SCFlow: Implicitly Learning Style and Content Disentanglement with Flow Models

Supplementary Material

A. Dataset Construction

We curate the original content images from Pexels[1], following standard web-scraping practices[2]. Since the Pexels images often have sparse or incomplete captions, we generate improved captions using LLaVA 1.5 [40].

For the style prompts, we select 51 artistic styles, such as *e.g.* Cyberpunk and Cubism, each accompanied by a brief explanatory description. The selection was guided by a few art experts, and the descriptions were refined with assistance from ChatGPT-4o [29].

During stylization, we minimize pixel-level constraints by conditioning on scribbles and applying a tailored guidance scale. We also reweight the style component to ensure strong adherence to the specified artistic style. To generate stylized images, we use ControlNet [81] with prompts in the format:

```
"An image depicting {content_caption},
    in the style of {style_prompt}"
```

We will publish the content captions and the style prompts together with the stylized images. An overview of the curated dataset can be found in Fig. S18. Although we use Pexels images as content, the construction pipeline can be easily adapted to other content sources, such as LAION [57, 58] or COYO [4].

B. Evaluation Details for other models

For CSD [65], there are two output heads for the style vector and the content vector. Hence, we denote them as CSD-C for content and CSD-S for styles. And we used them accordingly for our evaluation of content and styles.

For DEADiff [49], mean query embeddings can be extracted using a pre-trained Q-Former, with visual features corresponding to the prompt "content" or "style".

C. Visualization of Content and Style Proxies

We show the aggregated embeddings by averaging them across all predictions conditioned on either style or content, respectively (see Fig. S11 and Fig. S12). These aggregated embeddings can be considered as the style or content class proxies in the resulting space.

D. More Analysis on mean content and style

Our method produces disentangled representations $\bar{c}$ (content from style) and $\bar{s}$ (style from content), visualized in Figs. S16 and S17. Three key findings validate their independence and authenticity:

1. **Content-Style Independence.** The disentangled embeddings exhibit no dependence on their original counterparts. For content embeddings $\bar{c}$, cat images rendered in diverse artistic styles all yield consistent $\bar{c}$ representations,

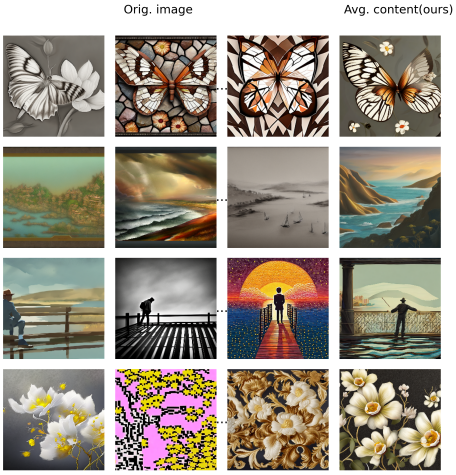


Figure S11. Visualization of proxy contents: The first three columns display a part of the mixed references I_{c_i,s_j}, while the last column shows the average content (ours).

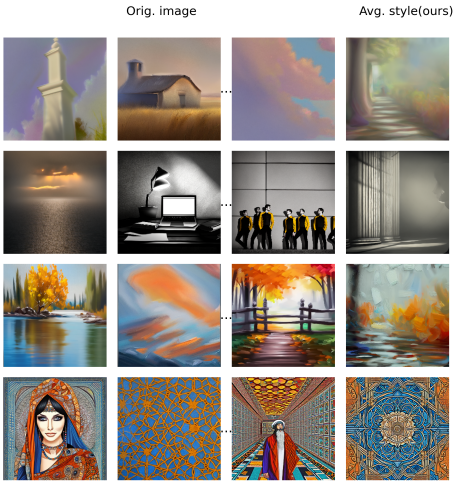


Figure S12. Visualization of proxy styles: The first three columns display a part of the mixed references I_{c_i,s_j}, while the last column shows the average style (ours).

[1] https://www.pexels.com/

[2] https://huggingface.co/datasets/opendiffusionai/pexels-photos-janpf

Figure S13. Generalization to textual condition during inference.

