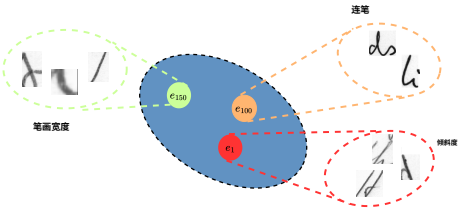
arXiv:2603.07543v1 [cs.CV] 8 Mar 2026

CONSTANT：通过局部块对比增强和风格感知量化实现高质量单样本手写生成

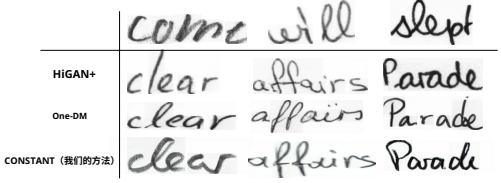
Anh-Duy Le[1], Van-Linh Pham[1], Thanh-NamVo[1],XuanToanMai[2],Tuan-AnhTran[1,2] 越南 Viettel 人工智能与数据服务中心[2] 越南胡志明市技术大学{leanhduy497, phamvanlinh143, thanhnam040501}@gmail.com, {mxtoan, trtanh}@hcmut.edu.vn

摘要

尽管近年来单样本风格化手写图像生成已取得令人瞩目的成果，但由于仅使用单张参考图像难以捕捉人类手写复杂而多样的特征，该任务仍然具有挑战性。现有方法仍难以生成视觉上美观且真实的手写图像，也难以适应复杂、未见过的书写者风格，难以分离不变风格特征（例如倾斜度、笔画宽度、曲率），同时忽略无关噪声。为解决这一问题，我们提出图像块对比增强与风格感知量化，通过去噪扩散（***CONSTANT***），这是一种新颖的基于扩散模型的单样本手写生成方法。*CONSTANT*依托三项关键创新：*1*）风格感知量化（*SAQ*）模块，将风格建模为捕捉不同概念的离散视觉令牌；*2*）对比目标，确保这些令牌在嵌入风格空间中充分分离且具有意义；*3*）基于潜在图像块的对比（$L_{LatentPCE}$）目标，通过在潜在空间中对齐生成特征与真实特征的多尺度空间局部块，有助于提升质量与局部结构。在涵盖英语、中文以及我们提出的面向越南语的*ViHTGen* 数据集等多语言基准数据集上的大量实验与分析表明，我们的方法在适应新的参考风格并生成高度精细图像方面优于最先进方法。代码可在*GitHub*



(a) 嵌入空间中的手写风格概念



(b) 我们的 CONSTANT 与.....之间的风格模仿能力其他方法（One-DM [4],HiGAN+ [12]）

图 1. 我们的方法利用向量量化的思想来学习多样化风格概念的嵌入空间，从而比其他方法更好地适应新风格。

1. 引言

手写合成是一项重要任务，其应用涵盖辅助技术、鲁棒认证的数据增强 [18] 以及文本识别系统 [20]。单样本方法通过单个风格示例生成手写体，对于用户可能只能提供一个参考样本的实际应用至关重要——样本。然而，这项任务仍极具挑战性。它要求仅凭一张图像，准确捕捉书写者独特且多变的风格，包括笔画宽度、曲率和倾斜度等特征，以及墨水密度等细微差异。尽管生成模型近年来取得了进展，现有方法仍难以从未见过的书写者的手写样本中提取这些**核心风格属性**，并生成**高质量**结果。因此，生成的图像在风格上往往并不完整；例如，模型或许能复现书写者的倾斜度，却无法捕捉其书写压力的变化。这引出了一个核心问题：如何有效地从单个手写样本中捕捉复杂的风格特征，从而生成多样、逼真且高质量的文本？

以往的手写合成方法大多依赖生成对抗网络（GAN） [2, 9, 11, 12, 21, 36]。尽管 GAN 能够生成多样化的样本，但往往难以生成逼真的图像，尤其是复杂书写风格的图像，而且

以训练不稳定而闻名。

近年来，扩散模型（DM） [7, 17, 39] 已成为一种强大的替代方案，在质量和多样性方面均展现出优于 GAN 的性能 [4, 32, 33, 53]。然而，仍存在一个关键局限：无法全面建模风格信息。例如，One-DM [4]，作为当前最先进水平的单样本扩散模型方法，采用了高频风格编码器，却忽略了笔画密度和颜色等重要特征；而其他方法 [2, 36]则采用计算开销巨大的基于 Transformer 的编码器 [47]。为提升质量，许多方法 [2, 33, 36]转而采用少样本生成（使用多张参考图像），但这种设置不如单样本方法实用。

在这项工作中，我们聚焦于提升单样本手写文本生成（HTG）的性能，并提出一种名为「补丁**对**比增强与**风**格感**知**量化**技**术」的新方法，通过去噪扩散实现（**CONSTANT**）。我们采用简单而有效的向量量化（VQ）方法，将复杂的手写风格信息表示为离散视觉令牌，其中每个令牌对应嵌入空间中的一个基本风格概念，类似于人们直观地对手写特征进行分类（见图1a）。与 One-DM 等依赖固定高频滤波器的方法不同 [4]，我们学习得到的「风格概念」更能稳健地从单张参考图像中捕捉核心特征，同时有效滤除偶然噪声（见图1b）。我们还引入风格对比增强目标 L_{SCE}，通过提升同一书写者的风格特征之间的相似性、区分不同书写者的风格特征，来优化潜在空间。为提升合成质量，并避免训练过程中出现模糊或目标输出与生成输出不一致等问题，我们提出潜在补丁对比增强（$L_{LatentPCE}$），这是一种超越 One-DM 所用标准去噪损失的辅助目标 [4]，它利用对比学习对齐生成图像与目标图像的潜在特征，并通过最大化对应目标图像块与生成图像块位置之间的互信息，锐化局部细节，从而获得更好的结果。与 [4, 12],等工作采用的多阶段训练不同，我们的整个流程在单个端到端阶段完成。为在越南语上评估我们的方法，我们专门创建了一个新的手写文本生成数据集，**ViHTGen**。我们在多种语言的多个数据集上测试了该方法，包括英语、中文以及我们提出的 ViHTGen，以展示其有效性。总而言之，我们的贡献如下：1）我们提出风格感知量化（SAQ）模块，将每个风格概念视为离散视觉令牌，以提取手写风格，从而实现风格表示的更清晰分离，并有助于减少复杂风格参考图像中提取的风格特征损失。图像。结合风格对比目标（L_{SCE}），该模块能够强化不同风格之间的差异。2）我们提出一种新的对比学习目标 $L_{LatentPCE}$，利用扩散模型中的补丁级潜在特征来细化局部细节，并保持生成图像的一致性。3）我们在多个语言数据集上进行了全面评估，包括我们新建的 ViHTGen 数据集，结果证实，我们的方法在视觉质量、风格质量和可读性方面达到了最先进性能。

2. 相关工作

2.1. 手写文本生成

早期方法 [2, 9, 11, 12, 21, 29, 36, 51]主要依赖生成对抗网络，因为它们能够对手写体固有的复杂数据分布进行建模。在生成对抗网络中，ScrabbleGAN [9]等单样本生成方法将随机噪声用作风格输入，以合成手写图像，因此无法控制书写者的风格。HiGAN [11] 提出了一种基于卷积神经网络的风格编码器，用于提取风格特征，并结合文本识别损失来提升质量。HiGAN+ [12]进一步改进了这类方法，引入了上下文损失和基于 PatchGAN 的 [5] 判别器；SmartPatch [29], 则采用了类似的轻量级补丁判别器。

近年来，该领域转向扩散模型，因为它们能够生成高质量且多样化的图像。WordStylist [32] 和 CTIG-DM [53] 等模型展示了扩散模型的潜力，但由于依赖预定义的书写者 ID，无法泛化到未见风格。类似地，以字体为重点的扩散模型，如 Diff-Font [14] 和 FontDiffuser [49]，将扩散与对比元素结合起来进行单样本风格适配，但其目标是结构化字形，而非变化多样的手写体。与本文最相关的单样本扩散模型 One-DM [4], 通过使用增强型风格编码器来解决这一问题；然而，它依赖预定义的拉普拉斯核来提取高频信息，因此可能不够稳健。我们通过两项关键贡献克服了这一局限。首先，我们的 SAQ 模块能够学习一种自适应的离散“风格概念”词汇表，使其能够更稳健地应对真实图像中多样且含噪的风格。其次，我们引入 $L_{LatentPCE}$ 目标，专门用于提升生成图像的局部细节与真实感。

另一些技术采用少样本策略来增强风格模仿。GANWriting [21] 使用15张参考图像提取风格特征。相比之下，HWT [2]采用Transformer 架构而非卷积神经网络，实现编码器-解码器 Transformer 框架以生成风格化文本。VATr [36] 与VATr++ [46] 对此加以改进，在 Transformer 解码器中将字符视为连续查询向量。DiffusionPen [33]利用扩散模型，

并结合度量学习目标以提升生成质量，但仍依赖五样本参考图像。与这些方法不同，我们的工作聚焦于单样本学习范式，该范式虽面临重大挑战，却具有更高的实际适用性。

2.2. 解耦表示学习（DRL）

DRL [48] 致力于创建能够识别并分离数据中潜在因素的模型，将其转化为不同且具有语义意义的表示。基于这一概念，包括 [10, 19]在内的一些研究进一步探索了特定的归纳偏置和模型架构，通过使用离散表示来增强 DRL，其中每个离散值都可被赋予一致的含义。这一概念已应用于文本图像超分辨率等任务。例如，[25] 采用汉字码本，而[3] 则训练了一个变分自编码器，使相似字母的表示彼此更接近。与这些研究不同，我们的方法旨在通过 VQ 利用离散表示来对手写体中的复杂风格概念进行建模，这比 [3, 25]的全局字符结构更为困难。

3. 方法

3.1. 问题定义

单样本手写生成模型的目的是以参考风格图像和文本内容信息为条件，生成任意手写文本图像。数学上，给定来自书写者W_s的参考图像 X_s，以及一个词语输入 $C = \{c_i | c_i \in \mathcal{A}\}_{i=1}^{L}$，其中 $\mathcal{A}$ 为字母表，L为词语输入的长度，我们的目标是生成新图像 X_t，下标“t”表示“目标”，该图像表示文本内容 C，且不依赖任何预定义语料库，同时具有与 X_s相同的手写风格。

