从生成的图像中识别提示的艺术家姓名

Grace Su[1] Sheng-Yu Wang[1] Aaron Hertzmann[2]
Eli Shechtman[2] Jun-Yan Zhu[1] Richard Zhang[2]

[1]Carnegie Mellon University [2]Adobe Research

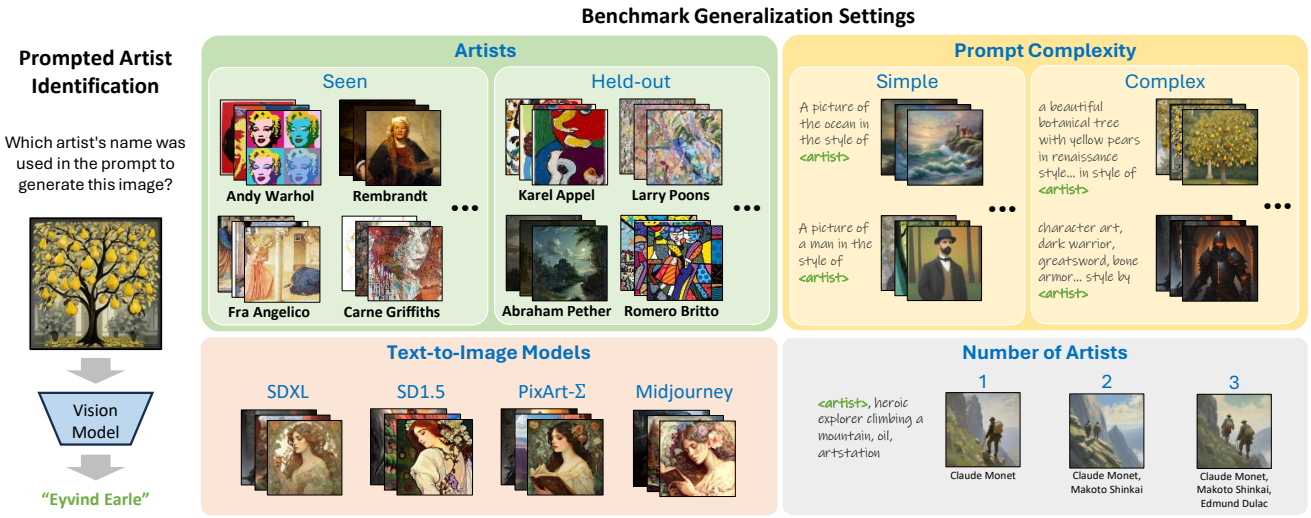

Figure 1. Prompted Artist Identification Benchmark . We introduce the first large-scale benchmark for identifying prompted artist names from generated images. The benchmark covers four axes of generalization that match realistic use cases: (1) Artists: we collect artists commonly used in prompts and simulate open-set artist classification by testing on artists not seen during training. (2) Prompt complexity: users describe images in many different ways, including short, simple prompts as well as more descriptive, complex prompts. (3) Text-to-image models: users can generate images using various text-to-image models, which have different training data and architectures that may affect the generated image's overall style. (4) Number of artists: users may include multiple artists in the prompt to mix styles, creating images that are not easily attributable to a single artist.

Abstract

文本生成图像模型常见且有争议的使用之一是通过明确指定艺术家来生成图片，例如“以 Greg Rutkowski 的风格”。我们引入了一个针对艺术家识别的基准：仅从图像预测哪个艺术家名字在提示中被引用。数据集包含 110 位艺术家的 195 万张图片，并跨越四种泛化设置：留存艺术家、提示复杂性增加、多艺术家提示以及不同的文本生成图像模型。我们评估特征相似性基线、对比风格描述符、数据归属方法、监督分类器和少样本原型网络。泛化模式多样：监督和少样本模型在已知艺术家和复杂提示上表现优异，而风格描述符在艺术家风格明显时转移更好；多艺术家提示仍然是最具挑战性的。我们的基准揭示了显著的潜力，并提供了一个公共测试平台以推进文本生成图像模型的负责任管理。我们发布数据集和基准以促进进一步研究：https://graceduansu.github.io/IdentifyingPromptedArtists 。

1. 引言

目前的文本到图像模型允许用户生成特定艺术风格的图像，通常是通过直接引用艺术家的名字，比如“以某位艺术家的风格”。一些在线艺术作品分享平台禁止上传此类图像 [1, 23] ，因为这可能对艺术家造成潜在伤害 [3, 25, 50] ，并且有可能产生与原始艺术家作品非常相似或无法区分的衍生作品。然而，如果没有访问原始提示信息，就不清楚如何检测此类违规行为。

在这项工作中，我们专注于自动分类在提示中直接使用了哪位艺术家的名字的问题。我们将这个问题称为从生成的图像中识别“提示的艺术家”。为了解决这

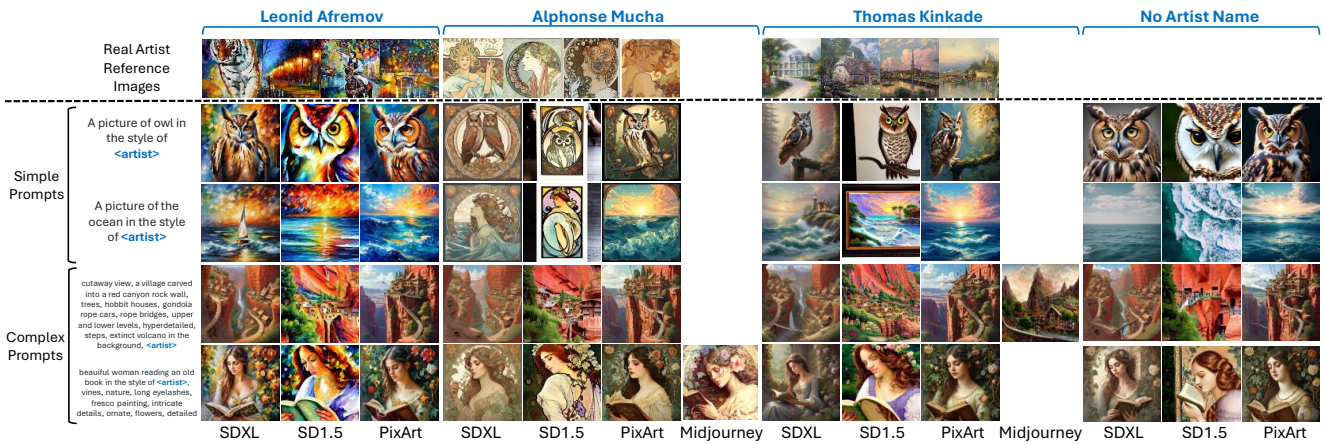

Figure 2. 提示画家识别数据集。我们构建了一个包含 195 万张图像的结构化数据集，用于基准测试不同方法从生成的图像中预测提示的艺术家名称。为了帮助分析给定艺术家名称的提示效果，我们查询一个文本到图像的模型，给出相同的内容提示（行），但插入不同的艺术家名称（列）。我们的数据集由 SDXL [58] 、SD1.5 [61] 、PixArt- Σ [11] 和 Midjourney [51] 生成的图像组成，包括复杂和简单的提示。每个艺术家的风格在提示变得更加复杂时往往会变得不太显著，特别是当提示要求额外的风格和形容词或指定内容可能不在艺术家作品的内容分布中（例如，第 2 行，“一个雕刻在红色峡谷岩壁上的村庄”）。艺术家风格在不同提示和模型中的可见性差异表明了提示艺术家识别任务的难度。

个问题，我们引入了一个大规模的基准测试，包含 195 万张标记的图像。我们的基准测试旨在评价跨越四个轴的泛化能力，这四个轴反映了在现实中使用文本生成图像模型以现有艺术家风格生成图像的应用，如图 1 所示。

(1) Artists. The set of artists that may be included in a prompt is open-ended. Any given model is likely to encounter images generated with artists that were not included in its training set, so we create a set of held-out artists in the benchmark.

(2) Prompt complexity: Users construct prompts in a variety of ways, from short, simple prompts to long, complex prompts that may specify additional styles, reducing the visibility of the artist's style in the resulting image. Thus, we manually design simple prompts (e.g., "picture of the ocean in the style of..."), along with long, complex prompts scraped from databases of real use-cases [77]. We then hold out a set of test prompts from the training set to evaluate generalization across unseen prompts and content. To reduce the association of a given artist name with specific content, we query the model using the same content prompt, with different artist names inserted into the prompt.

(3) Text-to-image models: Users can choose from a large selection of popular image generators. The complex interaction of the network architecture, learning algorithm, training images, and input prompts leads to varying representations of the same artist's style that may be challenging to identify. We collect training and evaluation set images from commonly-used models: SDXL [58], SD1.5 [61], PixArt- Σ [11], and Midjourney [51].

(4) Number of artists per prompt: Users may include multiple artists in one prompt, creating combined styles that can be challenging to associate with any one artist. To study these cases, we create subsets for images generated with 2 and 3 artists in our benchmark.

We evaluate a wide range of methods on our benchmark: feature similarity methods, contrastive style descriptors, data attribution methods, supervised classifiers, and few-shot prototypical networks. The degree of generalization varies across different methods. Our benchmark reveals substantial room for improvement across all generalization settings evaluated. Supervised and few-shot models trained on images generated with prompted artists perform better on seen artists and complex prompts. However, style descriptors trained on real artwork generalize better on simple prompts, where the artist's style is more apparent. We find that capturing and recognizing the representation of prompted artists, as learned and expressed by a generative model, is a related, yet distinct problem from style recognition of real artwork. Meanwhile, prompts referencing multiple artists continue to pose the greatest challenge. We release the benchmark and dataset to help advance the responsible moderation of AI-generated content at https://graceduansu.github.io/IdentifyingPromptedArtists.

2. 相关工作

生成图像检测。虽然生成的图像越来越逼真，但它们仍然留下可检测的痕迹，而这可以回溯到被操控的人脸 [18, 30, 32, 43, 62, 94] 、GANs [9, 48, 49, 53, 79, 88, 89] ，以及现在的扩散模型 [8, 15, 16, 20, 26, 54, 56, 70, 82, 93] 。尽管有这些进展，但对自然界中扰动的鲁棒性以及对未来生成器的泛化仍然是一个持续的挑战 [8, 79] 。在这里，我们研究一个不同的方面：我们不仅试图检测某物是否被生成，还研究导致其生成的属性，即是否有针对性地复制某个艺术家的风格。

数据归因。虽然大多数生成的图像是“新颖的”，并不是训练图像的精确复制，但所有图像都反映了训练集的某些特定元素。数据归因试图将合成图像与影响它的组成训练图像联系起来 [22, 80] ，假设训练图像集和训练过程是已知的，即一个封闭的世界。然而，对于我们的艺术家姓名分类任务，我们可能不了解训练集和学习算法。此外，现有的数据归因算法计算开销较大 [14, 22, 24, 28, 35, 36, 81, 92] 。相比之下，我们的基准测试重点是在艺术家姓名的影响是直接和有意的情况下寻找案例。我们展示了风格描述符和原型分类器在提示艺术家识别方面优于数据归因方法。

风格相似性。艺术风格分类方法通常是在真实艺术作品上训练的，提取或学习在相同风格图像中相似的图像特征 [31, 47, 64, 65, 67, 76] 。一些最新的研究还通过测量生成图像与艺术家原作之间的风格相似性，以确定生成的图像是否复制了艺术家的风格 [7, 37, 46, 52, 75] 。然而，when images are generated from complex prompts that obscure the artist's style, style similarity methods become less effective than a classifier trained directly on our dataset.

