# IntroStyle: Training-Free Introspective Style Attribution using Diffusion Features

Supplementary Material

## 1. Overview

In this supplementary material, we present further details about our methodology and experimental findings. Specifically, we provide an analysis of the hyper-parameter selection for IntroStyle features in Sections 2 and 3. Furthermore, we elaborate on our prompt engineering process utilizing ChatGPT for style isolation in the synthesis of ArtSplit dataset in Section 4. Finally, we present additional experimental results and analyses on both the WikiArt and ArtSplit datasets in Section 5. Our codes and ArtSplit dataset will be released.

## 2. Similarity Metrics

In this section, we include the details and formulate the Euclidean, Gram Matrices (using IntroStyle features), and Jensen-Shannon Divergence (JSD) metrics as discussed in section 3.3 of the main text.

The $L_2$ distance (Euclidean distance)

$$L_2(\mu_1, \mu_2)^2 = \|\mu_1 - \mu_2\|_2^2, \tag{1}$$

ignores covariance information and does not have an interpretation as a measure of similarity between probability distributions. This is also the case for the Frobenius norm between the Gram matrices, which is popular in the style transfer literature. It extracts deep features from the two images and then takes the outer product of their corresponding mean vectors $\mu_1$ and $\mu_2$,

$$\text{Gram}(\mu_1, \mu_2) = \|\mu_1\mu_1^T - \mu_2\mu_2^T\|_F. \tag{2}$$

Another popular similarity measure between distributions is the Kullback–Leibler (KL) divergence. For two multivariate Gaussians, it takes the form

$$\text{KL}((\mu_1, \Sigma_1)||(\mu_2, \Sigma_2)) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} + \text{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1)\right]. \tag{3}$$

Note that the KL divergence is not symmetric. To address this, the Jensen-Shannon Divergence (JSD) averages the KL divergence going in both directions,

$$\text{JSD}(I_1||I_2) = \frac{1}{2}\text{KL}(I_1||I_2) + \frac{1}{2}\text{KL}(I_2||I_1). \tag{4}$$

## 3. Timestep and Block Index

We study the choices of timestep $t$ and block indices $idx$ of IntroStyle by evaluating the image retrieval performance on the WikiArt Dataset. The results are presented in Figs. 1 and 2, showing that best performance is obtained with $t < 500$ and up-block indices $idx = 1$ or $idx = 2$, achieving a balanced trade-off between mAP@10 and recall@10.

| Block | mAP@$k$ 1 | mAP@$k$ 10 | mAP@$k$ 100 | Recall@$k$ 1 | Recall@$k$ 10 | Recall@$k$ 100 |
|---|---|---|---|---|---|---|
| $idx = 0$ | 0.947 | 0.925 | 0.791 | 0.947 | 0.970 | 0.991 |
| $idx = 1$ | 0.954 | 0.949 | 0.850 | 0.954 | 0.982 | 0.995 |
| $idx = 2$ | 0.950 | 0.946 | 0.820 | 0.950 | 0.964 | 0.987 |
| $idx = 3$ | 0.941 | 0.939 | 0.823 | 0.941 | 0.978 | 0.984 |

Table 1. DomainNet: Feature block index selection for $t = 25$.

| Metric | mAP@$k$ 1 | mAP@$k$ 10 | mAP@$k$ 100 | Recall@$k$ 1 | Recall@$k$ 10 | Recall@$k$ 100 |
|---|---|---|---|---|---|---|
| $L_2$ | 0.948 | 0.944 | 0.839 | 0.948 | 0.961 | 0.980 |
| $W_2$ | 0.954 | 0.949 | 0.850 | 0.954 | 0.982 | 0.995 |

Table 2. DomainNet: Comparison on different metrics.

