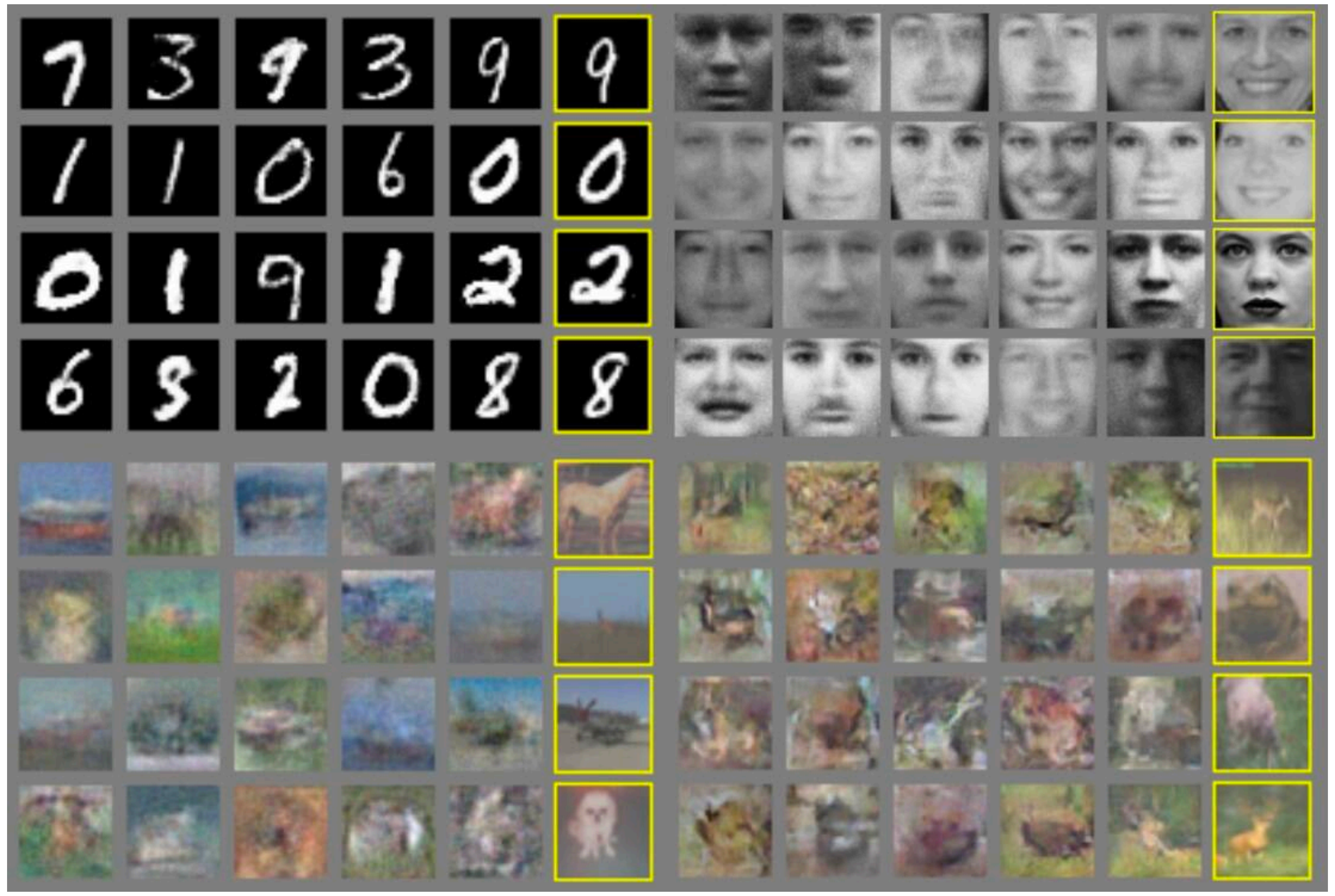
MLCollective DLCT | June 28

Measuring Style Similarity in Diffusion Models

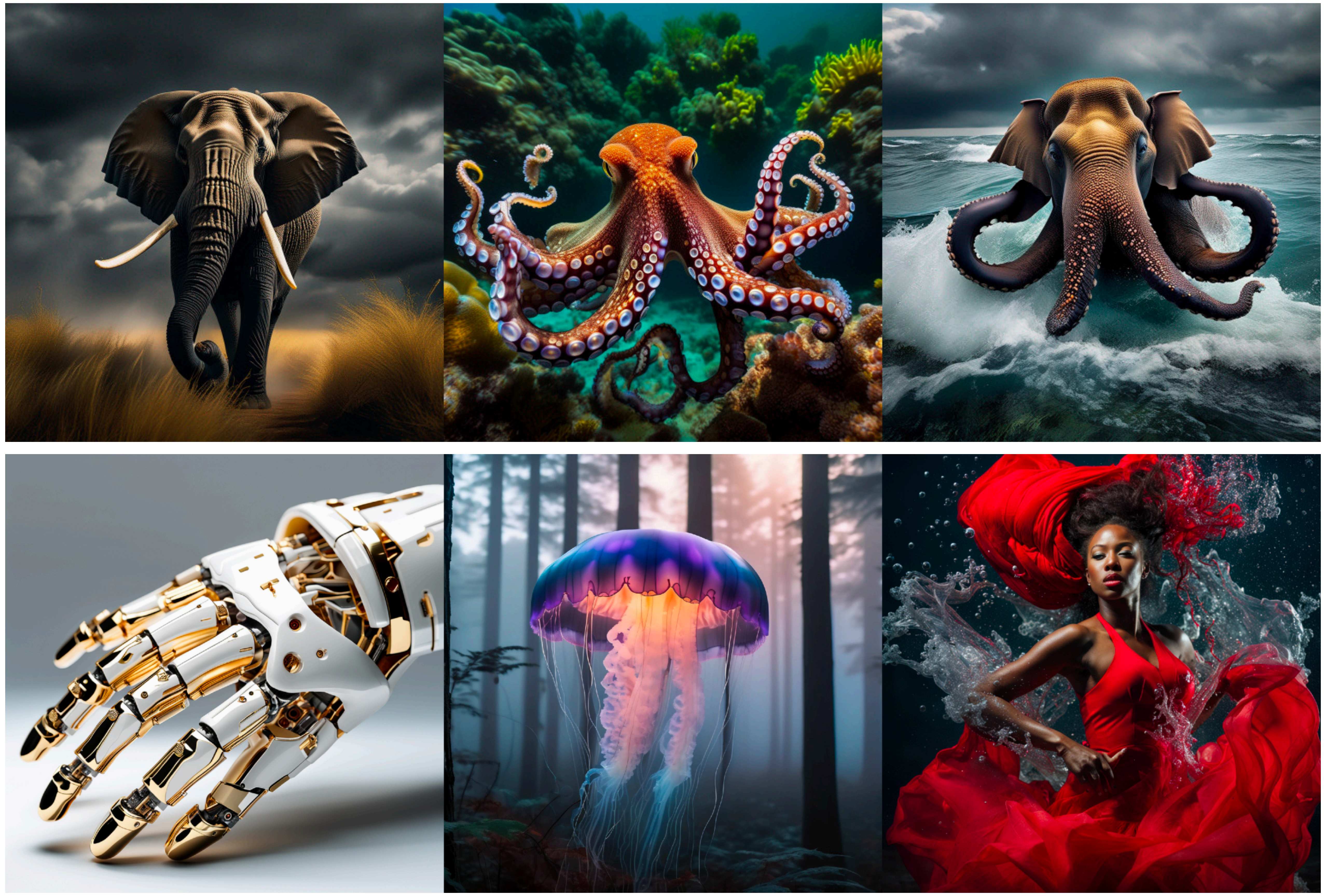
https://arxiv.org/abs/2404.01292

Gowthami Somepalli

We made incredible progress!



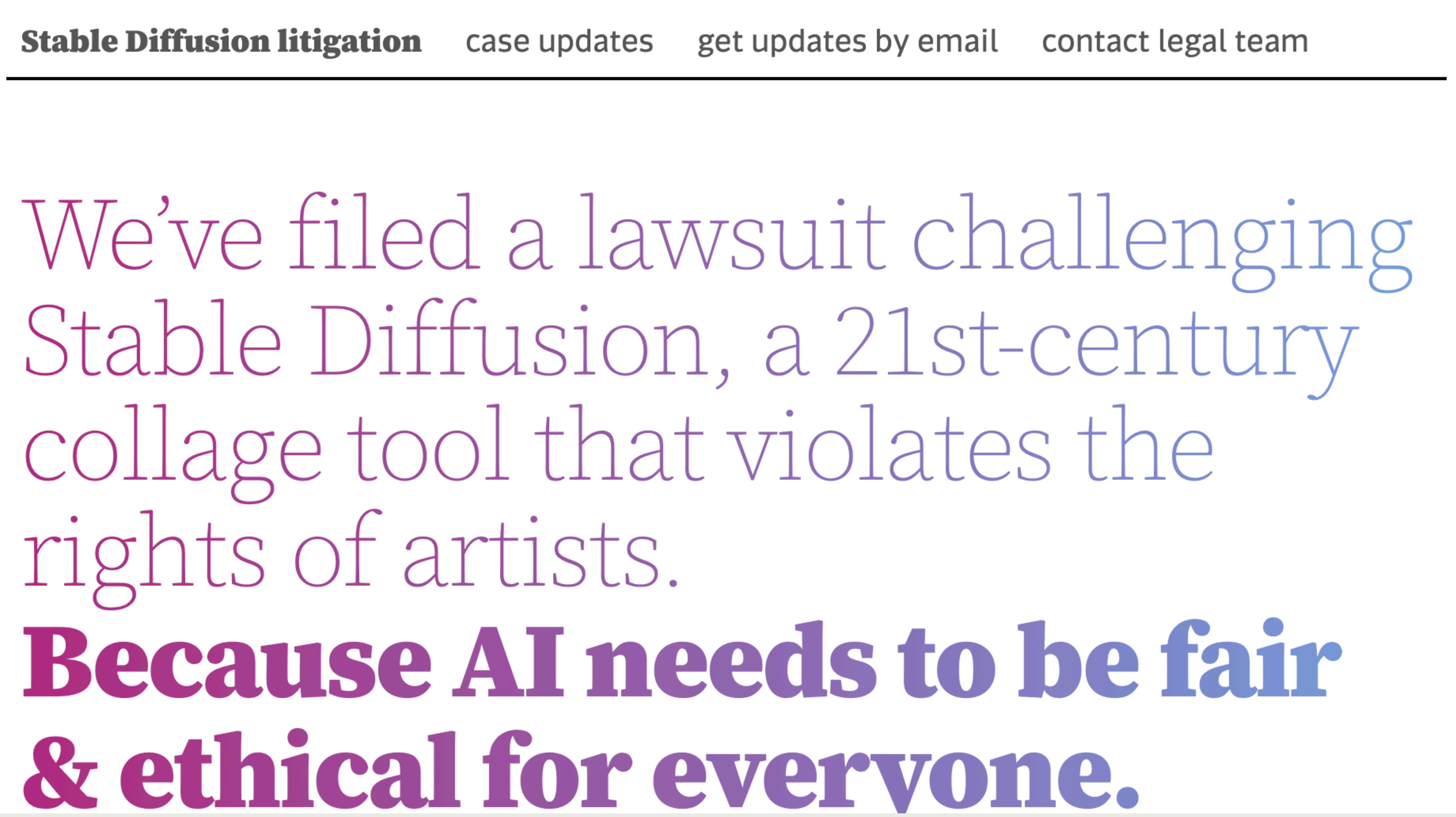
Goodfellow et al. GAN



Podell et al. SD-XL

2014 → 2023

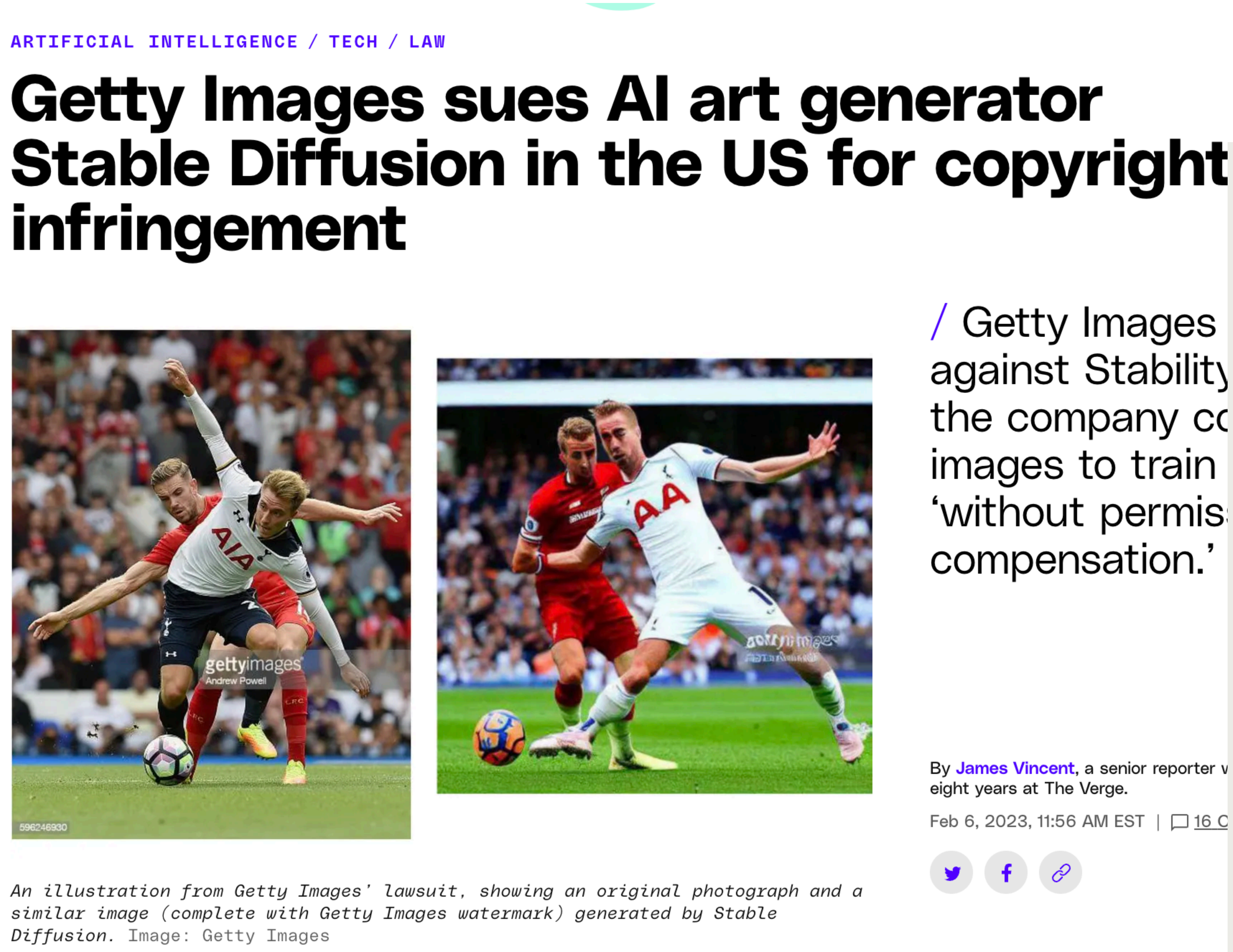
Memorization is problematic!