生成的图像-文本数据集。Many large-scale datasets containing paired prompts and generated images have been released, including DiffusionDB [83], JourneyDB [77], and TWIGMA [12], text-to-image model evaluation benchmarks [4, 27, 41, 66, 87], and user preference alignment datasets [13, 34, 78, 85]. Other efforts focus on user-generated art from text-to-image models and provide richer stylistic variety [71, 84, 91]. Yet none of these datasets targets images explicitly prompted with artist names. The closest work, by Leotta et al. [42], represents an early attempt to tackle the challenging problem of inferring artist names from generated images. They offer a dataset of 8,519 DALL · E 2 [60] images covering five artists. In contrast, we construct a large-scale benchmark of 1.95M images, which spans hundreds of artists, multiple generative models, diverse prompts, and varying numbers of artists, enabling comprehensive evaluation of open-set artist-name recognition.

3. 提示的艺术家识别基准

Our goal is to create a benchmark evaluating the generalization of prompted artist identification methods across four relevant axes to cover the various images generated from typical prompts from real-world text-to-image model users. Specifically, the benchmark's dataset includes seen vs. held-out artists (Section 3.1), prompt types of varying complexity (Section 3.2), different text-to-image models (Section ??), and different numbers of artists per prompt (Section 3.3). We refer to images generated with prompts invoking one or more artist names as "artist-prompted images."

3.1. 已见和未见艺术家

In the real world, users can reference an open-ended and continuously growing set of artists in image generation prompts, increasing the likelihood that a prompted artist identification model will encounter names not seen during training. While the training set may cover the most frequently used artists, the ability to generalize to held-out artists without retraining remains important. To ensure the benchmark includes the most relevant artists for real-world applications, we collect a list of artists most commonly used in text-to-image model prompts, then designate a set of held-out artists to evaluate how well vision models generalize to unseen artists.

具体而言，我们手动去重并筛选出 110 位艺术家，这些艺术家来自一个初始名单，其中包括在 Lexica.art 网站 [68, 75] 上由 Stable Diffusion 用户频繁提示的 400 位艺术家。对于这些艺术家，我们从 LAION-Styles [75] 中获得了 10k 张真实参考图像，LAION-Styles 是 LAION-5B [69] 的一个经过筛选的子集，专注于艺术风格归属的研究。为了评估不同方法对不同艺术家的泛化能力，我们使用 100 位艺术家的名字进行训练（“已见”），用 10 位不同的艺术家进行测试（“未见”）。因此，我们的真实艺术家图像数据集包括来自已见艺术家的 9907 张图像和来自未见艺术家的 860 张图像。We include the artist list, curation details, and artist name frequencies in our LAION subset in the supplement.

3.2. 变化的提示复杂性

文本到图像模型具有根据不同长度和复杂度的文本提示生成图像的灵活性。直观上，对于较短的提示，艺术家的名字在生成图像的风格中有更大的影响力，而较长且更复杂的提示可能会稀释这种效果。因此，我们研究了在以下两种提示类型中分类性能的差异。

简单的提示语。我们以“ 一幅 `<content>` 的画，风格为 `<artist>` ”的形式使用简单的提示语生成图像。为了多样性，我们使用从 ChatGPT [10] 采样的 500 种不同内容。图 2 展示了生成图像的例子；它们通常遵循一致的艺术风格。

复杂提示。在现实情况下，用户经常使用更复杂的提示生成图像，例如“ 双鱼座符号作为猫，采用阿尔丰斯 · 穆夏

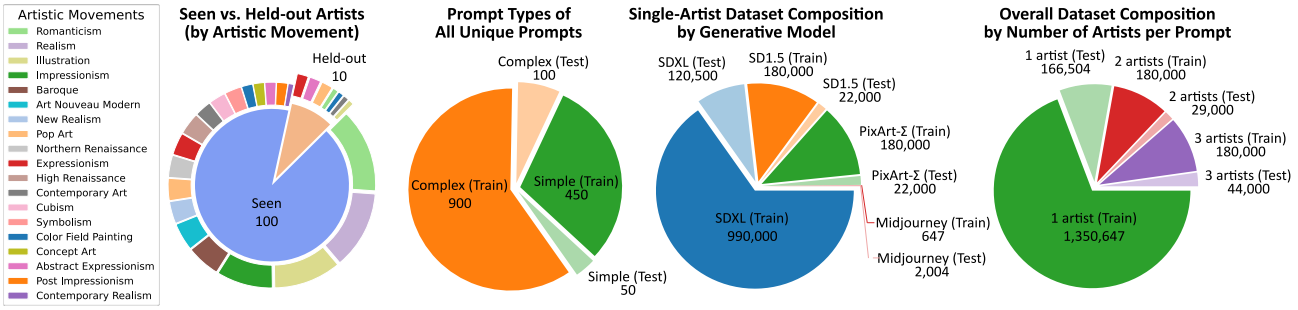

Figure 3. 数据集统计。提示的艺术家识别基准使用一个结构化的数据集，总共包含 195 万张图像，以评估视觉方法在四个广义轴上的性能。我们收集了 110 名被最频繁提示的艺术家，将其分为 100 名已知艺术家和 10 名保留艺术家（1st 图）。接下来，我们收集了 1000 条复杂提示和 500 条简单提示，其中插入了艺术家名称（2nd 图）。对于已知艺术家，我们使用一个单独的提示集进行测试。对于保留艺术家，我们进一步划分测试提示，生成在推断过程中使用的一组参考图像。然后，我们利用 SDXL [58] 、SD1.5 [61] 和 PixArt- Σ [11] 生成单个艺术家提示的图像，并收集 Midjourney 图像 [51] （3rd 图）。最后，我们通过生成带有 2 名艺术家和 3 名艺术家的 SDXL 图像的数据集，来评估方法在提示中泛化到多个艺术家的能力（4th 图）。

艺术风格，真实感 3D 渲染，注重细节，超现实主义 。” 上述提示取自 JourneyDB [77] ，这是一个包含真实用户提交给 Midjourney [51] 的文本提示的数据集。我们从 JourneyDB 中收集了 1000 个文本提示，每一个都提到了一位艺术家的名字。然后，我们用数据集列表中的 110 位艺术家中的一位替换掉艺术家的名字。例如，我们数据集中的前一个文本提示将会是 “双鱼座作为一只猫 <artist> 艺术风格真实的 3D 渲染注重细节超现实主义 。” 在图 2 中，我们观察到与简单提示相比，生成图像中艺术家的风格变得不那么突出，使得分类问题变得更加具有挑战性。

泛化到看不见的提示。为了评估不同艺术家分类方法对不同提示的泛化能力，我们使用 450 个简单提示和 900 个复杂提示进行训练，并保留 50 个简单提示和 100 个复杂提示用于测试。We include the list of prompts and processing procedure in the supplement.

策划生成器和图像。近年来，许多文本到图像的生成模型被开发出来，其开放源码的访问程度各不相同 [29, 41, 44, 45, 90] ，随着该领域的进步，更多的模型继续被发布。生成的图像可能来自这些模型中的任何一个，每个模型根据其训练数据和架构以不同的方式呈现提示和艺术家名称。因此，我们的基准测试还评估了提示艺术家识别方法在不同文本到图像模型中的泛化能力。为了策划一个以原始生成提示标记的大型结构化图像数据集，我们选择了近期流行的模型，这些模型 1）能够在直接提示艺术家名称时生成艺术风格，2）拥有开源权重或一个发布的大型图像-提示对数据集。

我们尝试使用最近的 SD3.5 [21] 和 Flux [39] 开源模型生成一些带有艺术家名字的图像作为提示，但发现即使提示很简单，使用艺术家名字对生成的图像风格通常没有什么效果。这可能是因为模型的训练数据集已经被过滤以排除有问题的内容 [2] 。虽然如果用户写出描述风格的提示，模型可能能够复制目标风格，但它们对艺术家名字的响应不如早期模型，如 SDXL [58] 。We show examples of their generated images in the supplement.

因此，对于每个艺术家和提示组合，我们使用开放权重模型 SDXL [58] 、SD1.5 [61] 和 PixArt- Σ [11] 以 2 个种子生成图像。对于使用复杂提示生成的 SDXL 图像，我们使用 10 个不同的生成种子。为了从 Midjourney [51] （一种闭源模型）收集图像，我们过滤 JourneyDB 数据集 [77] 以获得我们设定的已观察和保留的艺术家集。结果数据集由图 3 总结。Full dataset statistics tables are included in the supplement.

3.3. 每个提示多位艺术家

文本到图像模型用户不仅会提示单个艺术家的风格，还会通过在提示中包含多个艺术家名字来混合多位艺术家的风格。例如，在从 JourneyDB [77] 中收集的 Midjourney 用户的 10,000 个提示的随机样本中，我们发现 10.8 % 的提示包含多个艺术家名字，而包含 1 个艺术家名字的占 14.7 %，不包含任何艺术家名字的占 74.5 %。其分布在图 4 中可视化，呈长尾分布且严重偏斜，其中包含 2 个艺术家名字的提示占 4.8 %，包含 3 个艺术家名字的占 2.5 %，而包含 4 个或更多艺术家名字的提示占比为 1.3 % 或更少。

因此，我们的基准主要包含单艺术家提示的图像用于评估。然而，我们还通过生成一个包含 2 位艺术家提示的图像数据集，以及一个包含 3 位艺术家提示的图像数据集来评估提示的艺术家识别方法在多艺术家情境下的泛化能力。

生成多个艺术家风格提示的图像。为了策划一个多艺术家风格提示图像的数据集以评估在保留艺术家和不同提示复杂性上的泛化，我们稍微修改了 3.2 节中描述的程序。我们将多个艺术家名字插入到提示模板中，e.g.，“ 双鱼座星座作为一只猫，采用 <artist 1> 和 <artist 2> 艺术风格的真实 3D 渲染，注重细节，超级逼真 ”，对于 2 位艺术家和 3 位艺术家，随机抽取 100 个已见艺术家组合以及所有可能的保留艺术家组合，然后使用 SDXL

生成图像。第四个图表总结了多艺术家数据集，见图 3 。

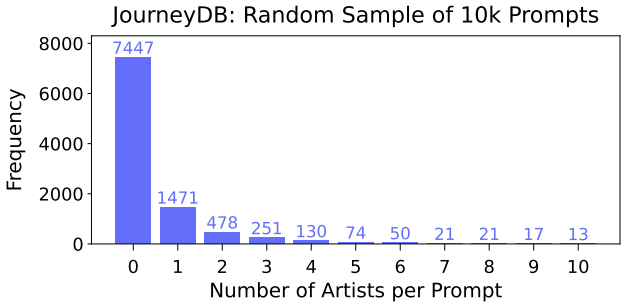

Figure 4. 关于每个提示中的艺术家人数的统计。为了理解文本到图像模型用户多频繁地提示多个艺术家，我们从 Midjourney 用户的 JourneyDB [77] 中随机抽样了 10,000 个提示，并可视化每个提示中艺术家人数的分布。虽然只有 14.7 % 的提示调用一个艺术家的名字，4.8 % 含有 2 个艺术家的名字，2.5 % 含有 3 个艺术家的名字。

在图 2 中，我们定性观察到，当提示变得更复杂时，生成的图像与艺术家原作风格的对齐程度降低。此外，在相同的提示下，生成图像的风格会根据所使用的文本到图像模型的不同而有所变化。因此，我们对 CLIP [59] 图像相似性得分进行简单的定量分析，发现有三个方面使得针对艺术家的提示识别任务在泛化上具有挑战性：1）当提示更复杂时，提示中艺术家名字对生成图像的影响被稀释。2）生成图像与艺术家平均艺术作品的对齐程度在不同的文本到图像模型之间存在差异。3）随着提示中艺术家数量的增加，将艺术家添加到提示中的效果减弱。

