查询。 [5] 的作者提出了相应策略，用于从候选查询的离散集合中选择关于理想点 a^* 尽可能包含新信息的查询。其中一种策略称为均值切分最大方差（MCMV）。直观地说，MCMV的目标是选择一个查询 $q = (a_p, a_n)$，使其沿着由后验分布 $p(a \mid Q_t)$确定的高方差轴，并且将后验分布的均值 $\mathbb{E}[a \mid Q_t]$一分为二。我们可以用法向量 $v = 2(a_p - a_n)$和标量偏移量 $\tau = |a_p|^2 - |a_n|^2$来表示平面，该平面将属性 a_p 和a_n 一分为二。我们可以计算查询方向上的后验分布方差，查询方向由 $\sigma_q = v\Sigma_a v^T$ 计算得到，其中 Σ_a 是后验分布中样本的协方差。切分均值意味着选择一个查询 $q = (a_p, a_n)$，使得由 v 和 τ 参数化的两个查询项之间的平面与分布均值相交。这可以计算为$\mu_q = (v^T\mathbb{E}[a \mid Q_t] - \tau)/|v|$。因此，最佳查询是使 $k_q\sigma_q - \lambda\mu_q$最小的查询，其中 k_q 是前述噪声常数， λ是一个超参数，用于平衡均值切分和最大方差标准的相对重要性。

在实践中，我们需要从属性空间中采样许多潜在查询，测量每个查询的 MCMV效用，并选择值最高的查询。随着我们

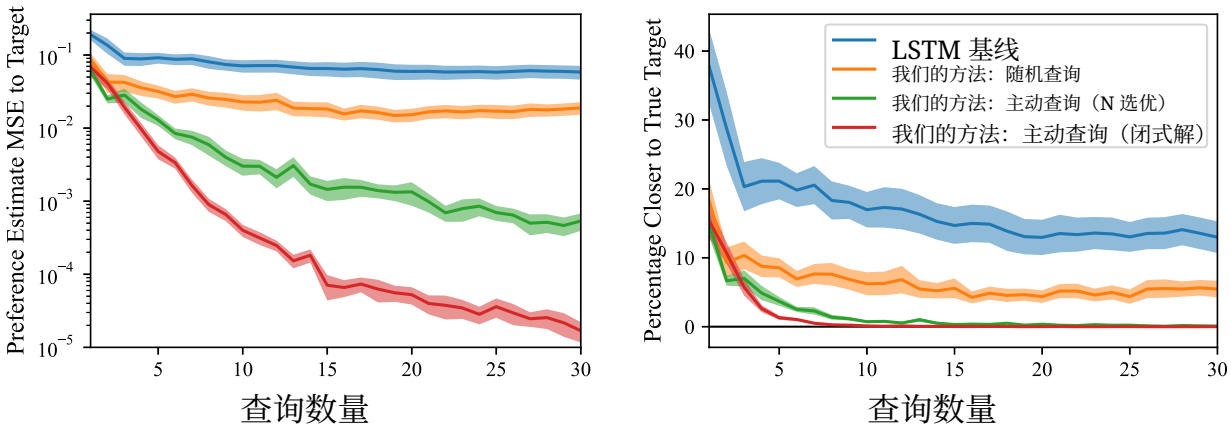


图 6. **比较从成对比较中估计用户偏好的各种方法的定量性能**。我们模拟预言机对一系列查询的响应，共进行 20 次试验，每次试验包含 30 次查询。我们在由人脸年龄和偏航角（左右姿态）构成的二维空间中进行搜索。（左）给定试验中，经过 t 次查询后偏好估计与真实属性 a^* 在属性空间中的均方误差。（右）在 1000 个属性向量中，相较于估计值、更接近目标属性的向量所占百分比。根据这两项指标，PrefGen 都能高效且精准地估计偏好。正如预期，主动查询相比随机查询提升了偏好估计的效率；而利用连续生成模型潜在空间的闭式方法，显著优于*N*选优 方法。

逐渐收敛到对预言机偏好的估计，性能便开始下降，因为相对于后验分布的体积而言，最佳查询离均值和最大方差轴越来越远（见图 6）。我们的生成模型框架使我们能够从属性空间中连续采样（见第 3.3 节）。因此，我们可以直接选择一个沿后验分布最大方差轴、且以均值为中心进行划分的查询，无需大量采样候选查询。我们只需选择一个查询，其中的点与均值等距，并位于后验分布的最大方差轴上 $p(a \mid Q_t)$。值得注意的是，这种方法未考虑可能影响不同查询预期信息量的噪声模型。我们将这种方法称为 主动闭式方法，将原始 MCMV 方法称为主动 *N*选优。

