Measuring Style Similarity in Diffusion Models
Appendix

A Human Evaluation

We used 30 human subjects (excluding authors) to evaluate human-level performance on the task of style matching. Participation in the study was voluntary, and none of the subjects had any prior familiarity with the task. The authors manually vetted the data presented to the subjects for the absence of any offensive or inappropriate visuals. The subjects were informed that their responses would be used to compare the human performance with ML models. We did not collect any personally identifiable information and secured an *exempt* status from IRB at our institute for this study.

B List of artists in style analysis

The following are the artists in the style analysis discussed in Sec. 2 - `roy lichtenstein, justin gerard, amedeo modigliani, leonid afremov, ferdinand knab, kay nielsen, gustave courbet, thomas eakins, ivan shishkin, viktor vasnetsov, ivan aivazovsky, frederic remington, frederic edwin church, marianne north, salvador dali, pablo picasso, robert delaunay, ivan bilibin, rembrandt, frans hals, dante gabriel rossetti, max ernst, diego rivera, andy warhol, wadim kashin, caspar david friedrich, jan matejko, albert bierstadt, vincent van gogh, cy twombly, amano, anton fadeev, gian lorenzo bernini, mark rothko, mikhail vrubel, hieronymus bosch, katsushika hokusai, alphonse mucha, winslow homer, george stubbs, taro yamamoto, richard hamilton, carne griffiths, edward hopper, jan van eyck, francis picabia, michelangelo, arkhip kuindzhi, isaac levitan, gustave dore, antoine blanchard, john collier, paul klee, caravaggio, m.c. escher, leonardo da vinci, alan bean, greg rutkowski, jean arp, marcel duchamp, thomas cole, takashi murakami, thomas kinkade, raphael, hubert robert, john singer sargent, fra angelico, gustav klimt, ruan jia, harry clarke, william turner, claude monet, gerhard richter, frank stella, francisco goya, giuseppe arcimboldo, otto dix, lucian freud, jamie wyeth, rene magritte, titian, john atkinson grimshaw, man ray, albert marquet, mary cassatt, georges seurat, fernando botero, martin johnson heade, william blake, ilya repin, john william waterhouse, edmund dulac, peter paul rubens, frank auerbach, frida kahlo`.

C Model Ablations

We conducted several ablations on the choice of the backbone for the model, the λ hyperparameter in the loss formulation and the temperature hyperparameter, τ. We present the results on the Wikiart dataset for various configurations in Tab. 4. In the first ablation, we can see that CLIP ViT-B as the initialization for CSD gives the best performance. In the second ablation, we study the loss hyperparameter λ. $\lambda = 0$ model is trained only on Multi-label Contrastive loss, while $\lambda = \infty$ refers to model with only SSL loss. We see the best outcome when we combine both losses in the training. In the final ablation we see the effect of temperature hyperparameter. We clearly see the best outcome at $\tau = 0.1$, which is conventionally the temperature used in other papers as well.

Table 4: Model Ablations: We present results on Wikiart dataset. All the models are trained for same number of iterations. The baseline hyperparameters are $\lambda = 0.2$ and $\tau = 0.1$ and backbone initialization with CLIP ViT-B. ★ refers to the CSD ViT-Base variant we discussed in the main paper.

		mAP@k			Recall@k		
Ablation	Variant	1	10	100	1	10	100
Architecture,pre-training style	SSCD RN-50	33.2	24.8	14.11	33.2	58.8	83.8
	CLIP RN-50	51.8	44.2	25.2	51.8	77.0	92.1
	DINO ViT-B	49.8	39.6	24.4	49.8	76.3	92.6
	CLIP ViT-B ★	56.2	46.1	28.7	56.2	80.3	93.6
Loss Hyperparameter	$\lambda = \infty$	49.1	40.2	23.9	49.1	70.3	85.5
	$\lambda = 1$	52.3	48.1	26.2	52.3	80.1	91.4
	$\lambda = 0.2$ ★	56.2	46.1	28.7	56.2	80.3	93.6
	$\lambda = 0.1$	54.9	45.5	28.3	54.9	78.2	92.1
	$\lambda = 0$	51.8	44.9	26.6	51.8	79.5	90.3
Temperature	$\tau = 0.01$	55.1	44.4	27.3	55.1	79.7	93.6
	$\tau = 0.1$ ★	56.2	46.1	28.7	56.2	80.3	93.6
	$\tau = 0.5$	42.3	39.9	20.3	42.3	70.6	90.1
	$\tau = 1$	36.2	28.7	18.0	36.2	64.4	86.3

