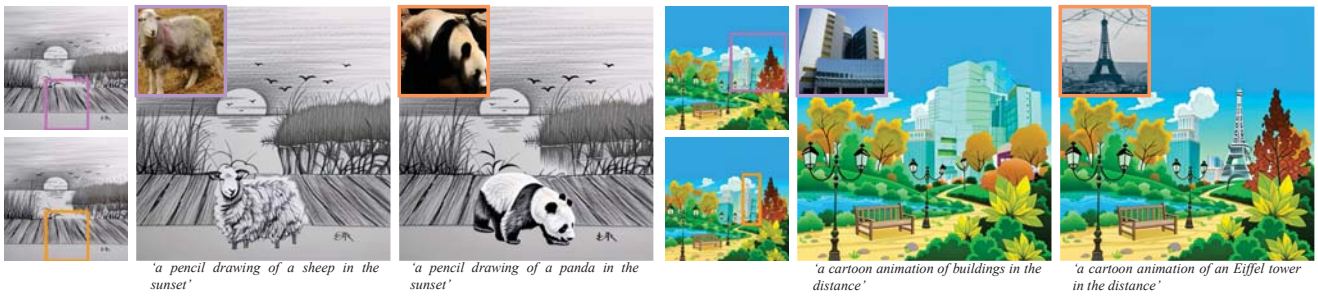
TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition

Shilin Lu[1] Yanzhu Liu[2] Adams Wai-Kin Kong[1]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Institute for Infocomm Research (I^{2}R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
shilin002@e.ntu.edu.sg, liu_yanzhu@i2r.a-star.edu.sg, adamskong@ntu.edu.sg

(a) Composition within the photorealism domain (b) Composition between real & oil painting domains



(c) Composition between real & sketchy painting domains (d) Composition between real & cartoon domains

Figure 1: Image composition aims to seamlessly blend distinct objects into a specific visual context. Our training-free framework equips attention-based text-driven diffusion models with the capability to achieve this task across various domains (a) photorealism, (b) oil painting, (c) sketching, and (d) cartoon animation, within 20 sampling steps.

Abstract

Text-driven diffusion models have exhibited impressive generative capabilities, enabling various image editing tasks. In this paper, we propose TF-ICON, a novel Training-Free Image COmpositioN framework that harnesses the power of text-driven diffusion models for cross-domain image-guided composition. This task aims to seamlessly integrate user-provided objects into a specific visual context. Current diffusion-based methods often involve costly instance-based optimization or finetuning of pretrained models on customized datasets, which can potentially undermine their rich prior. In contrast, TF-ICON can leverage off-the-shelf diffusion models to perform cross-domain image-guided composition without requiring additional training, finetuning, or optimization. Moreover, we introduce the exceptional prompt, which contains no information, to facilitate text-driven diffusion models in accurately inverting real images into latent representations, forming the basis for compositing. Our experiments show that equipping Stable Diffusion with the exceptional prompt outperforms state-of-the-art inversion methods on various datasets (CelebA-HQ, COCO, and ImageNet), and that TF-ICON surpasses prior baselines in versatile visual domains. Code is available at https://github.com/Shilin-LU/TF-ICON

1. Introduction

Image composition task involves incorporating unique objects from different photos to create a harmonious image within a specific visual context, a.k.a. image-guided composition. For instance, consider the scenario where one desires to incorporate a beloved panda into one's favorite artwork, *e.g.*, oil or sketchy painting. The objective is to create a new image where the panda blends seamlessly into the scene without altering the appearance of the panda and the background, just as an artist meticulously crafted this panda

for that artwork (See Figure 1). This task is inherently challenging, as it requires maintaining illumination consistency and preserving identifying features. The challenge is further compounded when the photos come from various domains.

While recently large-scale text-to-image models [10, 19, 51, 56, 59, 61, 78] have achieved remarkable success in text-driven image generation, the ambiguity inherent in natural language presents challenges in conveying precise and nuanced visual details, even with highly detailed text prompts. Although this challenge is effectively addressed by enabling personalized concept learning [24, 25, 34, 37, 60], these methods require costly instance-based optimization and are limited in generating concepts with specified backgrounds. Recent studies [67, 77] have shown that diffusion models can achieve image-guided composition by explicitly incorporating additional guiding images. However, these models are retrained from the pretrained diffusion model on tailored datasets, which can damage the rich prior of the model. As a result, these models have limited compositional abilities beyond their training domain and still require significant computational resources.

Given the wealth of large text-to-image models that have been trained on extensive language-image datasets, we pose a question: *how could these models be leveraged for image-guided composition without incurring costly training or finetuning, thereby avoiding damaging the diverse prior?* To answer it, we propose the Training-Free Image COmpositioN (TF-ICON) framework, which equips attention-based text-to-image diffusion models with the capability to perform image-guided composition without requiring additional training, fine-tuning, extra data, or optimization. To the best of our knowledge, this is the first training-free framework developed for image-guided composition. The framework is compatible with various diffusion model samplers, enabling completion within 20 steps, and harnesses rich semantic knowledge to facilitate image-guided compositions across diverse domains (see Figure 1).

Our approach constitutes an image-guided composition interface through denoising from a reliable starting latent code with the injection of composite self-attention maps. Finding the latent code that allows for reconstructing an input image while maintaining its editability, a.k.a. image inversion, is a challenging yet crucial step for state-of-the-art (SOTA) image editing frameworks involving real images [15, 26, 35, 38, 48, 53, 54, 70]. For diffusion models, while denoising diffusion implicit models (DDIM) inversion [65] has been effective for unconditional diffusion models, it falls short for text-driven diffusion models [26, 50, 70, 71]. To circumvent this, we introduce the exceptional prompt to accurately invert real images into latent codes upon pre-trained text-to-image models to serve for further composition generation. The accurate latent codes are composed as the starting noise for the diffusion process. Through the gradual injection of composite self-attention maps that are specifically designed to reflect the relations between guiding images, we are able to infuse contextual information from the background into the incorporated objects, which results in harmonious image-guided compositions.

To summarize, we make the following key contributions:

1. We demonstrate the superior performance of high-order diffusion ODE solvers compared to commonly used DDIM inversion for real image inversion.
2. We present an exceptional prompt that allows text-driven models to achieve accurate invertibility, laying a solid groundwork for subsequent editing. Experimental results show that it surpasses SOTA inversion methods on three vision datasets.
3. We propose the first training-free framework that enables cross-domain image-guided composition for attention-based diffusion models.
4. We demonstrate quantitatively and qualitatively that our framework outperforms prior baselines for image-guided composition.

2. Related Work

Image composition. Image composition is widely applied to electronic commerce, entertainment, and data augmentation [20, 43] for downstream tasks. It can be broadly categorized into two types: text-guided [5, 6, 11, 22, 45] and image-guided [8, 23, 39, 67, 76, 77, 79]. The former involves composing multiple objects specified by only a text prompt without limiting the appearance of objects, as long as their semantics align with the prompt. Despite the great successes of text-conditioned models, they are often prone to semantic errors [22, 56], especially when the text prompt involves multiple objects. These errors include attribute leakage, attribute interchange, and missing objects, which cause the generated images to critically different from the user's intention [22, 56]. As a result, extensive prompt engineering [74] is often necessary to achieve the desired results. In contrast to text-only guided composition, image-guided composition involves incorporating specific objects and scenarios from user-provided photos, potentially with the aid of a text prompt. However, due to the nature of involving additional real images, it is more challenging particularly when images from different visual domains. Conventionally, image-guided composition is divided into several sub-tasks [52], such as object placement [7, 12, 40, 69, 80], image blending [75, 81], image harmonization [14, 16, 30, 76], and shadow generation [29, 42, 63, 83], each of which is typically addressed by different models and pipelines.