Figure S14. Inference in-the-wild. We use ImageNet images as styleless inputs for the forward pass, with style references obtained online. After obtaining the merging results (3rd column), we further use the reverse mapping to map them back to our content and style (4th and 5th columns)

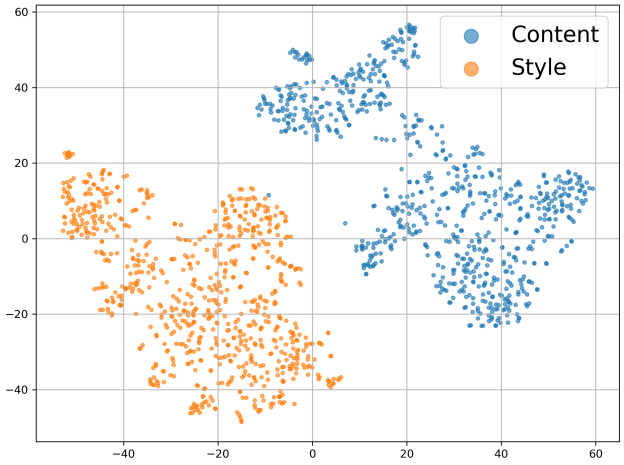


Figure S15. Content and style disentanglement shown by tSNE of WikiArt.

as do bicycles (Fig. S16, right). Similarly, style embeddings $\bar{s}$ remain stable across different content inputs – for instance, the same style emerges whether applied to cats or branches (Fig. S17, left).

2. **Intrinsic Signal Origin.** The $\bar{c}$ and $\bar{s}$ signals originate from our embeddings rather than unCLIP hallucinations. This is evidenced by two observations: (i) varying text prompts (different columns) produce negligible changes in $\bar{c}$ / $\bar{s}$, despite unCLIP's known prompt sensitivity; and (ii) our embedding signals consistently overpower prompt conditioning, maintaining semantic stability.
3. **Initial Noise Invariance.** When testing different initial noise seeds in unCLIP, we observe minor variations in image details (*e.g.*, object pose or texture) but noticeable consistency: $\bar{c}$ and $\bar{s}$ remain preserved across all noise configurations. This confirms their independence from generation artifacts.

Collectively, these results demonstrate that $\bar{c}$ and $\bar{s}$ capture intrinsic content/style properties rather than inversion artifacts or model biases from unCLIP.

E. Visualization of Unseen Styles and Contents

Unseen textual condition. Our model is trained and evaluated solely on CLIP image embeddings, **without** using any text descriptions. Nevertheless, thanks to the multi-modal alignment in CLIP space, our model is capable of taking text as style and content references (Fig. S13) to generate meaningful results.

Unseen constructed style and content. We curated an additional subset, similar to the main dataset, where both the style and content are never used during training or testing for our model. We visualize the corresponding forward and backward inference results. (see Fig. S19a and Fig. S19b)

Unseen real-world data from ImageNet and WikiArt. Although not trained for the style and content retrieval task, our model yields competitive performance, reflecting strong representation quality. Our primary goal is to introduce a new possibility for semantic disentanglement with generative models. As shown in Fig. S15, we achieve clear style-content separation on WikiArt (quantitative in Tab. S4). Fig. S14 further demonstrates successful forward and reverse inference on real photos (content ref.) and artworks (style ref.), showing generalization beyond synthetic data.

F. Comparison with Conventional Discriminative Objectives

In addition to CSD [64], we train two models with *Contrastive Loss* [7] and *InfoNCE* [48] on our dataset, using similar capacity, training settings, and evaluate on our test sets and real-world datasets using the same metrics. Except for a slight NMI on WikiArt, our method outperforms them across all settings (Tab. S4), confirming that our gains *do not stem solely from the dataset*. Importantly, our model learns a well-structured embedding space that enables style/content interpolation (Fig. 6 and Fig. 9) and avoids collapsing to mean interpretations, indicating strong generalization and balanced intra-/inter-variance. While simpler methods may work for single tasks, ours unifies merging and disentangling within a single framework.

