Training-and-Prompt-Free General Painterly Harmonization via Zero-Shot Disentenglement on Style and Content References

Teng-Fang Hsiao, Bo-Kai Ruan, Hong-Han Shuai

National Yang Ming Chiao Tung University, Taiwan
{tfhsiao.ee13, bkruan.ee11, hhshuai}@nycu.edu.tw

Abstract

Painterly image harmonization aims at seamlessly blending disparate visual elements within a single image. However, previous approaches often struggle due to limitations in training data or reliance on additional prompts, leading to inharmonious and content-disrupted output. To surmount these hurdles, we design a Training-and-prompt-Free General Painterly Harmonization method (TF-GPH). TF-GPH incorporates a novel "Similarity Disentangle Mask", which disentangles the foreground content and background image by redirecting their attention to corresponding reference images, enhancing the attention mechanism for multi-image inputs. Additionally, we propose a "Similarity Reweighting" mechanism to balance harmonization between stylization and content preservation. This mechanism minimizes content disruption by prioritizing the content-similar features within the given background style reference. Finally, we address the deficiencies in existing benchmarks by proposing novel range-based evaluation metrics and a new benchmark to better reflect real-world applications. Extensive experiments demonstrate the efficacy of our method in all benchmarks.

Code — https://github.com/BlueDyee/TF-GPH

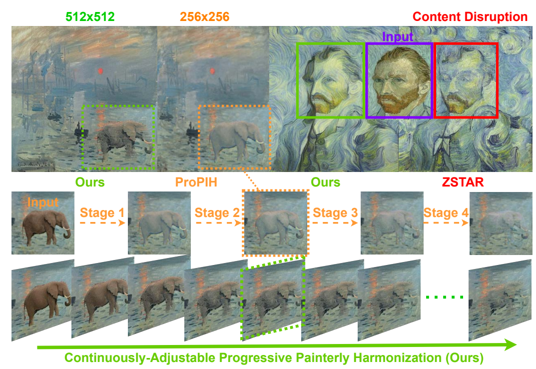


Figure 1: Our method overcomes the resolution and staged-progressive painterly harmonization limitations present in the SOTA method ProPIH (Niu et al. 2024b), where users are restricted to selecting stylization strength from one of four stages. In contrast, our approach offer continuously adjustable hyperparameters, allowing for more flexible stylization. Additionally, our method effectively mitigates content disruption issues, such as facial alterations, commonly seen in image-editing methods like ZSTAR (Deng et al. 2023).

Introduction

Image composition, which involves blending a foreground element from one image with a different background, often results in composite images with mismatched colors and illumination between the foreground and background. Image harmonization techniques have been developed to adjust the appearance of foreground for a seamless integration with the background (Tsai et al. 2017; Wu et al. 2019; Tan et al. 2023; Xing et al. 2022). A specialized area within this field, painterly image harmonization, focuses on integrating elements into paintings to enable artistic edits (Lu et al. 2023; Luan et al. 2018). For instance, ProPIH (Niu et al. 2024b), pioneers progressive painterly harmonization, training the model with different levels of harmonization, enhancing its applicability to real-world scenarios.

Despite notable advancements, current painterly image harmonization techniques still face challenges with generalizability, particularly when dealing with novel art styles or unique content compositions. One promising solution is to leverage insights from other image-editing methods. For instance, (Zhang et al. 2023; Cheng et al. 2023) suggest fine-tuning models to adapt to input styles. However, each styles require additional computational costs that are 10x times higher than a single inference. Alternatively, (Lu, Liu, and Kong 2023; Kwon and Ye 2022) propose text-guided editing strategies, but these approaches are limited by the difficulty of adequately describing complex visual styles through text alone. Recently, methods such as (Cao et al. 2023; Deng et al. 2023) explore training-free techniques. These methods utilize attention-sharing across images combined with techniques like AdaIN(Huang and Belongie 2017) to align content features with style references. While effective, this brute alignment lead to content disruption as shown in Fig. 1.

In this work, we present **TF-GPH**, an innovative diffusion pipeline that operates without additional training or prompts by leveraging the pretrained diffusion model (Rombach et al. 2022). TF-GPH solves a wider range of painterly

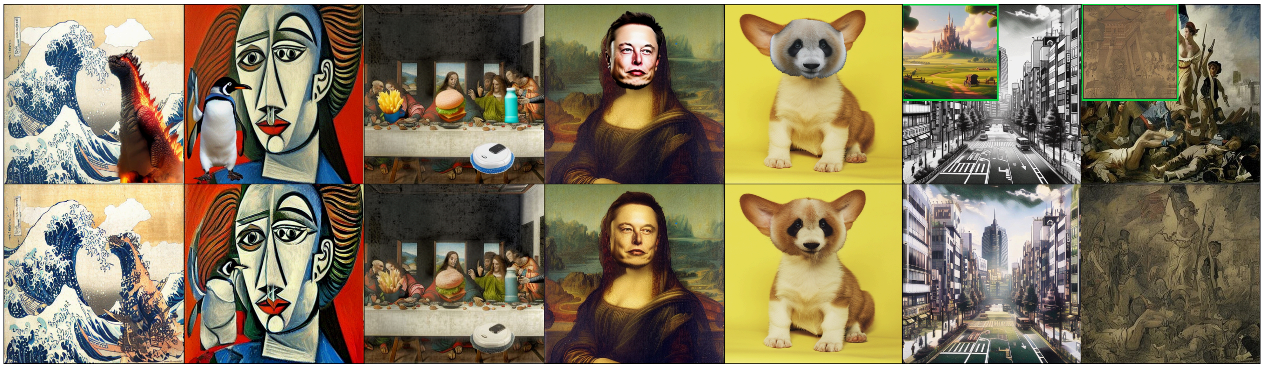


Figure 2: An example demonstrates three tasks in general painterly harmonization: Object Insertion (columns 1 to 3), Object Swapping (columns 4 and 5), and Style Transfer (columns 6 and 7). The top row features user-generated composite images, where green boxes highlighting the style reference of final two. The bottom row showcases the results using our method.