Image inversion. Extensive research has been conducted on image inversion for GANs, including latent-based op-

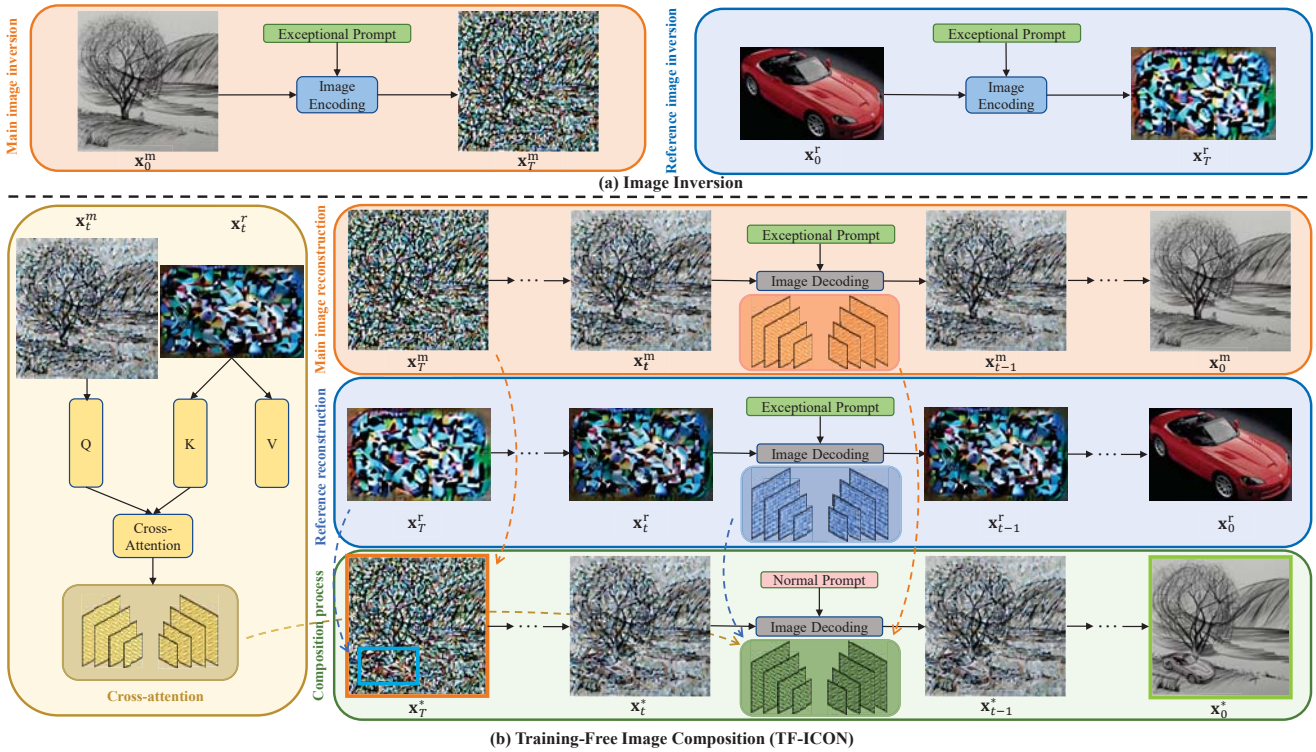


Figure 2: The proposed training-free image composition framework. (a) The exceptional prompt is used to invert the main and reference images into noises $\mathbf{x}_T^{\mathrm{m}}, \mathbf{x}_T^{\mathrm{r}}$, which are then composed to form the starting point $\mathbf{x}_T^*$ for the composition process. (b) Three constituents are composed for injecting into the composition process at early timesteps, including self-attention maps from the main and reference image reconstruction processes, as well as cross attention between the main and reference images. For better clarity and readability, the original main and reference images are shown in the pixel space instead of the VAE latent space, and the reference image is presented without resizing and zero-padding.

timization [1, 2, 33], encoders [3, 57, 68], and fine-tuning [4, 58]. For diffusion models, DDIM [65] is a widely used technique for inversion in image editing frameworks. However, in text-driven settings, DDIM leads to significant reconstruction distortion due to the instability resulting from classifier-free guidance (CFG) [18, 28]. Recently, null-text inversion [50] has been proposed to achieve accurate inversion by optimizing the unconditional prediction of the text-to-image model. It demonstrates promising results but requires instance-based optimization. Concurrently, EDICT [71] also achieves near-perfect inversion, albeit doubling the computation time of the diffusion process.

3. Preliminary

Diffusion probabilistic models (DPM) [18, 27, 59, 64] are generative models in which an image is generated by progressively denoising from Gaussian noise. The forward diffusion process gradually perturbs data with infinite noise scales, which can be modeled as the solution of a stochastic differential equation (SDE) $\{\mathbf{x}_t\}_{t=0}^T$. Formally, given a data sample $\mathbf{x}_0 \sim p_0 = p_{\text{data}}$, random noise is gradually injected, eventually resulting in a sample $\mathbf{x}_T$ which is typically distributed as a tractable prior p_T without any information of p_0, as described by the following SDE [36, 66]:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{1}$$

where $\mathbf{w}_t \in \mathbb{R}^d$ is the standard Wiener process (a.k.a., Brownian motion), and $\mathbf{f}(\cdot, t)$ and $g(t)$ are commonly designated as the drift and diffusion coefficient, respectively. On the other hand, the reverse diffusion process can be described by the reverse-time SDE from T to 0 [66]:

$$\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t, \tag{2}$$

where $\bar{\mathbf{w}}_t$ is the Wiener process in the reverse time. The score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ is the only unknown term and can be estimated by a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ whose parameter θ is optimized by a denoising objective [27, 66].

Upon attaining the trained model that predicts the score function accurately, it can be utilized to numerically solve the reverse SDE (Eq. (2)), enabling the generation of samples from a noise distribution. Song *et al*. [66] outline various methods, including Variance Exploding (VE), Variance Preserving (VP), and sub-VP SDE, for constructing SDEs that perturb the unknown data distribution into a fixed prior. In this work, we leverage the pre-trained text-to-image La-

tent Diffusion Model (LDM) [59], a.k.a. Stable Diffusion, which applies the VP SDE in the latent space.

4. Method

Our objective is to utilize a main (background) image $\mathbf{I}^{\text{m}}$, a reference (foreground) image $\mathbf{I}^{\text{r}}$, a text prompt $\mathcal{P}$, and a binary mask $\mathbf{M}^{\text{user}}$ which designates the region of interest within the main image, to generate a modified image $\mathbf{I}^*$. The resultant image $\mathbf{I}^*$ should contain the reference subject with identifying features within the mask, *i.e.* $id(\mathbf{I}^* \odot \mathbf{M}^{\text{user}}) \approx id(\mathbf{I}^{\text{r}})$, while concurrently ensuring that the complementing area closely resembles the main image, *i.e.* $\mathbf{I}^* \odot (\mathbf{1} - \mathbf{M}^{\text{user}}) \approx \mathbf{I}^{\text{m}} \odot (\mathbf{1} - \mathbf{M}^{\text{user}})$. Moreover, it is ideal for the transition between the areas inside and outside the mask to be imperceptible.