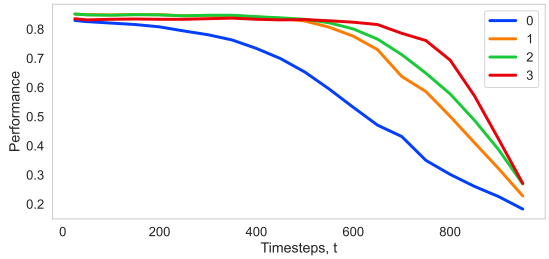



Figure 1. Precision (mAP@10) as a function of timestep $t$ for different up-block indices ($idx$) on the WikiArt dataset.

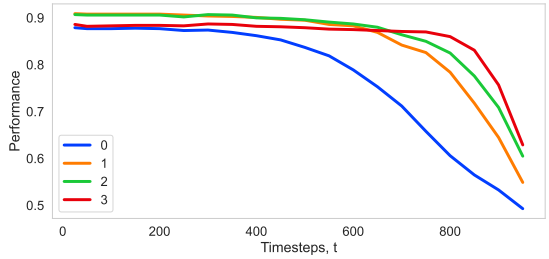



Figure 2. Recall (Recall@10 ) as a function of timestep $t$ for different up-block indices ($idx$) on the WikiArt dataset.

On DomainNet, we observe consistent trends: 2-Wasserstein ($W_2$) distance improved mAP@10 by 0.05 compared to $L_2$ (Fig. 1), features from block 1 led to a 0.03 improvement (Fig. 2), time-step vs performance curve is similar. For variance, varying seeds we obtained 0.002/0.001 for mAP/Recall @10 respectively.

We chose idx $= 1$ as our default configuration for computational efficiency, as shown in Table 3.

| Method | Parameters | Channel ($C$) |
|---|---|---|
| IntroStyle ($idx = 0$) | 548M | 1280 |
| IntroStyle ($idx = 1$) | 808M | 1280 |
| IntroStyle ($idx = 2$) | 881M | 640 |
| IntroStyle ($idx = 3$) | 900M | 320 |

Table 3. Comparison of the model size needed to compute IntroStyle features and channel size for different choices of the up-block index ($idx$).

## 4. Generating Prompts for ArtSplit dataset

As explained in the main text, we curated a comprehensive collection of prompt-image pairs representing 50 influential artists across various artistic movements and periods from the LAION Aesthetics Dataset: Albert Bierstadt, Alex Gray, Alphonse Mucha, Amedeo Modigliani, Antoine Blanchard, Arkhip Kuindzhi, Carne Griffiths, Claude Monet, Cy Twombly, Diego Rivera, Edmund Dulac, Edward Hopper, Francis Picabia, Frank Auerbach, Frida Kahlo, George Seurat, George Stubbs, Gustav Klimt, Gustave Dore, Harry Clark, Hubert Robert, Ilya Repin, Isaac Levitan, Jamie Wyeth, Jan Matejko, Jan Van Eyck, John Atkinson Grimshaw, John Collier, John William Waterhouse, Josephine Wall, Katsushika Hokusai, Leonid Afremov, Lucian Freud, M.C. Escher, Man Ray, Mark Rothko, Paul Klee, Peter Paul Rubens, Picasso, RenÃ© Magritte, Richard Hamilton, Robert Delaunay, Roy Lichtenstein, Takashi Murakami, Thomas Cole, Thomas Kinkade, Vincent Van Gogh, Wassily Kandinsky, William Turner, Winslow Homer. Using their seminal works as reference queries, we implemented a systematic prompt-generation strategy. For each artist, we derived a "style" prompt and used 2 of their paintings to generate "semantic" prompts, subsequently employing a Stable Diffusion v2.1 to synthesize 12 images per combination. This methodological approach yielded a richly diverse reference dataset comprising 60,000 images (50 artistic styles $\times$ 12 prompts $\times$ 100 semantic descriptions).

We leveraged the ChatGPT to generate both style and semantic prompts systematically.