Stable Diffusion litigation case updates get updates by email contact legal team

We've filed a lawsuit challenging Stable Diffusion, a 21st-century collage tool that violates the rights of artists.

Because AI needs to be fair & ethical for everyone.



ARTIFICIAL INTELLIGENCE / TECH / LAW

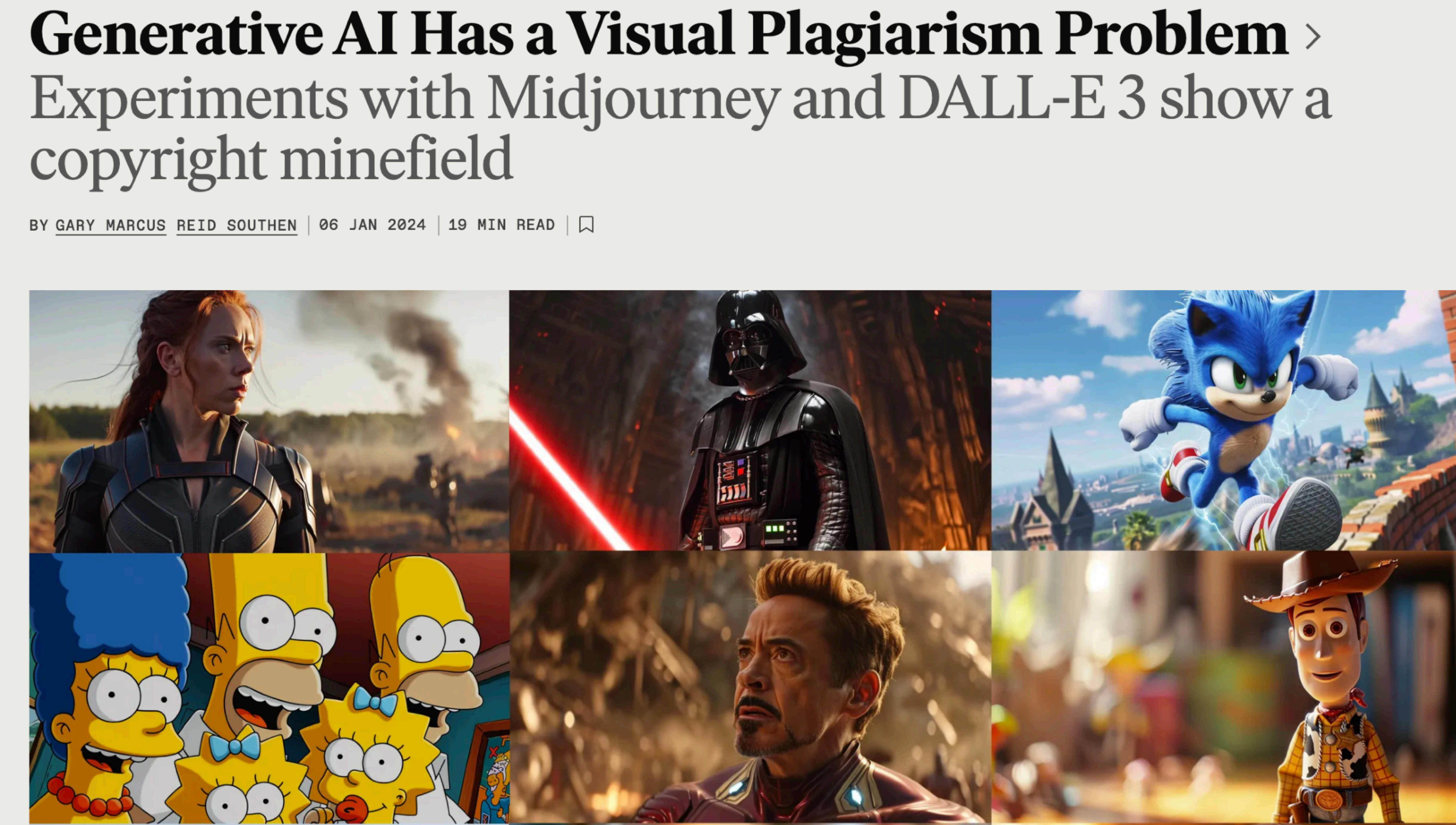
Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement

/ Getty Images against Stability the company co images to train 'without permis compensation.'

By James Vincent, a senior reporter v eight years at The Verge.

Feb 6, 2023, 11:56 AM EST | 16 C

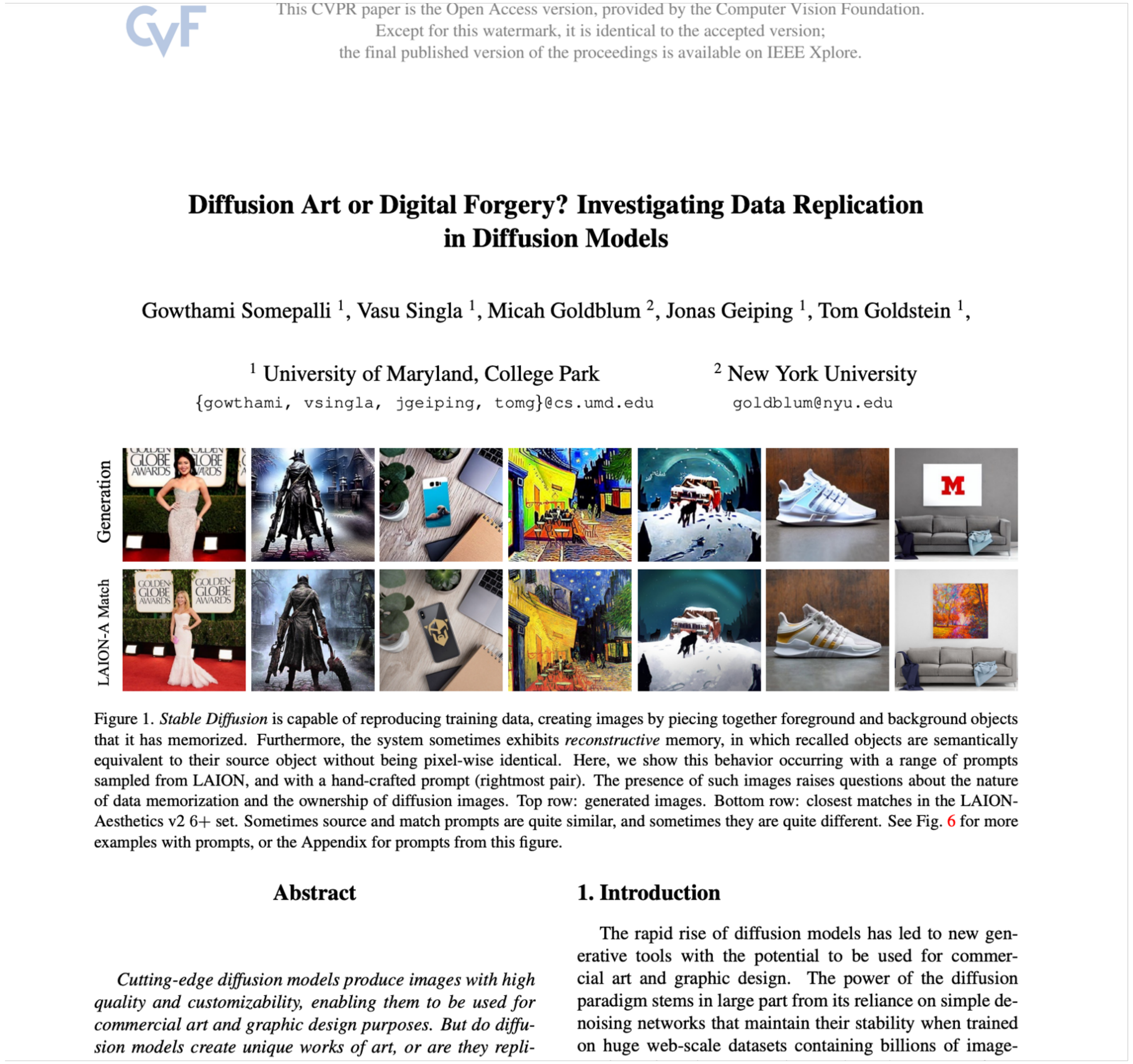
An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images



Generative AI Has a Visual Plagiarism Problem ›

Experiments with Midjourney and DALL-E 3 show a copyright minefield

BY GARY MARCUS REID SOUTHEN | 06 JAN 2024 | 19 MIN READ



Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

Gowthami Somepalli [1], Vasu Singla [1], Micah Goldblum [2], Jonas Geiping [1], Tom Goldstein [1],

[1] University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

[2] New York University
goldblum@nyu.edu

Figure 1. *Stable Diffusion* is capable of reproducing training data, creating images by piecing together foreground and background objects that it has memorized. Furthermore, the system sometimes exhibits *reconstructive* memory, in which recalled objects are semantically equivalent to their source object without being pixel-wise identical. Here, we show this behavior occurring with a range of prompts sampled from LAION, and with a hand-crafted prompt (rightmost pair). The presence of such images raises questions about the nature of data memorization and the ownership of diffusion images. Top row: generated images. Bottom row: closest matches in the LAION-Aesthetics v2 6+ set. Sometimes source and match prompts are quite similar, and sometimes they are quite different. See Fig. 6 for more examples with prompts, or the Appendix for prompts from this figure.

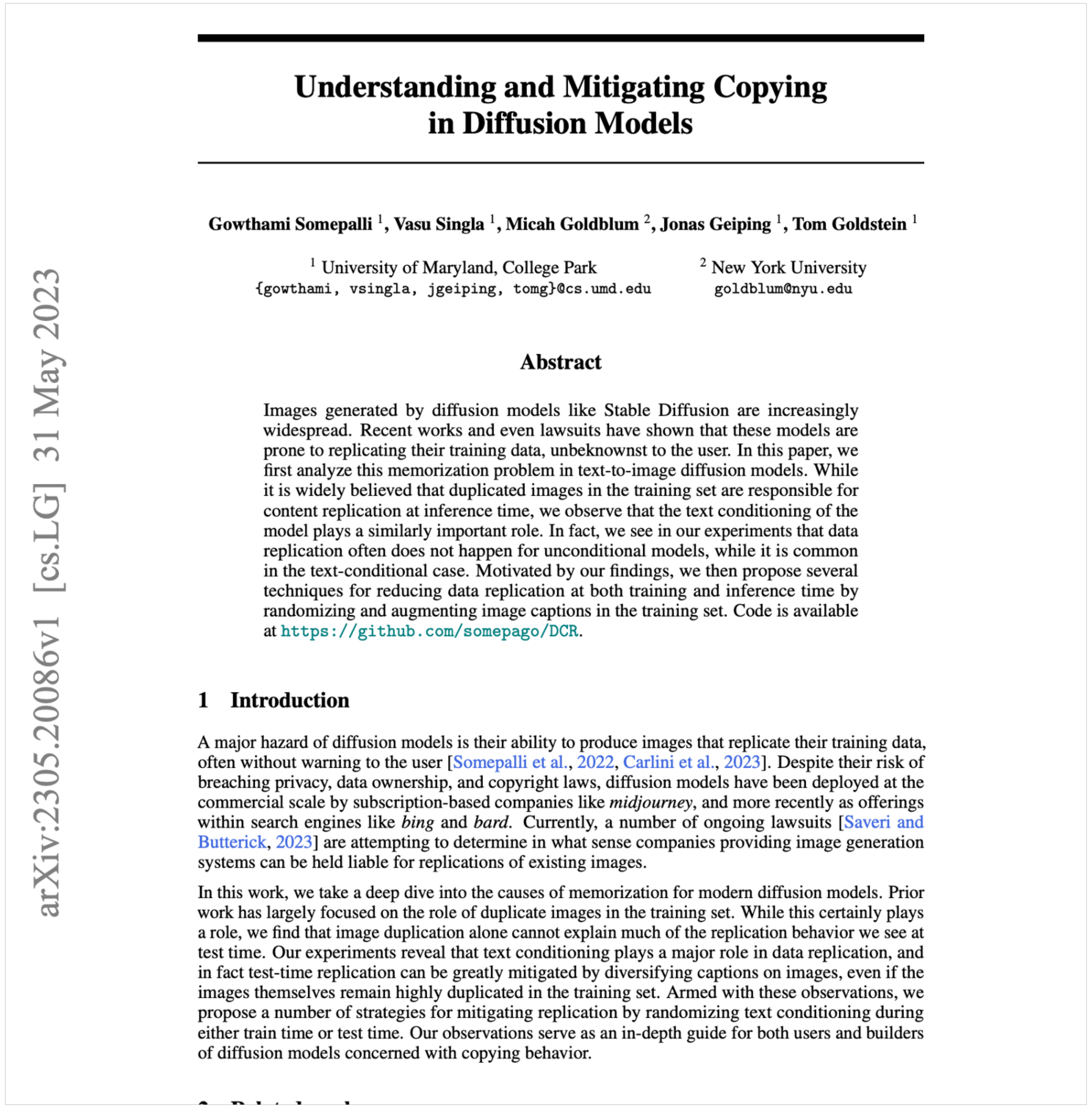
Abstract

Cutting-edge diffusion models produce images with high quality and customizability, enabling them to be used for commercial art and graphic design purposes. But do diffusion models create unique works of art, or are they repli-