D Analysis of other generative models.

We present mAP@1 for our feature extractor and CLIP model in Tab. 5 for two open-source T2I models Würstchen [44] and PixArt-α [9]. See Fig. 8 for sample generations and corresponding real image matches from the LAION. The artists are `Van Gogh`, `Afremov` and `Kandinsky`. **Our feature extractor outperforms CLIP on these 2 new generative models which we did not explore in the main text**.

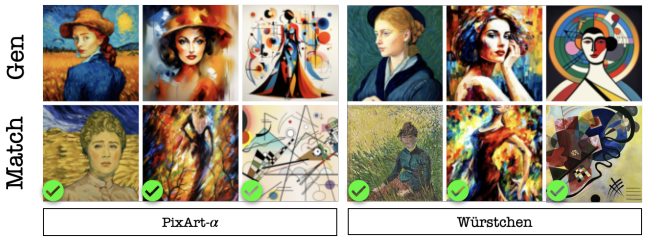


Fig. 8: A few generations and matches with our feature extractor

Table 5: CLIP vs Ours: mAP@1 on other OSS T2I diffusion models

Gen model	Model	Simple	Woman	Dog	House
PixArt	CLIP ViT-L	21.3	6.8	2.1	3.9
	Ours	**23.4**	**9.3**	**4.5**	**5.2**
Würstchen	CLIP ViT-L	24.6	7.8	3.2	4.2
	Ours	**27.9**	**10.2**	**5.7**	**6.4**

E Dataset Curation Details

E.1 Dataset Deduplication

During our initial retrieval experiments, we found out that most of the top-N nearest neighbors that were being returned with respect to a query image were essentially the same images. We performed deduplication of this dataset ($\sim$ 1.3 million images) by computing the SSCD [46] embedding of all the *Database* images and then computing the similarity of each image with all the other images and then recursively removing the duplicate images, keeping only one image in any connected graph. The labels corresponding to the removed images were merged with the labels of the master image such that the master image has labels of all its children, the children being removed from the *Database*. We considered 2 images to be the same if the inner product between their SSCD embeddings was greater than 0.8. After the filtering, we are left with $\sim$ 0.5 million images

E.2 Dataset Safety Filtering

A recent study [62] found that LAION dataset contains a small fraction of NSFW and CSAM content. To ensure the safety and integrity of our dataset, we ran all of ContraStyles through PhotoDNA API by Microsoft[7]. PhotoDNA is regarded an industry standard and can detect problematic content even if

[7] https://www.microsoft.com/en-us/PhotoDNA

the images are slightly altered. We also run an open-source NSFW classification model[8] to filter out any remaining problematic content.

F Stable Diffusion Analysis Extended

Content-constrained templates. The following templates are used in generating content-constrained SD synthetic datasets.