比较文本到图像模型的风格对齐。不同的文本到图像模型有不同的训练数据集和架构，这可能会影响生成图像的整体风格。因此，我们在我们的基准中，通过表格 1a 中比较不同文本到图像模型之间的图像相似性来比较文本到图像模型。对于每个以艺术家为提示的图像，我们计算其与使用相同种子和提示生成的图像之间的 CLIP 图像相似度，只不过去掉了艺术家的名字。对于所有模型，我们观察到，对于简单提示的 CLIP 相似度分数低于复杂提示，这表明在提示简单的情况下，艺术家名字对生成图像的影响更大。与 SDXL 和 SD1.5 相比，艺术家名字的提示在使用 PixArt 时对生成图像的影响也较小。

我们还计算了艺术家提示图像嵌入与真实艺术家图像平均嵌入之间的相似性，这就是艺术家的原型。当提示更复杂时，生成的图像与艺术家的真实风格的契合度较低。此外，PixArt 图像与艺术家真实风格的契合度低于 SDXL 和 SD1.5。

在提示中添加艺术家名字的影响。由于图像生成提示可能包含多个艺术家名字，我们研究了在提示中增加艺术家名字的数量的效果，以及识别所有艺术家的难度。在表格 1b 中，我们展示了随着提示中添加更多艺术家名字，相同种子和提示生成的每个图像之间的相似性增加，无论是对于复杂提示还是简单提示。我们可以在图 5 中定性地观察到每添加一个艺术家后的影响减少。这表明在一个提示中分类多个艺术家比分类单个艺术家更困难，因为随着更多艺术家名字的加入，每个艺术家名字的效应都被稀释了。

我们在提示的艺术家识别基准上评估了一系列具有公开实现的计算机视觉模型，并采用评价程序，在所有模型上使用相同的训练和测试集进行公平比较。

对生成图像进行分类的两种方法是：搜索相似的艺术家参考图像和训练前馈分类器。虽然从在大规模数据集上训练的模型中检索图像特征可以实现对未见过类别的泛化，但与前馈分类器相比，它们需要更多的运行时间和存储。因此，我们比较了几种预训练的基于检索的方法和在我们的艺术家识别数据集上训练的分类器。

- 对比风格描述符（CSD）[75] 是一种测量图像之间风格相似性的风格描述符。CSD 使用监督对比学习在来自 LAION-5B 的 ~500 k 精选真实艺术家图像上进行训练 [33] [69] 。
- DINOv2 [55] ，CLIP [59] ：常用于基于网络图像训练的自监督图像特征。
- 通过定制进行归因（AbC）[80] ：DINO（AbC）和 CLIP（AbC）是用于在定制的文本到图像扩散模型中归因训练数据的特征，分别是从 DINO [6] 和 CLIP [59] 微调而来的。它们是在从根据 ImageNet [17] 和艺术数据集定制的 SD1.5 模型合成的图像上训练的。
- 原型网络：我们基于原型网络 [72] 训练一个分类器，这是一种小样本学习方法，使得可以预测未见过的艺术家。在训练期间，它学习一个图像编码器，在嵌入空间中输出的特征接近于正确艺术家的原型，即艺术家的真实参考图像经过平均处理得到的预训练 CLIP 特征。在推理过程中，可以使用相同的原型特征来预测训练期间看到的艺术家标签集。我们还可以通过使用预留艺术家的参考图像作为原型特征来预测训练中未见过的艺术家。
- 基础分类器：我们在 CLIP 图像编码器上添加一个具有一层隐藏层的 MLP 作为分类头，然后在我们的数据集上训练所有层。

3.4. 评价框架

单一艺术家分类。我们评估每个模型对已知和未出现的艺术家的 top-1 分类准确性。对于每个测试集的我们的提示式艺术家识别基准，每个模型都被给予相同的训练/检索图片集以进行对照比较。训练图片集是从我们的 SDXL、SD1.5、PixArt 和 Midjourney 数据集中生成的所有已知艺术家的训练图片，并包含简单和复杂提示生成的图片。为了评估基于检索的方法对未出现艺术家的效果，我们使用生成的未出现提示的图片作为测试时的参考。对于普通分类器，由于模型不能应用于在训练时未见过的类别，我们只报告已知艺术家的分类结果。为了估计每次评估的统计显著性，我们通过艺术家姓名、提示模板和生成种子进行重采样，对每

(a) 文本到图像模型比较

Model	Content only vs. Artist-prompted		Artist-prompted vs. Real prototype	
	Simple	Complex	Simple	Complex
SDXL	0.571	0.738	0.570	0.476
PixArt	0.632	0.816	0.522	0.441
SD1.5	0.520	0.643	0.590	0.527
Midjourney	–	–	–	0.541

(b) 越来越多的艺术家

Prompt Type	Number of Artists Compared		
	0 vs. 1	1 vs. 2	2 vs. 3
Complex	0.738	0.811	0.866
Simple	0.571	0.711	0.782

Table 1. CLIP 平均图像相似性。(a) 对于每个文本到图像模型，我们计算没有艺术家名字生成的图像与使用一个艺术家名字生成的图像之间的平均 CLIP 相似性（前两列），以及使用艺术家名字生成的图像与该艺术家的真实原型之间的平均 CLIP 相似性（最后两列）。当提示词更复杂时，艺术家的影响和生成图像与艺术家真实风格的对齐度会降低。此外，PixArt 图像与艺术家真实风格的对齐度低于 SDXL 和 SD1.5 图像。(b) 给定相同的提示词和生成种子，我们计算使用 0、1、2 和 3 个艺术家名字生成的 SDXL 图像的平均 CLIP 相似性。在提示词中添加更多艺术家名字时，每个艺术家名字的影响会减弱。

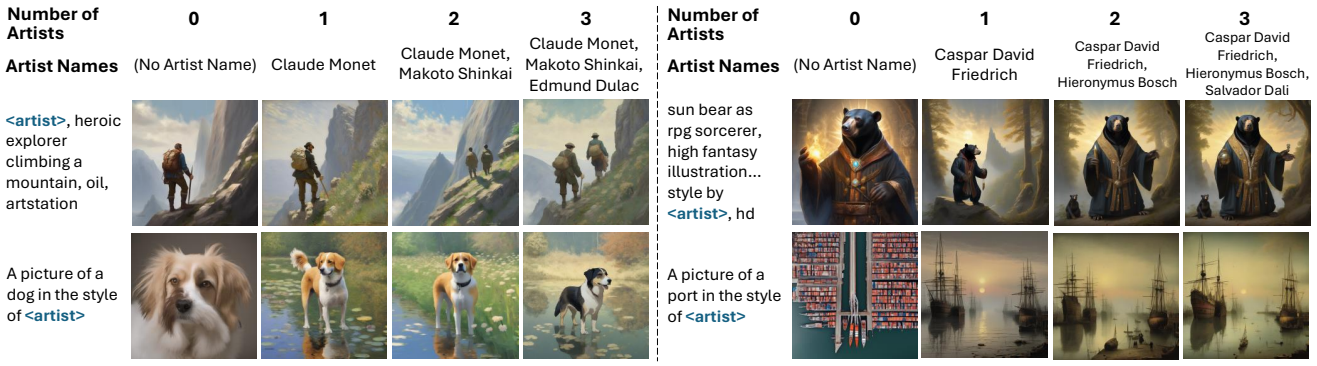

Figure 5. 多位艺术家驱动的图像。在示例图像中，我们观察到当添加更多艺术家的名字到提示词中时，每个艺术家的效果变得不那么明显。每一行展示了一组使用相同提示词和生成种子生成的图像，提示词中的艺术家数量从左到右逐渐增加。对于每个提示词，随着每位新增艺术家的加入，图像变化更小，而使用复杂提示生成的图像通常变化小于使用简单提示生成的图像。

个模型的预测进行引导 Bootstrap 方法，循环 2000 次。We plot our bootstrapping procedure's convergence at 2000 iterations in the supplemental material.

多艺术家分类。为了评估在多艺术家名称提示的图像上的艺术家分类性能，我们使用多标签分类指标：前 10 个唯一检索标签的排名平均准确率（mAP@10）。我们通过使用标签平滑的二元交叉熵损失，将原型网络的训练调整为多标签目标。训练数据集包括来自 SDXL 的所有单艺术家图像，以及所有 2 艺术家和 3 艺术家训练图像。为了评估基于检索的方法，我们仅使用单艺术家 SDXL 图像作为参考数据库。

我们分析了在我们基准的所有泛化设置下，各种视觉表现方法的提示艺术家分类性能。

3.5. 单一艺术家分类

我们在图 6 中呈现了所有方法在预设艺术家分类任务上的定量结果。虽然所有方法都超过了随机机会，但没有方法的准确率超过 91 %，这强调了该任务的内在困难。当提示的复杂度增加，以及在评价 PixArt 生成的图像时，与 SDXL 和 SD1.5 生成的图像相比，性能会持续下降。这表明，提示的复杂性和用于图像生成的文本到图像模型显著影响了分类的难度。

在所有泛化任务中，CSD、原型网络和基础分类器是性能最好的方法，其次是 CLIP、AbC 模型，最后是 DINOv2。这表明对于提示艺术家识别，风格描述符和训练好的分类器比现成的通用视觉表示和数据归属方法更有效。

原型网络在大多数评估数据集上优于 CSD，并且可以从生成的图像中学习艺术家风格的表示，这比在真实艺术作品上训练的 CSD 对复杂提示更具鲁棒性。同时，CSD 对持留的艺术家和简单提示的图像以及带有复杂提示和持留艺术家的 Midjourney 图像的泛化更好。这表明复杂提示生成的图像比简单提示更不符合艺术家的真实风格。

未见文本到图像模型的泛化。我们通过逐步增加训练图像集 [20]，包括来自额外生成器（SDXL → SD1.5 → PixArt → Midjourney）的图像，对表现最佳的方法——原型网络和 CSD——进行未见文本到图像模型的泛化评估。结果如图 7 所示。对于这两种方法，增加训练数据集并不会提升在未见文本到图像模型上的性能——只有当训练集中包含来自该生成器的图像时，性能才能在每个文本到图像模型上得到提升。这表明模型没有学习到可以跨文本到图像模型泛化的风格表示。值得注意的是，即使在训练集中加入 PixArt 图像后，PixArt 图像的性能提升也不明显，支持我们在第 ?? 节中的观察，即 PixArt 图像与艺术家的真实风格相比，较不契合 SDXL 或 SD1.5。