1. Introduction

The rapid rise of diffusion models has led to new generative tools with the potential to be used for commercial art and graphic design. The power of the diffusion paradigm stems in large part from its reliance on simple denoising networks that maintain their stability when trained on huge web-scale datasets containing billions of image-

CVPR'23



Understanding and Mitigating Copying in Diffusion Models

Gowthami Somepalli [1], Vasu Singla [1], Micah Goldblum [2], Jonas Geiping [1], Tom Goldstein [1]

[1] University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

[2] New York University
goldblum@nyu.edu

arXiv:2305.20086v1 [cs.LG] 31 May 2023

Abstract

Images generated by diffusion models like Stable Diffusion are increasingly widespread. Recent works and even lawsuits have shown that these models are prone to replicating their training data, unbeknownst to the user. In this paper, we first analyze this memorization problem in text-to-image diffusion models. While it is widely believed that duplicated images in the training set are responsible for content replication at inference time, we observe that the text conditioning of the model plays a similarly important role. In fact, we see in our experiments that data replication often does not happen for unconditional models, while it is common in the text-conditional case. Motivated by our findings, we then propose several techniques for reducing data replication at both training and inference time by randomizing and augmenting image captions in the training set. Code is available at https://github.com/somepago/DCR.

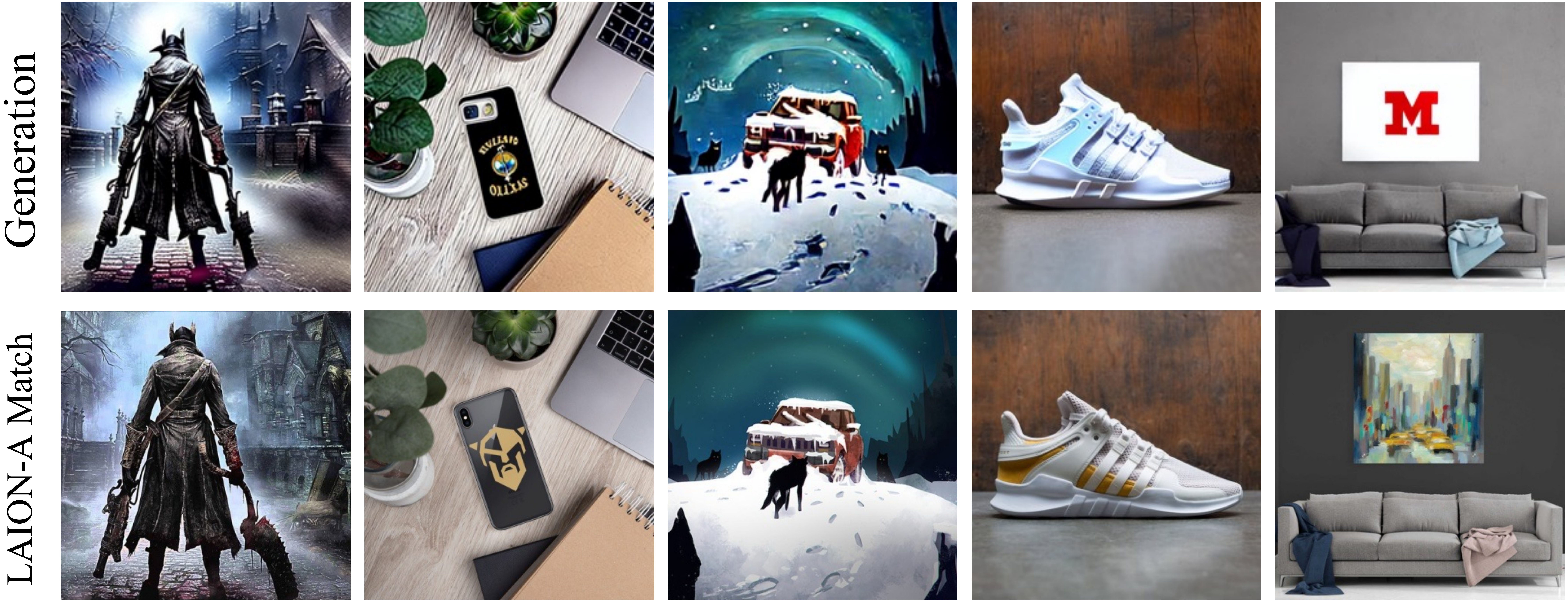
1 Introduction

A major hazard of diffusion models is their ability to produce images that replicate their training data, often without warning to the user [Somepalli et al., 2022, Carlini et al., 2023]. Despite their risk of breaching privacy, data ownership, and copyright laws, diffusion models have been deployed at the commercial scale by subscription-based companies like *midjourney*, and more recently as offerings within search engines like *bing* and *bard*. Currently, a number of ongoing lawsuits [Saveri and Butterick, 2023] are attempting to determine in what sense companies providing image generation systems can be held liable for replications of existing images.

In this work, we take a deep dive into the causes of memorization for modern diffusion models. Prior work has largely focused on the role of duplicate images in the training set. While this certainly plays a role, we find that image duplication alone cannot explain much of the replication behavior we see at test time. Our experiments reveal that text conditioning plays a major role in data replication, and in fact test-time replication can be greatly mitigated by diversifying captions on images, even if the images themselves remain highly duplicated in the training set. Armed with these observations, we propose a number of strategies for mitigating replication by randomizing text conditioning during either train time or test time. Our observations serve as an in-depth guide for both users and builders of diffusion models concerned with copying behavior.

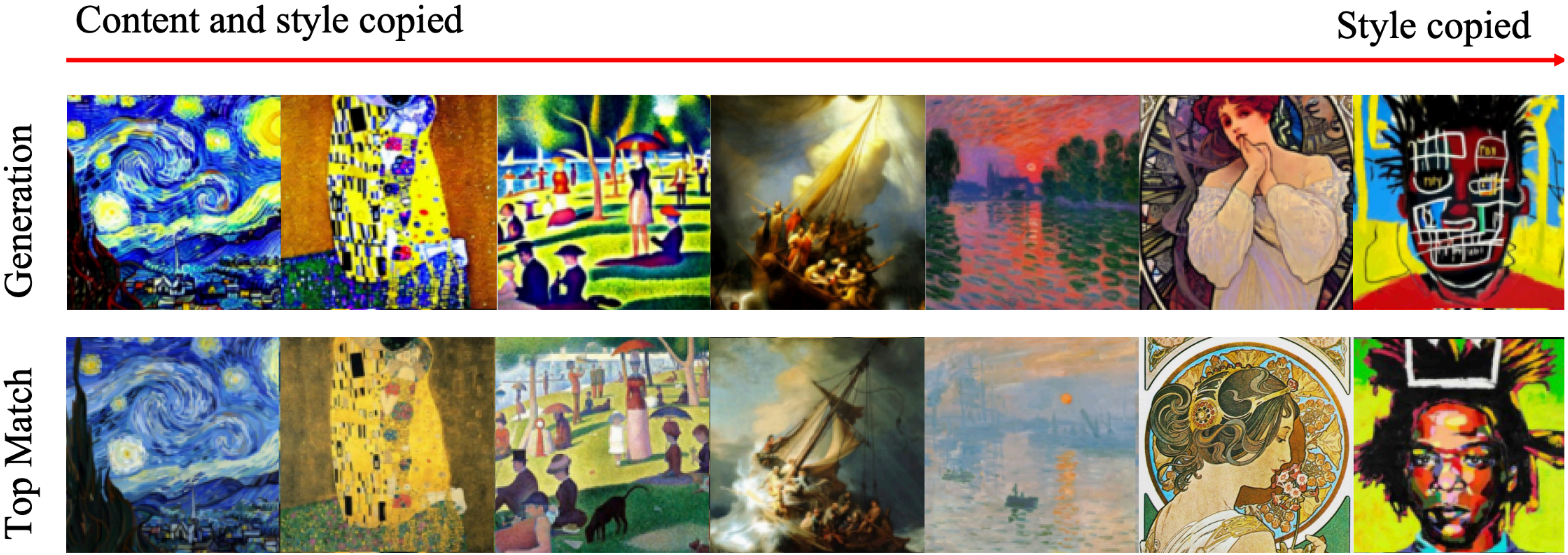
NeurIPS'23

Stable diffusion memorizes some training examples!



Somepalli et al. Diffusion Art or … (CVPR'23)

Sparks of Style Memorization? 🤔

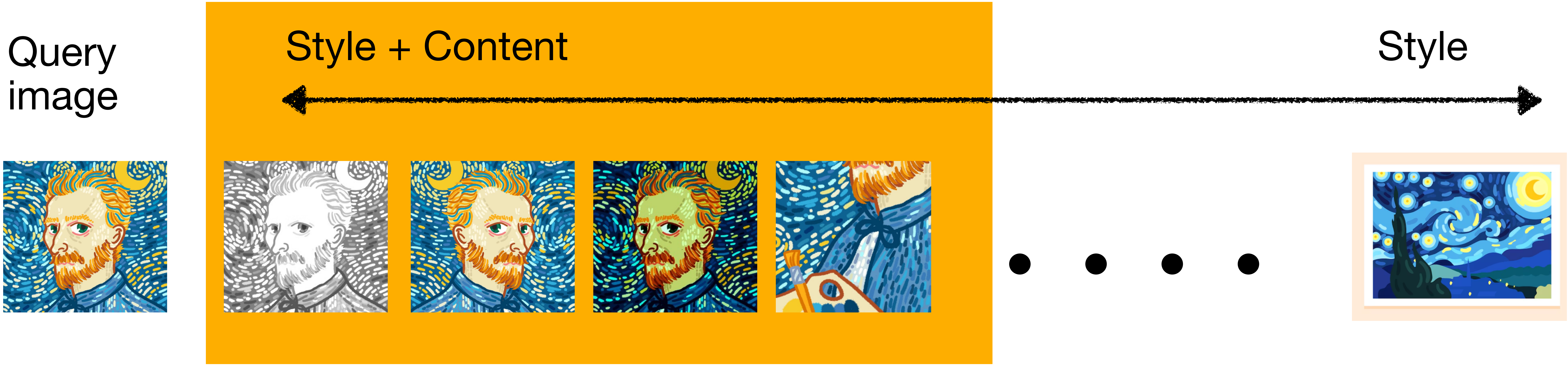


Somepalli et al. Diffusion Art or … (CVPR’23)

What is style?

…the collection of global characteristics of an image that are identified with an artist or artistic movement. These characteristics encompass various elements such as color usage, brushstroke techniques, composition, and perspective.