3.3. 控制生成图像的属性

本文探讨了对属性映射函数 $f(z_i,\ a)$进行建模的两种主要方法。需要注意的是，这两种方法只是将属性向量映射到潜向量（或直接映射到图像）的众多可能方式中的两种。

监督属性映射：从属性到潜在空间构建监督映射对于我们的第一种方法，我们训练从面部姿态和年龄等属性到生成器潜在空间的监督映射。为了实现这一方法，我们基于 GANControl [34], 展开工作，原作者训练了从属性到解耦 StyleGAN 的监督映射

潜在空间。这项工作表明，可以修改定量属性，而不会显著改变图像的底层身份。

CLIP 属性映射：使用 StyleCLIP 修改相对文本属性 在 StyleCLIP [25], 中，作者提出了一种方法，可以找到一个潜向量，使其生成具有特定属性的图像，同时不改变给定输入潜向量的身份。他们介绍了一种基于梯度的方法，用于寻找能够生成具有特定文本属性图像的潜向量。在 StyleCLIP 的基础上，我们对模型的相对属性进行编码，从而能够在两个文本提示之间进行插值：一个中性文本提示，例如“一个表情中性的人”，以及一个目标文本提示 “ 一个表情愤怒的人 ”。我们可以使用 CLIP 文本编码器 $\mathcal{T}(t)$ 将这些提示嵌入 CLIP [26]空间，分别得到中性向量和目标向量 t_n 与 t_t。余弦相似度接近 t_n 的图像会呈现中性表情，而更接近 t_t 的图像则会更符合目标提示。令 $\mathcal{V}(x)$ 表示图像 $x \in \mathcal{X}$的 CLIP 向量。我们可以通过计算 $\mathcal{V}(x)$ 与属性方向 $t_a = t_t - t_n$ 之间的余弦相似度，来预测属性强度（即 $\frac{\langle \mathcal{V}\ x\ \ t_a \rangle}{\|\mathcal{V}\ x\ \|\|t_a\|}$）。随后，我们只需选择某个 $\alpha \in [0, 1]$，并最小化当前预测的属性强度与 α之间的距离，就能修改图像的属性。我们采用梯度下降来实现这一点，正如 StyleCLIP 中所做的那样，并添加惩罚项以确保图像的身份保持不变。这种方法能够对相对属性的强度进行定量建模，而直接对文本提示进行编码（如 [28]中所做的那样）则难以实现这一点。

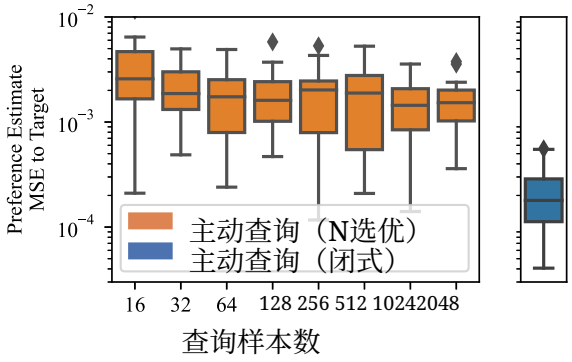


图7. **改进主动选择**。（左）使用*N*选优时，随着候选查询数量增加，30个查询响应下的偏好估计性能。（右）即使候选查询数量增加，*N*选优策略的属性估计误差仍高于所提出的闭式策略。

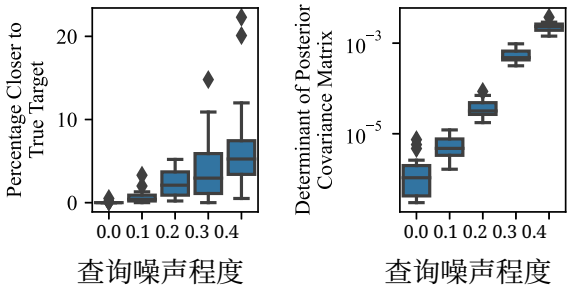


图8。**我们的方法对查询响应噪声的鲁棒性**。加性高斯噪声下的偏好估计。（左）随着噪声增加，性能逐渐下降，但仍保持在估计百分位的前列。（右）随着噪声增加，以属性后验协方差的行列式衡量的不确定性也随之增加。

4. 实验

我们的实验重点是评估所提出的方法能否有效地根据成对比较查询，估计用户在属性空间中的偏好，并有条件地生成符合这些偏好的图像。我们采用的一种核心评估方法是使用一个简单的预言机来模拟用户对查询的回答。对于每次偏好估计试验，我们从 $a^* \in \mathcal{A}$ 中均匀采样 $[0,1]^d$一个任意的目标属性向量 d，该向量对应于用户希望样本图像具备的图像属性。随后，我们依次提出若干成对比较查询，每个查询的形式为 $q = (a_p,\ a_n)$，且满足公式2。