我们进一步评估了所有方法在多艺术家分类任务中

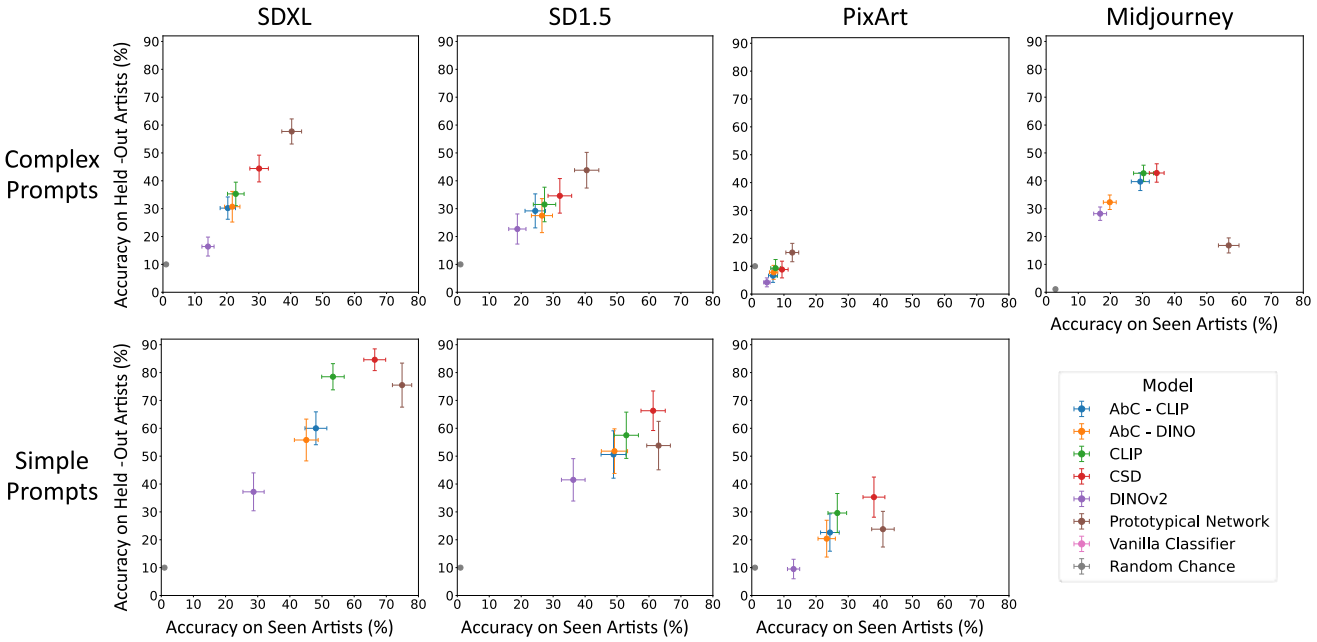

Figure 6. 单艺术家预测结果。我们比较了各种视觉表示方法的提示艺术家分类准确性。我们在已见过的艺术家集合（100 类分类，横轴）和未见艺术家（10 类分类，纵轴）上进行测试。我们还测试了使用不同文本到图像模型（SDXL、SD1.5、PixArt 和 Midjourney）生成的图像，以及复杂和简单的提示语。虽然原型网络 [72] 、训练过的普通分类器，以及 CSD [74] 超过了 CLIP [59] 、DINOv2 [6] 和 AbC [80] 方法，但在所有场景中达到高准确率仍然很困难。We include all numbers as tables in the supplemental material.

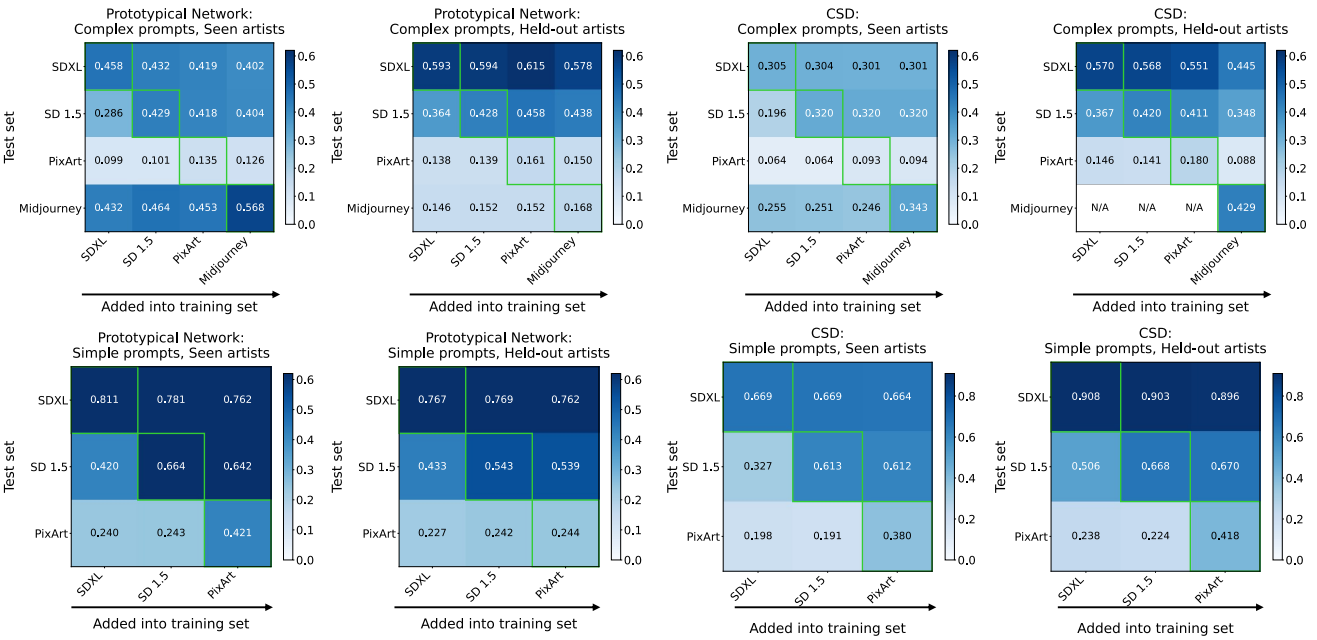

Figure 7. 对未知生成器的泛化。在表现最佳的方法中，原型网络和 CSD 方法，我们通过逐步增加训练图像集来评估在未知文本到图像模型上的表现。绿色对角线下方的单元格评估对未知领域的泛化。对于这两种方法，扩展训练数据集并不能提升在未知文本到图像模型上的表现；只有当训练数据包含由正在评估的特定模型生成的图像时，才会有性能提升。

的表现，其中输入图像由多个艺术家名称作为提示，并在图 8 中展示了结果。总体来看，相较于使用简单提示生成的图像，使用复杂提示生成的图像其性能往往会下降。在多艺术家数据集上，原型网络表现最好，这

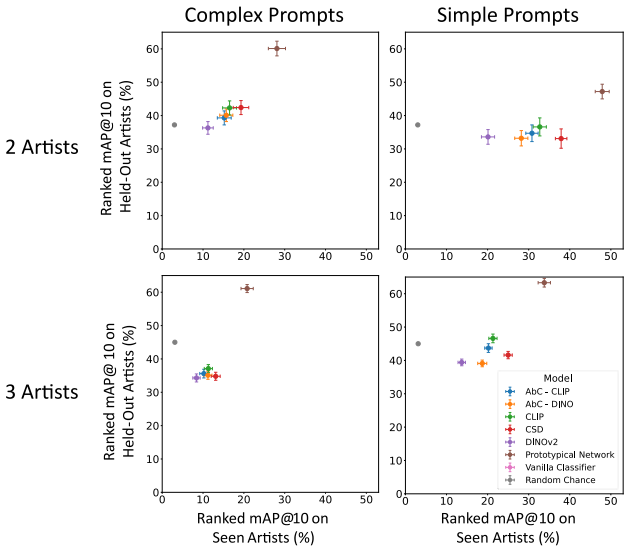

Figure 8. 多艺术家预测结果。我们在多艺术家分类任务中评估视觉表示方法，其中输入图像由多个艺术家的名字提示。我们测试使用 2 位艺术家和 3 位艺术家作为提示生成的 SDXL 图像，并报告已见艺术家（x 轴）和未见艺术家（y 轴）的排名 mAP@10。与简单提示生成的图像相比，所有方法在从复杂提示生成的图像上的表现通常有所下降。正如预期的那样，原型网络在所有数据集上的表现最高，因为它是在多个艺术家提示的图像上进行训练的。然而，其表现远未达到饱和。We include all numbers as tables in the supplemental material.

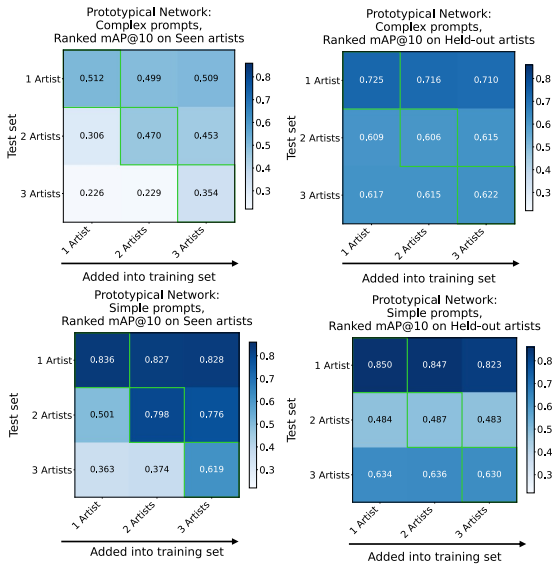

Figure 9. 对增加提示的艺术家数量的泛化。对于原型网络，我们通过逐步增加训练集的图像以包含多艺术家提示的图像来评估在提示中未见过的艺术家数量上的性能。绿色对角线以下的单元格评估对未见领域的泛化。增加训练集的大小并不能改善对未见艺术家数量的泛化。虽然使用与测试集相同数量艺术家的训练图像可以提高对已知艺术家的分类，但对未见艺术家的性能没有好处。

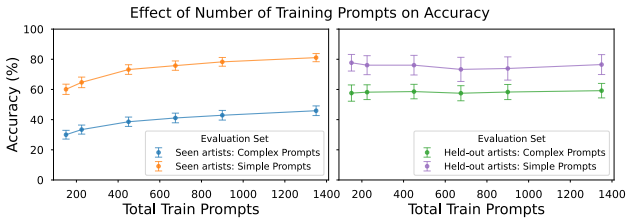

Figure 10. 关于训练提示数量的消融实验。我们绘制了典型网络在已知艺术家（左侧）和保留艺术家（右侧）上的分类准确率，作为训练集中提示数量的函数。随着训练提示的增多，已知艺术家的分类准确率稳步提升，但保留艺术家的表现在所有数据集大小上保持不变。典型网络能够学习一种艺术家风格的表示，这种表示对已知艺术家的未见提示可以泛化，但不能对保留艺术家的提示进行泛化。

是预料之中的，因为它是唯一在多艺术家提示图像上进行训练的方法，但其性能仍远未达到饱和状态。尽管未在多艺术家提示图像上训练，基础分类器是表现第二好的方法，其次是 CSD。其余基于检索的方法表现与 CSD 相似。

对增加数量的提议艺术家进行泛化。我们还研究了在图 9 中将多艺术家提示图像添加到原型网络的训练集中的影响。扩展训练集并不会导致对未见过的艺术家数量更好的泛化。用包含与测试集相同数量艺术家的图像进行训练可以改善已见艺术家的分类，但不能改善保留出的艺术家的分类。这表明，即使在多艺术家提示图像上进行训练，原型网络也无法学到一种能够泛化到未见过的艺术家数量和保留出的艺术家的风格表示。第 ?? 节中的图像相似性分析显示，用多个艺术家生成的图像彼此之间更为相似，这可能是导致缺乏泛化能力的原因。

训练数据集消融实验。为了评估训练数据集大小的影响，我们在将简单提示与复杂提示的比例固定为 1:2 的情况下，将原型网络的训练提示数从 150 变化到 1350。正如图 10 所示，随着训练提示数量的增加，对已知艺术家的分类准确性持续提高。相反，对未见过的艺术家的表现基本保持不变。这表明，虽然在更多独特提示上训练原型网络改善了模型对已知艺术家风格的表示，但它并不能提高对未知艺术家的泛化能力。

艺术家姓名检测。预测一幅图像是否完全以艺术家姓名为提示是另一个相关任务，这对于生成内容的审核应用非常有用。因此，我们对这一二元分类任务评估了几种表现优异的方法。我们从经过艺术家提示的 SDXL 子集中构建一个数据集，并用不包含艺术家姓名的提示来增强它，并如表 2 所示平衡测试集。

我们在 CLIP、CSD 和原型网络表示上以及一个完全微调的 CLIP 模型上评估线性探测分类器。表 3 中的结果显示，虽然所有方法都超过了随机概率，但即使经过检测任务的微调，仍没有任何方法达到完美的准确率。在简单提示上的性能始终较高，这与我们在表 ?? 中的观察一致，即简单提示在内容仅限和艺术家提示的图像之间产生更大的特征差异——这使得它们

比复杂提示更容易区分。值得注意的是，完全微调的 CLIP 模型在测试集上的表现不佳，这表明其易于过拟合。这可能是因为该任务本质上更困难，具有更高的类内变异性，因为模型必须学会区分被艺术家姓名提示的图像和未被提示的图像，而在每个类别中，视觉相似度都很低。