harmonization tasks, including object insertion, swapping, and style transfers, as illustrated in Fig. 2. In our pipeline, we modify the self-attention mechanism of the diffusion model and adopt the shared attention layer from (Hertz et al. 2024) to enable multi-image input (foreground content reference, background style reference, and composite image). While the shared attention layer can merge similar features across images, such as consecutive frames, the strong self-similarity of the composite image causes it to overlook dissimilar references in this task. To address this, we introduce a novel **similarity disentangle mask** within the shared attention layer. This mask applied before the softmax operation decouples the foreground and background features by redirecting the composite image's self-attention to the two reference images. This approach allows precise control over the foreground and background within the composite image by adjusting the attention to their respective references.

Moreover, to address the content disruptions caused by attention adjustments such as addition and AdaIN, as observed in prior work, we propose a similarity-based editing method termed **similarity reweighting**. This approach balances attention between content and style references by scaling similarity based on user specified hyperparameters. By prioritizing style features that closely match the content features the content disruption thus minimized. By integrating these two aforementioned adjustments into the existing image generation pipeline, we are able to perform image harmonization without requiring additional prompts or training. Additionally, this mechanism offering flexibility to tailor the output continuously from style-free to style-heavy, thereby accommodating various artistic preferences, as shown in Fig. 1.

Finally, a challenge in evaluating painterly harmonization is the limited diversity of test data styles (Tan et al. 2019), which are often restricted to those seen during training. This limitation fails to capture the wide variety of styles encountered in real-world scenarios, such as manga or cartoon. To address this issue, we introduce the "General Painterly Harmonization Benchmark" (GPH Benchmark). This benchmark encompasses three harmonization tasks—object insertion, object swapping, and style transfer—while incorporating diverse content and style references to ensure a comprehensive evaluation. Furthermore, existing metrics typically focus on either content or style similarity without adequately reflecting user preferences for different balances between stylization and content preservation. To bridge this gap, we propose range-based metrics, evaluating both the lower and upper bounds of stylization and content-preservation strength across the dataset. A wider range indicates greater flexibility and the adaptability to various scenarios.

Our contributions can be summarized as follows. 1) We introduce the **TF-GPH** framework, the first training-and-prompt-free pipeline using a diffusion model designed for general painterly harmonization. 2) Our proposed **similarity disentangle mask with similarity reweighting** not only shows promising results in painterly harmonization tasks, but also solves the content disruption issue of existing attention-based editing method. 3)We propose the **GPH Benchmark**, consisting of various data for real-world usage, together with a range-based metric to align model performance with user experience.

Related Work

Image Harmonization

Image Harmonization can be categorized into two main types: **Realistic Harmonization** and **Painterly Harmonization**. The former (Zhang et al. 2021; Cong et al. 2020; Lin et al. 2018; Chen et al. 2023) focus on seamlessly integrating objects into new backgrounds with consistent illumination, edge alignment, and shadow integrity. In contrast, Painterly Harmonization (Lu et al. 2023; Cao, Hong, and Niu 2023; Wu et al. 2019) aims to artistically blend objects into paintings, prioritizing stylistic coherence. Recently, ArtoPIH (Niu et al. 2024a) propose learning from painterly objects by using annotated objects within paintings as training data, ProPIH (Niu et al. 2024b) introduce a progressive learning approach. Despite these advancements, existing methods require training. Our proposed method, however, eliminates the need for training, enabling direct application to unseen styles and significantly enhancing the versatility of painterly image harmonization.

Attention-based Image Editing

Manipulation of attention layers within diffusion UNet architectures is a prevalent strategy in modern image editing techniques (Tumanyan et al. 2023; Chefer et al. 2023; Gu et al. 2024; Lu, Liu, and Kong 2023; Hertz et al. 2023). For instance, P2P (Hertz et al. 2023) utilizes prompt-driven cross-attention to modify images, while TF-ICON (Lu, Liu, and Kong 2023) integrates objects into backgrounds by constraining self-attention and cross-attention outputs with given mask. Despite their effectiveness, the reliance on descriptive prompts can be problematic when suitable prompts are not available. In contrast, our TF-GPH method function solely with image inputs, eliminating the need of prompts.