4.1. 使用监督属性映射的PrefGen

实现细节一种表示 f 的方法是通过监督学习训练一个神经网络，将属性映射到生成模型的潜在空间。我们采用GANControl中的方法 [34] 来构建监督映射，将人脸的姿态（偏航角、俯仰角和翻滚角）以及表观年龄映射到StyleGAN的潜在空间2 [15]，使用FFHQ人类属性数据集 [16]。在定量实验中，对于每种方法，我们进行20次偏好估计试验，每次包含30个查询，并且 $k_q = 10$。

PrefGen 的定量评估为衡量我们能否精准且高效地估计用户的目标属性 a^*，我们采用三个关键指标：（1）每次查询响应后，当前估计值与属性空间中真实属性之间的均方误差（MSE）；（2）比我们的估计值更接近目标属性的属性所占百分比 a^*；以及（3）估计值比负样本更接近正样本的成对比较约束所占百分比，其中目标属性 a^* 比负样本更接近正样本。图6 显示，我们的方法（主动采样和随机采样）都能收敛到与目标属性均方误差较低的偏好估计，并且其结果比很大比例的随机采样属性向量更接近目标属性。表4.1 进一步证实了这些结果，并显示*PrefGen*能够估计出一组满足很大比例约束的属性。附录C中的更多结果表明，在高维属性空间中进行搜索时也会呈现这一规律。

与基线方法的比较 我们实现了一种基线方法 [12], , 称之为 *LSTM* 基线。本文中，作者使用长短期记忆网络来估计一个潜向量，该向量可生成满足一系列成对比较约束的图像。在我们的方法中，我们将偏好估计与映射到生成模型潜空间分开。为便于比较，我们扩展了 LSTM 方法，使其能够估计满足给定约束集的低维属性向量。该基线方法的实现细节见附录 B。如图 6 和表 4.1所示，无论采用随机查询还是主动查询，我们的方法都能以更高的精度和效率估计目标属性。

PrefGen 的定性性能图4 展示了，随着向预言机提出更多查询，为给定偏好估计采样的图像如何发生变化。每次查询后，我们都会采样一个潜向量 $z_f = f(z_i, \mathbb{E}[a \mid Q_t])$，其属性对应于预期的用户偏好 $\mathbb{E}[a \mid Q_t]$，并生成相应的图像$G(z_f)$。从定性角度看，偏好样本的属性会逐渐收敛到目标图像的属性。图5 展示了我们*PrefGen* 过程的两次试验运行。从定性角度看，可以清楚地看到偏好估计正在收敛

	与真实属性的均方误差↓	更接近的百分比（%）↓	满足约束的百分比（%）↑
LSTM	5.86e-2 ± 5.60e-2	13.330 ± 10.390	83.83 ± 10.290
随机	1.89e-1 ± 1.64e-2	5.731 ± 5.209	97.17 ± 2.420
主动式（N选优）	5.37e-4 ± 6.41e-4	0.145 ± 0.139	98.50 ± 2.467
主动式（闭式）	**1.69e-5** ± **2.35e-5**	**0.003** ± **0.001**	**99.83** ± **0.007**

表 1。**不同方法下偏好估计的性能**。我们在 20 次实验试验中进行 30 次查询后，对不同的偏好估计方法进行定量评估。与真实值的均方误差衡量属性空间中最终偏好估计与真实值之间的均方误差。更接近的百分比（%）表示属性空间中比偏好估计更接近真实值的属性向量所占的百分比。满足约束的百分比（%）表示最终偏好估计满足所收集查询的百分比。对于所有指标，我们的方法（随机、主动式（*N*选优）和主动式（闭式））均优于 长短期记忆网络基线，并且主动式（闭式）在每项指标上都优于其他所有方法。

到目标属性，利用的是每次成对比较查询所提供的信息。

对查询响应噪声的鲁棒性在大多数实际应用场景中，各种噪声源可能导致预言机对查询的响应出现错误或不一致。因此，构建一个对这类噪声具有鲁棒性的系统是十分理想的。为评估 *PrefGen*对查询噪声的鲁棒性，我们有意向查询似然中加入高斯噪声，具体如下：$\sigma(k_q($ $||a-a_n||^2-||a-a_p||^2+n))$ 其中 $n\sim\mathcal{N}(0,\ \sigma_n)$，且 σ_n 表示我们加入的噪声程度。图 8 展示了在 20 次试验中进行 30 次查询后，从大型样本（1000 个）中比目标属性更接近目标属性的属性向量所占的百分比。对于每种噪声程度，我们都使用一组训练三元组上的最大似然估计来调节 k_q（详见附录 B）。随着噪声程度增加，我们注意到偏好估计的性能会下降，但更重要的是，它不会发生灾难性失效。此外，我们可以通过计算从$p(a\mid Q_t)$中采样所得样本的协方差矩阵的行列式，来量化后验分布中的不确定性。结果表明，随着噪声量增加，后验分布的不确定性也会增加。