3.2. 总体架构

我们的方法采用潜在扩散模型（LDMs） [39],与传统像素级模型相比，这类模型计算效率更高、占用内存更少 [7, 17]，而传统像素级模型会直接处理高分辨率图像。如图2a 所示，我们的方法以风格特征和文本特征作为潜在扩散模型的条件，这些特征分别由风格提取器和内容编码器提取。为了能够根据未见风格的参考样本合成手写图像，风格提取模块必须掌握手写领域中丰富多样的风格概念，并对每种风格概念进行独特表征。我们采用带有预定义码本嵌入的向量量化技术，将提取出的风格特征离散化为不同的离散视觉令牌；这一做法借鉴了许多离散表示学习领域的研究 [27, 35, 45], ，有助于我们的模型更有效地学习必要信息并忽略噪声信息。码本嵌入通过一个 L_{SAQ} 目标进行优化。

此外，为了更好地促使模型学习同一书写者不同样本之间的互信息，并抑制无关特征，我们采用 L_{SCE}。该模块在结合 SAQ 模块中的不同风格概念后接收全局风格表示，从而学习到更具判别性的风格嵌入空间，如图2b 所示。除了潜在扩散模型的去噪目标 $L_{denoising}$ 之外，[1], 我们的 $L_{LatentPCE}$ 有助于保留生成图像的视觉质量和细节，这一设计受到 [1]的启发。在文本内容编码方面，我们采用一个由 3 层 Transformer 构成的 [47]编码器。我们的内容编码器将每个输入字符独立表示为一个嵌入向量。具体而言，给定输入单词 $C = \{c_i\}_{i=1}^{L}$，我们的编码器会提取一个嵌入向量 $F_c \in \mathbb{R}^{L \times d}$，其中 d 为通道维度，这使我们的模型更加鲁棒，也能够更轻松地生成词汇表外词语。随后，将风格特征和内容特征作为上下文信息输入扩散模型内部的交叉注意力模块。总体而言，如图2a 所示，我们提出的方法通过以下目标的组合，以端到端方式进行联合优化：

$$L = L_{denoising} + \alpha \times (L_{LatentPCE} + L_{SCE} + L_{SAQ}) \quad (1)$$

其中，α 是所有受支持目标的权重。根据经验，我们选择 $\alpha = 0.1$。

3.3. 风格感知量化

如上文所述，我们提出了一种 SAQ 模块，用于从手写图像中提取紧凑且具有判别性的风格表示。我们的核心动机是超越单一、连续的风格向量，因为这类向量容易对**单张参考图像**中的噪声过拟合。为此，我们利用向量量化 [45]的离散表示能力，将手写体中丰富多样的风格变化所涵盖的不同风格概念，表示为码本嵌入中的离散嵌入令牌。我们提出的 SAQ 使用预训练的 InceptionV3 [42] 骨干网络，借助其从 ImageNet [6], 中提取多尺度特征的能力；这种能力已广泛应用于风格迁移 [13]。给定一张参考风格图像 $X_s \in \mathbb{R}^{H \times W \times 3}$，骨干网络首先生成一个特征图 $F \in \mathbb{R}^{h \times w \times d}$，其中 d 为特征维度。码本嵌入$E = \{e_i\}_{i=1}^{K} \in \mathbb{R}^{K \times D}$的长度为 K，代表一组不同的风格概念，$D$ 则为码本维度。该码本嵌入用于将特征图中的每个特征映射到一个特定的离散编码（即视觉令牌）。较大的 K 可以建模更细粒度的特征（例如细微的连笔变化），但可能会对噪声过拟合或增加内存需求；较小的值则可能过度简化风格，遗漏书写者特有的细微特征。令 $F_q = \{f_q^i\} \in \mathbb{R}^{h \times w \times d}$ 为量化后的

[1]关于潜在扩散模型的更多细节将在附录中提供

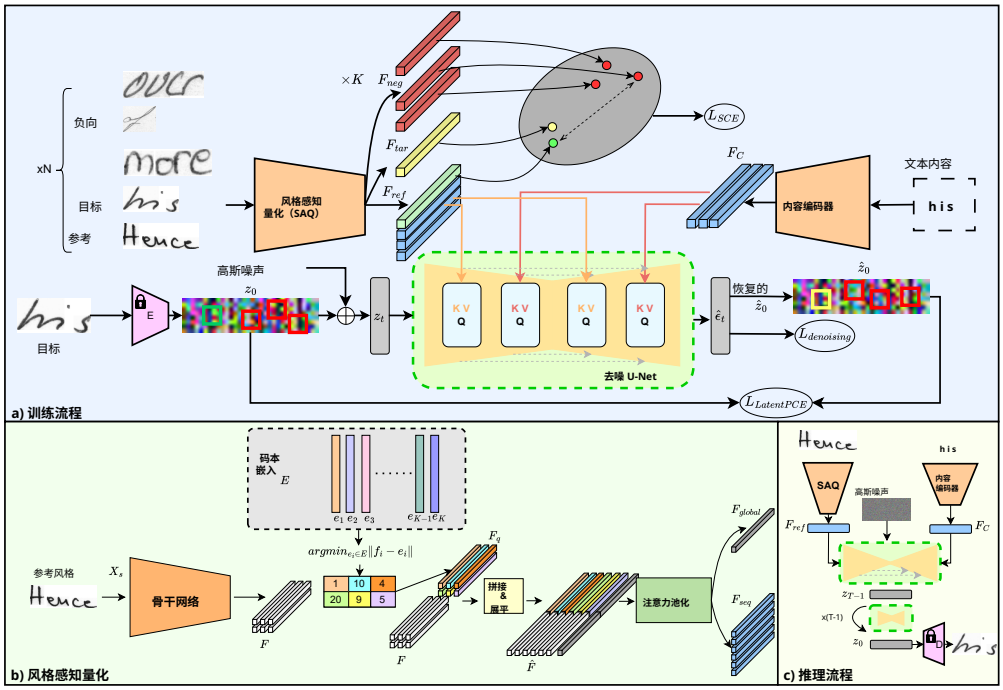


图2. **我们方法的总体架构**。a) 训练流程，模型使用以下目标进行优化： $L_{denoising}$、 L_{SCE} 和$L_{LatentPCE}$ 以及 L_{SAQ}，b) 我们的 SAQ 模块架构，包含 Inception-V3 骨干网络、码本嵌入以及一个注意力池化模块，c) 我们方法的采样流程。

特征图，其中每个向量为 f_q^i， F_q中的 $i \in [0, h \times w]$ 通过在码本 E 中查找与 f_i 最近的邻居来计算，表示为

$$f_q^i = argmin_{e_i \in E} \|f_i - e_i\|_2 \tag{2}$$

与其他先前方法类似 [27, 35, 45], 我们优化码本嵌入 E 使用向量量化损失 L_{vq} 和承诺损失 L_{cmt} 以确保码本学习到有意义的风格特征聚类。

$$L_{SAQ} = L_{vq} + \beta \times L_{cmt} \tag{3}$$

其中，损失项为 $L_{vq} = \|\text{sg}(F) - E\|_2^2$， $L_{cmt} = \|F - \text{sg}(E)\|_2^2$，其中 sg 表示停止梯度算子， β 是承诺损失的权重，我们选择$\beta = 0.25$ ，与 [45]相似。然而，我们注意到，直接使用量化特征图 F_q 并完全忽略来自 F 的所有信息，会导致忽略一些有用的局部细节信息，因为量化特征图只能捕捉从参考图像中提取时不应缺失的通用风格概念，而连续特征也有助于适应每种特定的、未见过的书写者风格。因此，我们将连续特征图 F 与量化后的版本 F_q 拼接，并将结果展平为 $\hat{F} \in \mathbb{R}^{T \times 2d}$，其中 $T = h \times w$ ，然后应用注意力池化模块

受.....启发 [38] 以融合来自两个特征图的信息。该模块对$\hat{F}$执行自注意力操作，生成两种类型的特征：1） $F_{global} \in \mathbb{R}^d$ 作为$\hat{F}$的全局表示 F，与 L_{SCE} 结合使用以增强书写者判别；2） $F_{seq} \in \mathbb{R}^{T \times d}$ 作为细化后的融合风格特征序列，为扩散模型提供上下文。这种混合方法兼顾了离散令牌的抽象性与连续特征的丰富性，使其能够稳健地应对手写图像的多样性。我们提出的 SAQ 总体架构如图2b 所示。

3.4. 风格对比增强

给定一个包含 N 个样本的训练批次 B ，其中包括来自不同书写者的 $N/2$对（参考图像、目标图像）。对于每一对（参考图像、目标图像），令目标（锚点）风格特征为 $F_{tar} \in \mathbb{R}^d$ ，索引为 j；参考风格特征为 $F_{ref} \in \mathbb{R}^d$ ，索引为 p。我们有 $K = N - 2$ 个负样本，将它们称为 $D = B \setminus \{j, p\}$，其中 $F_d \in \mathbb{R}^{Kxd}$是负样本的风格特征。F_{tar}、 F_{ref}和 F_d 均为由 SAQ 模块返回的全局风格特征 F_{global}。我们的风格对比增强（L_{SCE}）损失函数如下：

$$\ell_{SCE}(F_{tar}, F_{ref}, F_d) = \frac{-1}{N} \sum_{i \in B} \log \frac{\exp(\text{sim}(F_{tar}, F_{ref})/\tau)}{\exp(\text{sim}(F_{tar}, F_{ref})/\tau) + \sum_{d \in D(i)} \exp(\text{sim}(F_{tar}, F_d)/\tau)} \tag{4}$$

Where $\text{sim}((F_{tar}, F_{ref})/\tau)$ is the cosine similarity be-

参考图像与目标图像之间，以及 τ 一个温度超参数。更一般地，我们可以如[44], 中所述，将参考图像和目标图像互换角色，作为不同视图，其中参考图像是锚点，目标图像是正样本。因此，整体 L_{SCE} 如下：

$$L_{SCE} = \frac{1}{2}\ell_{SCE}\left(F_{tar}, F_{ref}, \mathrm{sg}\left(F_d\right)\right) + \frac{1}{2}\ell_{SCE}\left(F_{ref}, F_{tar}, \mathrm{sg}\left(F_d\right)\right) \quad (5)$$

这里，sg 表示停止梯度操作，该操作可防止梯度通过负样本传播，这是 [3] 提出的方法。

3.5. 潜在图块对比增强

在 HTG 中，生成逼真且细节丰富的输出至关重要。传统的回归损失（例如潜在扩散模型中的标准去噪损失）有助于生成全局连贯的结果，但往往难以捕捉精细细节，有时会导致输出模糊或过度平滑。我们提出的 $L_{LatentPCE}$ 是对去噪损失的补充，可确保局部一致性并增强感知真实感。与 HiGAN+ [12] 不同，后者通过像素级分块来增强真假样本分类；我们的 $L_{LatentPCE}$ 则在潜在空间中运作，直接优化生成图像的局部性信息。