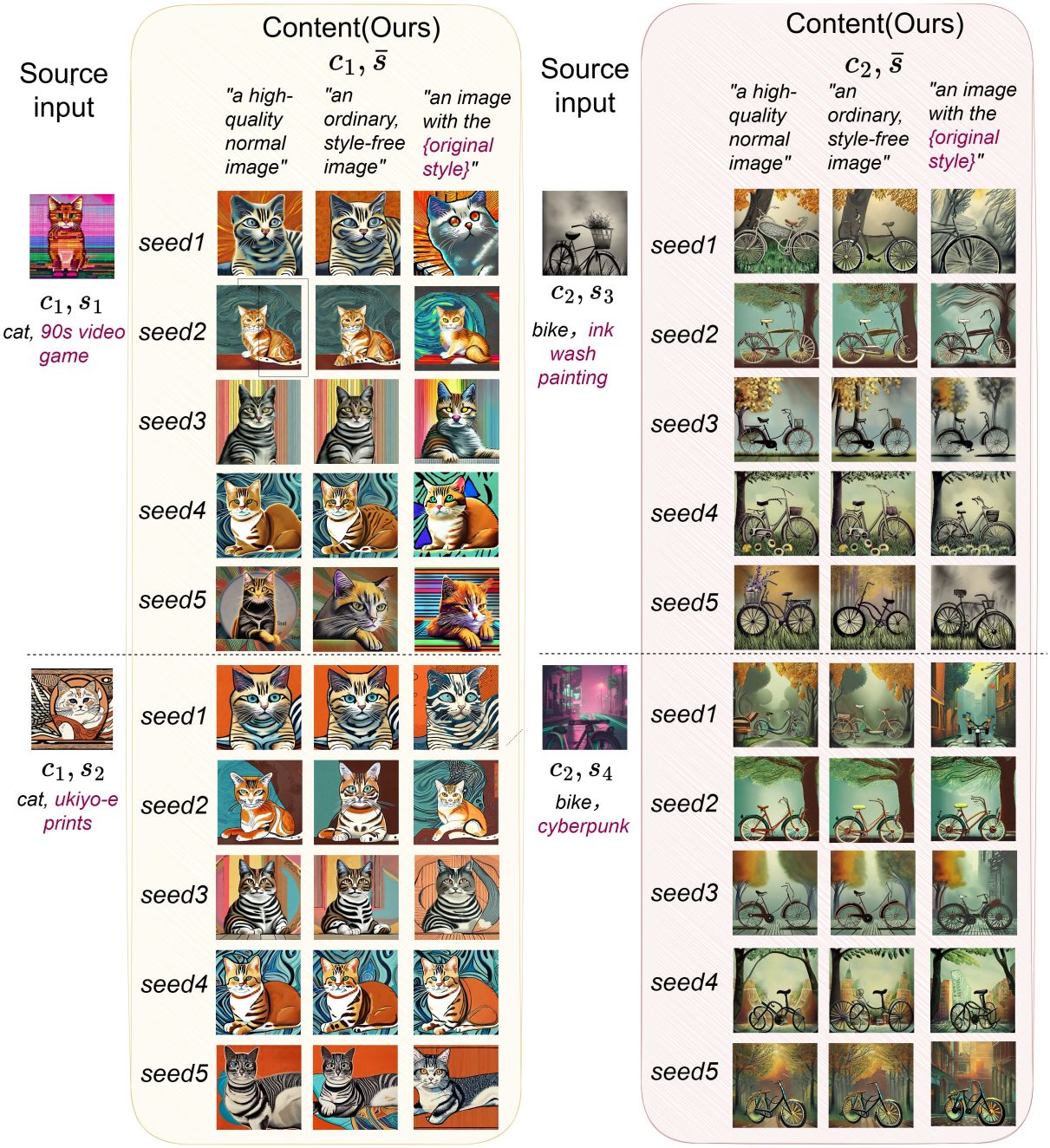


Figure S16. We use unCLIP to decode the same *content embeddings* generated by our method to pixel space, using different prompts and initial noises (denoted by seed).

Dataset	NMI Score↑			FDR↑		
	SCFlow	Contrastive [7]	InfoNCE [48]	SCFlow	Contrastive	InfoNCE
Our Styles	**0.8696**	0.2905	0.5904	**3.5184**	0.1102	0.3711
WikiArt	0.4010	0.4194	**0.4238**	**0.6474**	0.2923	0.2553
Our Contents	**0.8356**	0.4598	0.2327	**2.1693**	0.1799	0.0598
ImageNet	**0.9172**	0.7737	0.8194	**1.4264**	0.3529	0.2056

Table S4. Comparison to conventional discriminative approaches trained on our dataset.