闭式主动查询的有效性章节3.2 讨论了部署一种方法的重要性：该方法能够主动选择既能揭示预言机偏好、又能以较少响应有效进行估计的查询。图6清楚地表明，主动选择查询不仅大幅提高了偏好估计流程的效率，也提升了最终精度。此外，显然，所提出的闭式主动查询方法优于*N*选优方法。图7 显示，闭式随着候选样本数量的增加，仍持续优于*N*选优方法。

4.2. 具有文本属性的PrefGen

我们展示了如何使用 *PrefGen*编辑文本提示所描述的相对属性。我们在 CLIP 空间中找到了两个方向，分别对应属性“愤怒”和“红发”。对于“愤怒”t_n = ，我们使用“一个表情平静的人”和 t_p =“一个表情愤怒的人”；对于“红发”，我们使用 t_n = “一个深色头发的人”和 t_p =“一个红发的人”。图1 展示了*PrefGen* 在由这些属性构成的二维空间中的表现。我们不仅希望编辑 GAN 潜在空间中的向量，也希望能够编辑真实世界的图像。为此，我们采用 GAN反演技术，具体而言，使用 Encoder4Editing [38], 将图像嵌入 Style GAN2 模型的潜在空间，并编辑这些向量。

5. 讨论

在这项工作中，我们提出了*PrefGen*，一种利用成对比较信息控制生成模型的框架。我们通过在人脸编辑任务中使用 StyleGAN2，展示了这一方法的有效性。我们的方法既适用于有监督的属性映射，也适用于基于CLIP的文本属性映射。

我们实验方法的一个潜在局限在于，在实验中我们使用模拟预言机来代替用户本人对查询作出响应。然而，我们认为这是一个合理的假设，因为我们专门构造了与人类可解释特征相对应的属性嵌入，并且已有大量研究探讨了在这种情境下利用成对比较来建模用户偏好的有效性[9, 18]。另一个潜在局限是，我们使用 CLIP 空间中的线性方向来建模相对属性；然而，无法保证这些线性方向能够编码文本特征的相对强度。未来工作的一个潜在方向是更好地利用 CLIP 等联合文本图像模型来建模相对属性。

尽管生成模型具有潜在益处，但它们也会引发各种伦理问题；当应用于人物图像时，这些问题尤其突出。潜在的有害用途包括制造虚假信息 [31] 或“深度伪造”[20]。我们敦促应用这项技术的开发者，谨慎权衡其潜在益处与危害，并在适当情况下采取策略来减轻危害（例如， [39]）。

6. 致谢

我们衷心感谢佐治亚理工学院 SIP 实验室成员慷慨提供的支持与指导。本研究部分获得了佐治亚理工学院校长本科生研究奖（PURA）、美国国家科学基金会 CAREER 奖（CCF-1350954）、美国国家科学基金会本科生研究体验（REU）项目以及美国海军研究办公室资助的支持。我们还要感谢 Asa Harbin 协助撰写稿件。