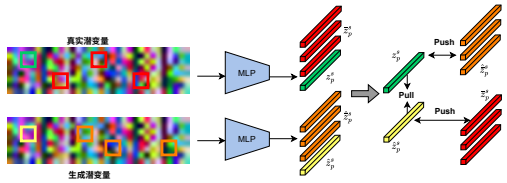


图 3. 我们的 $L_{LatentPCE}$ 目标的双向形式可视化，其中包含两个子对比损失。第一个损失接收（锚点、正样本、负样本），第二个损失接收（锚点、正样本、负样本）。

给定真实潜特征和生成潜变量，我们从相同的空间位置提取 y_p、$\hat{}y_p \in \mathbb{R}^{h_{patch} \times w_{patch} \times d_{patch}}$（例如图 3 中的绿色和黄色部分）。类似地，$\bar{y}_p$、$\hat{}\bar{y}_p$ 对应真实特征和生成特征中的其他空间图像块（例如图 3 中的红色和橙色部分）。这些图像块经过浅层多层感知机网络投影为特征嵌入 z_p、$\hat{}z_p \in \mathbb{R}^D$、$\bar{z}_p$、$\hat{}\bar{z}_p \in \mathbb{R}^{(H-1)\times D}$。其中，$D$ 为通道维度，H 为图像块总数。目标是确保来自相同空间位置的图像块在嵌入空间中彼此更接近，同时让来自不同位置的图像块彼此远离。此外，我们通过改变图像块大小，在多个尺度上提取图像块，以捕获不同空间分辨率下的局部信息。总之，我们在特定尺度级别 s 上的 $\ell_{LatentPCE}$ 可写为：

$$\ell_{LatentPCE}(z_p^s, \hat{}z_p^s, \hat{}\bar{z}_p^s) = -\frac{1}{H}\sum_{p=1}^{H}\log\frac{\exp(sim(z_p^s, \bar{z}_p^s)/\tau)\exp(sim(z_p^s, \hat{}z_p^s)/\tau)+\sum_{n=1}^{H_s-1}\exp(sim(z_p^s, \hat{}\bar{z}_p^s)/\tau)}{\exp(sim(z_p^s, \bar{z}_p^s)/\tau)} \quad (6)$$

与 L_{SCE}类似，通过交换锚点和正样本的角色，我们可以将 $L_{LatentPCE}$ 写成双向形式。因此，多尺度下的 $L_{LatentPCE}$ 目标可写为：

$$L_{LatentPCE} = \frac{1}{2S}\sum_{s=1}^{S}\ell_{LatentPCE}\left(z_p^s, \hat{}z_p^s, \mathrm{sg}(\hat{}\bar{z}_p^s)\right) + \frac{1}{2S}\ell_{LatentPCE}\left(\bar{z}_p^s, z_p^s, \mathrm{sg}(\hat{}\bar{z}_p^s)\right) \quad (7)$$

这里的 sg 表示一种类似于 L_{SCE} 的停止梯度操作，而 S 是尺度级别的数量。

4. 实验

4.1. 实验设置

4.1.1. 数据集

我们在三个数据集上评估所提出的方法，包括标准**IAM**数据集 [28], 、真实世界的复杂数据集**IMGUR5k** [24], 以及一个新的**IIIT-English-Word**数据集 [30]。**IAM** 数据集包含来自 500 位不同书写者的 62,857 张英语单词图像。与 GANWriting [22], 中的设置类似，我们使用 339 位书写者的数据进行训练，其余 161 位书写者的数据用于评估。**IMGUR5K** 数据集是一个真实世界的多源领域数据集，涵盖多种手写风格，包含来自 5305 位书写者的 135,375 个单词。**IIIT-English-Word** 数据集是一个新数据集，包含大量多样化的手写单词，样本超过 700000 个，来自 1,215 位书写者。*IIIT-English-Word*上的结果见附录。

4.1.2. 评估指标

为了评估所提出方法的视觉质量，并将其与其他最先进方法进行比较，我们采用 Fr´echet Inception Distance (FID) [15]。对于风格模仿质量，我们使用两项指标：手写距离（HWD） [37] ，通过欧氏距离衡量生成图像与真实图像之间细微几何特征的感知距离；以及书写者分类准确率（即 Acc_{Wid}），通过在我们生成的样本上测试一个用真实图像训练的分类器获得。对于可读性，我们采用 [34]中的设置，通过词错误率（WER）评估手写文本识别性能。最后，我们还通过一项用户研究评估感知质量和多样性，以补充这些定量结果；完整详情见附录。

4.1.3. 实现细节

在所有实验中，我们仅使用 PyTorch 1.12.1 进行单阶段训练，迭代 800000 次；IAM 数据集的批次大小为 64，IMGUR5K 和 IIIT-English-Word 数据集的批次大小为 128。所有图像均调整为固定高度 64，并保持纵横比。我们假设，复杂度更高的数据集——例如背景噪声更大、前景变化更多的数据集——需要

更大的码本大小 K，以捕获更多样的风格概念。根据经验，我们为 $K = 1024$ IAM 设置该值，为$K = 2048$ IMGUR5K/IIIT-English-Word 设置该值。我们在附录中提供了详细的消融实验，以验证这一设定在不同数据集上的有效性。训练过程使用 AdamW 优化器 [26], , $\beta_1 = 0.9$, $\beta_2 = 0.999$，初始学习率为 10^{-4}，并在单张 Nvidia V100-32GB GPU 上进行。训练阶段，我们遵循无分类器引导（CFG）的思路 [16]，以概率 $p = 0.2$随机丢弃条件特征。推理阶段，如图 2c 所示，我们采用 CFG 策略，指导尺度设为 7.5，并使用去噪扩散隐式模型（DDIM） [41] 进行 50 个采样步数 [2]。

4.1.4. 与其他最先进方法的比较

我们将我们的方法与多种最先进的手写文本生成（HTG）方法进行比较。基于生成对抗网络的方法包括 HiGAN [11], HiGAN+ [12]、HWT [2], 和 VATr [36]。对于基于扩散模型（DM）的方法，我们与当前最先进的 One-DM [4] 和 DiffusionPen [33]进行比较。其中，HiGAN、HiGAN+和 One-DM 是单样本手写文本生成方法，而 HWT、VATr 和 DiffusionPen 属于少样本方法 [3]。

4.2. 与最先进方法的比较

4.2.1. 定量比较

表 1展示了在 IAM 测试集上的定量结果。我们的方法在单样本和少样本方法中均取得最先进性能，HWD 和 FID 分别达到最佳的 0.74 和 10.20。这一结果优于排名第二的 HiGAN+ （HWD 为 0.89，FID 为 13.90）。作为一种基于扩散模型的单样本方法，我们的方法在 HWD 和 FID 得分上也优于 One-DM。在可读性方面，我们取得了最先进的 WER 0.22，略优于 DiffusionPen 的 0.23，并显著优于 One-DM 的 0.36。这些结果证明了我们 $L_{LatentPCE}$ （第 3.5 节）方法的有效性。此外，提出的 SAQ（第 3.3 节）模块提升了风格质量，作者分类准确率达到 69.43%，为此提供了有力佐证。

此外，我们还在四种不同场景下评估了我们的方法，包括 IV-S（词汇内，已见风格）、OOV-S（词汇外，已见风格）、IV-U（词汇内，未见风格）和OOV-U（词汇外，未见风格），其中OOV-U被认为是最困难的情形，其设置类似于[2, 22]。如表2所示，我们的方法在FID和HWD分数方面均展现出强劲性能。尽管DiffusionPen在基于GAN的方法中表现优异，尤其是在HWD上，凸显了扩散模型在风格适配方面的优势，但我们的方法达到了最先进的

方法	参考 HWD ↓	FID ↓		WER ↓	Acc_{Wid} ↑
HWT [2]	少样本	1.23	19.82	0.62	9.08
VATr [36]		1.13	16.30	0.51	49.81
DiffusionPen [33]		1.04	18.94	0.23	38.86
HiGAN+ [12]	单样本	0.89	13.90	0.56	55.20
HiGAN [11]		1.55	27.13	0.55	29.79
One-DM [4]		1.05	15.97	0.36	4.5
Ours		**0.74**	**10.20**	**0.22**	**69.43**

表 1. IAM测试数据集上最先进方法的比较。**粗体**数值表示最佳结果，下划线数值表示次佳结果。

在已见风格情形（IV-S和OOV-S）下的结果。尽管仅使用单张参考图像，我们在HWD和FID上也均显著领先于DiffusionPen。对于未见情形，我们的方法在IV-U和OOV-U的HWD上均优于次优方法（DiffusionPen），并在OOV-U的FID上超过One-DM达3%以上。这些结果清楚地证明了我们方法的有效性，尤其是SAQ 模块（第3.3节）在泛化到未见书写风格方面的有效性。

除 IAM 数据集外，我们还在更复杂的 IMGUR5K 数据集上评估了我们的方法。如表3所示，我们的方法在 IMGUR5K 测试集上取得了最先进的结果，HWD 分数为 0.99，优于 One-DM 的 1.22。这表明，与 One-DM 的高频编码器相比，我们的方法借助 SAQ 模块，在复现复杂手写风格方面具有更强的能力。

方法	IV-S		OOV-S		IV-U		OOV-U	
	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓
HWT [2]	2.30	135.51	2.30	146.16	2.35	138.39	2.36	148.75
VATr [36]	2.50	132.87	2.51	140.56	2.58	137.34	2.60	144.02
DiffusionPen [33]	1.14	91.20	1.16	97.65	1.52	112.87	1.65	122.52
HiGAN [11]	2.29	118.70	2.31	128.49	2.36	119.56	2.37	128.60
HiGAN+ [12]	1.76	117.76	1.79	122.56	1.78	117.63	1.81	124.38
One-DM [4]	1.95	104.04	1.99	107.81	1.94	117.74	1.99	121.94
我们的方法 (CONSTANT)	**0.96**	**89.88**	**0.94**	**96.13**	**1.61**	**112.03**	**1.63**	**118.10**

表 2. 四种场景下的 FID 和 HWD 分数评估：IV-S（词汇内，已见风格）、OOV-S（词汇外，已见风格）、IV-U（词汇内，未见风格）、OOV-U（词汇外，未见风格）。**粗体**值表示最佳结果，下划线值表示次佳结果。

方法	HWD ↓	FID ↓
HiGAN+ [12]	1.35	20.04
HiGAN [11]	1.55	17.58
One-DM [4]	1.22	18.94
Ours	**0.99**	**11.48**