Style Transfer

Style transfer aims to alter the style of a content image to match a specified style. Existing methods generally categorized into optimization-based and feedforward-based approaches. The former (Gatys et al. 2017; Li et al. 2017), refine the image by aligning it with features extracted from the style reference. For example, (Kwon and Ye 2022) utilizes a pre-trained CLIP model (Radford et al. 2021) for this purpose. In contrast, feedforward-based (Deng et al. 2022; Huang et al. 2023) involving VCT (Cheng et al. 2023) and InST (Zhang et al. 2023), which fine-tune models to integrate style into the model's architecture. Recently, attention-based techniques have been incorporated. For instance, the shared attention module introduced in (Hertz et al. 2024; Deng et al. 2023; Chung, Hyun, and Heo 2024) produces feature-consistent images by sharing attention across multiple images. However, these methods often suffer from content disruption due to the blending of unrelated features from different references. In contrast, TF-GPH minimizes content disruption and achieves superior performance across styles.

Method

Our research aims to facilitate a general form of painterly harmonization based on images only without the additional need for prompts, which can facilitate various applications, *i.e.*, object insertion, object swapping, and style transfer. Formally, given a foreground object image I^{f}, a background painting I^{b}, and I^{c}, which is the user-specified composition that guides the size and position of the foreground object on the background painting, the goal of painterly harmonization is to transfer the style from I^{b} to the object from I^{f} in I^{c} seamlessly, resulting a harmonized image I^{o}.

To address the challenge of painterly harmonization, we introduce a novel framework titled Training-and-Prompt-Free General Painterly Harmonization (TF-GPH), as depicted in Fig. 3. Specifically, the inputs—foreground I^{f}, background I^{b}, and composite I^{c}-are initially processed through an inversion mechanism equipped with either a null prompt embedding or a exceptional prompt embedding $\rho_{\text{exceptional}}$, which has demonstrated its ability for stabilizing inversion process (Lu, Liu, and Kong 2023). Subsequently, a denoising operation is applied concurrently to all three images, during which the composite image I^{c} is enriched with style attributes, producing harmonized output I^{o}

The core of our architecture is the **Similarity Disentangle Mask**, a novel attention mask designed to disentangle the features of the foreground object from the background image of I^{c} and link them to their corresponding references I^{f} and I^{b}. After disentanglement, we enhance the influence of the background style reference I^{b} on the pasted object through our **Similarity Reweighting** technique. This approach differs from existing attention-based editing techniques, which directly add (Hertz et al. 2023) or adjust the mean/variance of features (Hertz et al. 2024; Chung, Hyun, and Heo 2024), introducing disruption on semantic and structural details. By adjusting the similarity solely, we can minimize content disruption while applying the stylization effect, producing the final painterly harmonized output image I^{o}. Additionally, our framework is versatile enough to support not only painterly harmonization for object insertion—a traditional task of painterly harmonization—but also object swapping and style transfer. The former is viewed as a semantically richer variant of object insertion, and the latter as a broader aspect of the same. We summarize these related tasks under the term "General Painterly Harmonization".

Attention-based Diffusion UNet

In the framework of diffusion models, the attention mechanisms (Vaswani et al. 2017) are essential to capture characteristic details, facilitating both the elimination of noise and the enhancement of context information. Specifically, the self-attention module plays a vital role in synthesizing the output by internalizing the inherent data characteristics, while the cross-attention module is instrumental in incorporating contextual information from various modalities, *e.g.* text and audio, thus amplifying the conditional impact on the resultant images. Since our approach does not need an additional prompt to guide the fusion, we can simply utilize self-attention to ensure that the background style is harmonically fused into the composition image during denoising.

Given three input images I^{f}, I^{b}, and I^{c}, these images are first compressed by a VAE encoder (Rombach et al. 2022) into latent representations $z_0^{\mathrm{f}} \in \mathbb{R}^{w\times h\times d}$, $z_0^{\mathrm{b}} \in \mathbb{R}^{w\times h\times d}$, $z_0^{\mathrm{c}} \in \mathbb{R}^{w\times h\times d}$, respectively, where w and h denote the width and height of the latent shape, d is the feature channels and the subscript 0 denotes the initial timestep of the diffusion process. Next, we apply the DPM-Solver++ inversion to convert the initial latents z_0^{f}, z_0^{b}, and z_0^{c} to noisy latents z_T^{f}, z_T^{b}, and z_T^{c}. This preprocess enabling the image modification during subsequent reconstruction process.

Share-Attention Module

During the reconstruction process from the time step T to 0, we incorporate the style feature into z_t^{c} using the shared attention module. This module can be viewed as a more general form of the self-attention module, allowing for feature flow between input images. Specifically, the traditional self-attention module projects the input feature $z \in \mathbb{R}^{m\times d}$ of length $m = (w \cdot h)$ onto the corresponding $Q, K, V \in \mathbb{R}^{m\times d}$ through learned linear layers inside the original self-attention module and computes the atten-