参考文献

[1] R. Abdal, P. Zhu, N. J. Mitra 和 P. Wonka。StyleFlow：使用条件连续归一化流探索属性条件化的StyleGAN 生成图像。《ACM图形学汇刊》，40(3):1–21，2021年5月。1, 2, 3[2] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman 和 S. Belongie。广义非度量多维尺度分析。载于《人工智能与统计》，第11–18页。PMLR，2007年。3[3] M. Aggarwal 和 A. Fallah Tehrani。利用偏好学习对人类决策行为建模。《INFORMS计算杂志》，31(2):318–334，2019年。3[4] R. Atterer、M. Wnuk 和 A. Schmidt。了解用户的一举一动：用于网站可用性评估和隐式交互的用户活动跟踪。载于第15届万维网国际会议论文集，WWW '06，第203–212页，美国纽约州纽约，2006年。美国计算机协会。3[5] G. Canal、A. Massimino、M. Davenport 和 C. Rozell。通过含噪成对比较主动搜索嵌入。载于国际机器学习会议（ICML），美国加利福尼亚州长滩，2019年6月。2, 3, 4, 5[6] W. Chu 和 Z. Ghahramani。利用高斯过程进行偏好学习。载于第22届国际机器学习会议论文集，ICML '05，第137–144页，美国纽约州纽约，2005年。美国计算机协会。4[7] C. H. Coombs。无需度量单位的心理测量尺度。《心理学评论》，57(3):145–158，1950年。3, 4[8] M. A. Davenport。迷失方向：非度量三角测量与地标多维尺度分析。载于2013年第5届IEEE多传感器自适应处理计算进展国际研讨会（CAMSAP），第13–16页。IEEE，2013年。3, 4[9] H. A. David。《成对比较法》，第12卷。伦敦，1963年。3, 8[10] J. F¨urnkranz 和 E. H¨ullermeier。偏好学习与成对比较排序。载于《偏好学习》，第65–82页。施普林格，2010年。3[11] I. J. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio。生成对抗网络，2014年。2, 3[12] E. Heim。用于交互式图像生成的受约束生成对抗网络，2019年。4, 7, 12, 13[13] E. H¨ark¨onen、A. Hertzmann、J. Lehtinen 和 S. Paris。GANSpace：发现可解释的GAN控制。2020年。1, 2[14] K. G. Jamieson 和 R. D. Nowak。使用成对比较进行主动排序。载于《神经信息处理系统进展》，2011年。2, 3[15] T. Karras、S. Laine 和 T. Aila。生成对抗网络的基于风格的生成器架构，2018年。7[16] T. Karras、S. Laine 和 T. Aila。生成对抗网络的基于风格的生成器架构。载于录于IEEE/CVF计算机视觉与模式识别会议论文集，第4401–4410页，2019年。2, 7, 12[17] M. G. Kendall和B. B. Smith。论成对比较法。Biometrika，31(3/4):324–345，1940年。1, 2[18] A. Kovashka、D. Parikh和K. Grauman。WhittleSearch：基于相对属性反馈的图像搜索。载于IEEE计算机视觉与模式识别会议（CVPR）论文集，2012年6月。3, 8[19] T.-Y. Liu等。面向信息检索的学习排序。《信息检索基础与趋势》，3(3):225–331，2009年。3[20] M. Masood、M. Nawaz、K. M. Malik、A. Javed和A. Irtaza。深度伪造的生成与检测：最新进展、开放性挑战、对策及未来方向，2021年。8[21] A. Nichol、P. Dhariwal、A. Ramesh、P. Shyam、P. Mishkin、B. McGrew、I. Sutskever和M. Chen。Glide：基于文本引导扩散模型的逼真图像生成与编辑，2021年。3[22] W. Nie、A. Vahdat和A. Anandkumar。利用潜在空间能量模型实现可控和组合式生成。载于NeurIPS，2021年。1, 2, 3[23] A. Oussidi和A. Elhassouny。深度生成模型：综述。载于2018年国际智能系统与计算机视觉会议（ISCV），第1–8页，2018年。1[24] D. Parikh和K. Grauman。相对属性。载于2011年国际计算机视觉会议论文集，ICCV '11，第503–510页，美国，2011年。IEEE计算机学会。1, 2[25] O. Patashnik、Z. Wu、E. Shechtman、D. Cohen-Or和D. Lischinski。StyleCLIP：对StyleGAN图像进行文本驱动操控，2021年。2, 3, 6[26] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger和I. Sutskever。通过自然语言监督学习可迁移视觉模型，2021年。2, 3, 6[27] F. Radlinski和T. Joachims。通过点击数据主动探索学习排序。载于第13届ACM SIGKDD知识发现与数据挖掘国际会议论文集，KDD '07，第570–579页，纽约，美国，2007年。计算机协会。2, 3[28] A. Ramesh、M. Pavlov、G. Goh、S. Gray、C. Voss、A. Radford、M. Chen和I. Sutskever。零样本文本到图像生成，2021年。3, 6[29] R. Rombach、A. Blattmann、D. Lorenz、P. Esser和B. Ommer。利用潜在扩散模型进行高分辨率图像合成，2021年。3[30] F. Schroff、D. Kalenichenko和J. Philbin。FaceNet：用于人脸识别和聚类的统一嵌入。2015年IEEE计算机视觉与模式识别会议（CVPR），2015年6月。3[31] K. Sedova。人工智能与虚假信息宣传活动的未来：第二部分，威胁模型。安全与新兴技术中心，2021年。8[32] B. Settles。主动学习。《人工智能与机器学习综合讲座》，6(1):1–114，2012年。2, 5

[33] Y. Shen、J. Gu、X. Tang和B. Zhou。解读生成对抗网络的潜在空间以进行语义人脸编辑。载于CVPR，2020年。1, 2, 3[34] A. Shoshan、N. Bhonker、I. Kviatkovsky和G. Medioni。GAN-Control：显式可控生成对抗网络。载于IEEE/CVF国际计算机视觉会议论文集，第14083–14093页，2021年。1, 2, 3, 6, 7[35] K. K. Singh和Y. J. Lee。面向相对属性的端到端定位与排序。载于《计算机视觉—ECCV 2016：第14届欧洲计算机视觉会议，阿姆斯特丹，荷兰，2016年10月11日至14日，论文集》第VI 14部分，第753–769页。施普林格，2016年。3[36] O. Tamuz、C. Liu、S. Belongie、O. Shamir和A. T. Kalai。自适应学习群体核。载于国际机器学习会议（ICML），2011年。3[37] L. L. Thurstone。比较判断定律。《心理学评论》，34:278–286，1927年。1, 2, 3[38] O. Tov、Y. Alaluf、Y. Nitzan、O. Patashnik和D. Cohen-Or。为StyleGAN图像操控设计编码器，2021年。4, 8[39] J. D. Weisz、M. Muller、J. He和S. Houde。迈向生成式人工智能应用的通用设计原则，2023年。8[40] A. Xu和M. Davenport。从成对比较中同时学习偏好与度量。神经信息处理系统进展，33，2020年。3, 4[41] X. Yang、T. Zhang、C. Xu、S. Yan、M. S. Hossain和A. Ghoneim。深度相对属性。《IEEE多媒体汇刊》，18(9):1832–1842，2016年。3