表 3. IMGUR5K 测试集上的定量结果。

方法	FID ↓
基线	16.73
基线 + L_{cosine}	21.06
基线 + $LPIPS$	14.05
基线 + $L_{PatchL2}$	15.10
基线 + $L_{LatentPCE}$	**14.01**

表4. $L_{LatentPCE}$与其他辅助目标在FID方面的定量比较。

4.2.2. 定性比较

图 4 展示了定性比较，体现了我们的方法能够生成风格（倾斜度、笔画宽度和背景颜色）与参考图像 closely matchingrefer-

[2]有关我们实现和架构的更多详情，请参阅附录。[3]对于 IAM，我们在有可用预训练权重时使用这些权重；否则，我们会调整提供的源代码，在其他数据集上从头开始训练模型。

ence images. 相比之下，HiGAN 和 HiGAN+ 无法捕捉倾斜度的变化。尽管 HWT 和 VATr 使用的参考图像数量是我们方法的 15 倍，但它们难以准确适配字符颜色和形状，而且参考图像中较大的书写者内差异往往会导致结果不一致。虽然 One-DM 和 DiffusionPen 提取风格的效果优于其他基于GAN的方法，但在生成正确文本内容和足够的局部细节方面仍有不足。图 5 直观展示了该方法在 IMGUR5K 数据集上模仿风格（包括墨水颜色和字符形状）方面的优势。我们还可以看到，仅使用高频信息的 One-DM 能够生成合理的结果，但仍无法有效学习风格图像中的墨水颜色信息。

4.3. 消融实验

4.3.1. 所提组件的有效性

为验证我们的贡献，我们进行了累积消融实验，结果如表 5 所示。我们从“Base”模型开始，这是一个不含我们所提模块的标准条件LDM。第一行结果显示，该基线生成的文本清晰可辨，但较为模糊，风格保真度较差，FID为16.73。尽管 Base+SAQ 显著提升了视觉质量，使FID骤降至12.47。接下来，Base+SAQ +L_{SCE} 有助于将HWD分数从0.85提升至0.84，表明该损失成功促使模型形成更具判别性的风格空间。最后，Base+SAQ+L_{SCE}+$L_{LatentPCE}$ 进一步显著改善了局部细节，使FID降至出色的10.20，HWD降至0.74。视觉结果证明，所有组件协同作用，共同实现了我们的最终最先进水平性能。

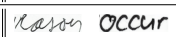
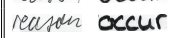
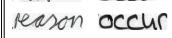
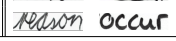
基础	SAQ	L_{SCE}	L_{PCE}	风格图像	FID ↓	HWD ↓
✓					16.73	0.87
✓	✓				12.47	0.85
✓	✓	✓			12.55	0.84
✓	✓	✓	✓		**10.20**	**0.74**

表 5. 从 FID 和 HWD 分数来看，我们的 SAQ、L_{SCE} 和 $L_{LatentPCE}$ 对基线模型的贡献。

4.3.2. $L_{LatentPCE}$ 的有效性

为验证我们提出的目标函数，我们将其与其他辅助目标进行比较，包括全局余弦相似度损失（L_{cosine}）、LPIPS [52]，和逐块L2（$L_{PatchL2}$）。如表 4所示，加入 $L_{LatentPCE}$后取得了最低的 FID（14.01），优于基线（16.73）和其他变体（例如， L_{cosine} 为 21.06，LPIPS 为 14.05， $L_{PatchL2}$ 为 15.10），证实了其在增强局部细节方面的优势。这源于最大化多尺度图像块之间的互信息；不同于 L_{cosine}，后者在没有局部约束的情况下优化整体特征相似度；LPIPS 和 $L_{PatchL2}$则尽管 LPIPS 二者都利用 VGG 预训练特征，依赖 L2 距离来缩小图像块差异，却未能促进判别性表征。

4.4. 向其他语言的泛化

为评估所提出方法的适应性和有效性，我们将实验范围从英语扩展到具有不同书写系统的语言，并考虑对中文和越南语进行实验。对于中文，我们使用CASIA数据库 [50]，其中包含 60 位书写者的数据，每位书写者有 3755 个汉字。对于越南语，我们构建了一个包含多种风格和多样背景的新数据集，名为ViHTGen。该数据集由从大学考试 [4]中收集的 50000 多张越南语手写图像组成。随后，我们将模型与One-DM [4] 方法进行比较。表6 显示，我们的方法在两项指标上均优于 One-DM，HWD 分数提升超过 10%，表明我们的方法比 One-DM 更能适应不同风格。从图6a 可以看出，我们的方法能够有效适应复杂风格，例如不同的笔画密度，同时保留字符内容。相比之下，One-DM 往往会生成模糊的细节，且难以在其生成的手写体中绘制足够的笔画。对于图6b 中的越南文字，我们的方法能够有效处理具有复杂字符形状和多样色彩的复杂背景，即使是数字字符也不例外。与此同时，One-DM 无法生成连贯的结果，其前景往往与背景混杂，表明它未能提取具有区分性的风格概念。附录中还展示了与其他方法进行比较的更多定量结果以及额外的定性结果。

方法	中文 HWD ↓	中文 FID ↓	越南语 HWD ↓	越南语 FID ↓
One-DM [4]	0.48	22.97	1.08	22.53
Our	**0.37**	**22.74**	**0.83**	**18.81**

表6. 在中文和越南文字上与 One-DM 进行的定量比较，指标为 FID 和 HWD。

4.5. 分析

4.5.1. 风格提取分析

为了更好地理解我们的 SAQ 模块的行为，我们在 SAQ 与基线 Inception-V3（第 3.3 节）之间进行可解释性分析，通过可视化风格特征与生成像素之间的交叉注意力分数，灵感来自 [43]。图 7 中得到的注意力图揭示了明显差异：我们的 SAQ 模块（A）产生清晰、局部化的图，其中每个特征对应特定的字符笔画。相比之下，基线的注意力（B）弥散且不聚焦，表明我们的

[4]有关 ViHTGen 的更多信息请参见附录

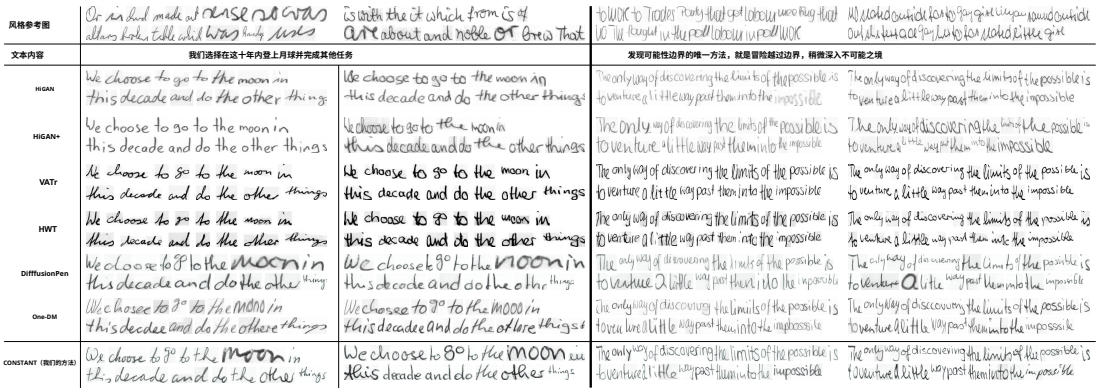


图 4. 使用来自 IAM 测试集的随机风格和两种不同的文本内容，将我们的方法与其他方法进行比较。



图 5. IMGUR5K测试集上的手写文本生成可视化。

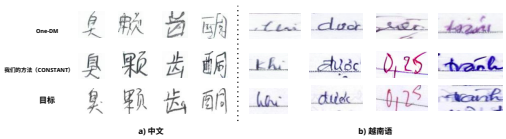


图 6. CONSTANT 与 One-DM 的比较 [4]中文和越南文字。

SAQ 模块成功学会将风格解耦为彼此不同且有意义的概念。

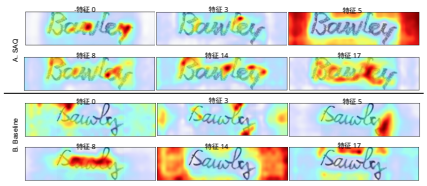


图 7. 我们的 CONSTANT 与基线模型之间的风格交叉注意力图可视化。

4.5.2. 领域泛化分析

为探究我们模型的跨领域泛化能力，我们进行了两项具有挑战性的实验。首先，我们使用来自 IMGUR5K 的风格参考图测试在 IAM 上训练的模型，以评估其对域外特征的鲁棒性。如图 8所示，模型成功保留了学习到的结构风格，同时正确忽略了墨水颜色等未见过的概念。其次，我们评估了在复杂领域上训练的模型能否利用干净的 IAM 参考图准确生成干净风格。图 9 显示，该模型能够准确生成干净风格，而不会“幻觉生成”其训练领域中的复杂风格。总的来说，这些情形展现了我们的方法在具有不同特征的领域之间进行泛化的强大能力。



图 8. 在 IMGUR5K 参考图像上使用在 IAM 上训练的模型生成的图像与真实图像的对比。



(a) 示例 1　　(b) 示例 2

图 9. 使用 IAM 参考图像对 IMGUR5K 模型进行领域自适应。

5. 结论

本文提出了 CONSTANT，这是一种新颖的扩散方法，可显著提升单样本手写文本生成。我们的方法采用风格感知量化（SAQ）模块，并通过对比损失（L_{SCE}）加以增强，借助向量量化这一创新应用，从单张参考图像中稳健地捕捉核心风格概念。为解决模糊等常见问题，我们还引入了多尺度块级（$L_{LatentPCE}$）目标，以细化局部细节。我们在多个多语言数据集上进行了大量实验，包括英语、中文以及新建的 ViHTGen 数据集，结果证实，我们的方法在视觉质量、风格保真度、可读性和跨领域泛化方面均表现出卓越性能。

CONSTANT：通过局部块对比增强和风格感知量化实现高质量单样本手写生成

补充材料

A. 引言

在本补充材料中，我们在第B节介绍潜在扩散模型的更多细节。在第C节中，我们开展用户研究，将最先进方法进行比较，以评估视觉效果和风格模仿质量。第D节详细介绍了我们的 ViHTGen 数据集。我们还在第E节更清晰地介绍实现细节。在第F节中，我们进一步介绍评估指标的更多细节。接下来，我们在第G节展示了 IIIT-English-Word 数据集上的实验结果。第H节展示了我们的方法与其他最先进方法在多语言泛化方面的更多结果。在第I节中，我们通过实验评估了本方法对提升手写文本识别性能的效果。此外，我们还在第J节研究了码本嵌入长度对模型性能的影响。第K节进一步介绍效率方面的细节，包括训练成本和推理时间。第L节讨论了局限性和未来工作。最后，我们展示了模型在多个数据集上的更多定性结果。