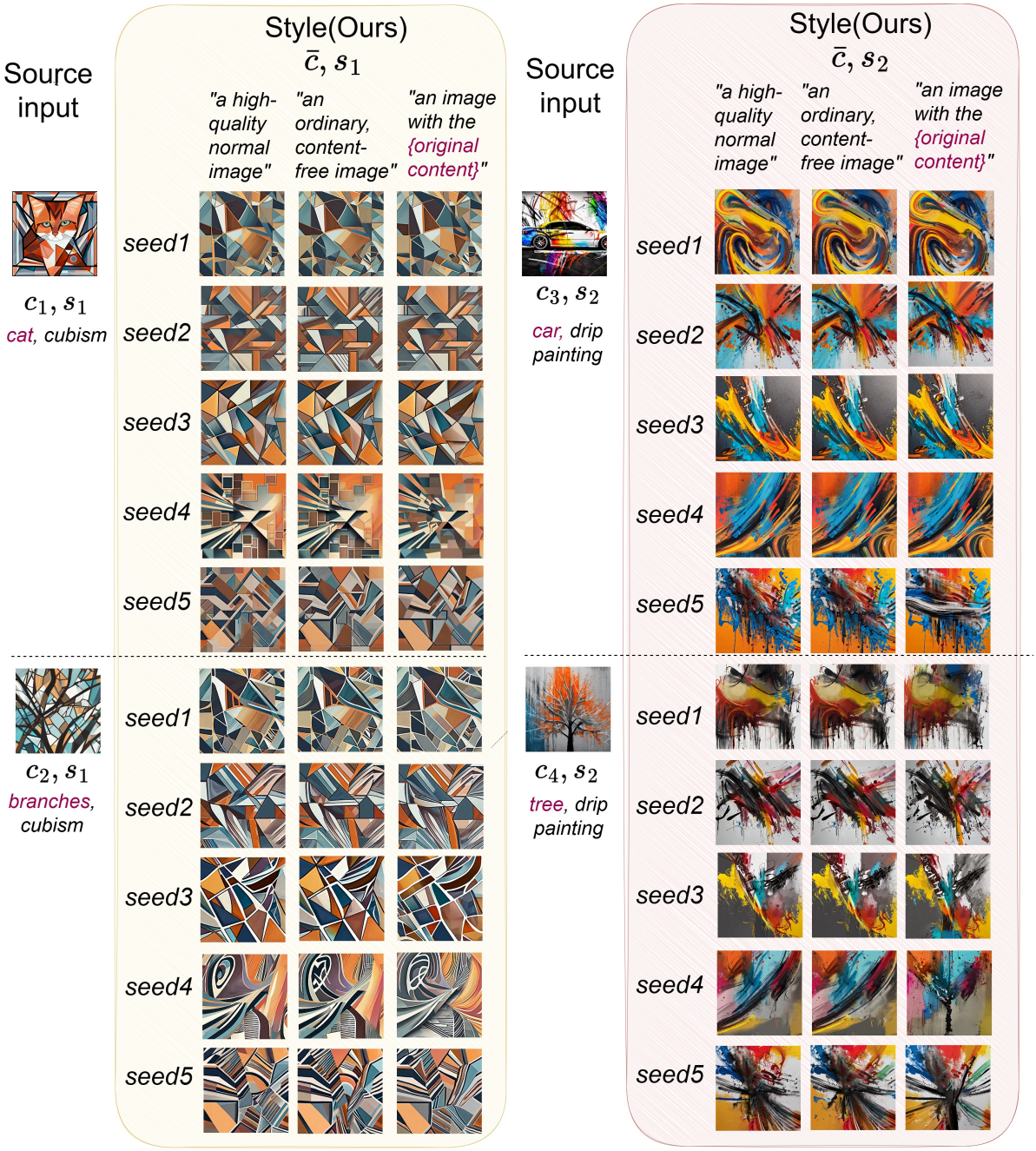


Figure S17. We use unCLIP to decode the same *style embeddings* generated by our method to pixel space, using different prompts and initial noises (denoted by seed).