检测非公有领域艺术家。触发艺术家识别的一个重要下游应用是检测不当使用，例如基于受版权保护的艺术家生成衍生作品。我们通过将 SDXL 触发的艺术家识别数据集分为两个元标签：公有领域或非公有领域（根据去世 95 年来定义），来评估各种方法在检测非公有领域艺术家姓名任务中的表现。我们在 100 种已知艺术家的测试集上进行测试，其中包含 55 位公有领域艺术家和 45 位非公有领域艺术家，并在表 4 中报告表现情况。所有方法均优于随机概率，但训练的分类方法——原型网络和普通分类器——表现最佳。这表明在给定的艺术家封闭集合中，通过艺术家触发的图像进行训练可以改善非公有领域艺术家的检测。

		Train		Test	
		Content only	Artist-prompted	Content only	Artist-prompted
Complex Prompts	# Seen artists	–	100	–	5
	# Held-out artists	–	0	–	5
	# Prompts	900	900	100	100
	# Seeds	10	10	10	1
Simple Prompts	# Seen artists	–	100	–	5
	# Held-out artists	–	0	–	5
	# Prompts	450	450	50	10
	# Seeds	2	2	2	1
	# Images	9,900	990,000	1,100	1,100

Table 2. 艺术家姓名检测数据集。为了评估不同方法在检测一张图像是否通过艺术家姓名提示生成的任务上，我们构建了一个数据集，其中包含通过艺术家姓名提示生成的 SDXL 图像以及没有艺术家姓名但内容相同的图像。训练集包含了我们基准中所有通过艺术家提示生成的 SDXL 图像，测试集则在只有内容和通过艺术家提示这两类图像之间保持平衡。

Method	Complex			Simple		
	Accuracy	AP	AUCROC	Accuracy	AP	AUCROC
Chance	50.0	–	–	50.0	–	–
CLIP - linear probe	61.6	79.9	79.5	82.5	97.7	97.9
CSD - linear probe	70.2	81.3	79.6	92.5	98.2	98.3
Proto. Net. - linear probe	70.7	83.6	80.8	91.0	97.3	97.0
Full CLIP fine-tune	67.3	78.6	79.3	84.5	96.4	97.1

Table 3. 艺术家名称检测结果。我们评估了在艺术家名称检测任务中表现最佳的提示艺术家分类方法，并报告了准确率、平均精度（AP）和接收器操作特征曲线下面积（AUCROC）。尽管所有方法都经过微调以用于检测任务，但没有一个方法能实现完美的准确性。在简单提示上性能始终较高，而完全微调的 CLIP 模型在测试集上的表现较差，表明其易受过拟合影响。这可能是因为艺术家名称检测任务本身更具挑战性。

在这项工作中，我们研究了在生成图像中识别提示的艺术家姓名的任务，并证明了这对当前视觉表示方法提出了显著的泛化挑战。我们建立了第一个适合该任务的大规模基准和数据集，包括 195 万张图像，并提示使用了 110 位艺术家的名字。该基准旨在探查四个泛化维度：未涉及的艺术家、不同提示复杂性、不同文本到图像模型，以及涉及多个艺术家的提示。利用这

Method	Complex			Simple		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Chance	45.0	–	–	45.0	–	–
CLIP	64.3 ± 1.6	63.7 ± 5.8	69.4 ± 2.8	81.0 ± 2.0	79.7 ± 4.2	84.0 ± 2.8
CSD	68.3 ± 1.7	67.0 ± 5.6	74.1 ± 2.7	86.5 ± 1.8	85.3 ± 3.4	88.9 ± 2.4
Proto. Net.	74.3 ± 2.0	70.3 ± 5.4	85.2 ± 2.3	91.2 ± 1.6	90.3 ± 2.8	92.8 ± 2.4
Classifier	74.4 ± 1.8	71.9 ± 5.3	81.2 ± 2.6	91.1 ± 1.5	89.9 ± 2.8	92.9 ± 2.1

Table 4. 非公共领域见过的艺术家的准确性。我们评估各种方法在检测提示中非公共领域艺术家名字存在的任务中的表现。我们在 100 个见过的艺术家测试集上进行测试，其中包含 55 个公共领域和 45 个非公共领域艺术家。训练过的分类器方法、原型网络和普通分类器表现最佳，表明在给定封闭艺术家集的情况下，基于生成图像的训练可以提高对非公共领域艺术家的检测。

个基准，我们评估了多样的视觉模型。总之，我们的发现是：这些发现强调了需要具备更好泛化能力的视觉模型来进行提示的艺术家识别。我们发布我们的基准和数据集，以促进能够负责任地管理 AI 生成内容的泛化视觉方法的发展。

局限性。我们将基准测试重点放在四个文本到图像模型上作为起点，但我们承认还有几个其他流行的闭源模型可以进行评估，而我们的 Midjourney [51] 评估规模有限。我们也没有评估使用个性化技术 [38, 63, 86] 生成的图像，这是文本到图像用户复制艺术家风格的另一种常见方式。最后，我们没有评估多模态大型语言模型的性能，因为许多模型缺乏关于艺术家的领域特定知识 [5] 。

致谢。我们感谢 Maxwell Jones 和 Kangle Deng 的有益讨论和建议。此项工作开始于 Grace Su 在 Adobe 实习期间。Grace Su 得到了 NSF 研究生研究奖学金（Grant No. DGE2140739）的支持。本项目部分由 NSF IIS-2403303、Adobe Research 和 Packard Fellowship 支持。

References

[1] Adobe. Updated artist name guidelines, 2023. Accessed: 2024-11-13.

[2] Stability AI. Stable safety, 2025. Accessed on April 16, 2025.

[3] Sarah Andersen. The alt-right manipulated my comic. then a.i. claimed it. The New York Times, 2022.

[4] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 20041–20053, 2023.

[5] Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. Gallerygpt: Analyzing paintings with large multimodal models. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 7734–7743, 2024.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand

Joulin. Emerging properties in self-supervised vision transformers. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[7] Stephen Casper, Daniel Filan, Gabriel Nicholas, Miles Brundage, and Matthew L. Schwartz. Measuring the success of diffusion models at imitating human artists. In Workshop on Generative AI and Law (GenLaw), International Conference on Machine Learning (ICML), 2023.

[8] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fake-Inversion: Learning to detect images from unseen models by inverting stable diffusion. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[9] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In European Conference on Computer Vision (ECCV), 2020.

[10] ChatGPT. https://chat.openai.com/chat, 2024.

[11] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In European Conference on Computer Vision (ECCV), 2024.

[12] Yiqun Chen and James Y Zou. Twigma: A dataset of ai-generated images with metadata from twitter. Conference on Neural Information Processing Systems (NeurIPS), 36:37748–37760, 2023.

[13] Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7727–7736, 2024.

[14] Sang Keun Choe et al. What is your data worth to gpt? llm-scale data valuation with influence functions. arXiv preprint arXiv:2405.13954, 2024.

[15] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

[16] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255. Ieee, 2009.

[18] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397, 2020.

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[20] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In IEEE International Conference on Computer Vision (ICCV) Workshop, 2023.

[21] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning, 2024.

[22] Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. arXiv preprint arXiv:2312.06205, 2023.

[23] Getty Images. Ai generation faqs, 2024. Accessed: 2024-11-13.

[24] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. arXiv preprint arXiv:2308.03296, 2023.

[25] Melissa Heikkilä. This artist is dominating ai-generated art. and he' s not happy about it. MIT Technology Review, 2022.

[26] Yan Hong, Jianfu Zhang, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, and Weiqiang Wang. Wildfake: A large-scale challenging dataset for ai-generated images detection. arXiv preprint arXiv:2402.11843, 2024.

[27] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[28] Masaru Isonuma and Ivan Titov. Unlearning traces the influential training data of language models. arXiv preprint arXiv:2401.15241, 2024.

[29] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. In Conference on Neural Information Processing Systems (NeurIPS), 2024.

[30] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[31] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. arXiv preprint arXiv:1311.3715, 2013.

[32] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In Conference on Neural Information Processing Systems (NeurIPS), 2021.

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Conference on Neural Information Processing Systems (NeurIPS), 2020.

[34] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. Conference on Neural Information Processing Systems (NeurIPS), 36:36652–36663, 2023.

[35] Myeongseob Ko, Feiyang Kang, Weiyan Shi, Ming Jin, Zhou Yu, and Ruoxi Jia. The mirrored influence hypothesis: Efficient data influence estimation by harnessing forward passes. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[36] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In International Conference on Machine Learning (ICML), 2017.

[37] Anand Kumar, Jiteng Mu, and Nuno Vasconcelos. Introstyle: Training-free introspective style attribution using diffusion features. arXiv preprint arXiv:2412.14432, 2024.

[38] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[39] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[40] LAION.ai. Safety review for laion 5b, 2023. Accessed on November 21, 2024.

[41] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. Conference on Neural Information Processing Systems (NeurIPS), 36:69981–70011, 2023.

[42] Roberto Leotta, Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Not with my name! inferring artists' names of input strings employed by diffusion models. In International Conference on Image Analysis and Processing, pages 364–375. Springer, 2023.

[43] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[44] Zhikai Li, Xuewen Liu, Dongrong Joe Fu, Jianquan Li, Qingyi Gu, Kurt Keutzer, and Zhen Dong. K-sort arena: Efficient and reliable benchmarking for generative models via k-wise human preferences. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[45] Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative ai: open-washing and the eu ai act. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pages 1774–1787, 2024.

[46] Chenxi Liu, Towaki Takikawa, and Alec Jacobson. A lora is worth a thousand pictures. arXiv preprint arXiv:2412.12048, 2024.

[47] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In Proceedings of the 25th ACM international conference on Multimedia, pages 1183–1191, 2017.

[48] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018.

[49] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019.

[50] Louise Matsakis. Artists rage against machines that mimic their work. Wired, 2023.

[51] Midjourney. https://midjourney.com, 2024.

[52] Mazda Moayeri, Samyadeep Basu, Sriram Balasubramanian, Priyatham Kattakinda, Atoosa Chegini, Robert Brauneis, and Soheil Feizi. Rethinking copyright infringements in the era of text-to-image generative models. In International Conference on Learning Representations (ICLR), 2025.

[53] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. In Electronic Imaging, 2019.

[54] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024.

[56] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[57] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 14532–14542, 2022.

[58] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In

International Conference on Learning Representations (ICLR), 2024.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), 2021.

[60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[62] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In IEEE International Conference on Computer Vision (ICCV), 2019.

[63] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 22500–22510, 2023.

[64] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. Aladin: All layer adaptive instance normalization for fine-grained style similarity. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[65] Matthia Sabatelli, Mike Kestemont, Walter Daelemans, and Pierre Geurts. Deep transfer learning for art classification problems. In Proceedings Of The European conference on computer vision (ECCV) workshops, 2018.

[66] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Conference on Neural Information Processing Systems (NeurIPS), 35:36479–36494, 2022.

[67] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855, 2015.

[68] Gustavo Santana. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts, 2024.

[69] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Conference on Neural Information Processing Systems (NeurIPS), 2022.

[70] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, 2023.

[71] Ravidu Suien Rammuni Silva, Ahmad Lotfi, Isibor Kennedy Ihianle, Golnaz Shahtahmassebi, and Jordan J Bird. Artbrain: An explainable end-to-end toolkit for classification and attribution of ai-generated art and style. arXiv preprint arXiv:2412.01512, 2024.

[72] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. Conference on Neural Information Processing Systems (NeurIPS), 2017.

[73] Artists Rights Society. Artists rights 101. Artists Rights Society, 2024.

[74] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[75] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Investigating style similarity in diffusion models. In European Conference on Computer Vision (ECCV), 2024.