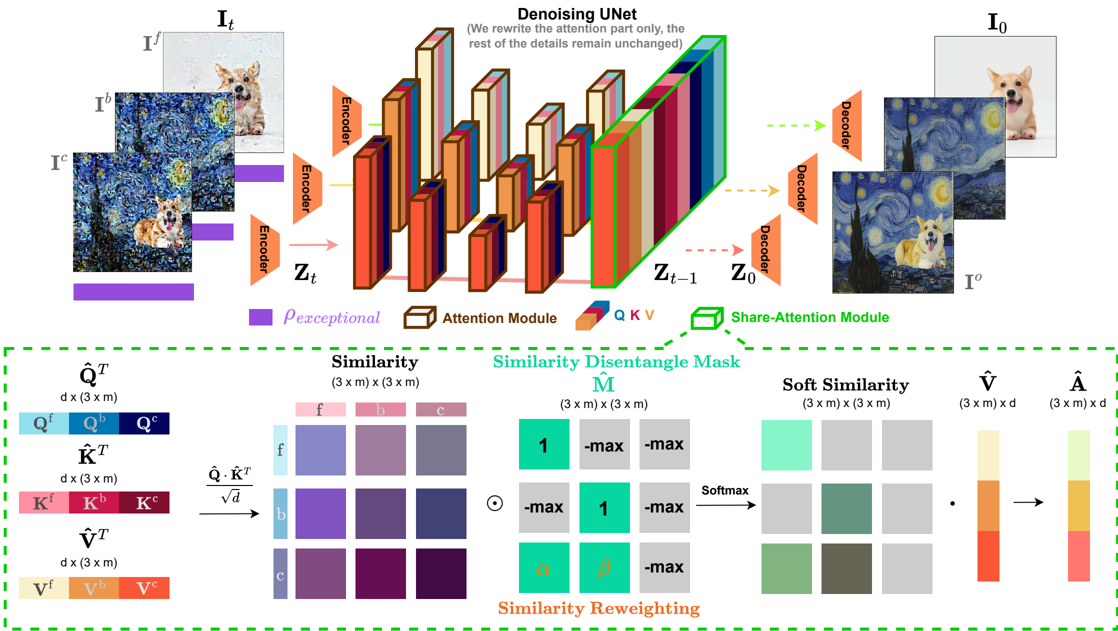


Figure 3: The architecture of our proposed TF-GPH method involves several stages. Initially, we feed the denoising U-Net with the inverse latent Z_t, and during the first $l < L_{\text{share}} - 1$ layers of the U-Net, the three latent representations, z_t^{f}, z_t^{b}, and z_t^{c}, are forwarded separately to the Attention Module. Afterward, they are fed into the Share-Attention Module (the blue part below), obtaining their image-wise attention via Eq. (2). In the end, the output harmonized image I^{o} is produced.

tion matrix $A \in \mathbb{R}^{m \times d}$ as follows.

$$A(Q, K, V) = \text{Softmax}\left(QK^T/\sqrt{d}\right)V, \tag{1}$$

To enable the flow of feature information between images during the attention operation, we should consider three images at the same time instead of processing each attention matrix independently. To create the query, key, and value from three different inputs, we first concatenate three input latents on the first dimension to form $Z_T \in \mathbb{R}^{(3 \cdot m) \times d} = [z_T^{\text{f}}, z_T^{\text{b}}, z_T^{\text{c}}]$. Then we project the latent Z_T into the corresponding $\hat{Q}, \hat{K}, \hat{V} \in \mathbb{R}^{(3 \cdot m) \times d}$.

Similarity Disentangle Mask

However, directly feeding $\hat{Q}, \hat{K}, \hat{V}$ into Eq. (1) may disrupt the features of z^{f} and z^{b} since the additional attention from other images making the latent differ from original reconstruction without attention from others. To keep the content of z^{f} and z^{b} intact for correctly guiding the harmonization of z^{c}, we propose a specially designed mask called **similarity disentangle mask** $\hat{M} \in \mathbb{R}^{(3 \cdot m) \times (3 \cdot m)}$ that allows z^{c} to utilize information from z^{f} and z^{b} while keeping z^{f} and z^{b} intact. The shared attention equation is thus calculated by:

$$\hat{A}(\hat{Q}, \hat{K}, \hat{V}) = \text{Softmax}\left(\hat{M} \odot (\hat{Q}\hat{K}^T)/\sqrt{d}\right)\hat{V}, \tag{2}$$

where $\odot$ denotes the Hadamard product. Afterward, Eq.(2) outputs the batch attention $\hat{A} \in \mathbb{R}^{(3 \cdot m) \times d}$ containing the intact A^{f}, A^{b}, and A^{c} guided by the features of z^{b} and z^{b}.

The specially designed $\hat{M}$ can be visualized as:

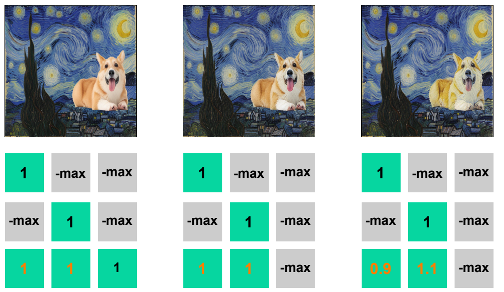
$$\hat{M} = \begin{bmatrix} 1 \cdot J & \nu \cdot J & \nu \cdot J \\ \nu \cdot J & 1 \cdot J & \nu \cdot J \\ \alpha \cdot J & \beta \cdot J & \gamma \cdot J \end{bmatrix}$$

Here, $J \in \mathbb{R}^{m \times m}$ is an all-one matrix, and $\nu = -\infty$ minimizes the similarity between Q and K on the corresponding entry, keeping A^{f}, A^{b} intact. While α, β, and γ control the attention of Q^{c} towards K^{f}, K^{b}, and K^{c}, respectively.. It is worth noting that when setting $\alpha = -\infty$, $\beta = -\infty$, and $\gamma = 1$, each row in Eq.(2) is equivalent to Eq. (1) as Q^{c}, Q^{f}, and Q^{b} can only attend to its counterpart K^{c}, K^{f}, and K^b without information from other images. Therefore, our proposed similarity disentangle mask can be viewed as an expansion of attention mechanism with adjustable entries controlling features sharing.