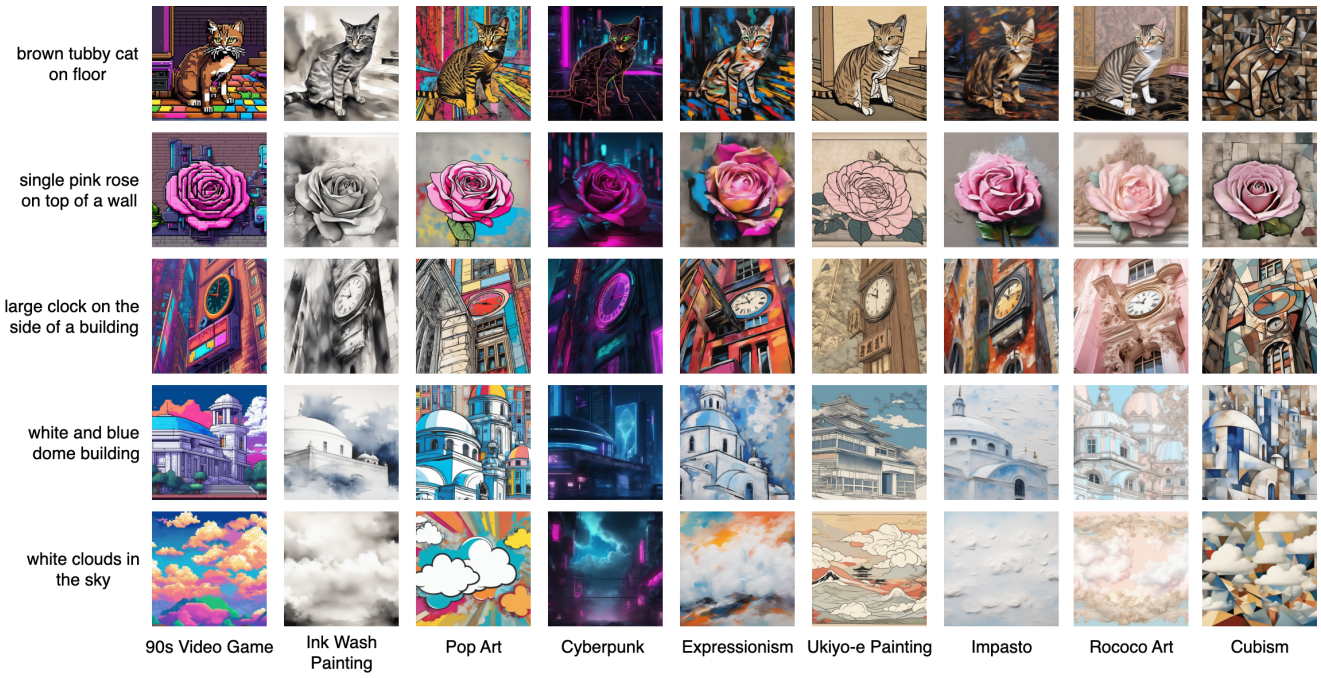
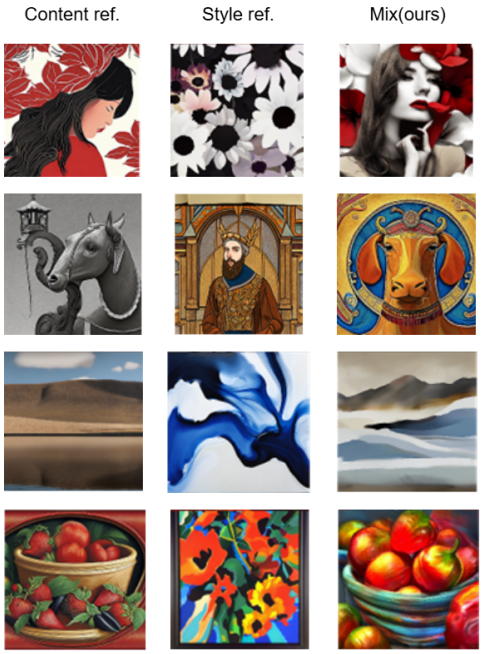
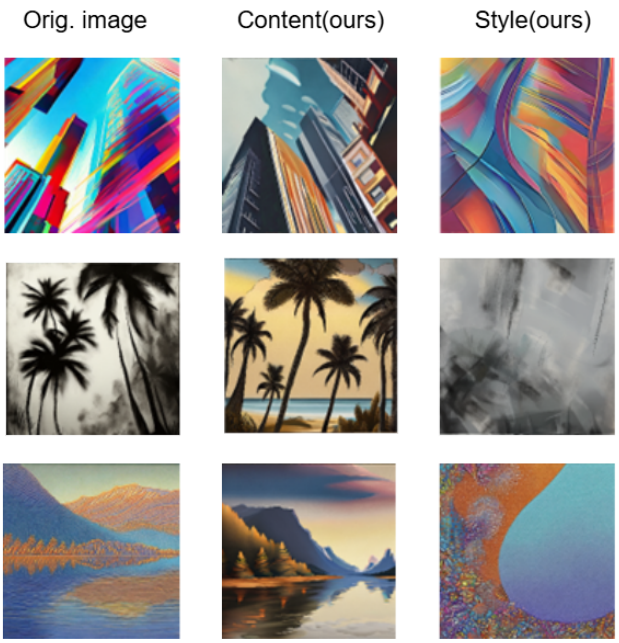


Figure S18. Overview of the curated dataset



(a) Mixing content and style references (unseen). The first and second columns show the content and style references, respectively; the third column shows the mixed results.



(b) Disentanglement of content and style from unseen images. The first column shows the original image, followed by the extracted content and style.

Figure S19. Visual results on unseen content and style inputs. Left: Mixing. Right: Disentanglement.