We propose a training-free framework that can make use of attention-based pre-trained text-to-image models to perform image-guided composition. To the best of our knowledge, it is the first training-free framework for image-guided composition, which can be accomplished within 20 steps of sampling. The framework is mainly comprised of two steps: **image inversion** (Section 4.1), and **composition generation** (Section 4.2), as shown in Figure 2. The full algorithm is presented in Appendix B.2.

4.1. Image Inversion with Exceptional Prompt

Achieving precise manipulation of real images often necessitates an accurate inversion process that identifies the corresponding latent representation, which not only provides editability for meaningful manipulation but also accurately reconstructs the input image [26, 68]. For diffusion models, optimal editability is typically characterized by a noise encoding that conforms to the ideal statistical properties of zero-mean, unit-variance Gaussian noise [53].

ODE inversion. Most diffusion frameworks for image editing [15, 26, 35, 38, 53, 70] use DDIM inversion to invert the real image into its latent representation. However, our findings suggest that this may not be the optimal choice for inverting real images. It has been proven that DDIM is a first-order discretization of the associated probability flow ordinary differential equations (ODE) of Eq. (2) [62, 65, 66]:

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t, \qquad (3)$$

which can be solved using $\epsilon_\theta(\mathbf{x}_t, t)$ and shares the consistent marginal probability distribution $\{p_t(\mathbf{x}_t)\}_{t=0}^T$ with Eq. (2). Various samplers [32, 44, 46, 47] have been developed for solving the diffusion ODE starting from noise $\mathbf{x}_T$ to achieve fast sampling (10∼20 steps). We offer the insight that utilizing these ODE solvers in turn as encoders starting from the real image $\mathbf{x}_0$ yields better latent representation $\mathbf{x}_T$, compared with those obtained through commonly used DDIM. A quantitative analysis is given in Appendix A.1. The enhanced alignment between the forward and backward ODE trajectories in the high-order DPM-Solver++ [47] implies that it is better suited for real image inversion. Thus, this paper employs it for all inversions of diffusion models.

(a) Original (b) CFG (c) Cond. (d) Uncond. (e) Ours

Figure 3: The real image reconstruction results using Stable Diffusion with (b) classifier-free guidance (CFG) output $\hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{E}, \varnothing)$; (c) conditional output $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{E})$; (d) unconditional output $\epsilon_\theta(\mathbf{x}_t, t, \varnothing)$; and (e) ours. The prompts for (b) and (c) are *'a photo of a singer'*, *'a photo of a car'*, and *'an oil painting'*. See Appendix A.3 for elaboration.

Exceptional prompt. In the unconditional setting $\epsilon_\theta(\mathbf{x}_t, t)$, solving the diffusion ODE (Eq. (3)) from 0 to T enables us to obtain the better latent code $\mathbf{x}_T$ for the real image $\mathbf{x}_0$. However, in the text-driven setting $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{E})$, existing image editing works [26, 50, 70, 71] have shown that the inversion process is prone to significant reconstruction errors, due to the instability induced by CFG [18, 28]:

$$\hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{E}, \varnothing) = s \cdot \epsilon_\theta(\mathbf{x}_t, t, \mathcal{E}) + (1-s) \cdot \epsilon_\theta(\mathbf{x}_t, t, \varnothing), \quad (4)$$

where $\varnothing = \psi(\text{""})$ and $\mathcal{E} = \psi(\mathcal{P})$ are embeddings of the null and normal prompt, and s is the guidance scale. Our experiments further reveal that even without CFG, both conditional output $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{E})$ and unconditional output $\epsilon_\theta(\mathbf{x}_t, t, \varnothing)$ of text-to-image diffusion models still produce large reconstruction errors, as depicted in Figure 3.

To achieve accurate inversion, we present a straightforward yet effective solution, namely *exceptional prompt*, $\mathcal{P}_{\text{exceptional}}$. Intuitively, any information contained within the input prompt can result in the deviation of the backward ODE trajectories from the forward trajectories. Hence, we remove all information by setting all token numbers to a common value and eliminating positional embeddings for the text prompt, as depicted in Figure 4. Importantly, the exceptional prompt is distinguished from the null prompt by its absence of special tokens, such as `[startoftext]`, `[endoftext]`, and `[pad]`, which still retain information. The exceptional prompt is applied only in image inversion but not in the composition process. The choice of the

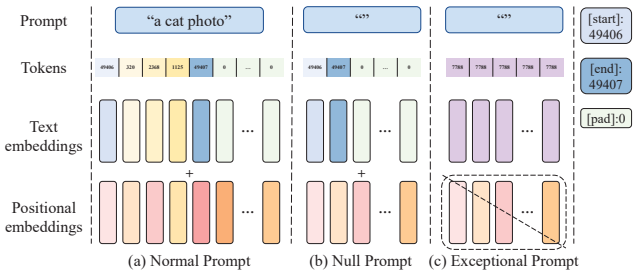


Figure 4: The illustration of comparison among (a) Normal Prompt, (b) Null Prompt, and (c) Exceptional Prompt.

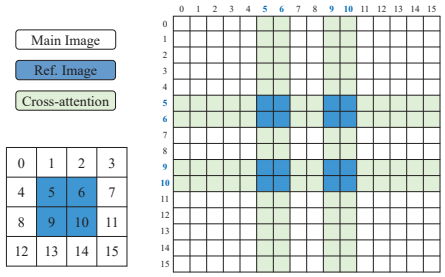


Figure 5: A toy example for attention composition.

token value does not significantly affect the inversion. The detailed analysis of the exceptional prompt and token value selection is provided in Appendix A.2 and A.4, respectively.

Figure 3 visually demonstrates the effectiveness of the exceptional prompt. Our results $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{W})$ with the exceptional prompt embedding $\mathcal{W} = \psi(\mathcal{P}_{\text{exceptional}})$ are more visually accurate than others. The quantitative experiments are shown in Section 5.1. All the results in Figure 3 are obtained by solving the forward and backward diffusion ODEs using the second-order DPM-Solver++ [47] in 20 steps.

4.2. Training-Free Image Composition

Upon equipping the accurate invertibility, image composition can be performed based on it. The composition process consists of two key components: **noise incorporation** and **composite self-attention maps injection**.

Noise incorporation. Before inverting images into noises, a simple preprocessing step is necessary for the reference image. Typically, only the foreground in the reference is desired for composition, so the preprocessing step involves using a pretrained segmentation model [84] to remove the background, resizing and repositioning the object to match the user's mask in the main image, and padding it with zeros to ensure it is the same size as the main image (See Appendix B.1 for visual illustration).