A. 方法详情

B. 实验详情

B.1. 选择最优噪声常数

偏好估计的性能对某些超参数较为敏感。其中一个参数是噪声常数 k_q，在我们用于估计用户偏好的贝叶斯方法中，该参数充当信噪比。对于某些实验，我们选择常数 $k_q = 10$，实验发现该常数通常能够带来较强的性能。当比较不同方法，且用户对查询的回答中的噪声程度保持不变时，这种做法是令人满意的（例如，图6所示的实验）。然而，在图8中，我们展示了一项实验结果，该实验研究了不同噪声程度对偏好估计性能和我们方法不确定性量化能力的影响。对于该实验，有必要针对每种噪声程度调节噪声常数 k_q。

我们针对每种噪声程度调节 k_q 的方法是生成一个小型数据集（≈ 1000个示例），其中包含满足 $|a_a - a_p| + n < |a_a - a_n|$的属性三元组$t = (a_a, a_p, a_n) \in T$，其中 $n \sim N(0, \sigma_n)$。在这里，我们生成了一个具有不同噪声程度的数据集，并以标准差 σ_n作为参数。我们希望针对每种噪声程度选择一个能够最大化似然的噪声常数k_q

$$\sum_{(a_a,a_p,a_n)\in T} \sigma(k_q(||a_a - a_n||^2 - ||a_a - a_p||^2)). \quad (5)$$

直观而言，随着数据集中的噪声程度增加，我们希望模型对其用户偏好估计的信心降低。具体做法是减小噪声常数 k_q，使 $\sigma(k_q(||a_a - a_n||^2 - ||a_a - a_p||^2))$的平均值更接近 0.5。这意味着，当我们减小 k_q 时，会降低任意给定查询的权重，从而导致收敛速度变慢，后验体积增大。

B.2. 偏好估计实验流程

给定一个生成模型 $G(z)$ (在我们的案例中是一个 Style- GAN2 [16]模型)，我们模拟用户对成对查询的回答，并尝试估计用户对一组属性 $\mathcal{A}$的偏好。设 T 为我们向用户呈现的查询数量。设 $a^* \in \mathcal{A}$ 为描述用户对一组属性 $\mathcal{A}$偏好的真实属性。我们的目标是近似 a^*。

每次试验都遵循以下流程：

1. 首先，我们从 $\mathcal{A}$中均匀采样一组目标属性 $a^* \in \mathcal{A}$。

2. 我们从 StyleGAN2 潜空间中随机采样一个初始潜向量 $z_i \in \mathcal{Z}$（实际上是 StyleGAN2 的 $\mathcal{W}$ space，但为方便起见，我们称其为 $\mathcal{Z}$。）

3. $Q_0 = \{\}$

4. 对于 T 成对查询：

   (a) 使用所需的采样方案（主动采样、随机采样等）采样一个属性成对查询 $q = (a_a, a_b)$。(b) 为该查询生成相应的图像集合 $q_x = (G(f(z_i, a_a))\ G(f(z_i, a_b)))$(c) 使用模拟预言机回答该查询。(d) $Q_t = Q_{t-1} \cup q$(e) 使用 Stan 建模语言[2] 从后验分布 $p(a \mid Q_t)$中进行马尔可夫链蒙特卡罗采样。(f) 计算预言机偏好的当前后验估计 $\mathbb{E}[a \mid Q_t]$。(g) 计算估计的潜向量 $z_t = f(z_i, \mathbb{E}[a \mid Q_t])$。(h) 生成估计图像 $x_t = G(z_t)$。(i) 根据 z_t、x_t 和 z_i计算所需的评估指标。

B.3. LSTM 基线实现细节

在核心论文中，我们实现了 [12], 的一个变体，并将其称为 LSTM 基线。这里我们介绍为确保公平比较而对其架构所做的调整。

原始方法 在 [12], 中，作者希望在给定一组两两约束的情况下，从生成对抗网络生成图像。这个动机与我们的非常相似，但他们的方法有很大不同。他们使用长短期记忆网络架构，根据输入潜向量 z_i 估计一个满足给定约束的潜向量 z_f。