- *Woman-constrained prompts:* (1) A painting of a woman in the style of <artist>, (2) A painting of a woman holding an umbrella in the style of <artist>, (3) A painting of a woman wearing a hat in the style of <artist>, (4) A painting of a woman holding a baby in the style of <artist>, (5) A painting of a woman reading a book in the style of <artist>
- *Dog-constrained prompts:* (1) A painting of a dog in the style of <artist>, (2) A painting of dog playing in the field in the style of <artist>, (3) A painting of a dog sleeping in the style of <artist>, (4) A painting of two dogs in the style of <artist>, (5) A portrait of a dog in the style of <artist>
- *House-constrained prompts:* (1) A painting of a house in the style of <artist>, (2) A painting of a house in a forest in the style of <artist>, (3) A painting of a house in a desert in the style of <artist>, (4) A painting of a house when it's raining in the style of <artist>, (5) A painting of a house on a crowded street in the style of <artist>

Fig 5 prompts. Please find the prompts to generate the images presented in Fig. 6. The prompts are from left to right in the order.

1. A painting of a dog in the style of **Van Gogh**
2. A painting of dog playing in the field in the style of **Georges Seurat**
3. A painting of a dog sleeping in the style of **Leonid Afremov**
4. A painting of a dog sleeping in the style of **Carne Griffiths**
5. A painting of a woman holding an umbrella in the style of **Katsushika Hokusai**
6. A painting of a woman holding an umbrella in the style of **Wassily Kandinsky**
7. A painting of a woman holding a baby in the style of **Amedeo Modigliani**
8. A painting of a woman holding a baby in the style of **Alex Gray**

First observations on synthetic datasets. When we scan through the generations, we find that for simple prompts the SD 2.1 model is borrowing the contents along with the artistic styles of an artist. For example, for the prompt `A painting in the style of Warhol` SD 2.1 always generated a version of Marilyn Diptych's painting. Similarly, all the prompts of `Gustav Klimt`

[8] https://github.com/GantMan/nsfw_model

generated “The Kiss" even at different random seeds. We observed that some artist names are strongly associated with certain images and we believe this is due to dataset memorization as also discussed in [58, 59].
Another interesting issue is, for certain artists, if the prompt content diverges too vastly from their conventional “content", the SD model completely ignores the content part sometimes and only generates the “content" typical of the artist. For example, even when the prompt is `A painting of a woman in the style of Thomas Kinkade`, the SD model still outputs an image with charming cottages, tranquil streams, or gardens. The SD model sometimes completely ignored the content element in the prompt.

F.1 Which styles do diffusion models most easily emulate?

In Fig. 2, we saw that General Style Similarity(GSS) and content-constrained style similarity scores are correlated, however, we see a few artists diverging away from the identity line. How far the score is diverging away from the identity line reflects how far out-of-distribution the ‘subject’ in the prompt is for that artist. This could serve as an indicator of the generalization capability of the artist’s style. We hypothesize that artists who painted diverse subjects may have styles that generalize better to out-of-distribution (OOD) objects. To that end, we computed intra-cluster style similarity among all the artists’ paintings. We plot the difference in General style similarity scores and content-constrained similarity scores and the artist-level intra-cluster similarity in Fig. 9. For this experiment, we selected ‘dog’ as the subject, a choice informed by the observation that many artists predominantly painted women, landscapes, or cityscapes. Thus, ‘dog’ represents a subject that is OOD, as evidenced by lower scores for the dog category compared to women or houses in Tab. 3. Additionally, we limited our analysis to artists with GSS greater than 0.7 to ensure the model’s proficiency in reproducing the artist’s style in an unconstrained scenario. We see a high correlation of 0.568 between these 2 variables. It seems painters of diverse subjects are more likely to have their style replicated for out of distribution objects.