[76] Gjorgji Strezoski and Marcel Worring. Omniart: a large-scale artistic benchmark. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(4):1–21, 2018.

[77] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. Conference on Neural Information Processing Systems (NeurIPS), 2024.

[78] Kailas Vodrahalli and James Zou. Artwhisperer: A dataset for characterizing human-ai interactions in artistic creations. In International Conference on Machine Learning, pages 49627–49654. PMLR, 2024.

[79] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot···for now. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[80] Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In IEEE International Conference on Computer Vision (ICCV), 2023.

[81] Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Data attribution for text-to-image models by unlearning synthesized images. In Conference on Neural Information Processing Systems (NeurIPS), 2024.

[82] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire

for diffusion-generated image detection. In IEEE International Conference on Computer Vision (ICCV), 2023.

[83] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In The Annual Meeting of the Association for Computational Linguistics (ACL), 2023.

[84] Yankun Wu, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In Proceedings of the 2023 ACM International conference on multimedia retrieval, pages 199–208, 2023.

[85] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. Conference on Neural Information Processing Systems (NeurIPS), 36: 15903–15935, 2023.

[86] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

[87] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research, 2022.

[88] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In IEEE International Conference on Computer Vision (ICCV), 2019.

[89] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In IEEE International Workshop on Information Forensics and Security (WIFS), 2019.

[90] Rui Zhao, Hangjie Yuan, Yujie Wei, Shiwei Zhang, Yuchao Gu, Lingmin Ran, Xiang Wang, Jay Zhangjie Wu, David Junhao Zhang, Yingya Zhang, et al. Evolvedirector: Approaching advanced text-to-image generation with large vision-language models. Conference on Neural Information Processing Systems (NeurIPS), 37, 2024.

[91] Matthew Zheng, Enis Simsar, Hidir Yesiltepe, Federico Tombari, Joel Simon, and Pinar Yanardag. Stylebreeder: Exploring and democratizing artistic styles through text-to-image models. In Conference on Neural Information Processing Systems (NeurIPS), 2024.

[92] Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. In International Conference on Learning Representations (ICLR), 2024.

[93] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In Conference on Neural Information Processing Systems (NeurIPS), 2023.

[94] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM International Conference on Multimedia, 2020.

补充材料

我们包含了额外的数据集整理细节（第 4 节），分类方法的训练细节（第 5 节），相关方法的实验和消融分析（第 ?? 节），以及额外的评估细节（第 6 节）。所有引用的 `.txt` 文件在 https://huggingface.co/datasets/cmu-gil/PromptedArtistIdentificationDataset/tree/main/metadata 发布的数据集中可获得。

4. 数据集整理详情

如在正文的第 3 节所述，为了收集文本图像生成模型中常用的艺术家名单，我们手动去重并从一个初始包含 400 个经常被 Stable Diffusion 用户在 Lexica.art 网站 [68, 75] 提示的艺术家列表中过滤我们看到过的艺术家。为了确保我们挑选出的 10 个待测艺术家在 CSD [75] 的训练过程中未被看到，我们在从用于训练 Midjourney 的 16000 个艺术家泄漏名单中过滤出至少有四张图片美学评分大于 4 的艺术家时，过滤 LAION-5B [69] 。

然后，我们将筛选出的艺术家与 CSD 数据集中说明文字进行交叉引用，以确保被保留的艺术家的名字不会出现在其中。我们在 seen_artists.txt 中包含了最终的 100 位已知艺术家的名单，在 保留艺术家.txt 中列出了 10 位保留艺术家，在 `public_domain_artists.txt` 中列出了 55 位属于公共领域的艺术家。

艺术家图像。我们从经过筛选的 LAION-5B [69] 中为每位艺术家收集真实的参考图像，该版本仅包括安全和合法的图像 [40] 。然后，我们遵循与 LAION-Styles [75] 相同的策展程序，筛选出预计美学评分为 6 或更高的图像、图像标题包含最终数据集的 3840 个风格标签的图像，以及 SSCD [57] 相似性小于 0.8 的图像。

多艺术家图像。为了生成包含 2 或 3 个艺术家名字的提示词的图像，我们将所有可能的未包含的艺术家组合和 100 个随机选择的已知艺术家组合插入到每个提示词中。我们在 `seen_artist_2combos_100samples.txt` 中包含了 2 个已知艺术家的 100 个随机组合，在 `seen_artist_3combos_100samples.txt` 中包含了 3 个已知艺术家的 100 个随机组合，在 `held_out_artist_all_2combos.txt` 中包含了所有 2 个未包含艺术家的组合，以及在 `held_out_artist_all_3combos.txt` 中包含了所有 3 个未包含艺术家的组合。

4.1. 提示管理与图像生成

简单提示。我们使用形式为 “ 一幅关于 `<content>` 的画，风格为 `<artist>` ” 的简单提示生成图像。为了多样性，我们从 ChatGPT 中随机抽取了 100 个主题。我们给 ChatGPT 的提示是 “提供一个以逗号分隔的 100 个在绘画中常见的主题列表 ”。我们手动检查并移除那些对于大多数艺术作品来说可能过于超出分布的主题或抽象概念，例如 “仙女” 或 “政治”，然后再次提示 ChatGPT 生成主题，直到我们总共整理出 100 个主题。

复杂提示。为了从真实的文本到图像模型用户那里整理一组复杂提示，我们从包含 ∼ 400 万提示-图像对的 JourneyDB [77] 训练数据集中进行过滤。我们首先通过指令大语言模型 Llama 3 8B Instruct [19] 获得只提到一个艺术家而不是多个艺术家的提示列表。具体来说，对于每个 JourneyDB 提示，我们给 Llama 3 以下指令：“ 仅输出以下文本中提到的艺术家的名字，作为一个 `JSON` 列表。不需要多说。”。使用大语言模型完成这项任务可以确保我们获得比通过硬字符串匹配到一个封闭的艺术家集合所能找到的更多提示。然后，我们移除 Llama 3 返回超过一个艺术家名字的提示。我们还移除少于 5 个单词或超过 77 个 CLIP [59] 文本标记的提示，以确保 SDXL [58] 能够根据完整提示为图像设定条件。最后，我们从数据集中采样 1000 个提示，并通过人工检查去除近似重复的提示。我们提供的提示最终列表包括训练和测试拆分。复杂提示在 `complex_prompts_train.txt` 和 `complex_prompts_test.txt` 中，而简单提示在 `simple_prompts_train.txt` 和 `simple_prompts_test.txt` 中。为了评估留存的艺术家，我们使用在 `complex_prompts_test.txt` 中前 5 个提示生成的图像作为复杂提示的测试时间参考图像，并使用在 `simple_prompts_test.txt` 中前 5 个提示生成的图像作为简单提示的测试时间参考图像。

图像生成。对于我们精心挑选列表中的每个艺术家和提示组合，我们使用开源文本到图像模型 SDXL [58] 、SD1.5 [61] 和 PixArt- Σ [11] 生成艺术家提示的图像。对于每个模型，我们使用其默认生成参数：SD1.5 的分辨率为 512×512 ，引导比例为 7.5，推理步数为 50；PixArt- Σ 的分辨率为 1024×1024 ，引导比例为 4.5，推理步数为 50；SDXL 的分辨率为 1024×1024 ，引导比例为 7.5，推理步数为 50。

真实艺术家图像的分布。图 11 显示了在我们包含 100 位艺术家的 LAION 数据集中，每位艺术家的名字在标题中出现的频率。我们比较了公共领域艺术家和非公共领域艺术家的频率，使用的指南是艺术家的最大可能美国版权期限为去世后 95 年 [73] 。我们发现 38.7 % 的图像来自 45 位非公共领域艺术家。如图 11 所示，公共和非公共领域艺术家的频率分布相似。我们将此数据集用作原型网络中艺术家风格的真实图像参考。

4.2. 用艺术家名字提示最近的文本到图像模型

我们尝试使用艺术家名字提示其他最近的文本到图像模型，特别是 Stable Diffusion 3.5 Large (SD3.5) [21] ，和 FLUX.1-dev [39] ，但发现在使用艺术家名字时，生成的图像风格往往几乎没有变化，即使提示词很简单。例如，在图 12 中，我们使用简单的提示词和数据集中训练集的艺术家生成图像。每个模型使用默认参数：SD3.5 使用 50 步推理，指导尺度为 7.0，而 FLUX.1-dev 使用 50 步推理，指导尺度为 3.5。对于每

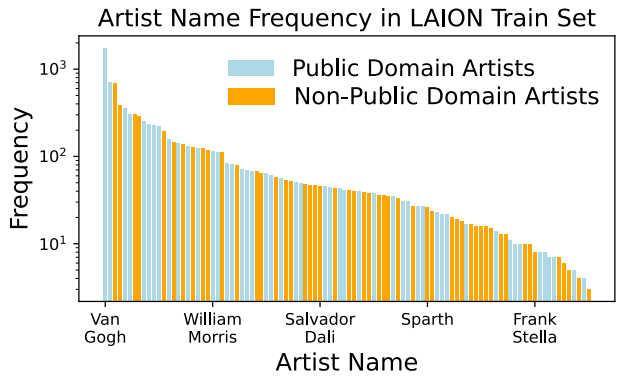

Figure 11. 艺术家统计。我们绘制了我们 LAION 训练集中每位艺术家名字的出现频率，其中艺术家被标记为公共领域（蓝色）或非公共领域（橙色），根据去世 95 年后的定义。在我们的 100 位已知艺术家中，有 45 位属于非公共领域，他们的图像占我们真实图像数据集的 38.7 %。

个提示词，模型生成的艺术家提示的图像几乎与没有艺术家名字提示的图像相同，都是写实风格。此外，艺术家提示的图像并没有与艺术家的实际风格一致。由于这些模型在提示艺术家名字时经常不能进行图像风格化，我们不在主要评估中包含这些模型。

4.3. 数据集统计表

根据数据集整理程序，我们创建了一个包含 195 万张图像的结构化数据集，以便在不同的泛化轴上评估视觉方法。单一艺术家的数据集汇总在表 5 中。对于每个艺术家和提示组合，我们使用开放权重模型 SDXL [58]、SD1.5 [61] 和 PixArt- Σ [11]，用 2 个随机种子生成图像。对于使用复杂提示生成的 SDXL 图像，我们使用 10 个不同的生成种子。为了从 Midjourney [51]（一个闭源模型）收集图像，我们从 JourneyDB 数据集中 [77] 过滤我们的一组已见与未见的艺术家。多艺术家数据集的汇总在表 6 中。我们对 2 位艺术家和 3 位艺术家组合的 100 个随机已见艺术家组合和所有可能的未见艺术家组合进行采样，并使用 SDXL 生成图像。

5. 训练详情

原型网络。我们通过对预训练的 CLIP [59] ViT-L/14 图像编码器进行端到端微调，为提示的艺术家识别训练原型网络 [72]。我们采用与 Wang et al. [79] 推荐的相同设置进行模糊和 JPEG 训练数据增强：图像以 10% 的概率被随机模糊和 JPEG 化。此外，还应用了随机调整大小裁剪和水平翻转。我们扫描学习率 $1e-7$、$1e-6$ 和 $1e-5$，以及训练周期数，从 1 到 5。我们选择表现最好的配置：以学习率 $1e-6$，温度 τ 为 0.07，和每个 GPU 的批处理大小为 128，通过 4 个 GPU 分布进行 1 个周期的微调。这导致有效的批处理大小为 512。计算环境为 4 个每个具有 40GB vRAM 的 A6000 GPU，并采用混合精度（fp16）训练。

Vanilla 分类器。我们在预训练的 CLIP [59] ViT-L/14 图像编码器上添加一个带有一层隐藏层的 MLP 作为分类头，然后在我们的数据集上训练所有层。训练设置与原型网络一致：以学习率 $1e-6$ 微调 1 个 epoch，有效批量大小为 512，并使用 fp16 混合精度训练。