Current feature extractors prioritize content



Query image

Style + Content

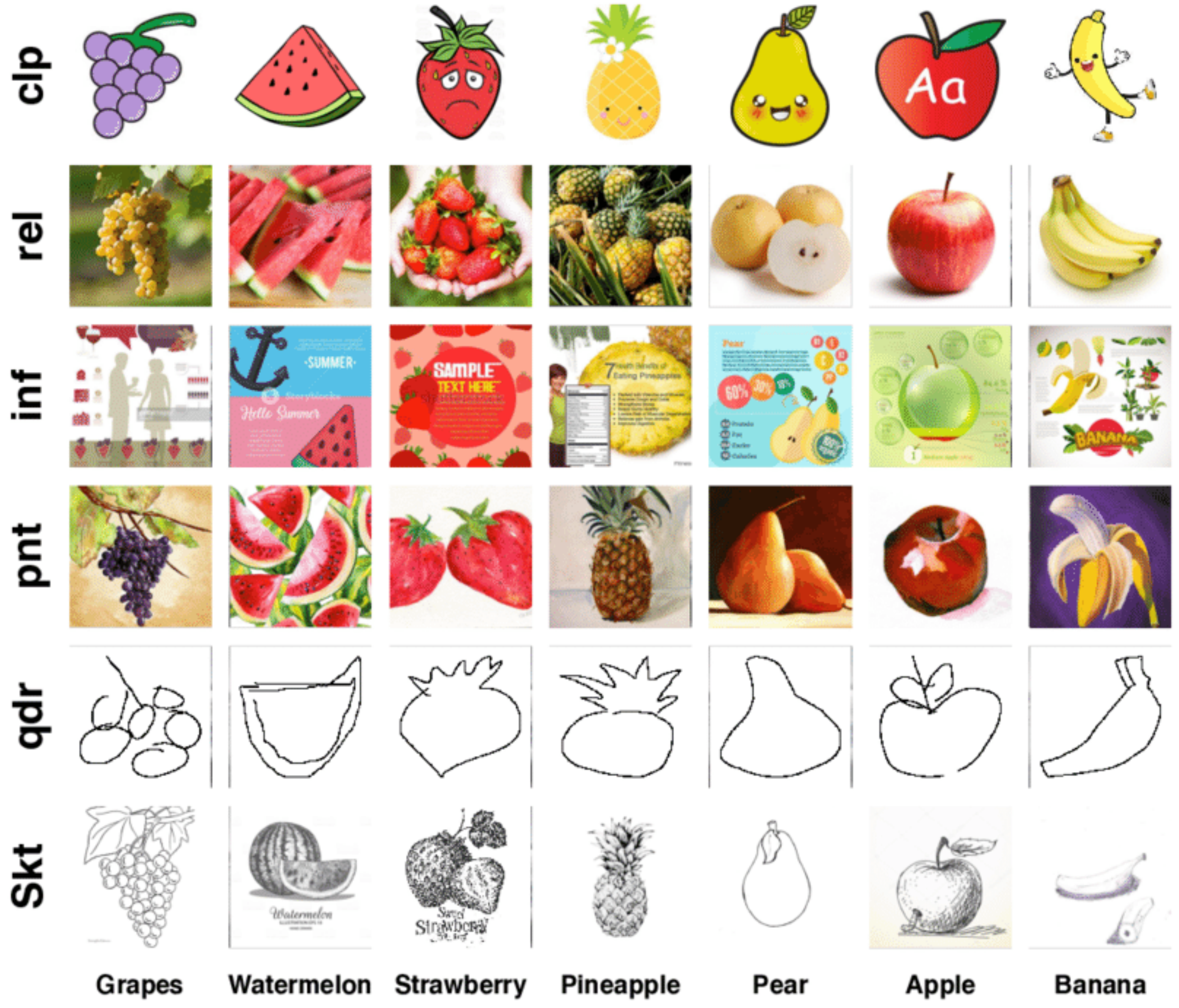
Style

Major challenges

- How to extract style?
- How to evaluate style?

Benchmark

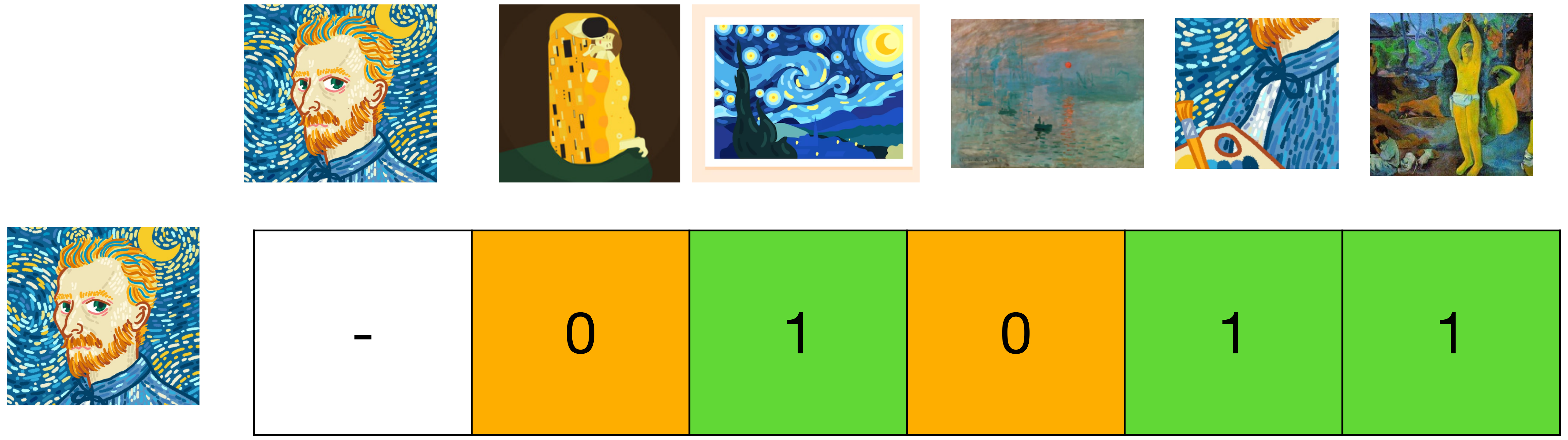
- WikiArt -
  - 80096 images - 1119 artists and 27 genres
  - Chance prob. - 0.09%
- DomainNet -
  - Six different domains: Clipart, Infograph, Painting, Real,Quickdraw, and Sketch.
  - Chance prob. - 20%



Contrastive Style Descriptors

Two main contributions

- Training loss
  - Multi-label CL
  - SSL w/o photometric augmentations (Gaussian Blur, Color Jitter etc)



-	0	1	0	1	1

Contrastive Style Descriptors

Two main contributions

- Style Dataset: LAION Styles
    - 511,921 images, 3840 style tags
    - (Deduped, overly represented tags removed)

Results

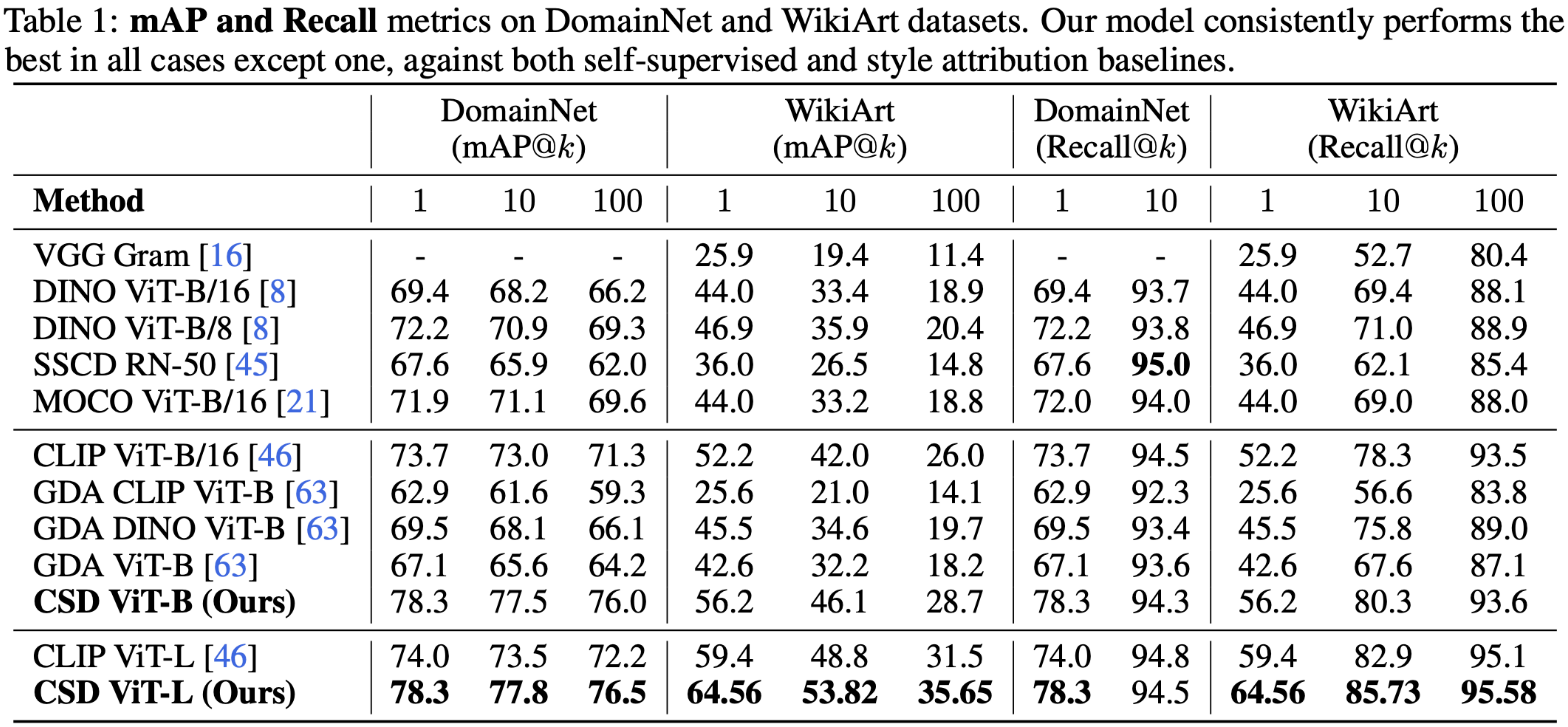
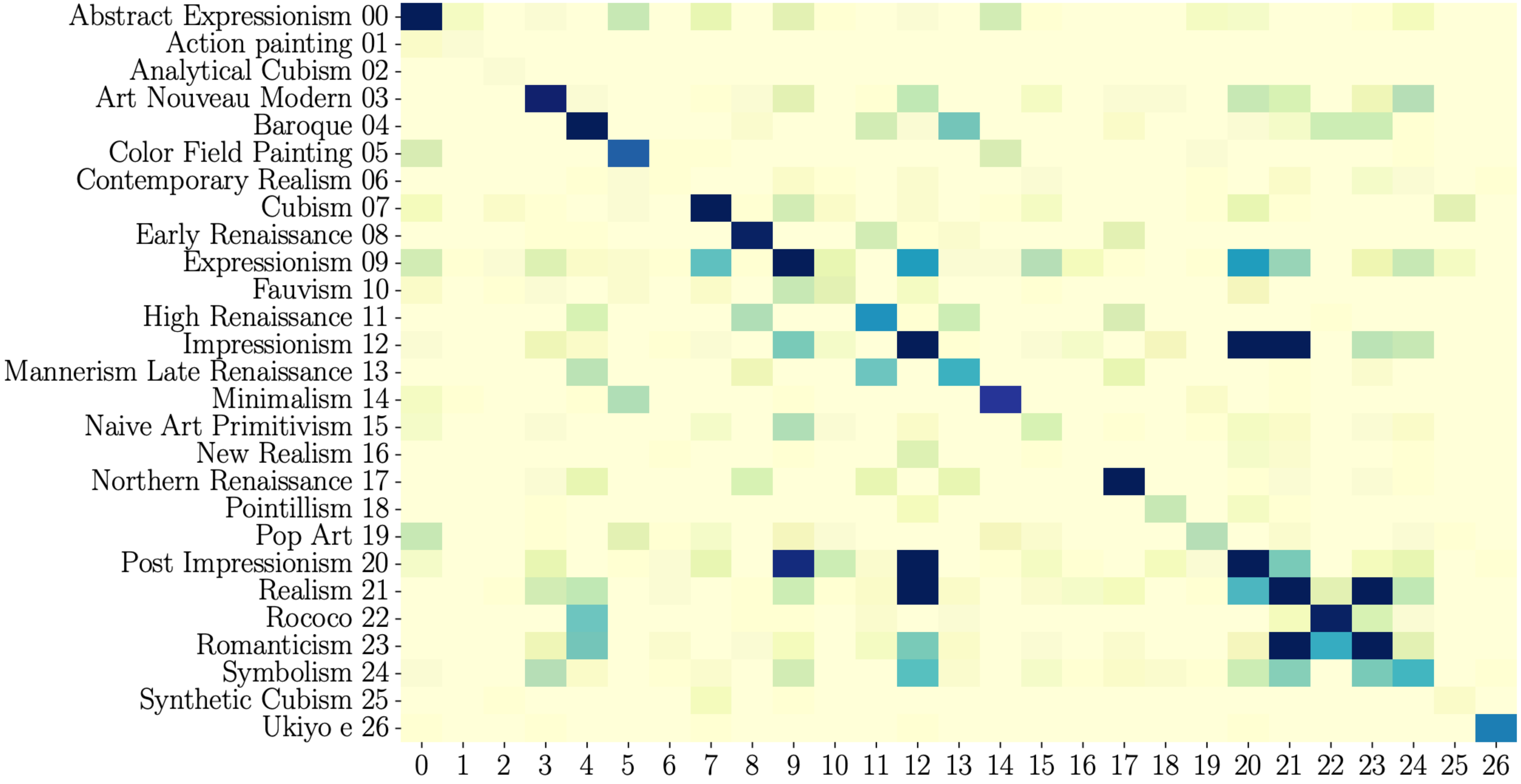


Table 1: **mAP and Recall** metrics on DomainNet and WikiArt datasets. Our model consistently performs the best in all cases except one, against both self-supervised and style attribution baselines.