Once the preprocessing is complete, the main and reference images are inverted to corresponding noises $\mathbf{x}_T^{\text{m}}$ and $\mathbf{x}_T^{\text{r}}$ by solving diffusion ODEs (Eq. (3)) from 0 to T with the exceptional prompt $\mathcal{P}_{\text{exceptional}}$. $\mathbf{x}_T^{\text{m}}$ and $\mathbf{x}_T^{\text{r}}$ are then merged with standard Gaussian noise $\mathbf{z}$ to create the starting point $\mathbf{x}_T^*$ for generating the composition. Formally, the incorporated noise $\mathbf{x}_T^*$ is calculated by

$$\mathbf{x}_T^* = \mathbf{x}_T^{\text{r}} \odot \mathbf{M}^{\text{seg}} + \mathbf{x}_T^{\text{m}} \odot (\mathbf{1} - \mathbf{M}^{\text{user}}) + \mathbf{z} \odot (\mathbf{M}^{\text{user}} \oplus \mathbf{M}^{\text{seg}}), \tag{5}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{M}^{\text{user}}$ is the user mask, $\mathbf{M}^{\text{seg}}$ is the segmentation mask for reference image, and $\mathbf{M}^{\text{user}} \oplus \mathbf{M}^{\text{seg}}$ is the XOR of them, which is the transition area. The incorporation of $\mathbf{z}$ enhances the smoothness of the transition between the regions inside and outside the user mask, by effectively leveraging the prior knowledge of the text-driven diffusion model to inpaint the transition area. Empirically, for cross-domain composition, incorporating the starting point in the noise space usually is more effective, while solely for photorealism, composing in the pixel/latent space and then inverting it as the starting point is more favorable.

Composite self-attention map injection. The incorporated noise $\mathbf{x}_T^*$ is employed as the starting point for solving the diffusion ODE from T to 0 with a normal prompt $\mathcal{P}$ to ultimately generate the composition. $\mathcal{P}$ is intended to assist in inpainting transition areas. However, relying solely on noise incorporation, the pretrained text-to-image model cannot preserve the appearance of the main and reference images effectively, as shown in Figure 7. To tackle this problem, we propose injecting composite self-attention maps in a specially designed manner, as the semantic information is basically retained within the rows and columns of self-attention maps (See Appendix E for visual illustrations).

The composite self-attention map comprises three constituents: two self-attention maps, $\mathbf{A}_{l,t}^{\text{m}}$ and $\mathbf{A}_{l,t}^{\text{r}}$, corresponding to the main and reference images, and a cross-attention map, $\mathbf{A}_{l,t}^{\text{cross}}$, calculated between them. The composition way is illustrated in Figure 5. To compose the reference image in the blue regions of Figure 5 (a), its self-attention map $\mathbf{A}_{l,t}^{\text{r}}$ should be placed in the corresponding blue regions of Figure 5 (b), since the 5th patch can only attend to the patches 6, 9, and 10 in the context of self-attention. The green regions in Figure 5 (b) should contain the cross-attention map, $\mathbf{A}_{l,t}^{\text{cross}}$, which infuses contextual information from the white regions into the blue regions. If the green regions are preserved as the self-attention of the white regions without replacement, the information stored there only reflects the relation between the original patches, such as the 0th and original 5th patches in the example of index (5,0) or (0,5). This results in a lack of surrounding information being provided to the new 5th patch, such as painting or sketching, which is necessary for seamless object transition to other domains (See ablation in Figure 7).

Essential constituents $\mathbf{A}_{l,t}^{\text{m}}, \mathbf{A}_{l,t}^{\text{r}}, \mathbf{A}_{l,t}^{\text{cross}}$ are calculated using self-attention modules of the pretrained Stable Diffusion. Typically, a self-attention module at layer l contains three projection matrices $\mathbf{W}_l^q, \mathbf{W}_l^k, \mathbf{W}_l^v$ in the same dimension $\mathbb{R}^{d \times d}$. Denote the spatial features of the main and reference image at timestep t and layer l as $\mathbf{f}_{l,t}^{\text{m}} \in \mathbb{R}^{(h \times w) \times d}$

and $\mathbf{f}^{\text{r}}_{l,t} \in \mathbb{R}^{(h' \times w') \times d}$, respectively, where $h' \times w'$ is the size of the reference image after resizing to match the size of the user mask. The queries, keys, and values for each self-attention module are obtained as:

$$\mathbf{q}^{\text{m}}_{l,t} = \mathbf{f}^{\text{m}}_{l,t}\mathbf{W}^{q}_{l}, \quad \mathbf{k}^{\text{m}}_{l,t} = \mathbf{f}^{\text{m}}_{l,t}\mathbf{W}^{k}_{l}, \quad \mathbf{v}^{\text{m}}_{l,t} = \mathbf{f}^{\text{m}}_{l,t}\mathbf{W}^{v}_{l}, \tag{6}$$

$$\mathbf{q}^{\text{r}}_{l,t} = \mathbf{f}^{\text{r}}_{l,t}\mathbf{W}^{q}_{l}, \quad \mathbf{k}^{\text{r}}_{l,t} = \mathbf{f}^{\text{r}}_{l,t}\mathbf{W}^{k}_{l}, \quad \mathbf{v}^{\text{r}}_{l,t} = \mathbf{f}^{\text{r}}_{l,t}\mathbf{W}^{v}_{l}, \tag{7}$$

where $\mathbf{q}^{\text{m}}_{l,t}$, $\mathbf{k}^{\text{m}}_{l,t}$, $\mathbf{v}^{\text{m}}_{l,t} \in \mathbb{R}^{(h \times w) \times d}$, and $\mathbf{q}^{\text{r}}_{l,t}$, $\mathbf{k}^{\text{r}}_{l,t}$, $\mathbf{v}^{\text{r}}_{l,t} \in \mathbb{R}^{(h' \times w') \times d}$. Thus, $\mathbf{A}^{\text{m}}_{l,t}$, $\mathbf{A}^{\text{r}}_{l,t}$, and $\mathbf{A}^{\text{cross}}_{l,t}$ are then calculated and composed as $\mathbf{A}^{*}_{l,t}$ for injection:

$$\mathbf{A}^{\text{m}}_{l,t} = \text{Softmax}\left(\mathbf{q}^{\text{m}}_{l,t} \cdot (\mathbf{k}^{\text{m}}_{l,t})^{\mathsf{T}} / \sqrt{d}\right), \tag{8}$$

$$\mathbf{A}^{\text{r}}_{l,t} = \text{Softmax}\left(\mathbf{q}^{\text{r}}_{l,t} \cdot (\mathbf{k}^{\text{r}}_{l,t})^{\mathsf{T}} / \sqrt{d}\right), \tag{9}$$

$$\mathbf{A}^{\text{cross}}_{l,t} = \text{Softmax}\left(\mathbf{q}^{\text{m}}_{l,t} \cdot (\mathbf{k}^{\text{r}}_{l,t})^{\mathsf{T}} / \sqrt{d}\right), \tag{10}$$

$$\mathbf{A}^{*}_{l,t} = \vartheta_{\text{compose}}(\mathbf{A}^{\text{m}}_{l,t}, \mathbf{A}^{\text{r}}_{l,t}, \mathbf{A}^{\text{cross}}_{l,t}), \tag{11}$$

where $\mathbf{A}^{\text{m}}_{l,t} \in \mathbb{R}^{(h \times w) \times (h \times w)}$, $\mathbf{A}^{\text{r}}_{l,t} \in \mathbb{R}^{(h' \times w') \times (h' \times w')}$, $\mathbf{A}^{\text{cross}}_{l,t} \in \mathbb{R}^{(h \times w) \times (h' \times w')}$, and $\vartheta_{\text{compose}}$ is the function to bulid composite self-attention maps $\mathbf{A}^{*}_{l,t}$ based on patch indices (Figure 5).