我们分析并验证了在主要论文的基准测试中评估的每种视觉表示方法的设计选择。所有实验都是在我们提示的艺术家识别数据集中使用的 SDXL 图像进行的。数据集包括所有复杂的提示和一部分 100 个简单的提示（90 个训练，10 个测试）。

5.1. 从真实图像中检索

基于检索的方法可以使用真实的艺术家图像或生成的图像作为参考数据库来分类由艺术家指示生成的图像。我们在基准测试中对这两种设置进行了比较，发现对于每种方法，从生成的图像中检索在大多数评估集上比从真实图像中检索效果更好，并且方法的最终性能排名基本保持不变。我们使用从 LAION-Styles 收集的真实艺术家图像，这是用于计算原型网络原型的同一组图像。见过的和未见过的艺术家的图像数量如表 7 所示。对于从生成的图像中检索，我们使用了来自我们的 SDXL、SD1.5、PixArt 和 Midjourney 数据集中所有生成的已见艺术家训练图像，包括简单和复杂提示的图像。在表 10 中，我们展示了对 SDXL 图像的结果，观察到对于所有方法，从真实图像中检索的性能低于从生成的图像中检索的性能，适用于除复杂提示图像和未见艺术家外的所有评估集。因此，我们在主要基准评估中专注于从生成图像中检索。

在主要的基准评估中，我们评估那些使用单个参考图像嵌入来进行检索的基于检索的方法。然而，原型网络通过原型对齐进行分类学习，在这种方法中，我们将艺术家的原型设置为该艺术家作品的真实参考图像的平均特征。

为了公平比较，我们在类似环境下测试基于检索的方法，即从一组真实图像的平均嵌入中进行检索，每位艺术家计算一个平均嵌入（与原型相同）。不同之处在于，这些方法没有经过微调以与原型对齐。我们评估这种对于 CSD [75] 和 CLIP [59] 的平均嵌入检索基线，并将结果与表 8 中的原型网络进行比较。我们发现原型网络在大多数类别中表现优于它们。

在我们对基于检索方法的基准测试的主要评估中，我们使用 k -NN 分类与 $k=1$，因为我们发现增加 k 并无帮助。

我们测试了 k -NN 分类中的多数投票方法，其中最终预测的类别标签是前 k 个最近邻居中的多数类别标签，如果有并列，则通过与查询嵌入最接近的已检索嵌入的多数类别标签来打破。我们评估了用于 k -NN 多数投票分类的多个 k 值，以研究 k 对艺术家分类准确率的影响，如图 ?? 所示。

首先，我们发现不同的 k 值下，各种方法的性能排名保持不变。此外，对于每种基于检索的方法，随着 k 增加，性能仅有少量变化。例外情况是，当图像由未包含在训练集中的艺术家使用简单的提示生成时，性能

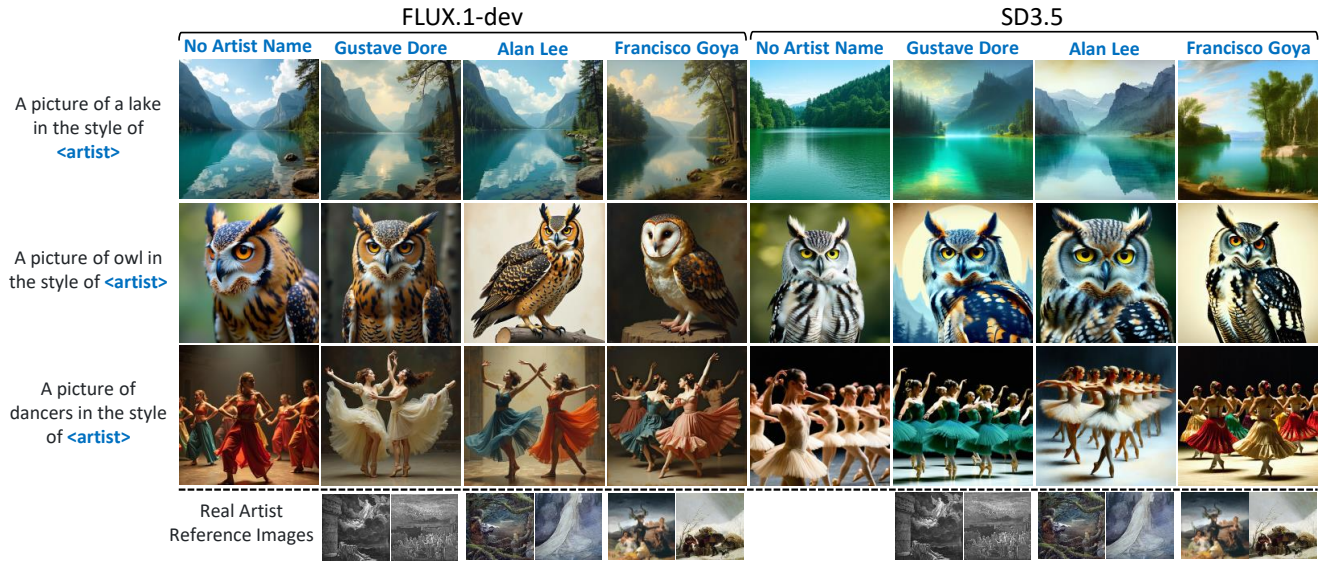

Figure 12. FLUX 和 SD3.5 示例。对于 FLUX.1-dev [39] 和 SD3.5 [21]，在提示中插入艺术家姓名对生成图像的风格通常几乎没有影响，与使用没有艺术家姓名的相同提示相比。大多数生成的图像都是写实的，并且生成的图像与艺术家的实际风格不匹配（底排）。

(a) SDXL

	Complex Prompts				Simple Prompts				Total Unique
	Seen artists		Held-out artists		Seen artists		Held-out artists		
	Train	Test	Test-Time Ref.	Test	Train	Test	Test-Time Ref.	Test	
# Artists	100	100	10	10	100	100	10	10	110
# Prompts	900	100	5	90	450	50	5	45	1,500
# Seeds	10	10	10	10	2	2	2	2	10
# Images	900,000	100,000	500	9,000	90,000	10,000	100	900	1,110,500

(b) SD1.5, PixArt

	Complex Prompts				Simple Prompts				Total Unique
	Seen artists		Held-out artists		Seen artists		Held-out artists		
	Train	Test	Test-Time Ref.	Test	Train	Test	Test-Time Ref.	Test	
# Artists	100	100	10	10	100	100	10	10	110
# Prompts	450	50	5	45	450	50	5	45	1,000
# Seeds	2	2	2	2	2	2	2	2	2
# Images	90,000	10,000	100	900	90,000	10,000	100	900	202,000

(c) Midjourney

	Complex Prompts				Total Unique
	Seen artists		Held-out artists		
	Train	Test	Test-Time Ref.	Test	
# Artists	35	35	95	95	130
# Prompts	647	658	649	697	2,651
# Images	647	658	649	697	2,651

Table 5. 单一艺术家数据集统计。提示的艺术家识别数据集包括使用 SDXL [58]、SD1.5 [61]、PixArt- Σ [11] 和 Midjourney [51] 生成的图像。如在第 3 节中所述，我们收集了 1000 个复杂提示，500 个简单提示，并在提示中使用了 110 个不同的艺术家名字。对于每个艺术家和提示组合，我们使用 10 个不同的种子生成 SDXL 复杂提示的图像，对于其他子集使用 2 个不同的种子。我们将 110 位艺术家分为 100 位已见艺术家和 10 位保留艺术家。对于已见艺术家，我们使用一个单独的提示集进行测试。对于保留艺术家，我们还额外划分测试提示图像供测试时参考。

在非常小的 k 时趋于饱和。

我们分析并验证了在主论文的基准测试中评估的分类方法的设计选择。表格 9 显示了来自我们的提示艺术家识别数据集的 SDXL 图像的结果，这些结果是从

(a) 2 位艺术家

	Complex Prompts				Simple Prompts				Total Unique
	Seen artists		Held-out artists		Seen artists		Held-out artists		
	Train	Test	Test-Time Ref.	Test	Train	Test	Test-Time Ref.	Test	
# Artists	100	100	10	10	100	100	10	10	110
# Prompts	450	50	5	45	450	50	5	45	1,000
# Seeds	2	2	2	2	2	2	2	2	2
# 2-artist combinations	100	100	45	45	100	100	45	45	145
# Images	90,000	10,000	450	4,050	90,000	10,000	450	4,050	209,000

(b) 3 位艺术家

	Complex Prompts				Simple Prompts				Total Unique
	Seen artists		Held-out artists		Seen artists		Held-out artists		
	Train	Test	Test-Time Ref.	Test	Train	Test	Test-Time Ref.	Test	
# Artists	100	100	10	10	100	100	10	10	110
# Prompts	450	50	5	45	450	50	5	45	1,000
# Seeds	2	2	2	2	2	2	2	2	2
# 3-artist combinations	100	100	120	120	100	100	120	120	220
# Images	90,000	10,000	1,200	10,800	90,000	10,000	1,200	10,800	224,000

Table 6. 多艺术家数据集统计。为了评估提示生成的包含多位艺术家的图像的艺术家识别，我们生成了由两个艺术家和三个艺术家提示生成的图像数据集。该数据集的生成过程与单个艺术家数据集相同，只是我们针对每个艺术家数量采样了 100 组艺术家组合。

	Seen artists	Held-out artists	Total
# Artists	100	10	110
# Images	9,907	860	10,767

Table 7. 真实艺术家图像数据集。我们使用了 [75] 的 LAION-Styles 的一个子集，其中包含 100 位已知艺术家和 10 位保留艺术家。然后，我们对各种视觉表示方法进行艺术家识别的基准测试，并使用此数据集作为每位艺术家风格的真实图像参考。

Method	Reference/ Prototype	Evaluation Set			
		Seen artists (100-way)		Held-out (10-way)	
		Complex	Simple	Complex	Simple
CLIP [59]	Artist Avg.	15.1	38.3	44.9	79.0
CSD [75]	Artist Avg.	19.7	48.1	56.8	92.0
Proto. Net. [72]	Artist Avg.	43.2	87.0	62.7	87.0

Table 8. 与从原型检索的基线进行比较。我们还评估了基线 CSD 和 CLIP 的平均嵌入检索，并将它们与原型网络进行比较。

复杂和简单提示评估集的平均值。

比较分类器类型。在表格 9a 中，我们训练并比较了不同的模型类型：普通分类、监督对比学习 [33] 和原型网络 [72] 。虽然分类器在已知艺术家分类中优于原型网络，但默认情况下，由于固定的标签集，它无法对未见过的艺术家进行分类。为了测试泛化能力，我们去掉分类器的最终线性层以提取用于检索的特征。我们发现从真实或生成的图像中进行检索，在未见过艺术家的分类中，效果比原型网络差。这表明使用参考图像的原型进行训练有助于提高泛化能力。

接下来，我们尝试通过运行监督对比学习 [33] 来进行度量学习，在生成的图像上进行实验，并为同一艺术家的所有图像使用相同的标签。我们发现这种方法未能实现强劲性能，因此将重点放在原型网络和普通分类上进行基准评价。

最近邻和原型的不同来源。原型网络也可以被用作特征提取器来进行检索，其性能与使用原型进行分类相当。我们将在表格 9c 中报告结果。此外，使用生成的图像作为原型会导致更差的性能，因此我们在基准评估中使用真实图像作为原型。我们假设使用不同模态（例如，真实艺术家图像）作为原型会促使编码器提取更多艺术家特异性的特征，从而帮助整体泛化能力。

哪些训练数据更有帮助？我们改变了原型网络的训练集，并在表格 9b 中报告了性能。首先，我们不是通过生成图像来训练艺术家的分类，而是使用真实的艺术家图像作为输入来训练原型网络。不使用合成图像进行训练大幅降低了性能。由于良好的分类性能需要在图像由复杂提示生成时也能提取与艺术家相关的线索，因此从生成图像中训练编码器可以帮助学习这些线索。