长短期记忆网络架构定义如下。该架构包含三个核心组件。第一个组件是一个 读取网络 $\phi(x)$，它将图像约束映射到语义空间；该语义空间的作用与我们的属性空间 $\mathcal{A}$ 类似。第二个组件是一个处理网络，它将所有约束整合为单一表征，其作用与我们的贝叶斯偏好估计过程类似。第三个组件是一个 写入网络，它将最终表征映射为图像，这正是我们的生成对抗网络生成器所做的。

$\{c_1, \ldots, c_n\}$ 是一组由读取网络生成的嵌入式约束。每个约束 $c_i = a_p \| a_n$对应一个查询 $q_i = (a_p, a_n)$，其中 $\|$ 表示向量拼接。作者描述了一种处理网络，该网络

[2] https://mc-stan.org/

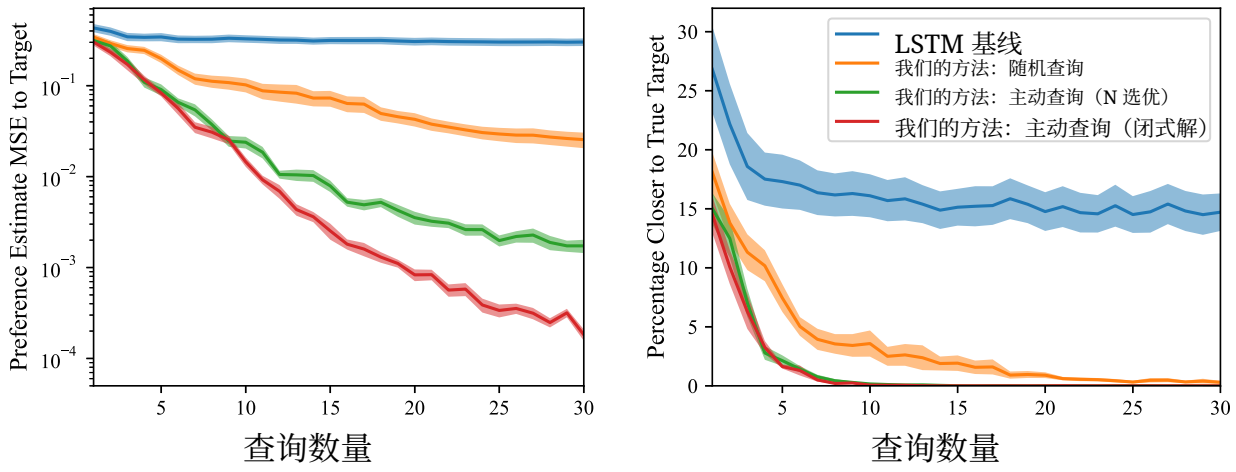


图 9. **比较通过成对比较估计用户偏好的各种方法的定量性能（四维偏航角、俯仰角、翻滚角和年龄）**。此图与核心论文中的图 6 相似。我们模拟预言机对一系列查询的响应，共进行 20 次试验，每次试验包含 30 次查询。我们在由人脸的年龄和姿态（偏航角、俯仰角和翻滚角）构成的四维空间中进行搜索。（左）给定试验中，经过 t 次查询后，偏好估计在属性空间中的均方误差，以及真实属性 a^*。（右）在包含 1000 个属性向量的一批样本中，与估计结果相比，更接近目标属性的属性向量所占百分比。根据这两项指标，PrefGen 都能够高效且精准地估计偏好。正如预期，主动查询相比随机查询提高了偏好估计的效率；而利用连续生成模型潜在空间的闭式 方法的性能显著优于*N* 选优方法。

接收这些约束和一个输入潜向量 z，并生成一个输出潜向量 z_f，该向量可输入生成器，以生成满足约束的图像。处理网络 的 LSTM 框架描述如下：

$$h_t = LSTM(z, h^*_{t-1}) \tag{6}$$

$$e_{i,t} = c_i \cdot b_t \tag{7}$$

$$a_{i,t} = \frac{exp(e_{i,t})}{\sum_j^n exp(e_{j,t})} \tag{8}$$

$$r_t = \sum_i^n a_{i,t} c_i \tag{9}$$

$$h^*_t = h_t \| r_t \tag{10}$$

如 [12] 输入 z 和前一次迭代的隐藏状态一起输入长短期记忆网络（LSTM）。随后，将下一个隐藏状态 h_t 与每个约束 c_i 进行点积运算，生成一个标量值 $e_{i,\ t}$ 对应每个约束。这些值随后输入 softmax 函数，得到标量 $a_{i,\ t}$ 并用于加权和。加权和用于组合这些约束。最终的 r_t 与 h_t 拼接，并作为下一次处理迭代的输入。重复此过程 p 步，最终生成 h^*_p，再输入全连接层以生成 s，并将其直接输入写入网络（生成器）。