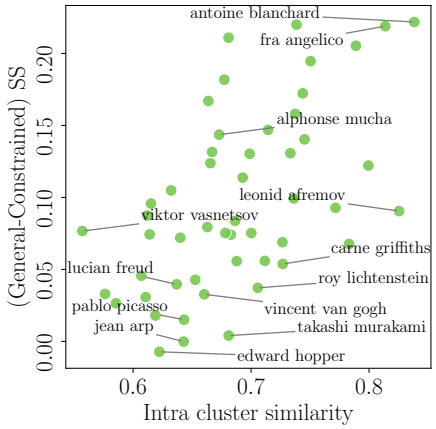
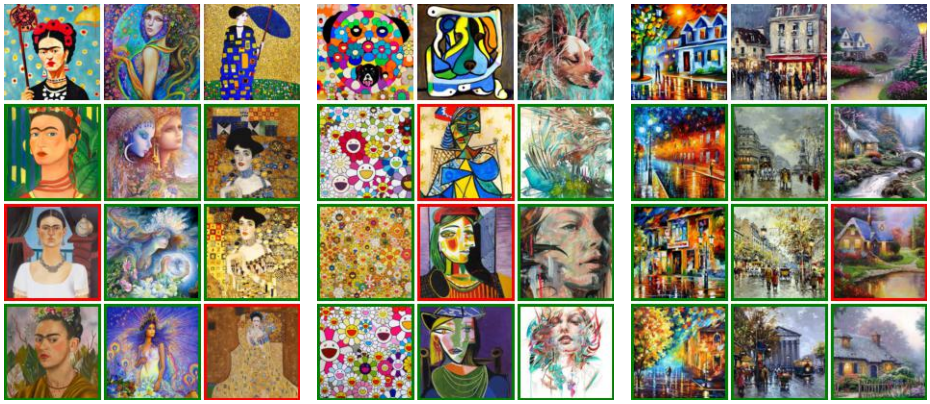


Fig. 9: Style generalization to new subjects: The X-axis represents the diversity of the artists’ paintings. The Y-axis shows the difference between General Style Similarity and Content-constrained Style Similarity on dog subjects.

In the top right corner of the figure, we note Antoine Blanchard, known for his Paris cityscapes, and Fra Angelico, who primarily focused on religious themes, including biblical scenes, saints, and other religious iconography. Conversely, in the lower left, we find Jean Arp and Pablo Picasso, whose work is characterized by abstract and non-traditional styles, encompassing a wide array of subjects. We conducted a qualitative verification to ascertain that style transfer was effective for artists located in the bottom right corner of the figure. Although this evaluation is not comprehensive, it serves as a preliminary investigation that may provide insights into the factors contributing to the generalization of style in diffusion models.

Some Caveats. We emphasize that the above results should be taken with a grain of salt. Firstly, the LAION dataset, which we used, is inherently noisy, despite the sanitization steps we implemented as outlined in Section 4. The captions within this dataset frequently have issues. For example, tags relating to an artist or style might be absent, leading to inaccuracies in our evaluation. Even when the model correctly maps to the appropriate images in the dataset, these missing tags can cause correct results to be wrongly categorized as incorrect. Secondly, the curated style list from CLIP Interrogator is noisy. There are frequent re-occurrences of the same artist with different spellings in the style list. For example, if `Van Gogh` *vs.*`Vincent Van Gogh` ended up as different 'style' classes, and led to a few meaningless "errors." Lastly, we assume that the model strictly adheres to the prompts during generation. However, our observations indicate that in a few cases, the SD model tends to ignore the style component of a prompt and focuses predominantly on the content. This divergence results in what would have been positive matches being classified as negatives. These factors collectively suggest that while the results are informative, they should be interpreted with an understanding of the underlying limitations and potential sources of error in the data and model behavior.

Qualitative results extended. We present a few content-constrained generations and their respective top-2 matches using CSD ViT-L feature extractor in Fig. 10. We provide extended versions of Fig. 5 in Fig. 11 and Fig. 12. These figures depict the results obtained from Stable Diffusion (SD) generations using both "simple" artist prompts and "user-generated" prompts, along with their respective top-10 matches. The first column corresponds to the SD generation, while the subsequent columns display the identified matches. To aid in the interpretation of these matches, we employed color-coded boxes to indicate the accuracy of the match. Specifically, green boxes represent true-positive matches, while red boxes indicate false-positive matches. However, it is important to note that the ground-truth labels assigned to the matched images may occasionally be incorrect because the ground-truth labels are generated from the LAION caption which may not always contain the artist's name. Our analysis reveals several instances of such mislabeling, particularly