B. 关于潜在扩散模型的更多细节

扩散模型（DMs）代表了生成建模的一项重大进展，在各类任务中往往超越生成对抗网络。从使用迭代去噪生成样本的DDPM [17], 开始，大量改进提升了质量与控制能力 [7, 39, 41]。无分类器引导（CFG） [16] 以及GLIDE [31] 中所见的多模态条件等技术进一步提升了性能，尤其是在文本到图像生成方面。

我们用于手写生成的扩散模型受到潜在扩散模型 [39], 的启发，它通过使用预训练的变分自编码器将手写图像X压缩为4维潜在空间表示 $z \in \mathbb{R}^{4\times W/8\times H/8}$，从而使采样过程能够在潜在空间中进行。与DDPM [17], 类似，所考虑的长度为T的前向过程构成一条马尔可夫链，该过程逐步向干净的潜在表示$\mathbf{z}_0 \sim q(\mathbf{z}_0)$添加高斯噪声，直至其变为纯噪声$p(\mathbf{z}_T) = \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathrm{I})$，每一步由转移概率 $q(z_t|z_{t-1})$定义。

反向过程旨在学习如何消除这种噪声添加过程，通过训练 U-Net 模型 [40] $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t,\ t,\ \boldsymbol{C},\ \boldsymbol{X}_s)$，使用均方误差损失预测添加的噪声，从而从纯噪声中生成样本：

$$L_{\text{denoising}} = E_{t\sim[1,T],z_0\sim q(z_0),\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{C}, \boldsymbol{X}_s\right)\|^2\right] \tag{1}$$

C. 用户研究

我们开展用户研究，包括用户偏好研究和用户合理性研究，以便更好地评估我们的方法，并与其他方法进行比较，包括 One-DM [4], Hi- GAN+ [12]、HWT [2], VATr [36], HiGAN [11],，使用 Type- Form [5] 调查平台。

C.1. 用户偏好研究

我们使用来自 IAM [28]数据集的未登录词语料库，对一组包含 30 条文本内容进行采样，并为每种方法生成 30 张图像。测试邀请了 28 名参与者，共收集到 840 份回答。我们请每位参与者从给定列表中选出与真实图像最相似的图像。如图 1所示，我们的方法获得了所有参与者最多的选择，回答占比超过 40.6%。

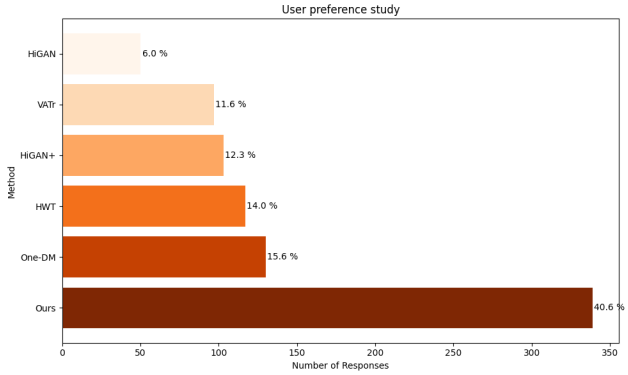


图 1. 用户偏好研究结果

C.2. 用户合理性研究

我们进行了实验，以研究我们的方法生成的图像是否与真实图像难以区分。每道题中，我们先向参与者展示来自同一书写者的 6 张真实图像示例，再请他们从总共 6 张图像中辨认出 3 张真实图像。这项研究共收集到 28 名参与者的 1680 份回答。结果如图2 所示，**准确率为 53.8%**，表明其接近随机分类。

D. ViHTGen 数据集

为模拟真实场景，我们构建了 **ViHT-Gen** 数据集，其中包含多种手写风格，背景复杂多样，

[5]https://www.typeform.com/

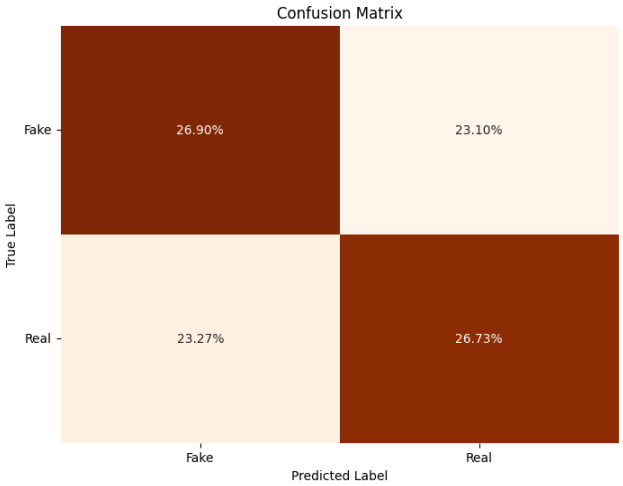


图2。用户合理性研究结果

这些数据集取自 200 多人撰写的 300 多份大学期末考试答卷。我们采用严格的半自动标注流程，以确保标签质量上乘。首先，使用 Google Cloud Vision API[6] 自动提取词实例边界框，随后进行人工核验。接着，采用先进的越南语视觉语言模型 VinTernVL-1B [8] 对每个词进行光学字符识别（OCR），并由人工检查和修正所有转写内容，以确保标签准确无误。

经过筛选后，最终数据集包含超过 50,000 张词级图像，我们将其划分为包含 42,000 张图像的训练集和包含 8,000 多张图像的测试集。如图 4 所示，我们还对文本长度和字符频率进行了统计分析。所得数据集是 HTG模型颇具挑战性的基准，包含多种笔画风格和复杂的噪声背景，如图 3 所示。表 1 还展示了我们的数据集与标准 IAM 数据集在多个方面的差异。为促进可复现性并推动未来研究，ViHTGen 数据集将在本文发表后公开。

统计	ViHTGen（我们的方法）	IAM
语言	越南语	英语
# 书写者	223	500
# 词实例	50000+	62857
# 唯一词语	4644	3332
书写风格复杂度（倾斜度、墨水颜色，笔画宽度，字符形状）	High	中
背景复杂度	High	Low
图像来源	大学考试	英文句子

表 1. 我们的 ViHTGen 与 IAM 数据集之间的比较。

E. 实现细节补充

E.1. 模型架构细节

我们的模型包含三个基本组件：1）基于潜在扩散的模型；2）用于风格特征提取的 SAQ 模块；3）文本编码器模块。

• **基于潜在扩散的模型**：我们遵循潜在扩散模型的架构 [39]，，该架构采用 U-Net [40] 模型，其中包含一个 ResNet块，后接一个空间 Transformer 块，用于融合上下文特征（此处为风格和文本特征）与输入特征。我们遵循 WordStylist [32] 的做法，减少 ResNet 块的数量以缩短训练时间；同时，将 Transformer 块中的头数设为 4，特征维度设为 512。

• **SAQ 模块**：我们提出的 SAQ 模块包含三个基本部分：Inception-V3 骨干网络、码本嵌入 E和 AttentionPool 融合模块。我们选择 Inception-V3，是因为它能够有效提取多尺度特征，风格迁移中也采用了这种做法 [13]。Inception-V3 的输出维度为 768。由于我们采用了论文正文中所述的结合离散特征与连续特征的混合方案，SAQ 中特征的输出维度 F 为 1536。最后，受 CLIP 框架启发 [38], 的 AttentionPool 模块通过自注意力操作，更好地融合来自离散特征和连续特征的信息。最终输出经线性层投影至 512 维，然后传入潜在扩散模型。

• **文本编码器模块**：文本编码器由 3 层 Transformer 块构成；每个块包含一个 MLP 块和一个维度为 512 的自注意力块。

E.2. $L_{LatentPCE}$ 实现细节

我们将 $L_{LatentPCE}$ 目标设计为一种在三个不同尺度上进行的多尺度对比损失。在每个尺度上，我们分别提取最多 256 个大小为 2×2、4×4 和 8×8 的图像块。随后将提取的图像块展平，并使用浅层多层感知机将其投影到 256 维嵌入空间。最终的 $L_{LatentPCE}$ 是三个尺度的对比损失的平均值。

[6]https://cloud.google.com/vision/docs

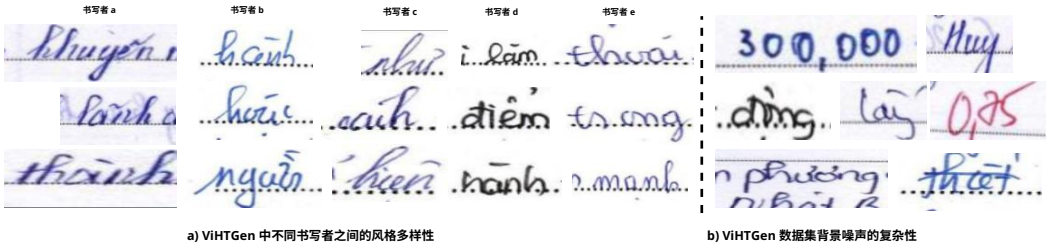
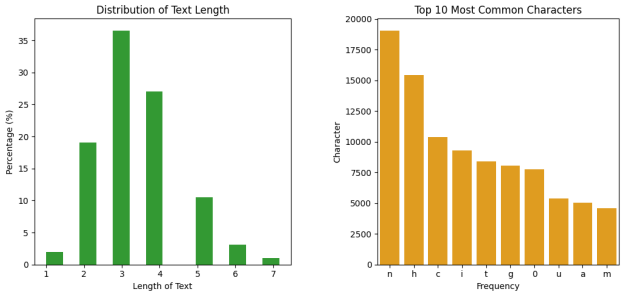


图 3. 我们的 ViHTGen 数据集概览。a) 不同书写者之间手写风格的多样性。b) 我们的数据集在背景和笔画形状两方面的复杂性。



(a) 文本长度的分布 (b) 最常见的字符

图 4. 对 ViHTGen 数据集的统计分析。(a) 不同文本长度图像所占的百分比。(b) 数据集中出现频率最高的 10 个字符的频率分布

F. 关于评估指标的更多细节

F.1. 基于OCR的指标

为了评估生成图像的可读性，我们遵循 [34]中的设置。核心思路是在生成图像上训练光学字符识别（OCR）模型，然后测试其在真实图像上的表现。成功的 HTG 模型应能生成样本，使 OCR 模型在真实数据上取得较低的词错误率（WER）。我们采用 [23]中的序列到序列 OCR 架构。在评估中，我们首先使用为 IAM 测试集生成的图像训练 OCR 模型。模型训练 200,000 次迭代，批量大小为 128。训练完成后，我们在真实的 IAM 测试集上测试 OCR 模型，并计算 WER。