As a result, three diffusion ODEs are solved simultaneously from T to 0. As depicted in Figure 2 (b), ODEs start from the accurate inverted noises $\mathbf{x}^{\text{r}}_{T}, \mathbf{x}^{\text{m}}_{T}$, and the interpolated noise $\mathbf{x}^{*}_{T}$, respectively. The first two ODEs are solved using the exceptional prompt $\mathcal{P}_{\text{exceptional}}$ to progressively reconstruct the main and reference, thus allowing for the precise retention of $\mathbf{A}^{\text{m}}_{l,t}$, $\mathbf{A}^{\text{r}}_{l,t}$, and $\mathbf{A}^{\text{cross}}_{l,t}$ at each time step t. These attention maps are then composed and injected into the third ODE for generating a natural and cohesive composition with a normal prompt $\mathcal{P}$.

To balance the generation of high-level context and finer details [13, 38], we set a threshold τ_A to determine the time steps for injecting composite self-attention maps in the early stage ($t \in [T \times \tau_A, T]$) and allow the model to explore ODE trajectories through a normal prompt $\mathcal{P}$ in the later stage ($t \in [0, T \times \tau_A]$), guided by the prior of the pre-trained model. However, this freedom, without the imposition of attention injection constraints, often results in deviations from the desired background (see Figure 7). Thus, similar to [6], we set an additional threshold, denoted as τ_B, which regulates the trajectory rectification process. This process entails replacing the regions outside the user mask with the reconstructed main image at various time steps, *i.e.*, $\hat{\mathbf{x}}^{*}_{t} = \mathbf{x}^{*}_{t} \odot \mathbf{M}^{\text{user}} + \mathbf{x}^{\text{m}}_{t} \odot (\mathbf{1} - \mathbf{M}^{\text{user}})$, where $t \in [T \times \tau_B, T]$. Note that only preserving the background at the final step can lead to noticeable artifacts, as shown in Appendix B.3.

5. Experiments

This section consists of two sets of experiments. The first set assesses the effectiveness of the exceptional prompt (Section 5.1). The second set evaluates our image composition framework qualitatively and quantitatively (Section 5.2), followed by an ablation study (Section 5.3).

Table 1: The reconstruction comparison on CelebA-HQ.

	Method	MAE ↓	LPIPS ↓	SSIM ↑
Optimization	I2S [1]	0.064	0.134	0.872
	PTI [58]	0.062	0.132	0.877
Encoder	pSp [57]	0.079	0.169	0.793
	e4e [68]	0.092	0.221	0.742
	ReStyle w/ pSp [3]	0.073	0.145	0.823
	ReStyle w/ e4e [3]	0.089	0.202	0.758
	HFGI w/ e4e [73]	0.062	0.127	0.877
Diffusion	SD w/ CFG	0.134	0.340	0.637
	SD w/ Cond.	0.126	0.308	0.654
	SD w/ Uncond.	0.126	0.304	0.655
	DiffusionCLIP [35]	0.020	0.073	0.914
	Ours	**0.019**	**0.047**	**0.918**
Upper Bound	VQAE [21]	0.018	0.043	0.919

Table 2: The further reconstruction comparison on COCO and ImageNet. *: an upper bound.

Method	MSCOCO (5000)			ImageNet (3000)		
	MAE ↓	LPIPS ↓	SSIM ↑	MAE ↓	LPIPS ↓	SSIM ↑
SD w/ CFG	0.150	0.458	0.568	0.132	0.496	0.575
SD w/ Cond.	0.122	0.359	0.633	0.109	0.389	0.645
SD w/ Uncond.	0.120	0.363	0.636	0.114	0.406	0.635
Ours	**0.030**	**0.073**	**0.868**	**0.033**	**0.087**	**0.852**
VQAE* [21]	0.030	0.069	0.870	0.032	0.084	0.854

5.1. Image Reconstruction

To assess the effectiveness of the exceptional prompt, we compared its performance with SOTA GAN [1, 3, 57, 58, 68, 73] and diffusion [35] inversion methods on the CelebA-HQ [31], following the same setting as described in [35, 73]. Additionally, we conducted experiments on the ImageNet [17] and COCO [41] with Stable Diffusion to further validate our findings. Our results (Tables 1 and 2) show that the exceptional prompt is highly effective in producing reconstructions that closely approximate the upper bound established by the vector quantized autoencoder (VQAE) [21] across all metrics, including MAE, LPIPS [82], and SSIM. The qualitative comparison is shown in Figure 3. All results of Stable Diffusion, including ours, are sampled by the second-order DPM-Solver++ [47]. Experimental settings are detailed in Appendix B.4.

5.2. Image Composition Comparisons

Test benchmark. As there is currently no benchmark for testing cross-domain image-guided composition as a whole, we developed a test benchmark containing 332 samples. Each sample in the benchmark consists of a main (background) image, a reference (foreground) image, a user mask, and a text prompt. The main images comprise

Figure 6: Qualitative comparison with SOTA and concurrent baselines in image-guided composition for sketching, photorealism, painting, and cartoon animation domains. Additional results are available in Appendix H.2.

Table 3: Quantitative evaluation results for image composition in the photorealism domain.

Method	$\text{LPIPS}_{(\text{BG})} \downarrow$	$\text{LPIPS}_{(\text{FG})} \downarrow$	$\text{CLIP}_{(\text{Image})} \uparrow$	$\text{CLIP}_{(\text{Text})} \uparrow$
SDEdit (0.4) [48]	0.35	0.62	80.56	27.73
SDEdit (0.6) [48]	0.42	0.66	77.68	27.98
Blended [5]	0.11	0.77	73.25	25.19
Paint [77]	0.13	0.73	80.26	25.92
DIB [81]	0.11	0.63	77.57	26.84
Ours	**0.10**	**0.60**	**82.86**	**28.11**

four visual domains: photorealism, pencil sketching, oil painting, and cartoon animation. All reference images are from the photorealism domain as the reference requires segmentation models, which are generally more effective in this domain. Further details are available in Appendix G.

Qualitative comparisons. Our qualitative comparisons are performed across four visual domains, employing SOTA and concurrent baselines that are applicable to image-guided composition, including Deep Image Blending (DIB) [81], DCCF [76], Blended Diffusion [5], Textual Inversion [24], Paint by Example [77], and SDEdit [48] under two different noising levels. As shown in Figure 6, our framework is capable of seamlessly composing objects into various domains while maintaining their identities. In contrast, DIB and DCCF fall short in processing the transition areas, leading to noticeable artifacts. Blended Diffusion's foreground generation and Textual Inversion's background generation rely solely on text prompts, causing deviations from the user's intention. While Paint by Example effectively composes images within its photorealistic training domain, it struggles to adapt to other domains. Additionally, SDEdit with fewer timesteps is suitable for image composition in terms of preserving the identifying features of the reference, but the background is changed. See Appendix H.2 for additional comparisons.

Quantitative analysis. The baselines are primarily trained in the photorealism domain, where the objective metrics are more effective; therefore, we focused our quantitative comparison within this domain and relied on user study for comparison in other domains. We assess the same baselines as in the qualitative comparison, with the exception of Textual Inversion [24], which involves instance-based optimization, and DCCF [76], which is used for harmonizing images after copy-and-paste operations. Four metrics are considered: (1) $\text{LPIPS}_{(\text{BG})}$ [82] measures the background consistency, (2) $\text{LPIPS}_{(\text{FG})}$ [82] evaluates the low-level similarity between the edited region and the reference foreground, (3) $\text{CLIP}_{(\text{Image})}$ [55] evaluates the semantic similarity between the edited region and the reference in the CLIP embedding space, and (4) $\text{CLIP}_{(\text{Text})}$ [55] measures the semantic alignment between the text prompt and the resultant image. As presented in Table 3, our method outperforms all baselines. We achieve well preservation

Table 4: User study: higher score, better ranking. P: photorealism; O: oil painting; S: sketchy painting; C: cartoon.