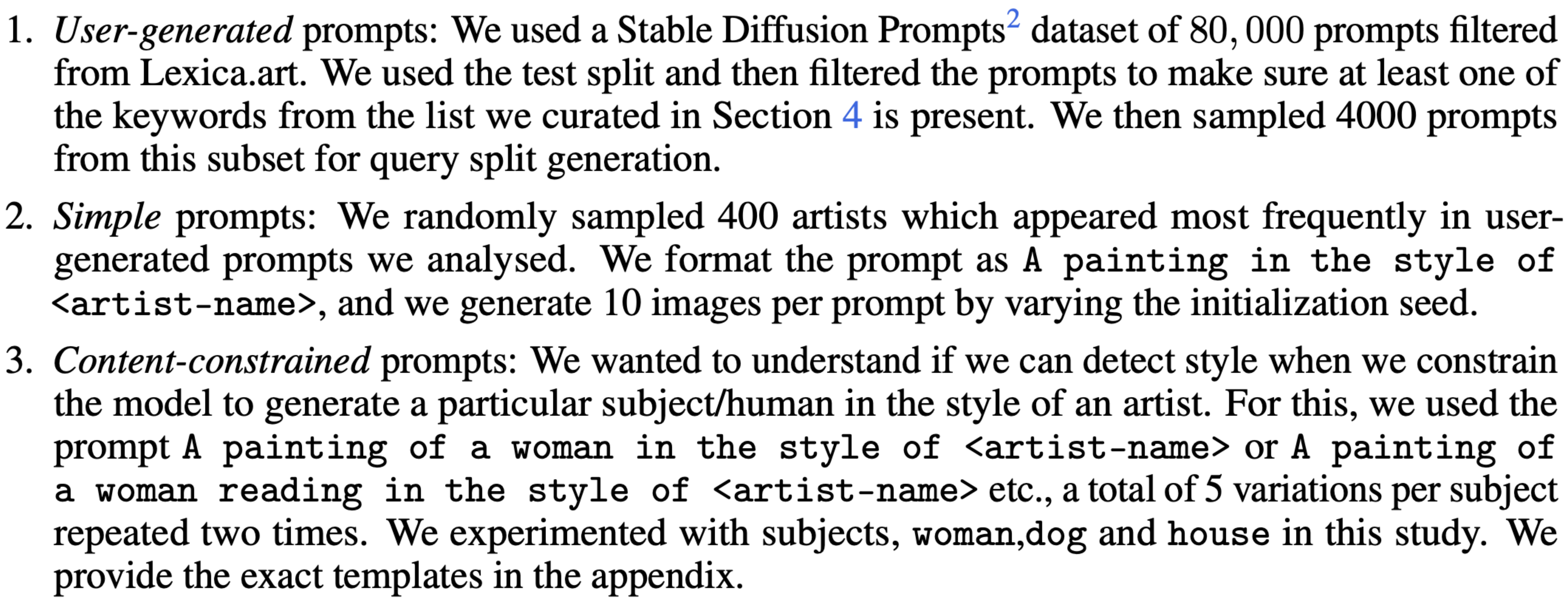
	DomainNet (mAP@k)			WikiArt (mAP@k)			DomainNet (Recall@k)		WikiArt (Recall@k)		
Method	1	10	100	1	10	100	1	10	1	10	100
VGG Gram [16]	-	-	-	25.9	19.4	11.4	-	-	25.9	52.7	80.4
DINO ViT-B/16 [8]	69.4	68.2	66.2	44.0	33.4	18.9	69.4	93.7	44.0	69.4	88.1
DINO ViT-B/8 [8]	72.2	70.9	69.3	46.9	35.9	20.4	72.2	93.8	46.9	71.0	88.9
SSCD RN-50 [45]	67.6	65.9	62.0	36.0	26.5	14.8	67.6	**95.0**	36.0	62.1	85.4
MOCO ViT-B/16 [21]	71.9	71.1	69.6	44.0	33.2	18.8	72.0	94.0	44.0	69.0	88.0
CLIP ViT-B/16 [46]	73.7	73.0	71.3	52.2	42.0	26.0	73.7	94.5	52.2	78.3	93.5
GDA CLIP ViT-B [63]	62.9	61.6	59.3	25.6	21.0	14.1	62.9	92.3	25.6	56.6	83.8
GDA DINO ViT-B [63]	69.5	68.1	66.1	45.5	34.6	19.7	69.5	93.4	45.5	75.8	89.0
GDA ViT-B [63]	67.1	65.6	64.2	42.6	32.2	18.2	67.1	93.6	42.6	67.6	87.1
CSD ViT-B (Ours)	78.3	77.5	76.0	56.2	46.1	28.7	78.3	94.3	56.2	80.3	93.6
CLIP ViT-L [46]	74.0	73.5	72.2	59.4	48.8	31.5	74.0	94.8	59.4	82.9	95.1
CSD ViT-L (Ours)	**78.3**	**77.8**	**76.5**	**64.56**	**53.82**	**35.65**	**78.3**	94.5	**64.56**	**85.73**	**95.58**

Error Analysis on WikiArt



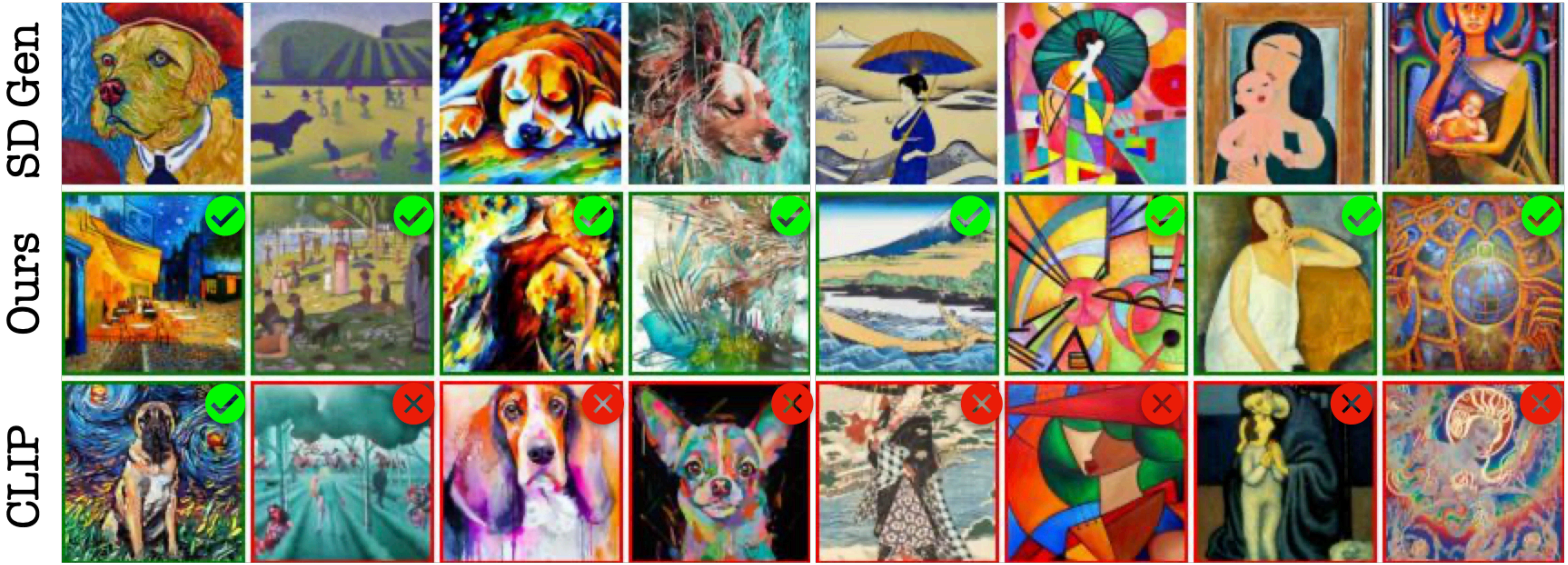
Stable Diffusion Study

Synth data creation

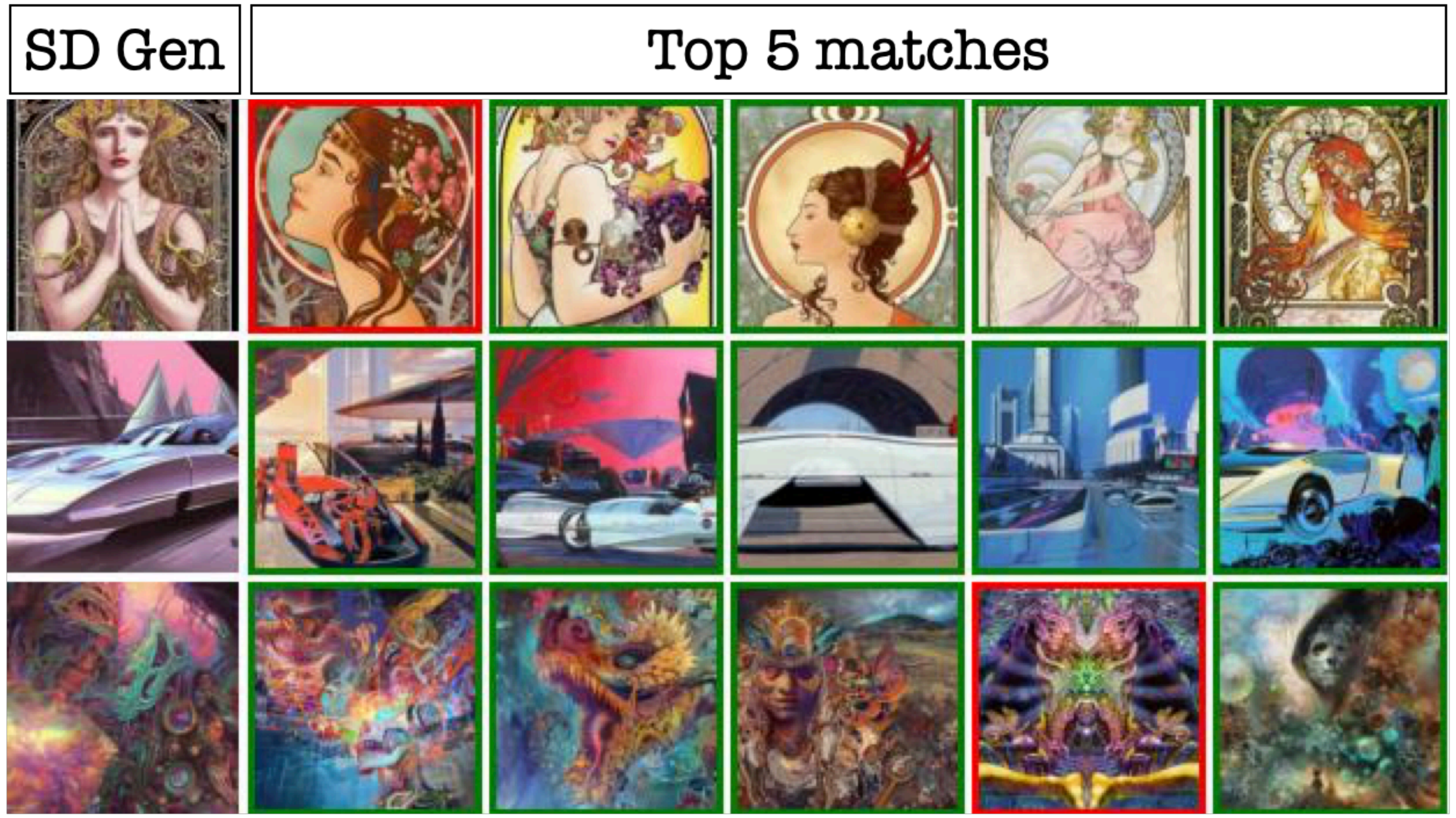


1. *User-generated* prompts: We used a Stable Diffusion Prompts[2] dataset of $80,000$ prompts filtered from Lexica.art. We used the test split and then filtered the prompts to make sure at least one of the keywords from the list we curated in Section 4 is present. We then sampled 4000 prompts from this subset for query split generation.
2. *Simple* prompts: We randomly sampled 400 artists which appeared most frequently in user-generated prompts we analysed. We format the prompt as `A painting in the style of <artist-name>`, and we generate 10 images per prompt by varying the initialization seed.
3. *Content-constrained* prompts: We wanted to understand if we can detect style when we constrain the model to generate a particular subject/human in the style of an artist. For this, we used the prompt `A painting of a woman in the style of <artist-name>` or `A painting of a woman reading in the style of <artist-name>` etc., a total of 5 variations per subject repeated two times. We experimented with subjects, `woman`,`dog` and `house` in this study. We provide the exact templates in the appendix.