Furthermore, to completely disentangle the features related to the object reference z^{f} from z^{b}, we set the entry γ to $-\infty$, which blocks the functionality of K^{c} and V^{c}. By this means, we can control the features related to the pasted-foreground object within z^{c} by modifying entry α, which control the influence of z^{f}, and similarly, control the background features within z^{c} related to z^{b} by adjusting its corresponding entry β. As shown in Fig. 4(b), the output remains nearly the same to Fig. 4(a) validating that the features of z^{c} can be totally controlled by two references z^{f} and z^{b}

Similarity Reweighting

Another intriguing observation from Fig. 4(a) and (b) is that the output image only changes slightly even when the pasted



(a) All attention (b) Share w/o reweighting (c) Share w/ reweighting

Figure 4: Comparisons of different attention strategy with corresponding similarity mask (read with Fig. 3).

"corgi" in I^{c} has a different resolution compared to the "corgi" in I^f. We infer that the robustness of the pretrained diffusion model enables it to capture high-level semantic and structural information despite minor disturbances, such as differences in scale and position. Consequently, the self-attention layer can withstand these perturbations, producing results that remain similar to the original input.

To determine the perturbation that can break the self-attention robustness while still generating high-quality results, a simple yet effective idea of attention injection has been widely adopted by previous research (Gu et al. 2024; Hertz et al. 2023; Lu, Liu, and Kong 2023; Tumanyan et al. 2023). This approach introduces strong perturbations to the attention mechanism by directly appending either prompt-guided cross-attention output or image-guided self-attention output computed with other images. Another common strategy is applying the "AdaIN" technique to different components. For example, (Chung, Hyun, and Heo 2024) compute the mean and variance of z^{b}, then normalize z^{c} with these computed values, or (Hertz et al. 2024) perform AdaIN normalization on K^{b} and K^{c}. Although these direct modifications to z^{c} can produce exaggerated image editing effects, they also disrupt semantic details and structural coherence.

In contrast to the aforementioned strategies, we argue that certain attributes crucial to content identity should not be entirely replaced by style features, as discussed in (Saini, Pham, and Shrivastava 2022). For example, the yellow hue of a corgi is an essential part of its identity and should be preserved rather than changed to the global background color tone such as blue or black. Instead, integrating the yellow color from the background style into the corgi would better maintain its content identity as shown in Fig. 4(c). To achieve this, we prioritize high-similarity style features that potentially possess content-related attributes such as color, texture, or semantics. Instead of evenly scaling the attention output as in (Deng et al. 2023), scaling the similarity has a different effect due to the softmax process involved. By scaling the input similarity before applying softmax, high-similarity features are amplified while low-similarity features are diminished in the final attention output. This approach helps minimize content disruption during stylization. Without loss of generality, we place a higher tendency on style reference z^{b} by setting β to 1.1, and a minor preference on content preservation related to z^{f} by setting α to 0.9, our designed TF-GPH achieves remarkable painterly harmonization effects without losing content structure and background consistency. The overall algorithm and visualization can be found in Appendix.

Experiments

Setup. We employ the Stable Diffusion model (Rombach et al. 2022) as the pretrained backbone and utilize DPM Solver++ as the scheduler with a total of 25 steps for both inversion and reconstruction. Specifically, we first resize the input images I^{f}, I^{b}, and I^{c} to 512×512, and encode them into corresponding z_0^{f}, z_0^{b}, and z_0^{c}. Afterward, we take these latents with prompt embedding $\rho_{exceptional}$ as the input during both inversion and reconstruction stage. As for the rest of setting, we refer these hyperparameters (T_{share}, L_{share}, α, β) as "inference-time-adjustable hyperparameters" since they can be flexibly adjusted to modulate the strength of style according to different use cases during the inference process, we leaves remain setting details in the Appendix.

Datasets. We generalize the computational metrics and benchmarks from various image editing methods including "Painterly image harmonization", "Prompt-based Image Composition", and "Style Transfer". Additionally, we examined different approaches on our proposed "General Painterly Harmonization" achieved by the General Painterly Harmonization Benchmark (GPH Benchmark). This benchmark generalizes real-world usage scenarios of the aforementioned methods including "Object Insertion", "Object Swapping" and "Style Transfer", providing a more practical benchmark and aims to mitigate the shortcomings of existing benchmarks such as WikiArt combined with COCO (Tan et al. 2019; Lin et al. 2014) and the TF-ICON Benchmark (Lu, Liu, and Kong 2023). Details and experiment of these datasets can be found in the Appendix.

LPIPS and CLIP regarding computation metrics. In our evaluation, we use LPIPS (Zhang et al. 2018) and CLIP (Radford et al. 2021) metrics, abbreviated as LP and CP respectively. LPIPS is sensitive to low-level visual features, while CLIP excels in capturing high-level semantic features. In TF-ICON benchmark, these two metrics are leveraged to assess content preservation and stylized performance, where LP_{fg} and CP_{img} are calculated to measure the content consistency and image semantic similarity, respectively. Moreover, LP_{bg} is also used to measure background consistency before and after harmonization. And CP_{dir} (Gal et al. 2022) to calculate the alignment level between the feature shift direction of pasted object and the background. Finally, we adopt CP_{st} (Cheng et al. 2023) to measure the feature similarity of harmonized images and style references.