Method	P & P	P & O	P & S	P & C	Total
Blended [5]	1.807	2.314	2.680	2.100	2.093
SDEdit (0.6) [48]	2.368	3.063	2.409	2.713	2.485
Paint [77]	2.879	2.306	2.043	2.673	2.666
DCCF [76]	3.838	3.237	3.297	3.470	3.583
Ours	**4.108**	**4.080**	**4.571**	**4.043**	**4.175**

Table 5: Ablation study: quantitative comparison of various variants of our framework.

Config	$\text{LPIPS}_{(\text{BG})} \downarrow$	$\text{LPIPS}_{(\text{FG})} \downarrow$	$\text{CLIP}_{(\text{Image})} \uparrow$	$\text{CLIP}_{(\text{Text})} \uparrow$
Baseline	0.34	0.65	75.13	**29.00**
+ $\mathcal{P}_{\text{exceptional}}$	0.32	0.64	78.54	28.23
+ SA injection	0.25	0.61	81.64	28.58
+ CA injection	0.26	0.60	81.63	28.52
+ Background	**0.10**	**0.60**	**82.86**	28.11

of the background, high object correspondence in both low-level and high-level feature spaces, as well as a high degree of alignment with the text prompt.

User study. We conducted a user study to compare image composition baselines across domains. We recruited 50 participants via Amazon and tasked them with completing 40 ranking questions. Each question consists of 5 options, generated using distinct methods. Details are available in Appendix D. The ranking criteria comprehensively considered foreground preservation, background consistency, seamless composition, and text alignment. The results are listed in Table 4, where the domain information is presented in the format of 'foreground domain & background domain', *e.g.*, photorealism & oil painting. Our method was favored by most participants across different domains.

5.3. Ablation Study

We ablate our key design choices in the following cases: (1) Baseline, where the composition is generated by solving the diffusion ODE from T to 0 using DPM-Solver++ without any injection. The starting point is composed by inverted noises under the normal prompt; (2) The exceptional prompt is applied to obtain accurate inverted noises; (3) The self-attention maps $\mathbf{A}^{\text{m}}_{l,t}$ and $\mathbf{A}^{\text{r}}_{l,t}$ are composed and then injected; (4) The cross-attention $\mathbf{A}^{\text{cross}}_{l,t}$ between the main and reference images is further composed for injection; (5) The background preservation is applied.

Table 5 presents the quantitative results, which indicate that the complete algorithm outperforms other variants in all metrics except for $\text{CLIP}_{(\text{Text})}$. Notably, the baseline achieves the best $\text{CLIP}_{(\text{Text})}$ as it generates compositions solely relying on the normal prompt without any extra constraint. While metrics alone may not reveal the complete effectiveness of cross-attention injection, Figure 7 illustrates that the interactions between the main and reference images

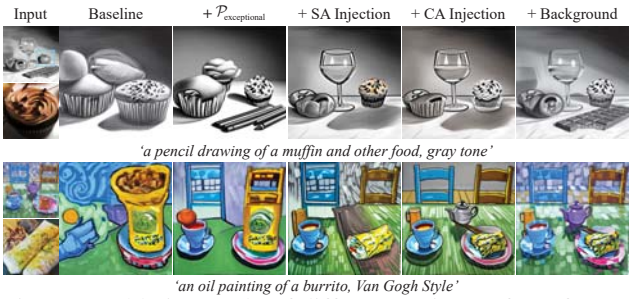


Figure 7: Ablation study of different variants of our framework. SA: self-attention. CA: cross-attention.

are highly beneficial to both foreground and background in terms of preserving appearance and switching domains. Note that preserving the background at different noise levels affects both background and foreground, which is distinct from preserving solely at the final step (Appendix B.3). Additional ablation results are shown in Appendix H.2.

6. Limitations and Future Work

The primary limitation of our work lies in its inability to generate an object view that critically differs from the given reference. As a result, the choice of reference image may be restricted at times. This is because the model relies on self-attention maps to provide layout and appearance information, which in turn constrains the development of alternative views. While introducing a loose self-attention injection can generate different views, this often compromises the preservation of the object's appearance. To overcome this, further research could explore utilizing personalized concept learning techniques, such as Textual Inversion [24], to encode identity information in the text prompt through special embeddings. Alternatively, utilizing NeRF-relevant techniques [9, 49, 72] can generate other views of the object for a specific scene, but this can require expensive training. Furthermore, due to the fact that our approach relies on Stable Diffusion, it inherits its shortcomings and biases, which may result in producing artifacts in certain scenarios.

7. Conclusion

We introduced a method that leverages the high-order ODE solver with the exceptional prompt to achieve precise inversion of real images, which serves as a foundation for further manipulation. Building upon this, we propose a novel training-free framework, TF-ICON, that enables attention-based text-to-image diffusion models to perform image-guided composition across different domains. Our experimental results demonstrate that our approach outperforms the SOTA baselines for both image inversion and composition. We believe that image composition has the potential to become an essential tool for content creators, offering significant benefits for downstream applications in various industries.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3, 6

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 3

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 3, 6

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 18511–18521, 2022. 3

[5] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 2, 7, 8

[6] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 6

[7] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10):2570–2585, 2020. 2

[8] Andrew Brown, Cheng-Yang Fu, Omkar Parkhi, Tamara L Berg, and Andrea Vedaldi. End-to-end visual editing with a generatively pre-trained artist. *arXiv preprint arXiv:2205.01668*, 2022. 2

[9] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Consistent single-view perpetual view generation with conditional diffusion models. *arXiv preprint arXiv:2211.12131*, 2022. 8

[10] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[11] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023. 2

[12] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2019. 2

[13] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 6

[14] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020. 2

[15] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2, 4

[16] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 4

[19] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2

[20] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 2

[21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 6

[22] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 2

[23] Oran Gafni and Lior Wolf. Wish you were here: Context-aware human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7840–7849, 2020. 2

[24] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 7, 8

[25] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder

for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2

[26] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 4

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 4

[29] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 914–922, 2022. 2

[30] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021. 2

[31] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[32] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 4

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[34] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[35] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2, 4, 6

[36] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

[37] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 2

[38] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2, 4, 6

[39] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. 2023. 2

[40] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 2

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[42] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8139–8148, 2020. 2

[43] Liu Liu, Zhenchen Liu, Bo Zhang, Jiangtong Li, Li Niu, Qingyang Liu, and Liqing Zhang. Opa: object placement assessment dataset. *arXiv preprint arXiv:2107.01889*, 2021. 2

[44] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 4

[45] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 2

[46] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 4

[47] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 4, 5, 6

[48] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 7, 8

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 8

[50] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 3, 4

[51] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[52] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021. 2

[53] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 2, 4

[54] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[57] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 3, 6

[58] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 3, 6

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4

[60] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[62] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 4

[63] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4380–4390, 2021. 2

[64] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 4

[66] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3, 4

[67] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing. *arXiv preprint arXiv:2212.00932*, 2022. 2

[68] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3, 4, 6

[69] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 461–470, 2019. 2

[70] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 2, 4

[71] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022. 2, 3, 4

[72] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 8

[73] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 6

[74] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022. 2

[75] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019. 2

[76] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *European Conference on Computer Vision*, pages 300–316. Springer, 2022. 2, 7, 8

[77] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 2, 7, 8

[78] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[79] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep image compositing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 365–374, 2021. 2

[80] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *European Conference on Computer Vision*, pages 566–581. Springer, 2020. 2

[81] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 231–240, 2020. 2, 7

[82] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6, 7

[83] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5(1):105–115, 2019. 2

[84] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 5

Appendix

A. Additional Analysis

A.1. Better Latent Code with ODE Solvers

In Section 4.1, we argue that the utilization of diffusion ODE solvers [6, 10, 11, 12] as encoders, commencing from the real image $\mathbf{x}_0$, results in better latent representation $\mathbf{x}_T$ as compared to those acquired via the commonly used DDIM [17]. This improvement is attributed to the better alignment between the forward and backward ODE trajectories produced by higher-order ODE solvers.