(a) A woman in style of <Y> (b) A dog in style of <Y> (c) A house in style of <Y>

Fig. 10: Top row: Images generated by Stable Diffusion [50] using the prompt - `A <X> in the style of <Y>` where `X` comes from the set {`woman`, `dog`, `house`} and `Y` in order are `Frida Kahlo, Josephine Wall, Gustav Klimt, Takashi Murakami, Picasso, Carne Griffiths, Leonid Afremov, Antoine Blanchard, Thomas Kinkade`. **Next two rows:** top three *style* neighbors of the generated images from the LAION aesthetics datasets [55] as predicted by our model. The green and red box around the image indicate whether it was a true or false positive prediction based on whether the caption of the LAION image contained the name of the artist Y (used to generate the images).

evident in Fig. 12. Notably, numerous images that bear striking stylistic resemblance to the generated images are erroneously labeled as false positives. These findings underscore the challenges involved in accurately assessing style copying in SD and emphasize the need for further exploration and refinement of evaluation methods.

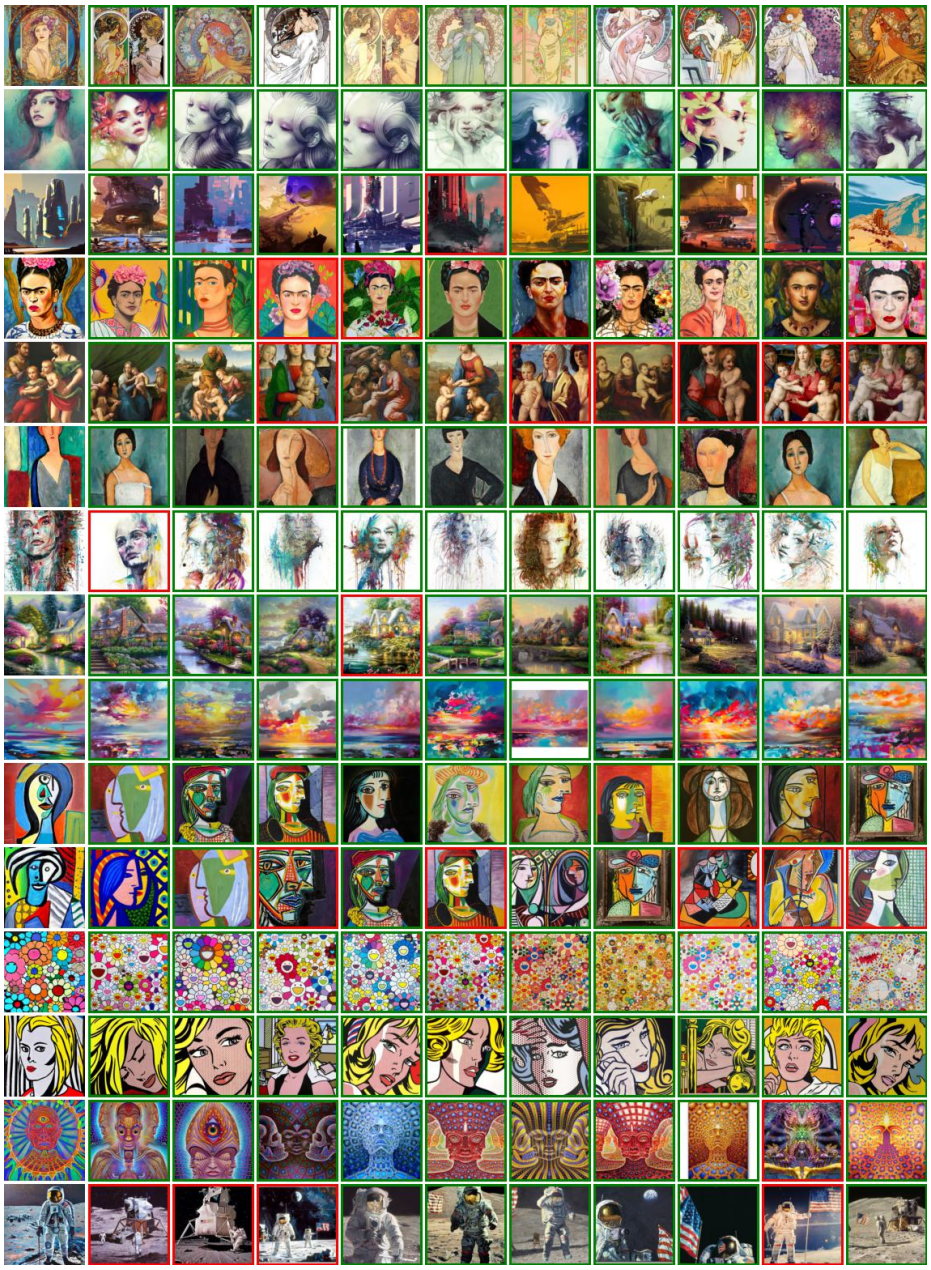


Fig. 11: Simple caption generations and matches: First column is SD generation, and the rest of the columns are top-10 matches in the ContraStyles database. The green box represents the correct match while the red box represents the incorrect match.

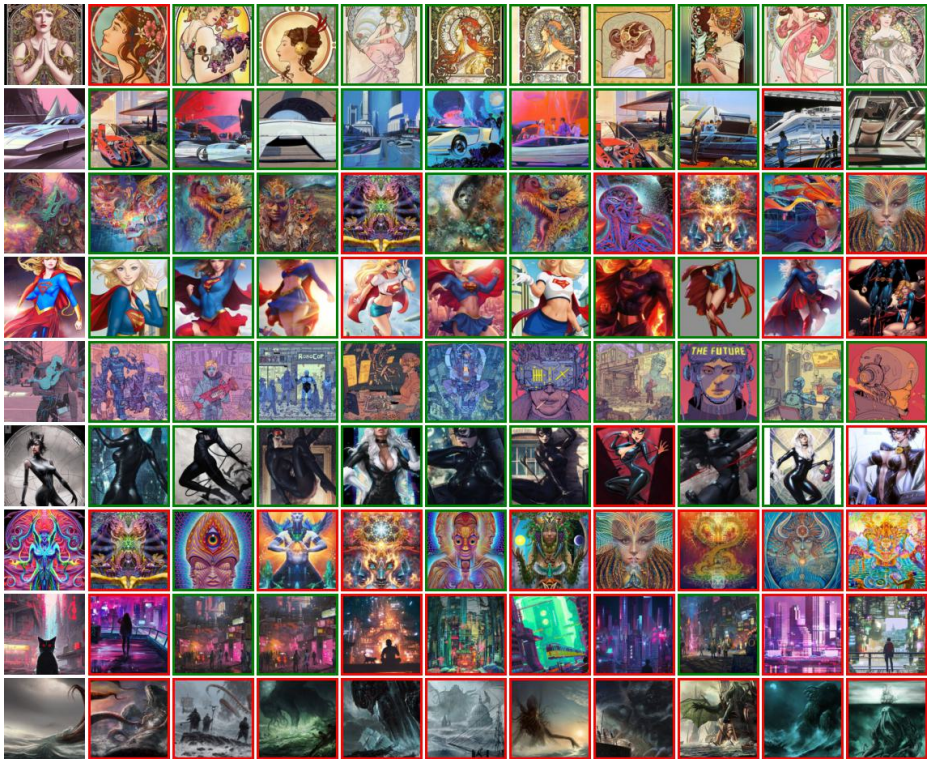


Fig. 12: ***User-generated*** **caption generations and matches:** First column is SD generation, and the rest of the columns are top-10 matches in the ContraStyles database. The green box represents the correct match while the red box represents the incorrect match.