F.2. 书写者分类指标

除了评估可读性外，我们还使用作者分类器模型评估风格模仿能力。这种评估策略已在以往的研究中采用 [32, 33]。在我们的工作中，我们使用在 ImageNet 上预训练的 ResNet18 模型作为基础架构。为了训练分类器，我们将真实的 IAM 测试集按 80/20 的比例划分为训练集和验证集。随后，我们使用训练好的模型评估 IAM 测试集生成结果上的书写者分类准确率

，并以此衡量我们的 HTG 模型的风格模仿能力。

G. 在 IIIT-English-Word 数据集上的实验

我们在 IIIT-English-Word 数据集上进行实验 [30]以比较我们的 CONSTANT 模型与 One-DM 的性能 [4]，指标为 FID 和 HWD 分数。定量结果见表 2 ，定性结果见图5

方法	HWD ↓	FID ↓
One-DM [4]	1.22	17.74
Ours	**0.73**	**10.22**

表2。IIIT-English-Word测试集上的定量结果。

H. 多语言泛化的更多结果

为更好地展示我们的方法在其他语言上的有效性，我们进行了额外实验，将我们的方法与最先进水平（SOTA）的竞争方法进行比较。在该实验中，我们使用DiffusionPen [33]作为基线模型。由于DiffusionPen是少样本模型，我们在单样本设置下对其进行训练，以确保公平比较。与主论文中针对One-DM的评估类似，我们在中文和越南语数据集上将我们的方法与DiffusionPen进行比较。如表 3 和图 6所示，与DiffusionPen相比，我们的方法在两种语言上的视觉质量和风格适应能力方面都取得了更好的结果。这也展现了我们的方法在单样本设置下的稳健性，而DiffusionPen模型在适应该设置时的性能较低（例如，其中文FID分数为54.07，而我们的方法为22.74）。

I. 手写文本识别改进

我们通过实验评估增加生成数据的数量后手写文本识别模型的性能提升

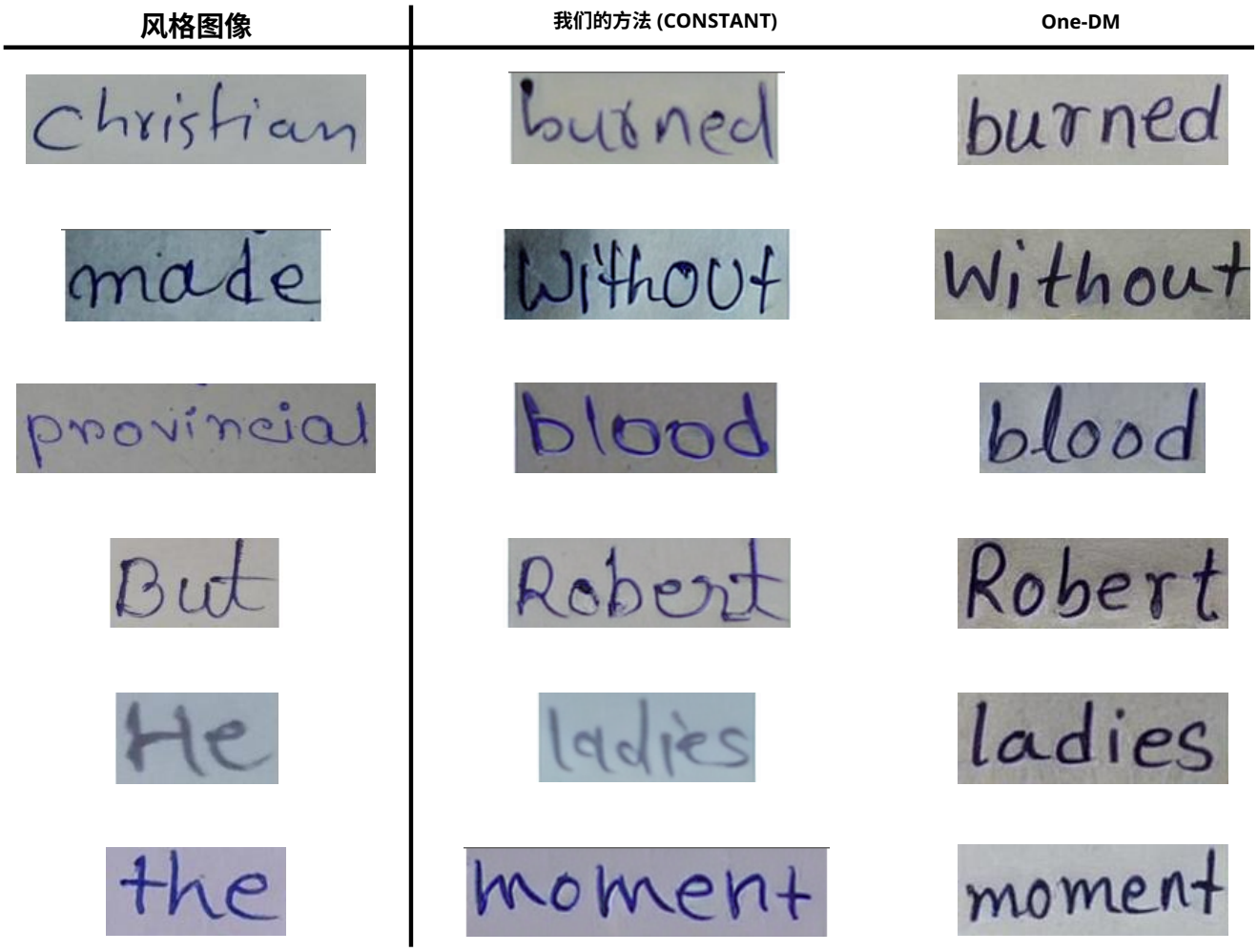


图 5. 我们的方法与 One-DM 的比较 [4] 在 IIIT-English-Word 上。

方法	中文 HWD ↓	中文 FID ↓	越南语 HWD ↓	越南语 FID ↓
DiffusionPen [33]	0.57	54.07	1.05	23.59
Our	**0.37**	**22.74**	**0.83**	**18.81**

表 3. 与 One-DM 在中文和越南文字上关于 FID 和 HWD 的定量比较。



图 6. CONSTANT 与 Diffusion- Pen [33] 在中文文字上的定性结果。

对训练集的影响。实验包括使用真实的 IAM 训练集进行训练，并逐步增加使用我们的方法生成的手写图像数量，每次增加 10 万张。

数据来源	准确率
真实数据	81.76
真实数据 + 100K	82.13
真实数据 + 200K	**82.96**

表 4. 生成结果数据有助于提升 IAM 测试集上的光学字符识别性能

J. 码本嵌入长度消融实验

我们进行了消融研究，以探究 SAQ 中码本嵌入长度的影响，并验证

关于其与数据集复杂度相关性的假设。研究在三个风格复杂度各异的数据集上，改变码本大小（K）：风格相对简单的 IAM 数据集，以及视觉风格更复杂的 IMGUR5K 和 IIIT-English-Word 数据集。我们评估了码本大小 K=512、K=1024和 K=2048。如表 5所示，结果与我们的假设一致。对于简单的 IAM 数据集，最大的码本（K=2048）略微降低了性能（FID 14.10，HWD 0.86），这表明出现了过拟合。相比之下，对于 IMGUR5K 和 IIIT-English-Word 中更为复杂的风格，较大的 K=2048 码本取得了最佳性能，使模型能够捕捉更丰富多样的风格特征。这些结果有力地印证了我们在论文正文中提出的假设。

码本长度	IAM		IMGUR5K		IIIT-English-Word	
	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓
512	0.86	13.80	1.02	13.37	0.75	14.02
1024	**0.84**	**12.46**	1.03	12.9	0.77	12.61
2048	0.86	14.10	**0.99**	**11.48**	**0.73**	**10.22**

表 5. SAQ 模块中码本嵌入长度的消融研究

K. 效率分析

为进一步说明我们方法的训练成本和效率，我们从训练成本方面将其与 One-DM 进行比较。表 6 详细列出了我们的方法与基线 One-DM 模型在训练和推理成本方面的信息。两种方法的实验均在同一台 NVIDIA V100 机器上进行。

评价标准	CONSTANT	One-DM
模型参数-ters	124×10^6	185×10^6
训练显存	6.82GB	18.66GB
推理时间[7]	1.25 秒/样本	1.85 秒/样本
推理显存	2.1GB	2.9GB

表6. CONSTANT与One-DM的训练和推理成本对比。

L. 局限性与未来工作

尽管取得了最先进的结果，我们的方法仍存在局限性。当参考图像过于模糊或具有7复杂背景时，模型的风格提取可能会受到影响（图 a）。同样，对于过于复杂或几乎难以辨认的手写体，我们的模型可能会优先保证内容可读性，而非精确模仿风格，如图7所示。另一项局限是，码本大小是通过经验确定的，对于风格复杂度不同的数据集而言，这一大小可能并非最优。未来的工作将着重于改进对高度艺术性文本的风格提取，并探索生成更长文本的方法
行文本。