This claim is supported by the experimental results presented in Figure 8. Specifically, we performed image inversion on the COCO2017 [9] validation set of 5000 images using DDIM, as well as the second and third order DPM-Solver++ [12]. This involved encoding real images into noises and subsequently decoding the noises back into real images, with both procedures consisting of 50 steps. The forward and backward intermediate states were preserved as $\{\mathbf{x}_0^{\text{enc}}, \mathbf{x}_1^{\text{enc}}, \cdots, \mathbf{x}_{50}^{\text{enc}}\}$ and $\{\mathbf{x}_0^{\text{dec}}, \mathbf{x}_1^{\text{dec}}, \cdots, \mathbf{x}_{50}^{\text{dec}}\}$, respectively. L1 and L2 distances between the forward and backward processes' intermediates were computed at each time step. Figure 8 presents the average values of the distances.

The experimental results demonstrate that the higher-order DPM-Solver++ exhibits a smaller difference between the forward and backward intermediates, signifying better alignment between the forward and backward trajectories, in comparison to DDIM, which is equivalent to a first-order solver. Furthermore, the experimental results suggest that an increase in the order of the ODE solver does not lead to additional improvement in alignment.

Figure 9 presents a visual comparison between image compositions achieved through the utilization of high-order DPM solvers and DDIM inversion. Due to subpar alignment between forward and backward intermediates, The inversion codes of DDIM yield blurred images when using the same sampling steps (20 steps).

Given that Lu *et al*. [12] has established the suitability of the second-order DPM-Solver++ for guided sampling[1], we employed the second-order one for all the experiments conducted with TF-ICON.

A.2. Exceptional Prompt Analysis

Denote the image features as $\mathbf{f} \in \mathbb{R}^{s \times d_1}$ and the embedding of the exceptional prompt as $\mathbf{T} \in \mathbb{R}^{l \times d_2}$, where d_1, d_2 denote the dim of the image and text embeddings, $s = h \times w$ is the res of the latent space, and l is the maximum length of the prompt. By assigning the same value to all tokens and discarding the positional embeddings, each row of $\mathbf{T}$ is identical. In a cross-attention module, we have

[1] https://github.com/LuChengTHU/dpm-solver

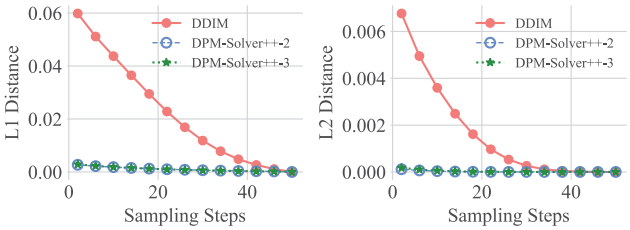


Figure 8: The comparison of the alignment of forward and backward trajectories from DDIM inversion and high-order DPM-Solver++. L1 and L2 distances were computed at each time step between the forward and backward intermediates, and then averaged over 5000 images. The curves representing the second and third order DPM-Solver++ are almost overlapping. Please zoom in for a closer look.

$\mathbf{W}^q \in \mathbb{R}^{d_1 \times d_1}, \mathbf{W}^k, \mathbf{W}^v \in \mathbb{R}^{d_2 \times d_1}$ and $\mathbf{q} = \mathbf{f} \cdot \mathbf{W}^q, \mathbf{k} = \mathbf{T} \cdot \mathbf{W}^k, \mathbf{v} = \mathbf{T} \cdot \mathbf{W}^v$. When a matrix with identical rows multiplies another matrix, the resultant matrix also exhibits identical rows. Thus, $\mathbf{k}, \mathbf{v}$ have identical rows, and $\mathbf{q} \cdot \mathbf{k}^\mathsf{T}$ has identical columns. Applying the softmax row-wise to $\mathbf{q} \cdot \mathbf{k}^\mathsf{T}$ generates a constant attention map $\mathbf{A} = \frac{1}{l} \cdot \mathbf{1}_{s \times l}$. The output $\mathbf{o} = \mathbf{A} \cdot \mathbf{v}$ hence exhibits identical rows and is then added to the input, *i.e*., $\mathbf{f} + \mathbf{o}$, before moving to the next layer. Each row of $\mathbf{f}_{s \times d_1}$ is the embedding of each patch. In the exceptional prompt, all patch embeddings experience a consistent directional movement, but normal and null prompts with varying row vectors cause embeddings to move in various directions, thereby disrupting the image pattern.

A.3. Elaboration of Inversion Results

Two specific points in Figure 3 require attention. Firstly, it is true that CFG typically amplifies instability, resulting in subpar metrics (Tables 1 and 2), while satisfactory reconstruction from CFG output is possible, albeit less common, even with only DDIM (Figure 10 in [4] and 3rd row of Figure 3). Secondly, the unconditional output does not necessarily outperform CFG or conditional one, as the unconditional/null prompt contains special symbols (Figure 4), which also add information and lead to inconsistent directional shifts in image embeddings (See Section A.2). Thus, the unconditional output may perform poorly than others (4th and 6th row of Figure 17). Figure 3 shows unconditional (1st row), conditional (2nd row), or CFG (3rd row) output can yield the best reconstruction among them.

A.4. Token Value Analysis

In Section 4.1, we contend that the choice of token value has no significant impact on the inversion performance. To justify this, we uniformly sampled 100 token values from the set of 49407 values and employed them as the common token value in the exceptional prompt $\mathcal{P}_{\text{exceptional}}$. All experimental results are obtained using Stable Diffusion [14] with

Figure 9: The visual comparison between image compositions achieved through the utilization of high-order DPM solvers++ and DDIM inversion. The image compositions resulting from DDIM inversion exhibit more blurring when compared to those generated by high-order DPM solvers++ employing the same 20-step sampling process. Augmenting the solver's order does not result in noteworthy visual enhancements.

the exceptional prompt (100 different token values), sampled through the second-order DPM-Solver++ in 50 steps. Three metrics, namely MAE, LPIPS, and SSIM, were used to assess the inversion performance.

The experimental results for four randomly sampled images from the COCO2017 validation set are shown in Figure 10. The top row of Figure 10 displays a magnified view of a specific area from the second row. Notably, for a single image, each token value produces nearly identical inversion performance, with only minor fluctuations occurring within a narrow range.