作者使用三元组损失和对抗训练流程，以端到端的方式训练这一框架。

我们的实验改动 在我们的论文中，无需端到端地训练系统。相反，我们将流程拆分为两个关键部分。首先，我们使用贝叶斯方法估计用户在属性空间中的偏好。其次，我们概述一种通用流程，用于修改潜向量，使其生成符合这些偏好的图像。我们的贝叶斯偏好估计流程与 [12] 中的处理网络功能非常相似。唯一的关键区别在于，[12] 中的处理网络会生成一个向量 s，并将其直接输入写入网络（生成器）。这就要求长短期记忆网络既要根据一组约束对用户偏好进行建模，也要理解生成器潜在空间的结构。我们的偏好估计流程不依赖于任何特定的潜在表示，只需在低维属性空间（相较于生成对抗网络潜在空间）中找到一个满足一组约束的向量。为便于比较，我们对长短期记忆网络框架进行修改，使其只需根据一组约束生成一个低维属性向量。然后，我们可以将其输入属性映射函数 $f(z_i,\ a)$，以生成满足约束的潜向量。这样，我们便能直接比较贝叶斯偏好估计方法与长短期记忆网络方法的性能。

对于此架构，我们唯一的改动是将零向量作为输入馈入长短期记忆网络（LSTM），而不是潜向量 z，同时仍向长短期记忆网络提供一组约束。随后，我们将长短期记忆网络的最终输出视为属性向量 a。



图10. **使用 CLIP 的 PrefGen**。这里展示了 PrefGen 在由文本属性“红发”和“愤怒”构成的二维空间中的两个试验示例。

C. 更多结果

C.1. 监督映射：面部姿态与年龄偏好估计

定性结果：面部姿态与年龄偏好估计在图5 中，我们展示了偏好估计过程两次试验运行的定性示例。对于每次试验，我们展示了呈现给预言机的查询、初始图像、目标图像，以及我们的预言机偏好估计随时间变化的样本。由于篇幅有限，我们无法纳入更多示例，但这里展示了另外几个示例。如图11所示。

定量结果：面部姿态与年龄偏好估计在核心论文中，我们展示了多项定量结果，证明了我们的方法能够有效估计用户在由人脸左右朝向和人脸年龄构成的二维属性空间中的偏好。然而，我们的方法能够泛化到包含更多属性的空间也十分重要。这里我们展示了在由面部年龄、偏航角、俯仰角和翻滚角构成的四维空间中的对应结果。这些结果见图9。

C.2. 基于 CLIP 的映射

我们使用 CLIP 推断出与不同相对文本属性相对应的方向。在图1 中，我们展示了 PrefGen 使用汤姆·克鲁斯的编码图像进行的一次试验。在图10中，我们展示了另外两次具有相同特征（即“红发”和“愤怒”）的试验示例，其起始图像由随机潜向量生成。我们还可以操控不同类型的相对文本属性。图12展示了另外几种相对属性的操控结果。

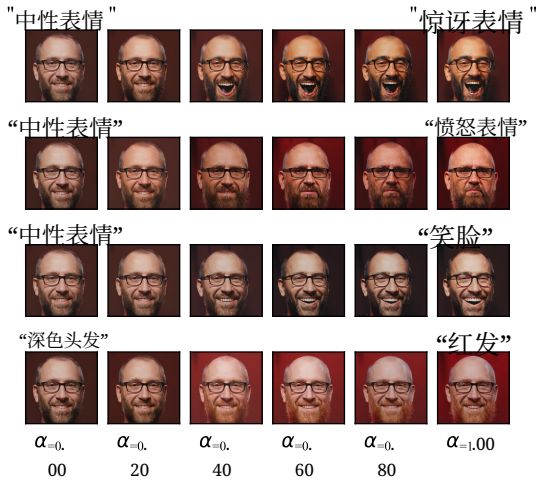


图12. **使用 CLIP 建模相对属性**。我们可以通过推断 CLIP 空间中与人脸相应属性对应的方向，来改变人脸的某些相对属性。随后，我们可以对图像的潜向量采用基于梯度的方法，直到生成的图像嵌入到 CLIP 向量中，且沿推断出的相对属性方向达到特定强度。



图 11. **更多姿态和年龄偏好估计示例**。这里展示了更多类似于图 5 的偏好估计过程示例。我们展示了 PrefGen 在由人脸的偏航角、俯仰角、翻滚角和年龄构成的四维空间中进行的 6 次试验。我们展示了呈现给预言机的每个查询，以及每次查询后从后验分布中采样生成的图像。样本图像通常会逐渐趋近于目标。