Range-based evaluation. While metrics such as LPIPS and CLIP are useful for assessing content fidelity and stylization in image harmonization, they can sometimes **emphasize either too much content preservation or excessive stylization**, resulting inharmonious image. Therefore, we argue that an effective pipeline should offer users the flexibility of balancing between stylization intensity and content integrity by adjusting hyperparameters. To measure this capability, we suggest defining upper and lower bounds for con-

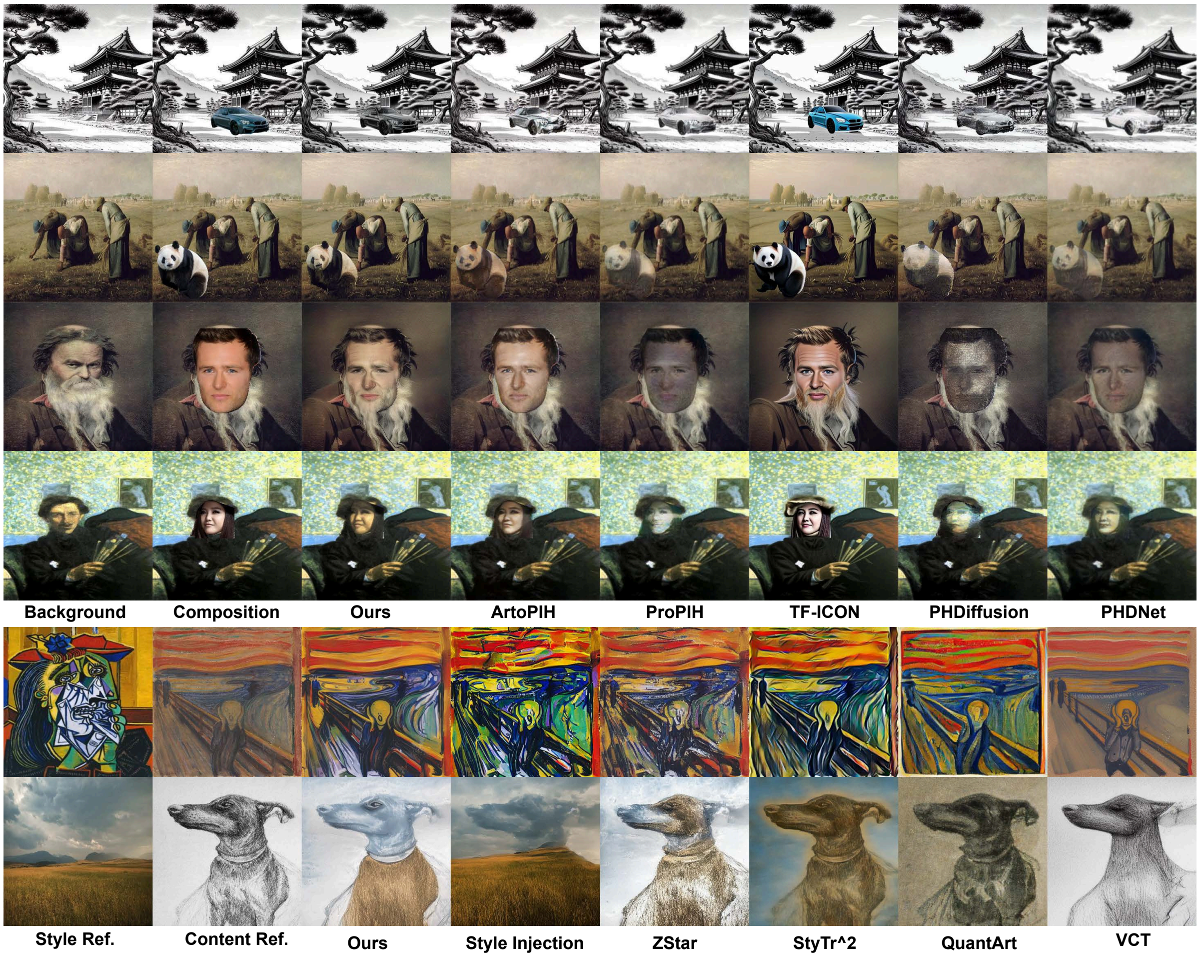


Figure 5: Qualitative result of object insertion (rows 1 and 2), object swapping (rows 3 and 4), and style transfer (rows 5 and 6)

tent preservation and stylization, which can serve as indicators of a method's adaptability across different harmonization scenarios. The corresponding upper and lower settings for the baselines are provided in the Appendix.

Baselines. We compared our proposed TF-GPH with different state-of-the-art methods on various tasks for the comprehensive assessment. For "Painterly Image Harmonization", we incorporate ArtoPIH (Niu et al. 2024a), ProPIH (Niu et al. 2024b), PHDiffusion (Lu et al. 2023), and PHDNet (Cao, Hong, and Niu 2023). For "Prompt-Based Image Composition", we use TF-ICON (Lu, Liu, and Kong 2023). In the "Style Transfer" category, we evaluate Style Injection (StyleID) (Chung, Hyun, and Heo 2024), ZSTAR (Deng et al. 2023), StyTr2 (Deng et al. 2022), QuantArt (Huang et al. 2023), and VCT (Cheng et al. 2023). For models designed for 256x256 resolutions, e.g. ProPIH, we resized the output to 512x512 for high-resolution evaluation. Comparisons of 256x256 resolution are in the Appendix.