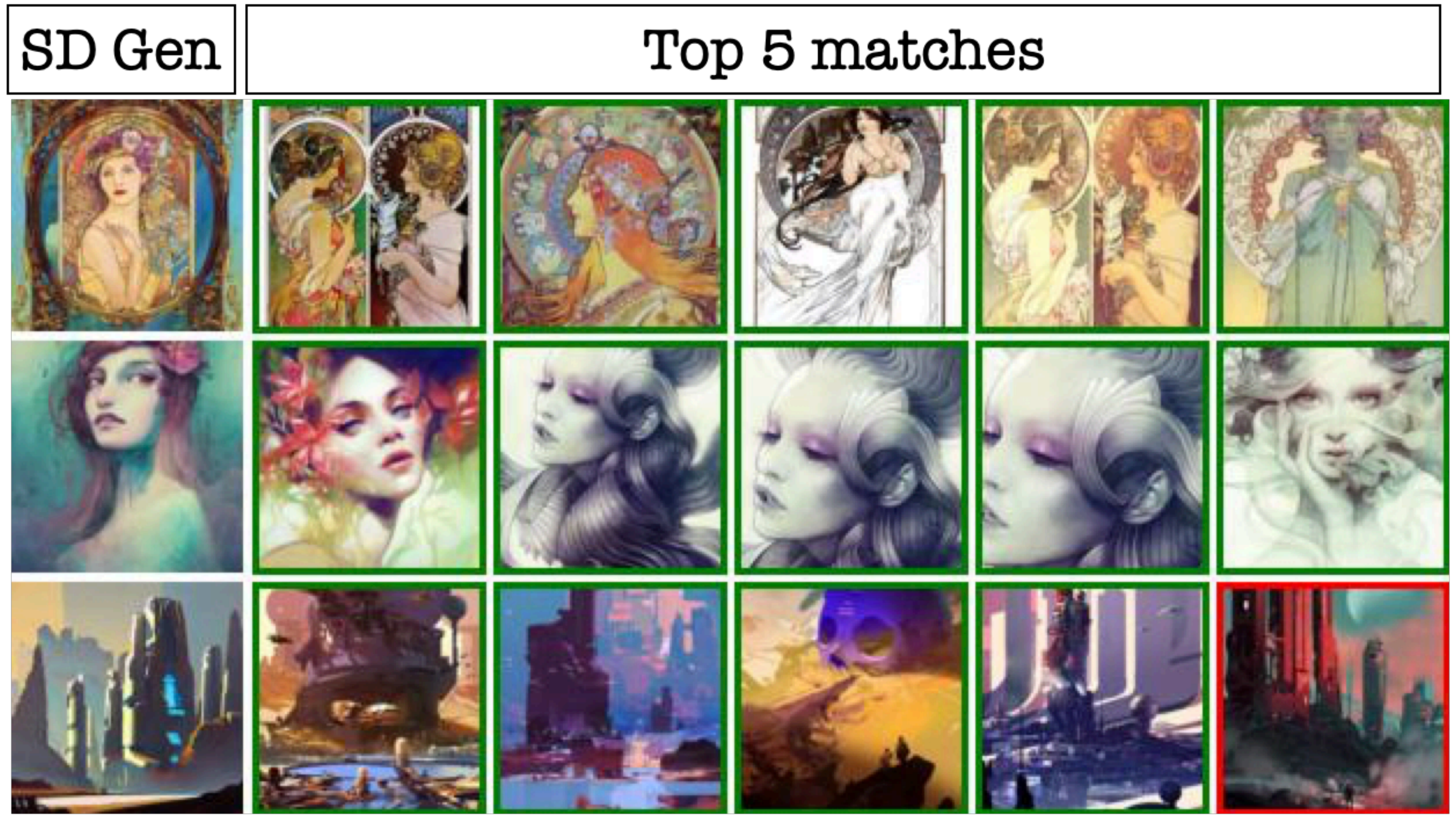
Style Matches



Style Matches



User-generated prompts



Simple-generated prompts

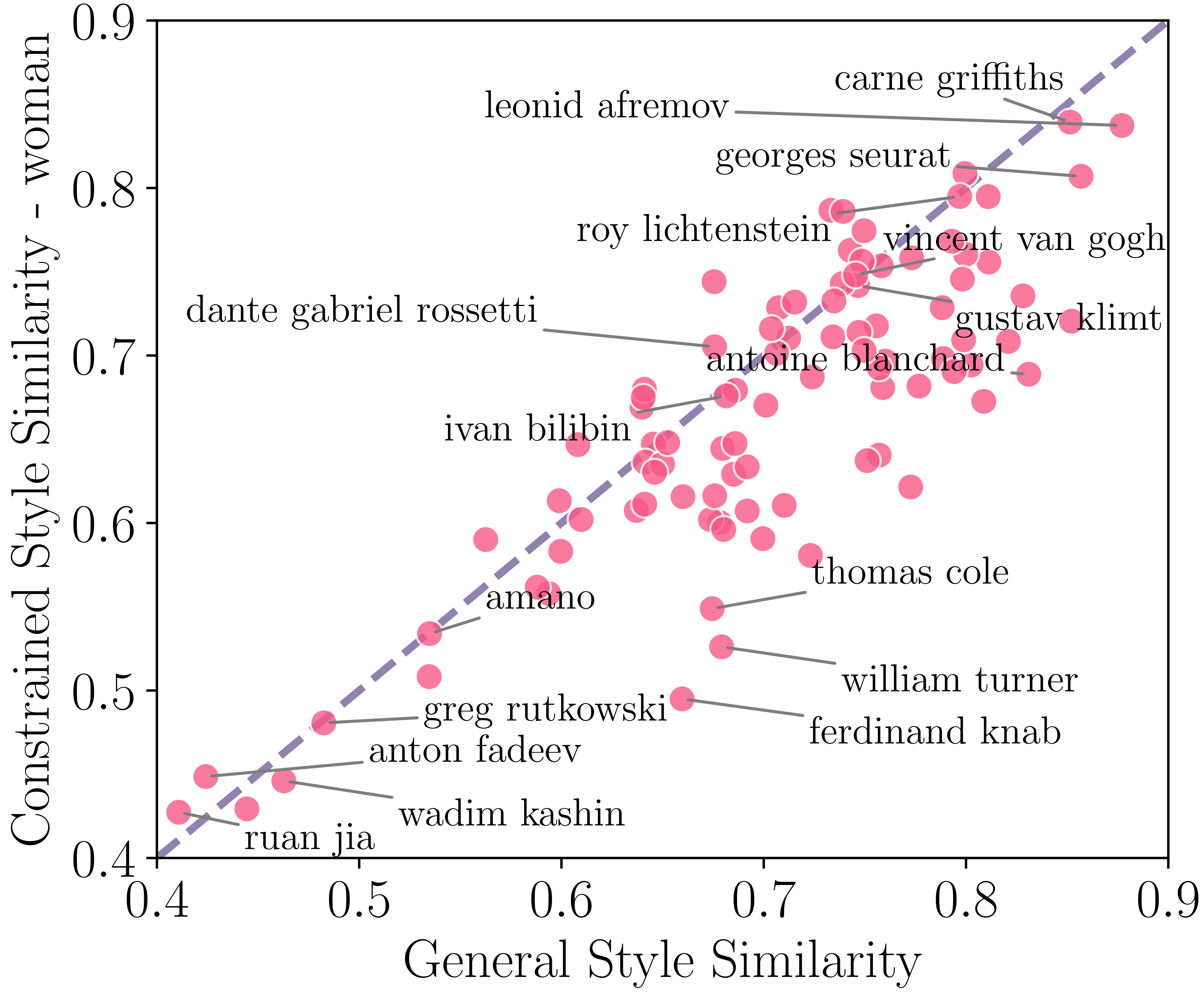
Can we predict if SD knows a style?

$$p = \Sigma_i C(o_i)$$

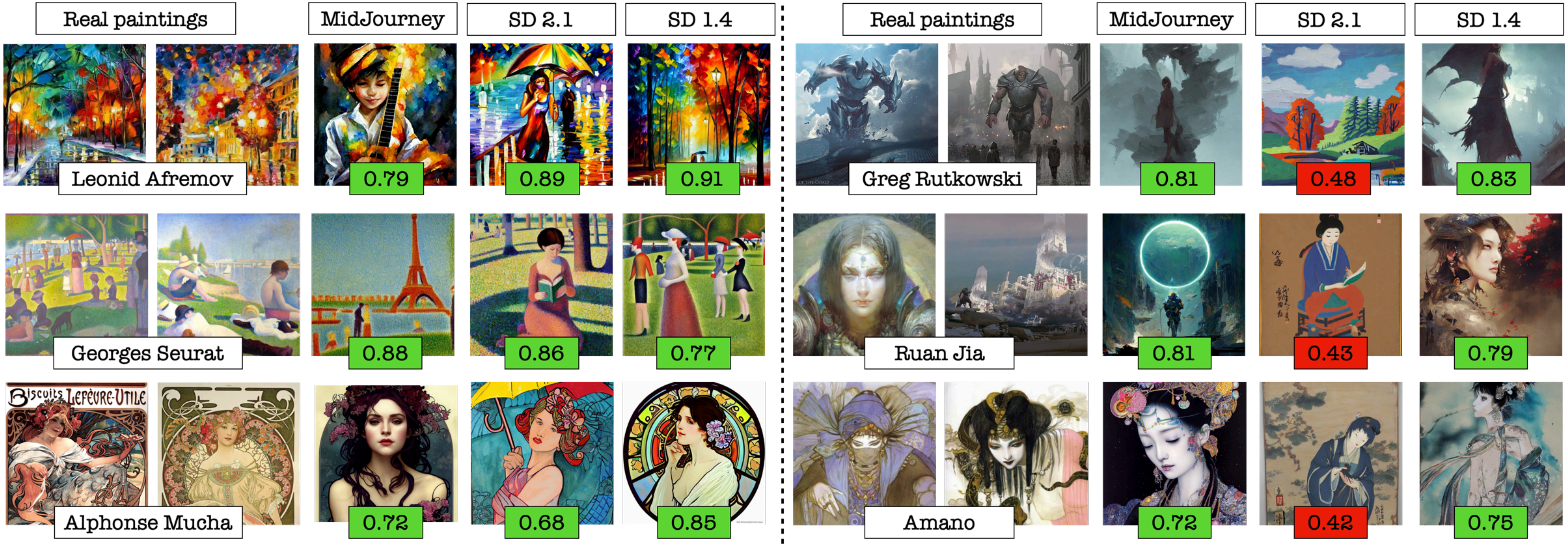
$$q = \Sigma_i C(g_i)$$

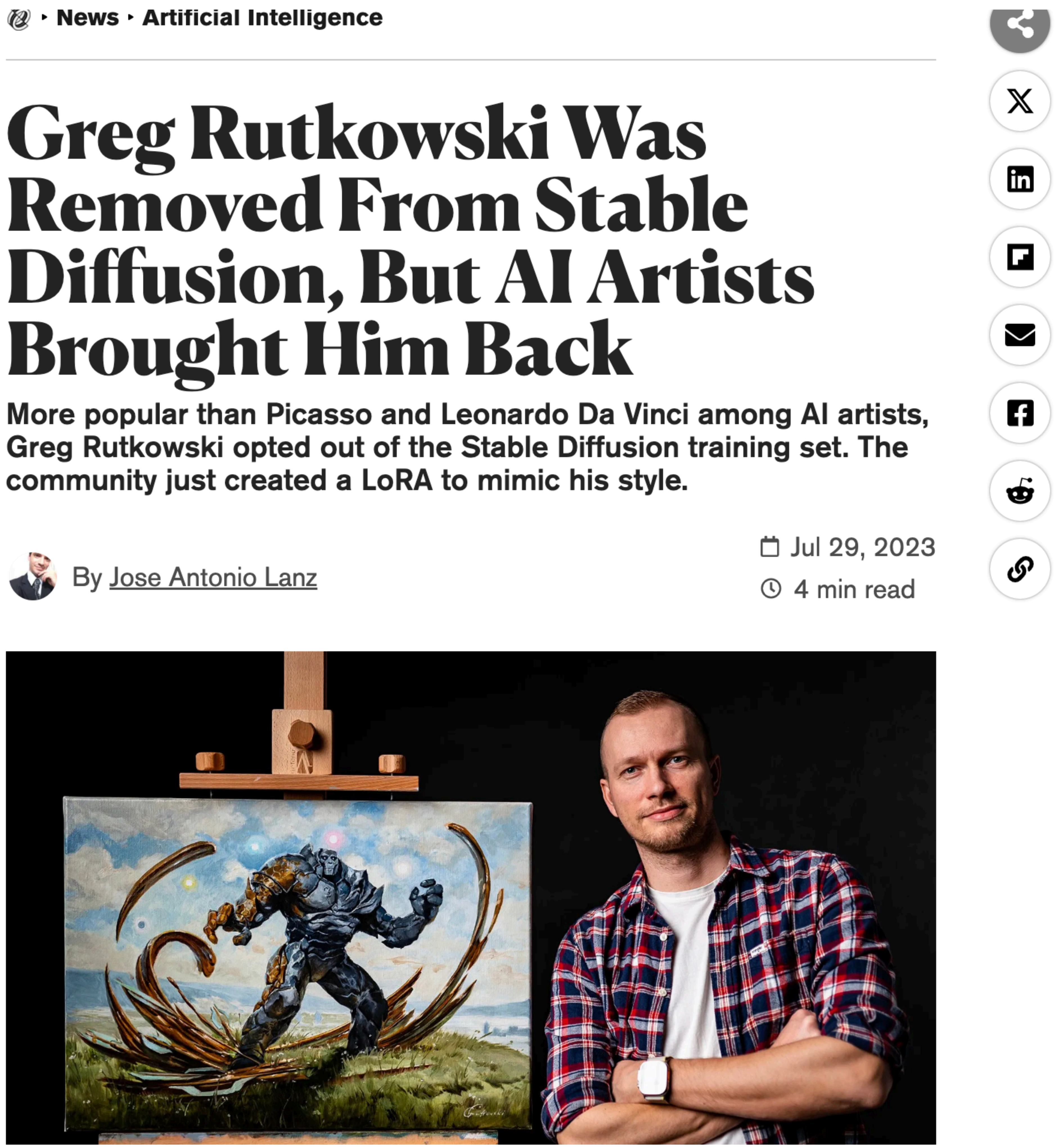
$$GSS = p \,.\, q$$

o - original images
g - generated images w/o constraint
C - feature extractor



Can we predict if SD knows a style?