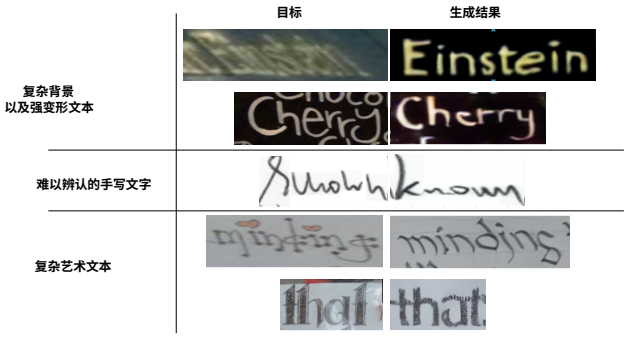


图 7. 一些失败案例的可视化。

M. 更多定性结果

[7]采样步数少于50步

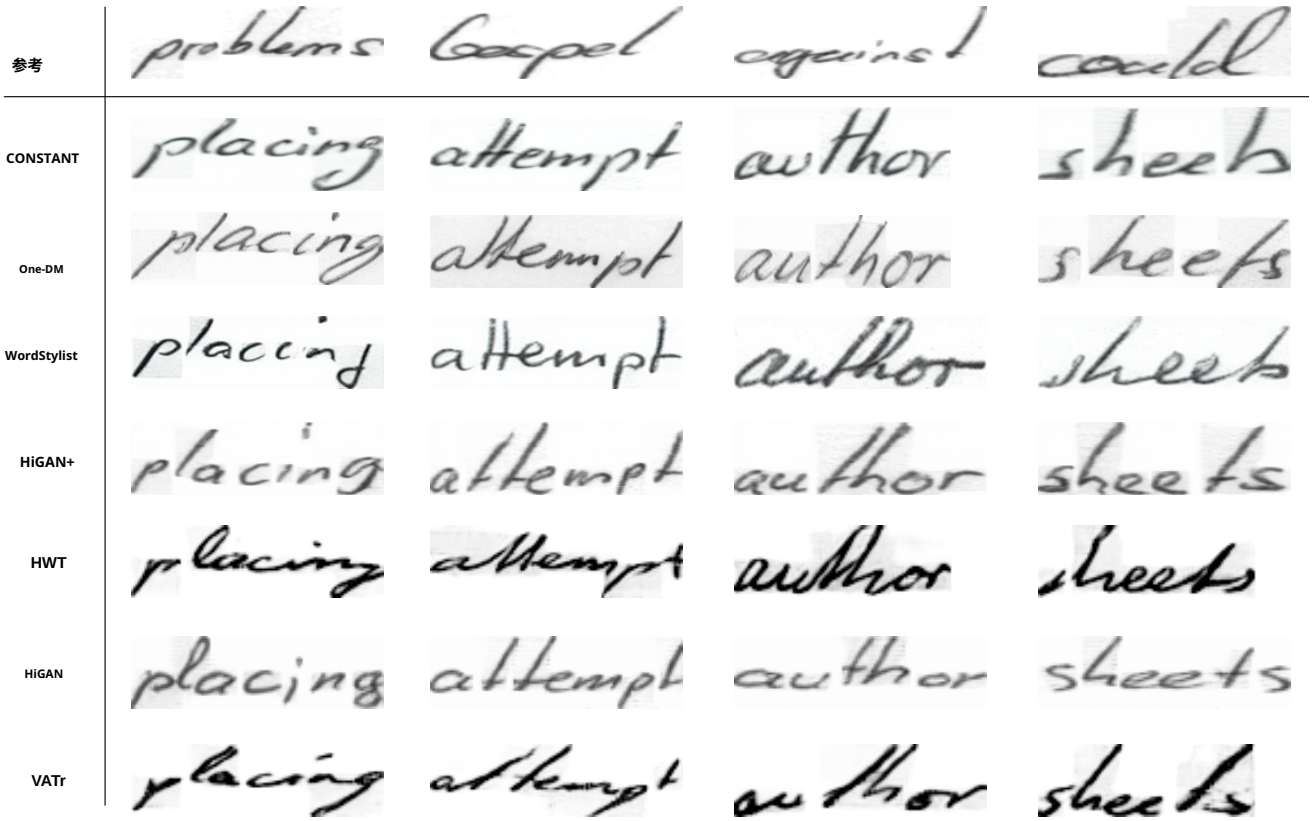


图 8. 在 IAM 测试数据集上，不同方法对任意文本内容的可视化结果。

致谢

我们感谢胡志明市理工大学（HCMUT）、VNU-HCM 对本研究的支持。

参考文献

[1] Alex Andonian、Taesung Park、Bryan Russell、Phillip Isola、朱俊彦和Richard Zhang。用于图像预测的对比特征损失。发表于*IEEE/CVF*国际计算机视觉会议论文集，第1934–1943页，2021。[2] Ankan Kumar Bhunia、Salman Khan、Hisham Cholakkal、Rao Muhammad Anwer、Fahad Shahbaz Khan和Mubarak Shah。手写体Transformer。发表于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，第1086–1094页，2021。[3] 陈新雷和何恺明。探索简单的孪生表示学习。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15750–15758页，2021。[4] 戴刚、张一凡、柯曲辉、郭强亚和黄双平。One-DM：用于手写文本生成的单样本扩散模仿器。发表于欧洲计算机视觉会议，第410–427页。施普林格，2025。[5] Ugur Demir 和 Gozde Unal。基于块的图像修复与生成对抗网络。*arXiv* 预印本 *arXiv:1803.07422*，2018年。[6] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。载于*2009*年 *IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[7] Prafulla Dhariwal 和 Alexander Nichol。扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展，34:8780–8794，2021。[8] Khang T Doan、Bao G Huynh、Dung T Hoang、Thuc D Pham、Nhat H Pham、Quan Nguyen、Bang Q Vo 和 Suong N Hoang。Vintern-1b：面向越南语的高效多模态大语言模型。*arXiv* 预印本 *arXiv:2408.12480*，2024。[9] Sharon Fogel、Hadar Averbuch-Elor、Sarel Cohen、Shai Mazor 和 Roee Litman。ScrabbleGAN：半监督可变长度手写文本生成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4324–4333页，2020年。[10] David Friede、Christian Reimers、Heiner Stuckenschmidt 和 Mathias Niepert。学习解耦的离散表示。载于机器学习与数据库知识发现联合欧洲会议，第593–609页。施普林格，2023年。[11] Ji Gan 和 Weiqiang Wang。HiGAN：手写体模仿，

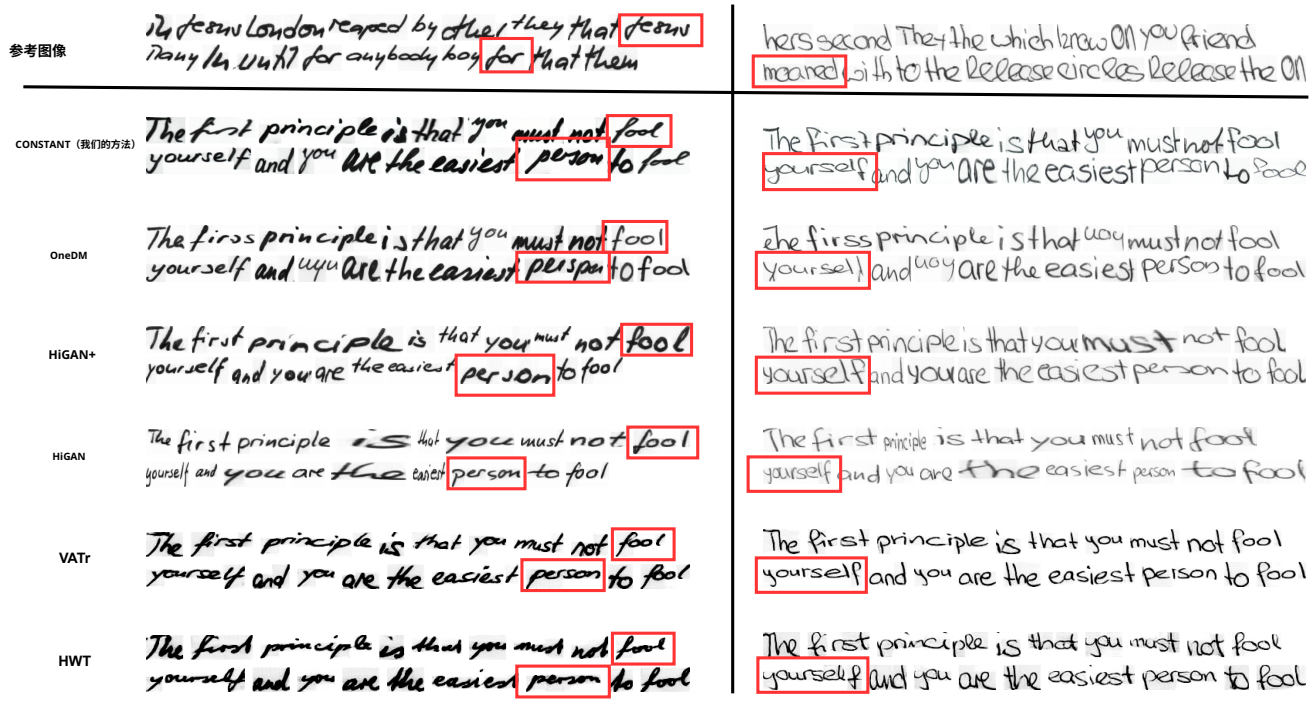


图 9. 我们的 CONSTANT 与其他竞争方法的可视化，其中红色方框展示了不同方法对参考图像的风格适应能力。

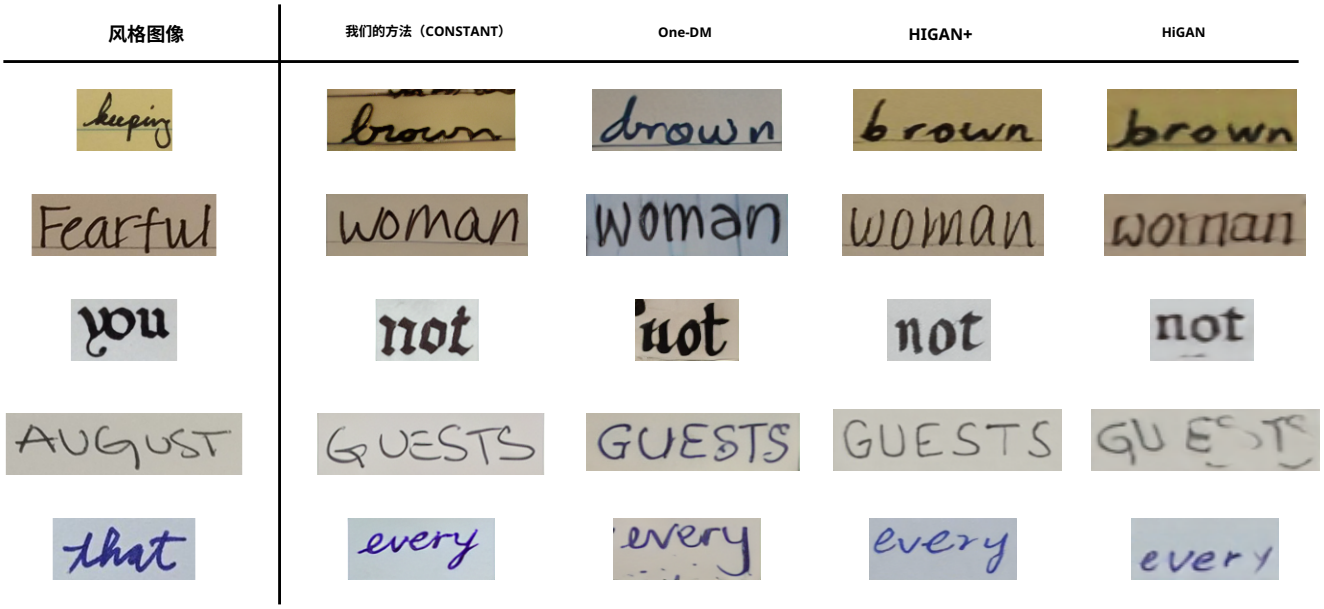


图 10. IMGUR5K 数据集上的其他示例。

以任意长度文本为条件并解耦风格。*AAAI*人工智能会议论文集，35(9):7484–7492，2021。1，2，6

[12] Ji Gan、Weiqiang Wang、Jiaxu Leng 和 Xinbo Gao。Hi- gan+：采用解耦表征的手写模仿生成对抗网络——表征。*ACM* 图形学汇刊，42(1)，2022。1，2，5，6