Qualitative Comparison

For qualitative comparison, TF-GPH showcases remarkable capabilities in our proposed GPH Benchmark as depicted in Fig. 5, ranging from low-level texture harmonization such as transitioning to singular colors and color matching (rows 1 and 2) to high-level semantic harmonization such as extending the skin color of the replaced man onto the pasted face or redrawing the covered beard along with the chin line of the pasted face (rows 3 and 4). Although ArtoPIH and ProPIH are able to achieve low-level texture harmonization, they struggle with high-level semantic blending, such as face recovery in the row 3 of Fig. 5, due to training data limitation. This highlights how our similarity reweighting technique effectively leverages the characteristics of the diffusion model to achieve both texture and semantic harmonization with image-wise attention.

Moreover, our proposed TF-GPH demonstrates exceptional performance in style transfer (row 5,6 in Fig. 5). Our method outperforms others in stylizing original content while maintaining high image quality, effectively mitigating the common issue of content disruption seen in other attention-based methods. For instance, our approach preserves content coherence more accurately than StyleID and ZSTAR as shown in row 5. Furthermore, our model excels in blending photographic features into sketches, where other

	Painterly Harmonization (512x512)					Style Transfer (512x512)			
	Ours†	ArtoPIH	ProPIH†	TF-ICON†	PHDiff†	Ours†	StyleID†	Z-STAR†	StyTr^2
$Venue$	-	AAAI'24	AAAI'24	ICCV'23	MM'23	-	CVPR'24	CVPR'24	CVPR'22
$LP_{\text{bg}} \downarrow$	**0.11**/0.12	0.25	0.31/0.31	0.20/0.36	0.12/0.12	0.72/**0.55**	0.60/0.56	0.70/0.63	0.61
$LP_{\text{fg}} \downarrow$	**0.10**/0.32	0.37	0.34/0.42	0.32/0.36	0.10/0.39	**0.11**/0.45	0.36/0.48	0.15/0.37	0.40
$CP_{\text{img}} \uparrow$	**95.42**/78.63	84.96	87.78/77.24	85.35/82.05	95.13/73.65	**96.43**/69.57	83.26/69.80	92.31/77.20	83.57
$CP_{\text{st}} \uparrow$	47.50/**56.37**	49.67	47.87/51.19	47.66/47.40	47.64/55.96	57.97/**78.60**	67.70/77.47	59.61/69.50	63.28
$CP_{\text{dir}} \uparrow$	0.11/11.69	5.40	2.83/10.09	2.96/4.63	0.35/**15.39**	3.97/**51.59**	26.96/50.24	9.76/34.83	22.08

Table 1: Quantitative results of GPH-Benchmark († represents the method with inference-time-adjustable hyperparameters. The left side of / represents content emphasis strategy, while the right side of / represents stylized emphasis strategy.)

methods fail, as depicted in the row 6. This illustrates that our similarity disentangle mask not only preserves content information effectively but also extracts style features robustly, even in scenarios like photography.

Quantitative Results

Tab. 1 presents the quantitative results of the GPH Benchmark. The performance of TF-GPH consistently surpasses that of existing evaluation criteria on different benchmarks. Our similarity disentangle mask significantly improves reference preservation compared to prompt-based editing methods such as TF-ICON, as well as traditional harmonization methods like ArtoPIH and ProPIH, achieving the lowest LP_{bg} and LP_{fg} values while also demonstrating superior stylization with the highest CP_{st}.

Moreover, TF-GPH employs a novel similarity-based editing technique that consistently outperforms existing attention-based methods, such as StyleID, in both content preservation metrics (LP_{fg}, CP_{img}) and stylization metrics (CP_{st}, CP_{dir}). Additionally, the wide content preservation and stylization range of TF-GPH confirm the potential of our inference-time-adjustable hyperparameters, which can accommodate various preferences.

We also conduct user preference studies, which are viewed more reliable (Podell et al. 2023). The study encompasses two tasks: Style Transfer and Painterly Harmonization, which includes object insertion and swapping. For each task, we recruited 20 participants, each asked with responding to 20 image pairs. Participants were instructed to compared the generated images based on three criteria: (1) Content Consistency, (2) Style Similarity, and (3) Visual Quality. We provide the results in Fig. 6, where TF-GPH achieving the highest preference in overall quality and content consistency, along with competitive style similarity. These results validate our hypothesis that visual quality transcends mere content preservation or style strength.