News › Artificial Intelligence

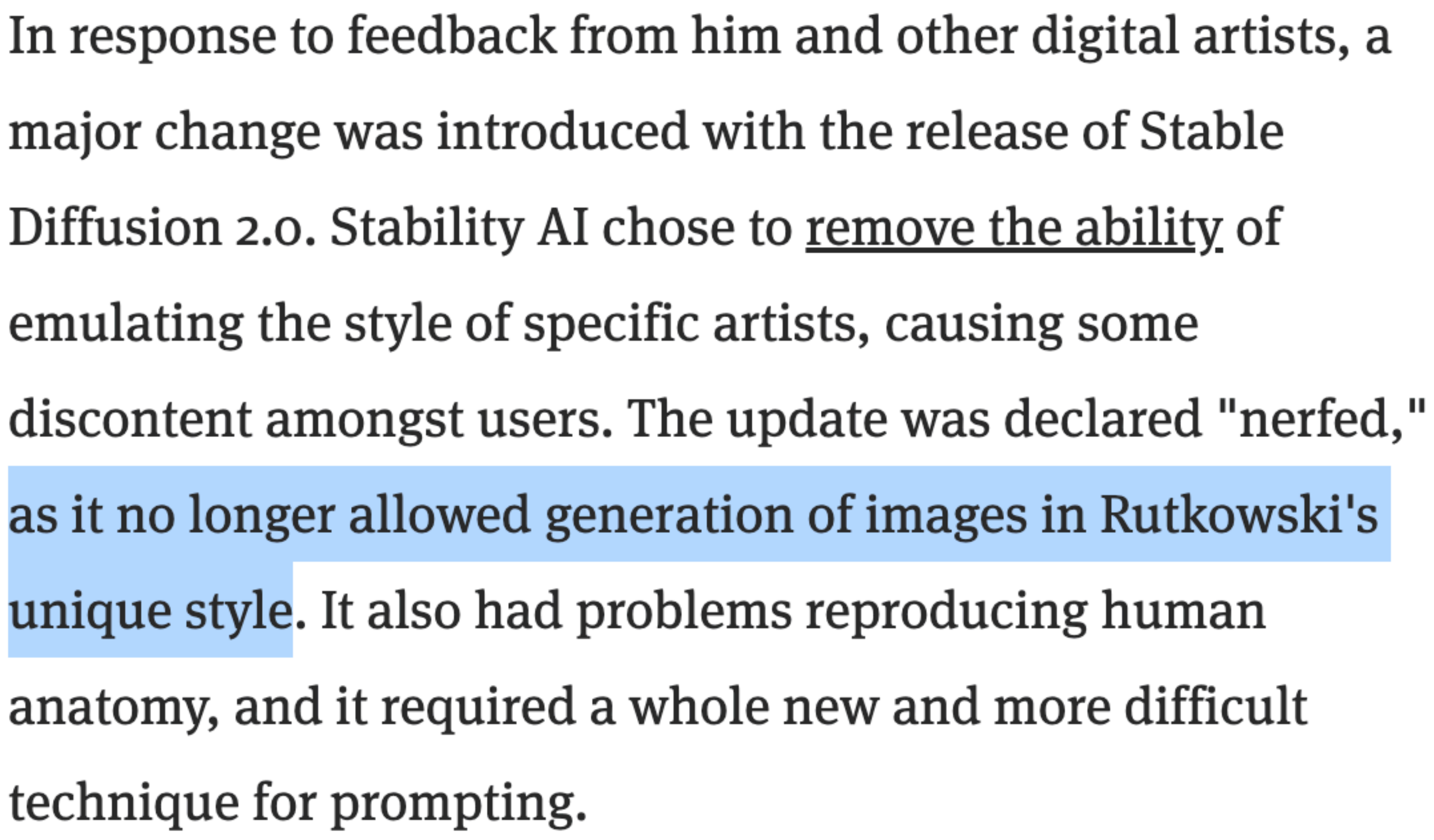
Greg Rutkowski Was Removed From Stable Diffusion, But AI Artists Brought Him Back

More popular than Picasso and Leonardo Da Vinci among AI artists, Greg Rutkowski opted out of the Stable Diffusion training set. The community just created a LoRA to mimic his style.

By Jose Antonio Lanz

Jul 29, 2023

4 min read



In response to feedback from him and other digital artists, a major change was introduced with the release of Stable Diffusion 2.0. Stability AI chose to remove the ability of emulating the style of specific artists, causing some discontent amongst users. The update was declared "nerfed," as it no longer allowed generation of images in Rutkowski's unique style. It also had problems reproducing human anatomy, and it required a whole new and more difficult technique for prompting.

https://decrypt.co/150575/greg-rutkowski-removed-from-stable-diffusion-but-brought-back-by-ai-artists

Potential directions

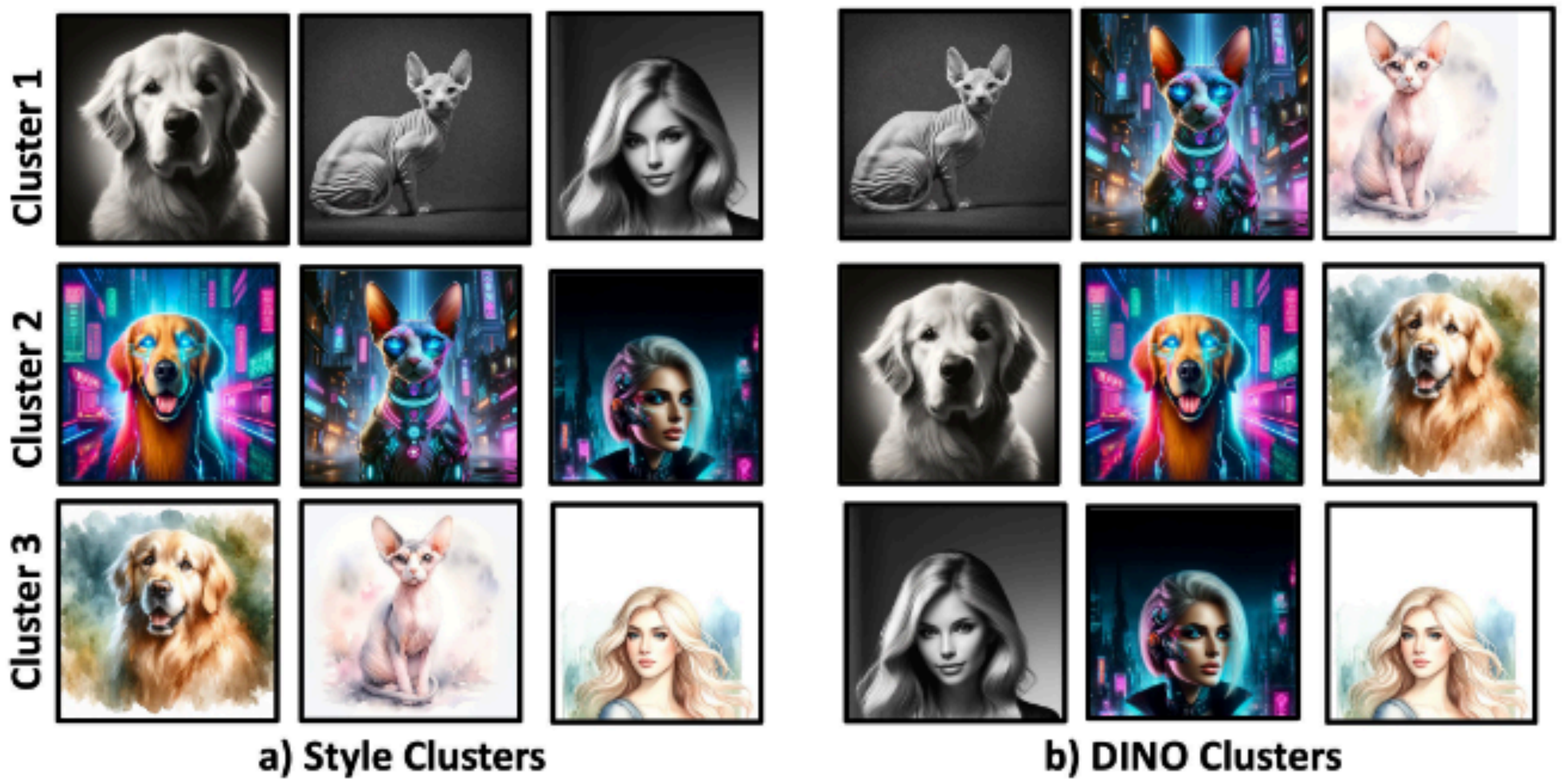
- Further clean up the dataset
- More exhaustive arch search or improve the training
- Style definition is flawed. Maybe train on movements
- Style detection in more complicated prompts
    - We observed style doesn’t always transfer

Downstream applications

- Use CSD to improve style transfer in diffusion
  - RB-Modulation paper - https://arxiv.org/abs/2405.17401
- Dataset cleanup and curation
  - StyleBreeder - https://arxiv.org/abs/2406.14599

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8

(a) User-generated images from 10 random clusters



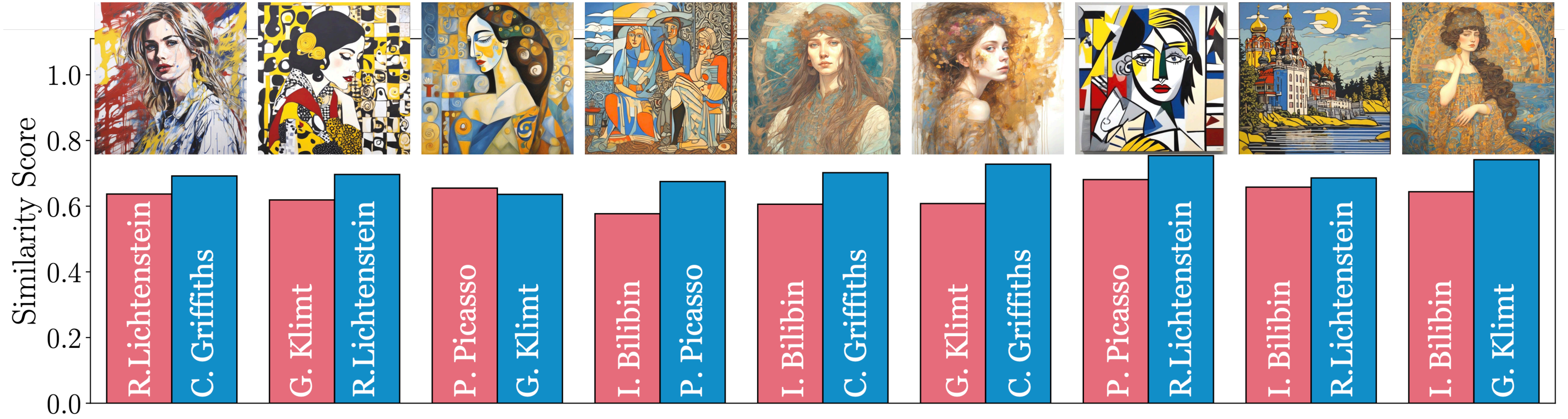
(b) Style-based vs. traditional clustering

Thank you!