**Style Prompt**. To create descriptions that effectively described the artistic style, we crafted a base system prompt as follows:

> *"You are a prompt generator for Stable Diffusion 2.1, and you are tasked with generating the style description of an artist.*
> *KEEP EVERYTHING SFW (SAFE FOR WORK).*
> *Only print the prompt without any other information and each prompt should not be more than 25 words. "*

The user prompt template is as follows:

> *"Create a prompt to generate an image in the style of [artist] and also include the artist's name"*

with the variable inputs based on the chosen artist, with artist being the variable inputs. An example response by ChatGPT, for Vincent Van Gogh, is shown in Fig. 3.

**Semantic Prompt**. The semantic prompts are created for a given artist's painting by providing descriptions of the contents of the image without any style information. The base system prompt is as follows:

> *"You are a prompt generator for Stable Diffusion 2.1, and you are tasked with generating a prompt with the semantics of a given painting.*
> *DO NOT HAVE ANY DESCRIPTION OF THE STYLE. KEEP EVERYTHING SFW (SAFE FOR WORK)*
> *Only print the prompt without any other information and each prompt should not be more than 25 words."*

And the user prompt template is as follows:

> *"Create a prompt to generate an image with the semantics of artists's painting"*

With the variable inputs based on the chosen artist's painting, with artist and painting being the chosen artist's painting. An example response by ChatGPT, for Vincent Van Gogh's Starry Night, is shown in Fig. 4.

**User**

Create a prompt to generate an image in Vincent Van Gogh's style and include the artist's name.

**ChatGPT**

Van Gogh-inspired painting, swirling brushstrokes, vivid warm colors, emotive expression, night sky, Starry Night atmosphere, Vincent Van Gogh style.

Figure 3. Style prompt generation using ChatGPT.

## 5. Results on WikiArt and ArtSplit Datasets

Figs. 6 and 7 show retrieval results of the proposed IntroStyle method on WikiArt and Fig 5 compares our

User

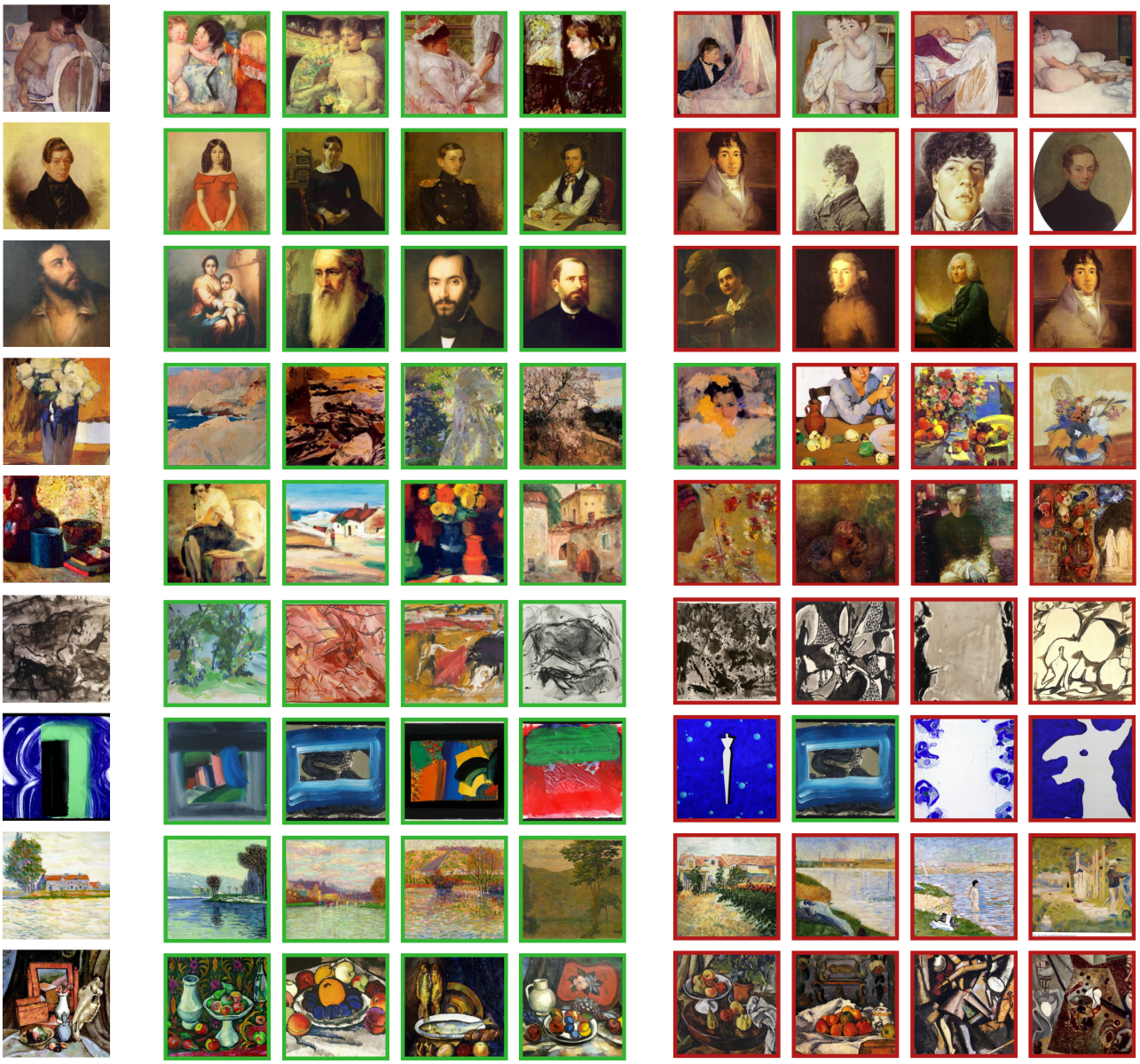
Create a prompt to generate an image with the semantics of Vincent Van Gogh's Starry Night.