[13] Golnaz Ghiasi、Honglak Lee、Manjunath Kudlur、Vincent Dumoulin和Jonathon Shlens。探索实时任意神经艺术风格化网络的结构。 *arXiv*预印本 *arXiv:1705.06830*，2017。3，2

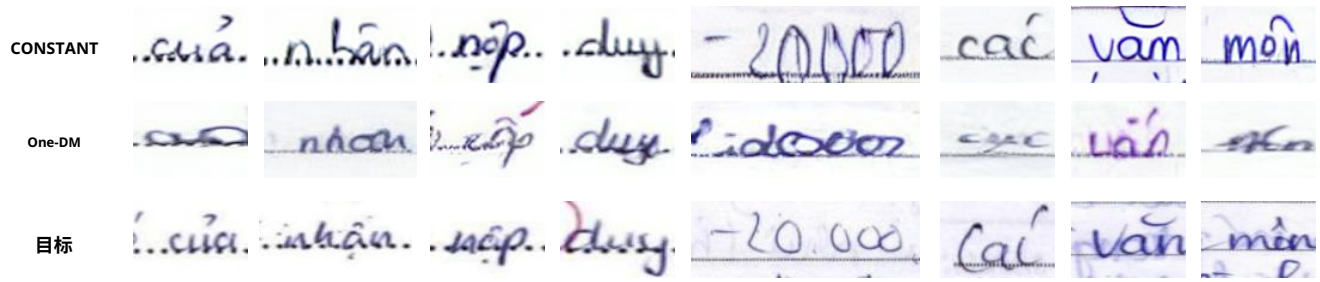


图 11. ViHTGen 数据集上的其他示例。

[14] Haibin He、Xinyuan Chen、Chaoyue Wang、Juhua Liu、Bo Du、Dacheng Tao和Qiao Yu。Diff-Font：用于稳健单样本字体生成的扩散模型。《国际计算机视觉期刊》，132(11):5372–5386，2024。2[15]Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler和Sepp Hochreiter。由双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡。《神经信息处理系统进展》，30，2017。5[16] Jonathan Ho和Tim Salimans。无分类器扩散引导。arXiv预印本 arXiv:2207.12598，2022。6，1[17] Jonathan Ho、Ajay Jain和Pieter Abbeel。去噪扩散概率模型。收录于《神经信息处理系统进展》，第6840–6851页。Curran Associates, Inc.，2020。2，3，1[18] Dong-Jin Hong、Won-Du Chang和Eui-Young Cha。使用带辅助分类过程的去噪扩散概率模型生成手写签名。《应用科学》，14(22):10233，2024。1[19] Kyle Hsu、William Dorrell、James Whittington、Jiajun Wu和Chelsea Finn。通过潜在量化实现解耦。《神经信息处理系统进展》，36:45463–45488，2023。3[20] R Reeve Ingle、Yasuhisa Fujii、Thomas Deselaers、Jonathan Baccash和Ashok C Popat。一种可扩展的手写文本识别系统。收录于2019年国际文档分析与识别会议（ICDAR），第17–24页。IEEE，2019。1[21] Lei Kang、Pau Riba、Yaxing Wang、Marc¸al Rusi˜nol、Alicia Forn´es和Mauricio Villegas。GANWriting：风格化手写单词图像的内容条件生成。收录于计算机视觉—ECCV 2020，第273–289页，Cham，2020。Springer International Publishing。1，2[22] Lei Kang、Pau Riba、Yaxing Wang、Marc¸al Rusinol、Alicia Forn´es和Mauricio Villegas。GANWriting：风格化手写单词图像的内容条件生成。收录于计算机视觉—ECCV 2020：第16届欧洲会议，英国格拉斯哥，2020年8月23日至28日，论文集，第XXIII部分16，第273–289页。施普林格，2020。5，6[23] Dmitrijs Kass和Ekta Vats。AttentionHTR：基于注意力编码器-解码器网络的手写文本识别。收录于文档分析系统国际研讨会，第507–522页。施普林格，2022。3[24] Praveen Krishnan、Rama Kovvuri、Guan Pang、Boris Vassilev, 和 Tal Hassner。Textstylebrush：从单个示例迁移文本美学，2021 5。[25] Xiaoming Li、Wangmeng Zuo 和 Chen Change Loy。学习用于盲文本图像超分辨率的生成结构先验。载于IEEE/CVF计算机视觉与模式识别会议论文集，第 10103–10113 页，2023 年。3[26] I Loshchilov。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101，2017。6[27] Chengzhi Mao、Lu Jiang、Mostafa Dehghani、Carl Vondrick、Rahul Sukthankar 和 Irfan Essa。离散表示增强视觉变换器的鲁棒性。arXiv 预印本 arXiv:2111.10493，2021。3, 4[28] U-V Marti 和 Horst Bunke。IAM 数据库：用于离线手写识别的英语句子数据库。《国际文档分析与识别期刊》，5：39–46，2002。5, 1[29] Alexander Mattick、Martin Mayr、Mathias Seuret、Andreas Maier 和 Vincent Christlein。SmartPatch：利用图块判别器改进手写单词模仿。载于《文档分析与识别—ICDAR 2021》，第 268–283 页，Cham，2021。Springer International Publishing。2[30] Ajoy Mondal、Krishna Tulsyan 和 CV Jawahar。弥合离线英语手写文本识别资源方面的差距。载于国际文档分析与识别会议，第 413–428 页。施普林格，2024。5, 3[31] Alexander Quinn Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob Mcgrew、Ilya Sutskever 和 Mark Chen。GLIDE：利用文本引导扩散模型实现逼真图像生成与编辑。载于国际机器学习会议，第 16784–16804 页。PMLR，2022。1[32]Konstantina Nikolaidou、George Retsinas、Vincent Christlein、Mathias Seuret、Giorgos Sfikas、Elisa Barney Smith、Hamam Mokayed 和 Marcus Liwicki。WordStylist：利用潜在扩散模型生成风格化逐字手写文本。载于国际文档分析与识别会议，第 384–401 页。施普林格，2023。2, 3[33] Konstantina Nikolaidou、George Retsinas、Giorgos Sfikas 和 Marcus Liwicki。DiffusionPen：探索对手写文本生成风格的控制。arXiv 预印本 arXiv:2409.06065，2024。2, 6, 3, 4[34]Konstantina Nikolaidou、George Retsinas、Giorgos Sfikas 和 Marcus Liwicki。重新思考 HTG 评估

：弥合生成与识别之间的差距。arXiv 预印本 arXiv:2409.02683，2024。5, 3[35] Zhiliang Peng、Li Dong、Hangbo Bao、Qixiang Ye 和 Furu Wei。Beit v2：利用向量量化视觉分词器进行掩码图像建模。arXiv 预印本 arXiv:2208.06366，2022。3, 4[36] Vittorio Pippi、Silvia Cascianelli 和 Rita Cucchiara。基于视觉原型的手写文本生成。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 22458–22467 页，2023。1, 2, 6[37] Vittorio Pippi、Fabio Quattrini、Silvia Cascianelli 和 Rita Cucchiara。HWD：一种用于风格化手写文本生成的新型评估分数。arXiv 预印本 arXiv:2310 20316，2023 5 . . [38] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021。4, 2[39] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利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。2, 3, 1[40] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-net：用于生物医学图像分割的卷积网络。载于《医学图像计算与计算机辅助干预—MICCAI 2015：第 18 届国际会议，德国慕尼黑，2015年10月5日至9日，论文集》，第三部分 *18*， 234 241 2015 1, 2 页码 – 。施普林格，。[41] Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。载于国际学习表征会议，2021。6, 1[42] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。重新思考计算机视觉中的 Inception 架构。载于 IEEE 计算机视觉与模式识别会议论文集，第 2818–2826 页，2016。3[43] Raphael Tang、Linqing Liu、Akshat Pandey、Zhiying Jiang、Gefei Yang、Karun Kumar、Pontus Stenetorp、Jimmy Lin 和 Ferhan Ture。DAAM 究竟是什么：利用交叉注意力解释稳定扩散。arXiv 预印本 arXiv:2210 04885，2022 7 . . [44] Yonglong Tian、Dilip Krishnan 和 Phillip Isola。对比式多视图编码。载于《计算机视觉—ECCV 2020：第 16 届欧洲会议，英国格拉斯哥，2020年8月》，第 XI 部分论文集，第 23 16, 776 794 th 页 – 。施普林格，2020。5[45] Aaron Van Den Oord、Oriol Vinyals 等。神经离散表征学习。《神经信息处理系统进展》，第 30 卷，2017。3, 4[46] Bram Vanherle、Vittorio Pippi、Silvia Cascianelli、Nick Michiels、Frank Van Reeth 和 Rita Cucchiara。Vatr++：为手写文本生成谨慎选择词语。《IEEE模式分析与机器智能汇刊》，2024。2 [47] A Vaswani。《Attention is all you need》。神经信息处理系统进展，2017年。2、3[48] Xin Wang、Hong Chen、Si'ao Tang、Zihao Wu 和 Wenwu Zhu。《解耦表示学习》。*IEEE*模式分析与机器智能汇刊，2024年。3[49] Zhenhua Yang、Dezhi Peng、Yuxin Kong、Yuyi Zhang、Cong Yao 和 Lianwen Jin。《FontDiffuser：通过结合多尺度内容聚合与风格对比学习的去噪扩散实现单样本字体生成》。载于*AAAI*人工智能会议论文集，第6603–6611页，2024年2–。[50] Fei Yin、Qiu-Feng Wang、Xu-Yao Zhang 和 Cheng-Lin Liu。《ICDAR 2013中文手写体识别竞赛》。载于*2013*年第*12*届国际文档分析与识别会议，第1464–1470页。IEEE，2013年。7[51] Jan Zdenek 和 Hideki Nakayama。《JokerGAN：具有文本行感知能力的高效手写文本生成模型》。载于第*29*届*ACM*国际多媒体会议论文集，第5655–5663页，美国纽约州纽约，2021年。计算机协会。2[52] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。《深层特征作为感知度量的非凡有效性》。载于*IEEE*计算机视觉与模式识别会议论文集，第586–595页，2018年。7[53] Yuanzhi Zhu、Zhaohai Li、Tianwei Wang、Mengchao He 和 Cong Yao。《基于扩散模型的条件文本图像生成》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第14235–14245页，2023年。2