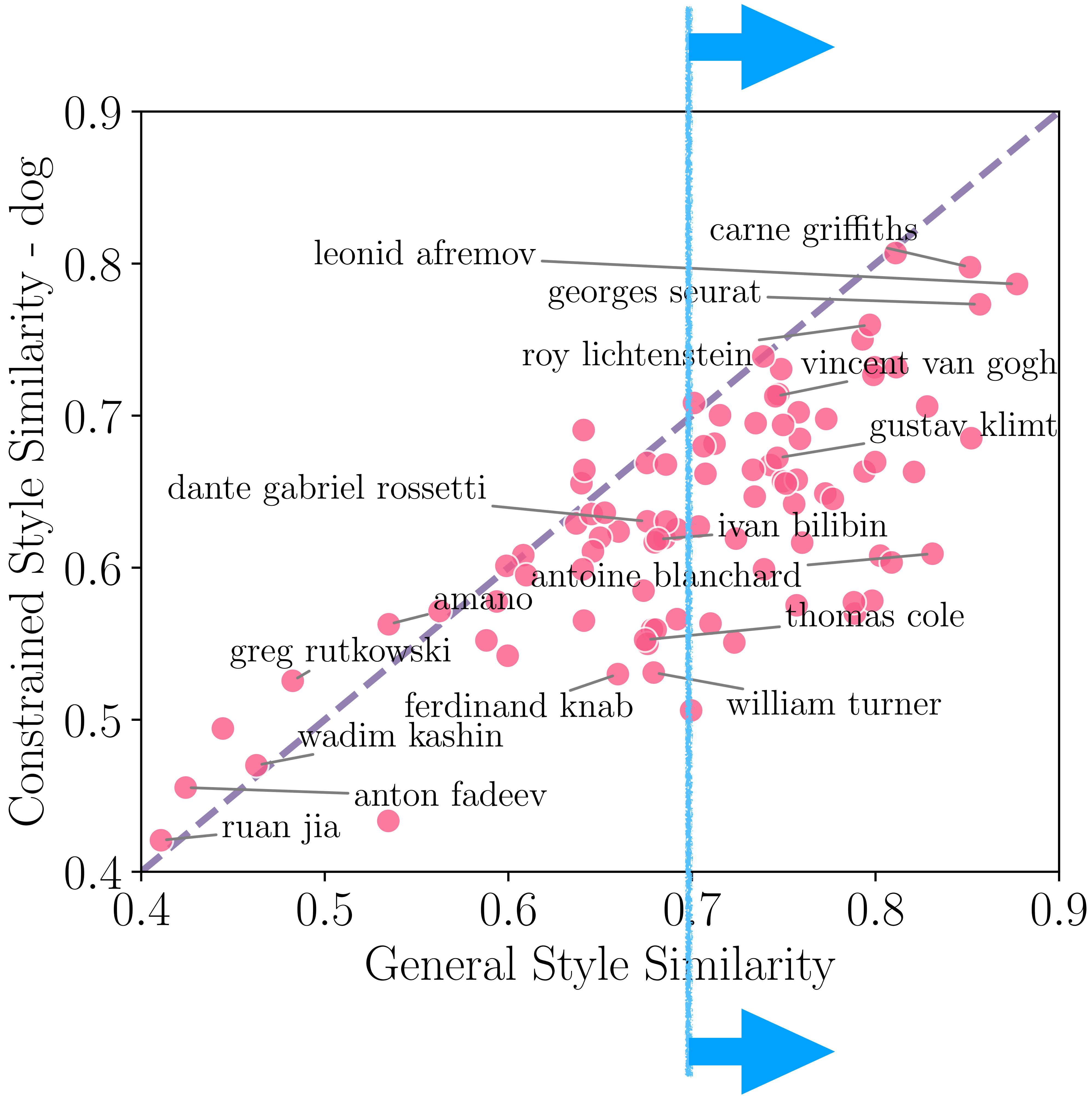
somepago.github.io

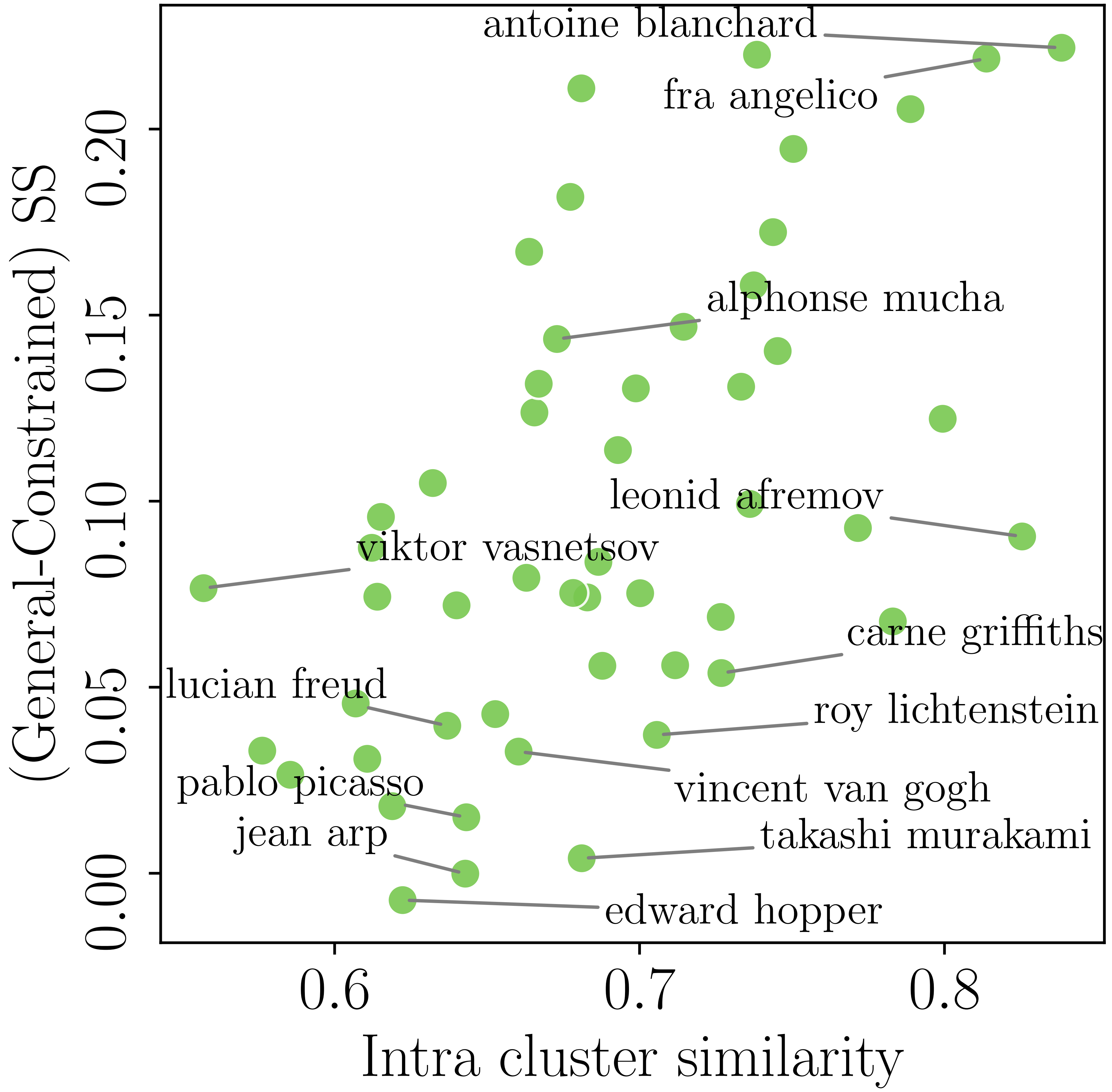
🐦 gowthami_s

Prefer one artist over the other?

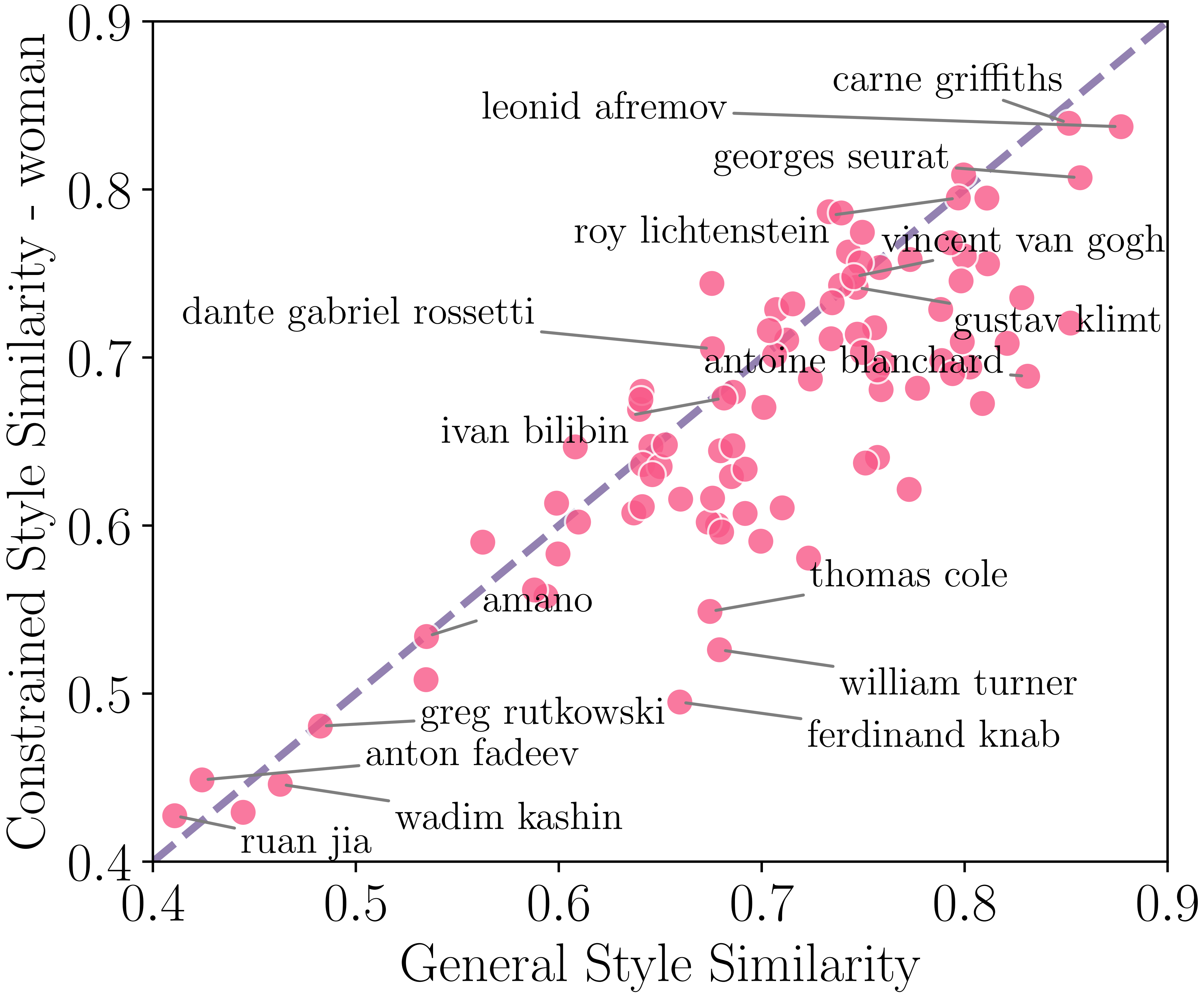


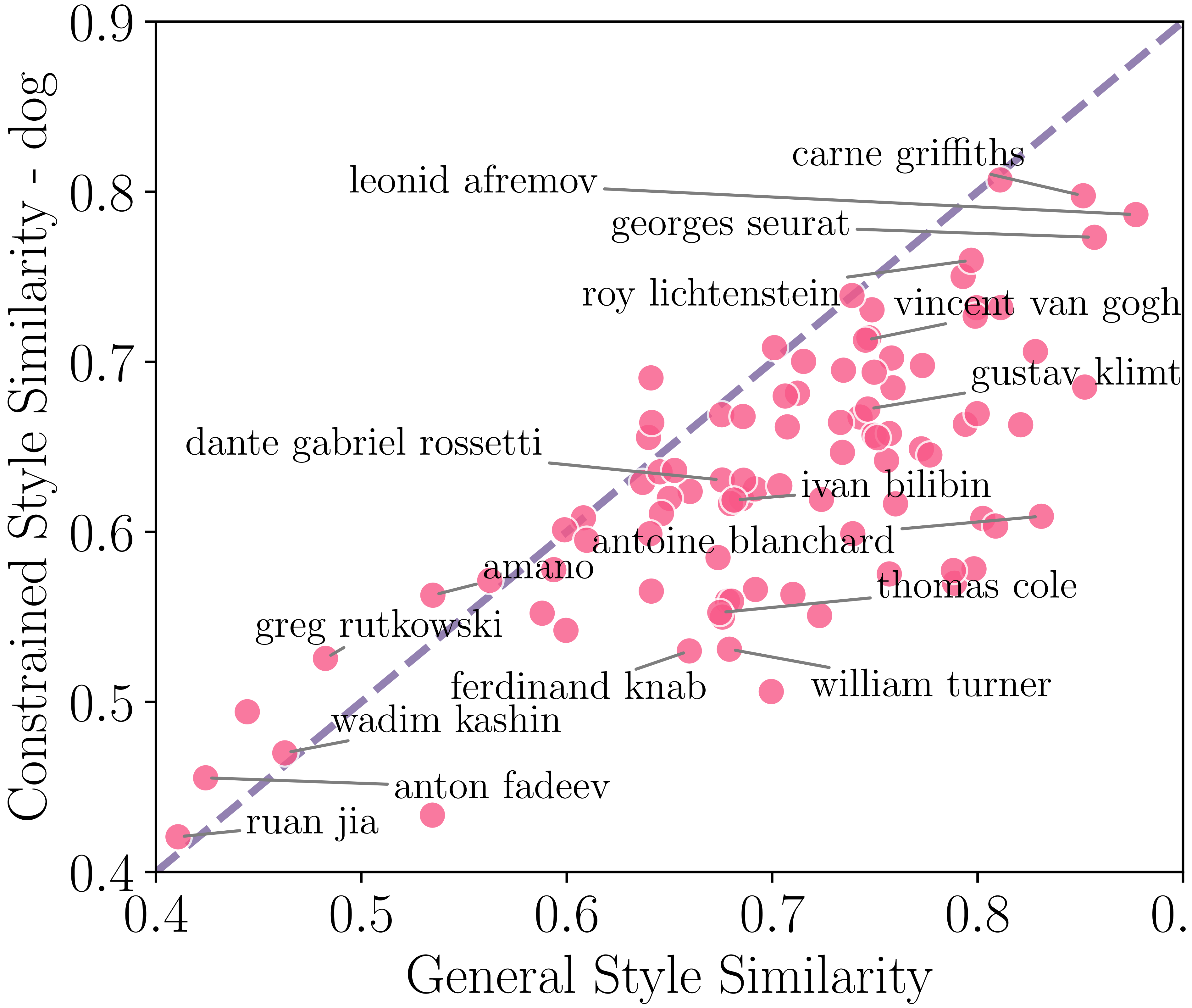
Can we predict which styles generalize?





Can we predict which styles generalize?


Constrained Style Similarity - woman
General Style Similarity
carne griffiths
leonid afremov
georges seurat
roy lichtenstein
vincent van gogh
dante gabriel rossetti
gustav klimt
antoine blanchard
ivan bilibin
thomas cole
amano
william turner
greg rutkowski
ferdinand knab
anton fadeev
wadim kashin
ruan jia


Constrained Style Similarity - dog
General Style Similarity
carne griffiths
leonid afremov
georges seurat
roy lichtenstein
vincent van gogh
gustav klimt
dante gabriel rossetti
ivan bilibin
antoine blanchard
amano
thomas cole
greg rutkowski
ferdinand knab
william turner
wadim kashin
anton fadeev
ruan jia