ChatGPT

A night sky filled with vibrant stars above a quiet, small village nestled beneath rolling hills.

Figure 4. Semantic prompt generation using ChatGPT.

retrievals with CSD. Remarkably, `IntroStyle` can identify images of the same style despite a great diversity of content. Note that the images retrieved rarely present images of similar content, indicating the robustness of the proposed approach in focusing on styles. Furthermore, we show more retrieval results in Figs. 8 and 9 for the `ArtSplit` Dataset, where our method works effectively and achieves high retrieval accuracy over state-of-the-art method (CSD).

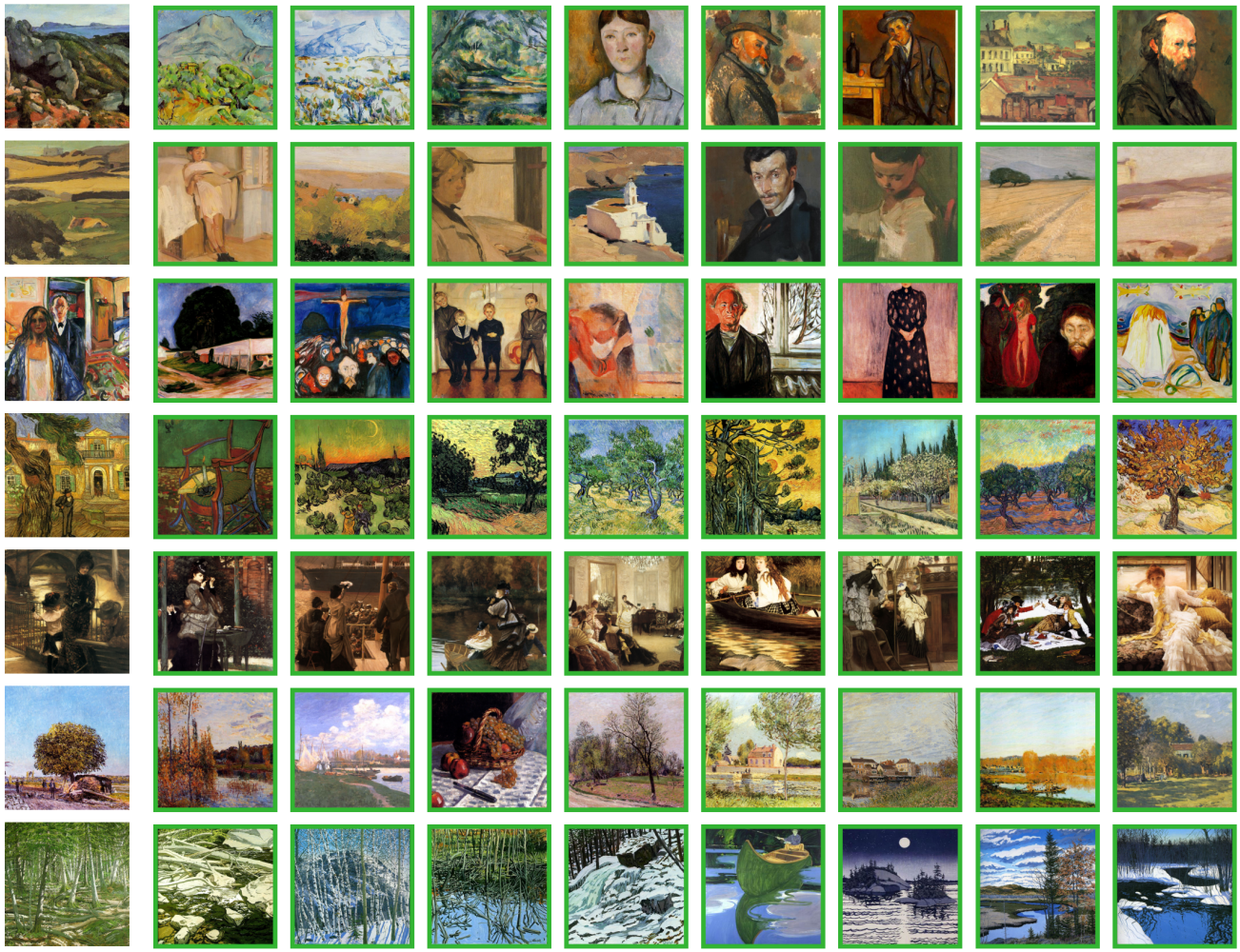


(a) Queries (b) `IntroStyle` (ours) (c) CSD [**?** ]

Figure 5. Image Retrieval Results on WikiArt Dataset for `IntroStyle`, with images ranked highest to lowest from left to right compared with CSD. Green colors indicate correct and red for incorrect retrievals.

(a) Queries (b) `IntroStyle` (ours)

Figure 6. Additional Image Retrieval on WikiArt Dataset for `IntroStyle` with images ranked highest to lowest from left to right. Green colors indicate correct and red for incorrect retrievals.



(a) Queries (b) `IntroStyle` (ours)

Figure 7. Additional Image Retrieval on WikiArt Dataset for `IntroStyle` with images ranked highest to lowest from left to right. Green colors indicate correct and red for incorrect retrievals.

(a) Queries (b) `IntroStyle` (ours) (c) CSD [**?** ]

Figure 8. Additional Image Retrieval Results for Style-based Evaluation on `ArtSplit` Dataset, with images ranked highest to lowest from left to right. The images are filtered to a fixed semantic. Green colors indicate correct and red for incorrect retrievals.

(a) Queries (b) `IntroStyle` (ours) (c) CSD [**?** ]

Figure 9. Additional Image Retrieval Results for Semantic-based Evaluation on `ArtSplit` Dataset, with images ranked highest to lowest from left to right. Green colors indicate correct and red for incorrect retrievals.