Ablation Study

Tab. 2 reveals the impact of components within TF-GPH. Simply applying reconstruction to the composite image is ineffective at harmonizing pasted object into background. By contrast, when we incorporating similarity disentangle mask, we perfectly disentangle the attention of z^{c} to the two other image latents z^{f}, z^{b} and reach **nearly no reconstruction loss**. Furthermore, the integration of the similarity reweighting strategy significantly improves the stylization

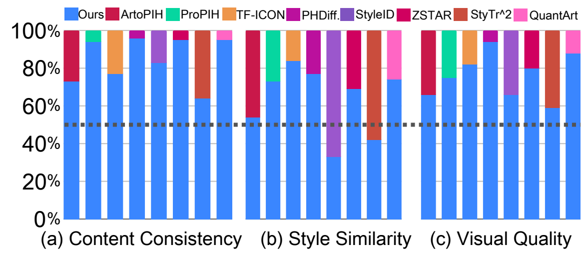


Figure 6: User preferencestudy result.

Metrics	$LP_{\text{bg}} \downarrow$	$LP_{\text{fg}} \downarrow$	$CP_{\text{st}} \uparrow$	$CP_{\text{dir}} \uparrow$
Reconstruction	**0.11**	**0.09**	47.50	0.11
$+SDM$	0.11	0.10	47.50	0.18
$+SDM+SR$	/0.12	/0.32	**/56.37**	**/11.69**
Reconstruction	0.72	**0.11**	57.97	3.97
$+SDM$	0.69	0.12	59.05	5.12
$+SDM+SR$	**/0.56**	/0.45	**/78.60**	**/51.59**

Table 2: Ablation study on TF-GPH's components in painterly harmonization (upper) and style transfer (bottom) on GPH Benchmark. (We abbreviate "Similarity Disentangle Mask" and "Similarity Reweighting" as "SDM" and "SR"). Because the +SDM+SR is the stylization emphasis strategy of +SDM, we put its stylization upper bound here.

indices CP_{st} and CP_{dir} across both tasks, demonstrating its effectiveness in encoding cross-image information into the composite image.

Conclusion

In this work, we introduce a novel **similarity disentangle mask**, faciliatating the utilization of attention from different images. Furthermore, we devised the **similarity reweighting** technique capable of controlling the attention strength of reference images without the need for fine-tuning or prompt. Based on them, we propose **TF-GPH** to perform a more general form of painterly harmonization. Also, we construct the **GPH Benchmark** with **range-based evaluation** aiming to mitigate the current shortage of evaluations for image editing. Both human and quantitative evaluations show that TF-GPH produces more harmonious results, which should benefit future research in image editing.

Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan under Grants NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

Cao, J.; Hong, Y.; and Niu, L. 2023. Painterly image harmonization in dual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 268–276.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, H.; Gu, Z.; Li, Y.; Lan, J.; Meng, C.; Wang, W.; and Li, H. 2023. Hierarchical dynamic image harmonization. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1422–1430.

Cheng, B.; Liu, Z.; Peng, Y.; and Lin, Y. 2023. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22736–22746.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

Cong, W.; Zhang, J.; Niu, L.; Liu, L.; Ling, Z.; Li, W.; and Zhang, L. 2020. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8394–8403.

Deng, Y.; He, X.; Tang, F.; and Dong, W. 2023. Z* : Zero-shot Style Transfer via Attention Rearrangement. *arXiv preprint arXiv:2311.16491*.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11326–11336.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gatys, L. A.; Ecker, A. S.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3985–3993.

Gu, J.; Wang, Y.; Zhao, N.; Fu, T.-J.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.; et al. 2024. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems*, 36.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations*.

Hertz, A.; Voynov, A.; Fruchter, S.; and Cohen-Or, D. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4775–4785.

Huang, S.; An, J.; Wei, D.; Luo, J.; and Pfister, H. 2023. QuantArt: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5947–5956.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Kwon, G.; and Ye, J. C. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18062–18071.

Li, Y.; Wang, N.; Liu, J.; and Hou, X. 2017. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2230–2236.

Lin, C.-H.; Yumer, E.; Wang, O.; Shechtman, E.; and Lucey, S. 2018. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9455–9464.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lu, L.; Li, J.; Cao, J.; Niu, L.; and Zhang, L. 2023. Painterly image harmonization using diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 233–241.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.

Luan, F.; Paris, S.; Shechtman, E.; and Bala, K. 2018. Deep painterly harmonization. In *Computer graphics forum*, volume 37, 95–106. Wiley Online Library.

Niu, L.; Cao, J.; Hong, Y.; and Zhang, L. 2024a. Painterly Image Harmonization by Learning from Painterly Objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4343–4351.

Niu, L.; Hong, Y.; Cao, J.; and Zhang, L. 2024b. Progressive Painterly Image Harmonization from Low-Level Styles to High-Level Styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4352–4360.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saini, N.; Pham, K.; and Shrivastava, A. 2022. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13658–13667.

Tan, L.; Li, J.; Niu, L.; and Zhang, L. 2023. Deep image harmonization in dual color spaces. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2159–2167.

Tan, W. R.; Chan, C. S.; Aguirre, H.; and Tanaka, K. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409.

Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; and Yang, M.-H. 2017. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3789–3797.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1921–1930.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, H.; Zheng, S.; Zhang, J.; and Huang, K. 2019. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, 2487–2495.

Xing, Y.; Li, Y.; Wang, X.; Zhu, Y.; and Chen, Q. 2022. Composite photograph harmonization with complete background cues. In *Proceedings of the 30th ACM international conference on multimedia*, 2296–2304.

Zhang, H.; Zhang, J.; Perazzi, F.; Lin, Z.; and Patel, V. M. 2021. Deep image compositing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 365–374.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10146–10156.