我们通过仅使用带有简单提示词生成的图像训练一个原型网络来进一步验证这一假设。虽然这一变体比仅用真实图像训练的网络表现更好，但其性能仍然不如使用带有复杂提示词生成的图像进行训练的网络。另一方面，仅使用复杂提示词进行训练所得到的性能与同时使用简单和复杂提示词的数据集进行训练的性能相似。

(a) 模型类型

Model		Evaluation set	
Training	Testing	Seen Artists (100-way)	Held-out (10-way)
Vanilla Classifier	Classifier	53.4	–
	NN - Real	24.4	45.0
	NN - Gen.	51.3	58.0
Supervised Contrastive Learning [33]	Classifier	49.3	–
	NN - Real	24.4	21.6
	NN - Gen.	47.5	29.6
Proto. Net. [72]	Proto. - Real	43.9	63.0

(b) 训练集

Model	Evaluation set	
	Seen artists (100-way)	Held-out (10-way)
Real	16.5	49.7
Simple only	21.1	55.2
Complex only	43.6	62.5
Simple + Complex	43.9	63.0

(c) 原型 vs. 最近邻居

	Reference		Evaluation set	
Method	Real	Gen.	Seen artists (100-way)	Held-out (10-way)
NN - Real	✓		26.7	54.3
NN - Gen.		✓	43.9	62.3
Proto. - Gen.		✓	23.0	61.2
Proto. - Real	✓		43.9	63.0

Table 9. 分类方法实验。在主论文中，我们对两种分类方法进行了基准测试：原型网络和普通分类法。在这里，我们测试更多的分类器方法变体和训练数据集，并报告在 SDXL 复杂和简单提示评估集上的平均准确率。(a) 我们测试 **模型类型** ，比较普通分类法、监督对比学习和原型网络。由于分类器有固定的标签集，我们在测试未见过的艺术家时使用其特征空间上的最近邻。监督对比学习的表现比原型网络和普通分类法都要差，因此我们没有在基准评估中包括它。(b) 我们测试了原型网络的 **训练集** 。从真实图像学习不如从生成图像学习有效。使用简单和复杂的提示比单独使用其中一个要优，证明它们提供了互补的信息。(c) 对于原型网络，我们比较了使用 **原型与最近邻居的比较** 。当使用学习到的原型网络表示方式进行最近邻时，性能略有下降。使用生成图像作为原型，而不是真实图像，也会降低性能，因此我们在主要实施中使用真实图像作为原型。

6. 评估细节

6.1. 自举过程

为了估计我们基准测试中每个评估的统计显著性，我们使用了一种自举程序。我们通过重新采样评价图像来自举每个模型的预测，其中包括重复 2000 次的画家姓名、提示模板和生成种子。我们通过绘制标准误差的收敛图（如图 ?? 所示）选择了 2000 次迭代。

给定原型网络对 SDXL 图像和简单提示的艺术家分类预测，我们绘制了分类准确性标准误差与自举迭代次数的关系。自举迭代次数从 50 增加到 4000，每次增加 50 次。在大约 2000 次迭代时，标准误差保持在数据点的 90 % 范围内，表明标准误差在此点收敛。因此，在我们的主要基准评估中，我们使用 2000 次自举迭代。

我们在第 ?? 节中提供了主要基准评估的定量结果，如下表所示。单艺术家评估结果包括表 10 中的 SDXL 图像，表 11 中的 SD1.5 图像，表 12 中的 PixArt- Σ 图像，以及表 13 中的 Midjourney 图像。多艺术家评估结果包括表 14 中的两位艺术家提示的图像和表 15 中的三位艺术家提示的图像。

Family	Method	Architecture	Reference/ Prototype images		Evaluation set			
					Seen artists (100-way)		Held-out (10-way)	
			Real	Gen.	Complex	Simple	Complex	Simple
Chance	–	–	–	–	1.0	1.0	10.0	10.0
Retrieval-based methods	AbC - DINO [80]	ViT-B/16	✓		7.2 ± 1.0	13.9 ± 2.0	33.9 ± 6.4	47.7 ± 9.4
				✓	21.7 ± 2.4	45.1 ± 3.7	30.7 ± 5.5	55.8 ± 7.5
	AbC - CLIP [80]	ViT-B/16	✓		10.5 ± 1.5	20.4 ± 2.7	35.4 ± 6.3	52.3 ± 8.4
				✓	20.3 ± 2.4	48.1 ± 3.4	30.2 ± 4.0	60.0 ± 5.9
	DINOv2 [55]	ViT-L/14	✓		6.0 ± 1.0	11.0 ± 1.9	29.6 ± 5.5	44.6 ± 8.2
				✓	14.1 ± 1.9	28.7 ± 3.3	16.4 ± 3.4	37.2 ± 6.8
	CLIP [59]	ViT-L/14	✓		12.4 ± 1.7	24.9 ± 2.9	41.1 ± 6.8	62.7 ± 8.3
				✓	22.8 ± 2.6	53.4 ± 3.5	35.3 ± 4.2	78.5 ± 4.7
	CSD [75]	ViT-L/14	✓		15.3 ± 2.0	31.9 ± 3.3	47.4 ± 7.0	74.3 ± 6.6
				✓	30.1 ± 2.9	66.4 ± 3.4	44.4 ± 4.8	84.6 ± 3.9
Fine-tuned classifiers	Prototypical Network [72]	ViT-L/14	✓		40.3 ± 3.1	74.9 ± 3.0	57.7 ± 4.5	75.5 ± 7.9
	Vanilla Classifier	ViT-L/14		✓	40.7 ± 3.2	76.4 ± 3.0	–	–

Table 10. 在 SDXL 图像上的分类准确性。我们比较了基于检索的方法和在我们的提示艺术家识别数据集上的 SDXL 图像上微调的分类器。尽管基于检索的方法可以使用真实或生成的图像作为参考数据库，我们发现除了复杂提示和保留的艺术家之外，在所有评估集上，从生成图像检索在所有方法中都优于从真实图像检索。因此，在我们的主要基准评估中，我们专注于从生成图像中进行检索。

Family	Method	Evaluation set			
		Seen artists (100-way)		Held-out (10-way)	
		Complex	Simple	Complex	Simple
Chance	–	1.0	1.0	10.0	10.0
Retrieval-based methods	AbC - DINO [80]	26.5 ± 3.3	49.2 ± 4.1	27.5 ± 6.1	51.8 ± 8.0
	AbC - CLIP [80]	24.4 ± 3.2	48.9 ± 3.9	29.2 ± 6.1	50.6 ± 8.5
	DINOv2 [55]	18.8 ± 2.7	36.3 ± 3.7	22.7 ± 5.4	41.5 ± 7.6
	CLIP [59]	27.3 ± 3.5	52.9 ± 3.8	31.5 ± 6.2	57.5 ± 8.3
	CSD [75]	32.1 ± 3.7	61.3 ± 3.8	34.6 ± 6.2	66.3 ± 7.1
Fine-tuned classifiers	Prototypical Network [72]	40.5 ± 3.8	63.0 ± 3.7	43.8 ± 6.4	53.8 ± 8.7
	Vanilla Classifier	40.2 ± 3.8	65.2 ± 3.7	–	–

Table 11. SD1.5 图像的分类准确率。

Family	Method	Evaluation set			
		Seen artists (100-way)		Held-out (10-way)	
		Complex	Simple	Complex	Simple
Chance	–	1.0	1.0	10.0	10.0
Retrieval-based methods	AbC - DINO [80]	6.8 ± 1.3	23.3 ± 2.7	7.8 ± 2.5	20.4 ± 6.6
	AbC - CLIP [80]	6.6 ± 1.4	24.3 ± 2.9	6.6 ± 2.4	22.6 ± 6.7
	DINOv2 [55]	4.7 ± 1.0	13.0 ± 1.9	4.2 ± 1.6	9.5 ± 3.5
	CLIP [59]	7.4 ± 1.5	26.6 ± 2.9	9.3 ± 3.1	29.6 ± 7.0
	CSD [75]	9.4 ± 1.9	38.0 ± 3.4	8.8 ± 3.0	35.3 ± 7.2
Fine-tuned classifiers	Prototypical Network [72]	12.6 ± 2.0	40.8 ± 3.5	14.9 ± 3.3	23.8 ± 6.4
	Vanilla Classifier	12.9 ± 2.0	41.9 ± 3.5	–	–

Table 12. 在 PixArt- Σ 图像上的分类准确率。

		Evaluation set	
		Seen artists (34-way)	Held-out (96-way)
Family	Method	Complex	Complex
Chance	–	2.9	1.0
Retrieval-based methods	AbC - DINO [80]	19.8 ± 2.0	32.3 ± 2.6
	AbC - CLIP [80]	29.3 ± 2.8	39.7 ± 3.2
	DINOv2 [55]	16.8 ± 2.0	28.2 ± 2.4
	CLIP [59]	30.3 ± 3.1	42.7 ± 2.9
	CSD [75]	34.4 ± 2.3	42.8 ± 3.3
Fine-tuned classifiers	Prototypical Network [72]	56.8 ± 3.2	16.8 ± 2.7
	Vanilla Classifier	57.6 ± 2.7	–

Table 13. Midjourney 图像的分类准确率。

		Evaluation set			
		Seen artists (100-way)		Held-out (10-way)	
Family	Method	Complex	Simple	Complex	Simple
Chance	–	3.0	3.0	37.2	37.2
Retrieval-based methods	AbC - DINO [80]	15.7 ± 1.6	28.2 ± 1.6	40.1 ± 1.9	33.2 ± 2.3
	AbC - CLIP [80]	15.2 ± 1.7	30.8 ± 1.5	39.3 ± 2.1	34.7 ± 2.5
	DINOv2 [55]	11.2 ± 1.3	20.1 ± 1.6	36.3 ± 1.9	33.6 ± 2.2
	CLIP [59]	16.5 ± 1.7	32.7 ± 1.6	42.3 ± 2.1	36.6 ± 2.7
	CSD [75]	19.3 ± 1.9	37.9 ± 1.4	42.4 ± 2.1	33.1 ± 2.9
Fine-tuned classifiers	Prototypical Network [72]	45.1 ± 3.2	77.6 ± 2.1	61.5 ± 2.3	48.3 ± 2.4
	Vanilla Classifier	28.5 ± 2.1	49.3 ± 1.8	–	–

Table 14. 在 SDXL 2 艺术家提示图像上的排名 mAP@10 艺术家检索

		Evaluation set			
		Seen artists (100-way)		Held-out (10-way)	
Family	Method	Complex	Simple	Complex	Simple
Chance	–	3.1	3.1	45.0	45.0
Retrieval-based methods	AbC - DINO [80]	11.2 ± 1.0	20.4 ± 1.0	35.1 ± 1.2	39.1 ± 1.0
	AbC - CLIP [80]	10.2 ± 1.0	20.2 ± 0.9	35.6 ± 1.3	43.6 ± 1.3
	DINOv2 [55]	8.4 ± 0.9	13.7 ± 0.9	34.3 ± 1.3	39.3 ± 1.1
	CLIP [59]	11.3 ± 1.1	21.3 ± 1.0	37.1 ± 1.2	46.6 ± 1.3
	CSD [75]	13.2 ± 1.1	25.0 ± 1.0	34.8 ± 1.2	41.6 ± 1.1
Fine-tuned classifiers	Prototypical Network [72]	35.3 ± 2.7	61.9 ± 2.1	62.2 ± 1.3	63.0 ± 1.4
	Vanilla Classifier	20.7 ± 1.6	34.2 ± 1.4	–	–

Table 15. 在 SDXL 3-艺术家提示图像上的艺术家检索排名 mAP@10。