Table 6: Means and standard deviations of metrics among the reconstruction results of 100 tokens, averaged over 150 images randomly sampled from the COCO.

	MAE	LPIPS	SSIM
Mean	0.0323	0.0703	0.8560
Standard Deviation	9.22×10^{-5}	9.29×10^{-4}	3.81×10^{-4}

Furthermore, we randomly sampled 150 images from the COCO2017 validation set. For each image, we calculated the means and standard deviations of the three metrics among the reconstruction results of the 100 tokens. The metrics were averaged over 150 images, as listed in Table 6. Importantly, the average standard deviations of all metrics for the reconstructions of different tokens are remarkably low, indicating that the selection of token values does not significantly affect the performance of inversion.

B. Implementation Details

B.1. Preprocessing

Figure 11 illustrates the preprocessing process. Typically, only the foreground in the reference image is desired for composition, so a pretrained segmentation model [21] is utilized to segment the object from the background. Next, the extracted object is resized and repositioned to correspond with the user's mask in the main image. Finally, zero padding is applied to the object to ensure it is the same size as the main image.

B.2. Algorithm and Running Time

Algorithm 1 describes the pseudocode of the proposed training-free image composition framework (TF-ICON). The synthesis time for a single image using one A100 GPU card is around 8 seconds, depending on the size of the user mask and reference image.

B.3. Background Preservation

As discussed in Sections **??** and **??**, preserving the background during denoising should be done gradually at different levels of noise. Preserving the background only at the final time step may result in noticeable artifacts. Figure 12 provides a comparison between the naïve implementation, which preserves the background only at the final step, and our implementation, which follows a gradual way. The naïve implementation results in obvious artifacts, while ours successfully produces high-quality results.

The rationale behind this phenomenon is that when two noisy images are blended at a certain noise level, the resulting image may lie outside the targeted manifold. The subsequent steps of diffusion can rectify this issue by moving

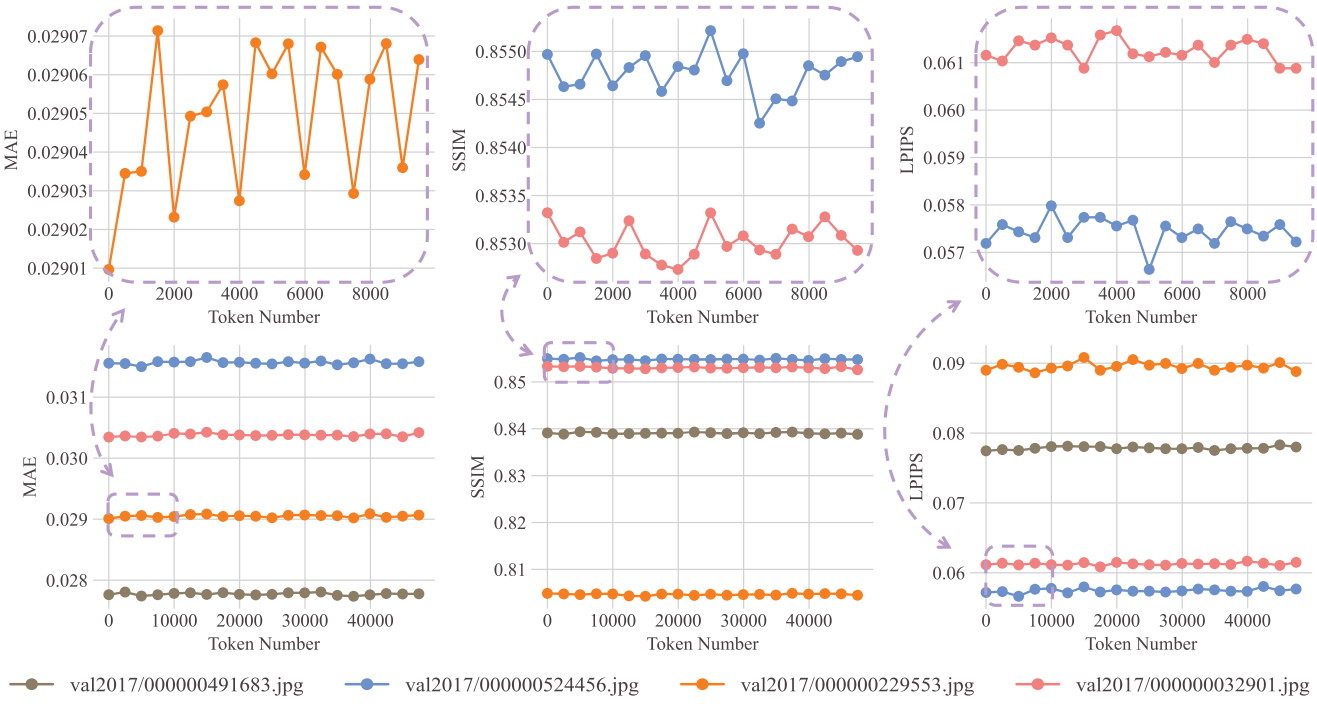


Figure 10: The analysis of the impact of the common token values in the exceptional prompt. The first row displays a magnified view of an area from the second row. For each image randomly sampled from the COCO, the exceptional prompt is applied with 100 uniformly sampled token values on Stable Diffusion to perform image inversion. The inversion metrics, including MAE, SSIM, and LPIPS, exhibit negligible variations as the token value is modified.

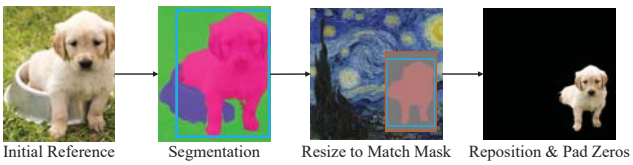


Figure 11: The preprocessing pipeline for the reference image. (1) The centralized reference image is initially processed by a pretrained segmentation model; (2) the segmented object region is then extracted, and its dimension is adjusted to match the size of the user mask; (3) the resized image is finally repositioned and padded with zeroes to match the main image's dimension.

it toward the next level manifold, thereby gradually improving the coherence of the image. However, if the blending is only performed at the final step in a simplistic manner, the image cannot be corrected any further.

B.4. Experimental Settings and Hyperparameters

Image Reconstruction. To conduct inversion experiments on the CelebA-HQ [5] (*i.e.*, Table 1), we followed the experimental settings outlined in [7, 18]. The first 1500 images from the CelebA-HQ were inverted, and the quality of

Figure 12: The comparison between two implementations of background preservation. Naïve implementation only preserves the background at the final step, while ours gradually blends background information at various time steps.

reconstruction from the inverted latent was evaluated using MAE, LPIPS, and SSIM metrics. All Stable Diffusion re-

Algorithm 1 Training-Free Image Composition

Input: The embeddings of the normal prompt and the exceptional prompt $\mathcal{E} = \psi(\mathcal{P})$ and $\mathcal{W} = \psi(\mathcal{P}_{\text{exceptional}})$, the main image $\mathbf{I}^{\text{m}}$, the reference image $\mathbf{I}^{\text{r}}$, the user mask $\mathbf{M}^{\text{user}}$, the segmentation mask $\mathbf{M}^{\text{seg}}$, thresholds τ_A, τ_B
Output: The composition result $\mathbf{I}^*$

// Step 1: Starting Point Incorporation
$\mathbf{x}_0^{\text{m}} = \text{VQ-Encoder}\,(\mathbf{I}^{\text{m}})$; $\mathbf{x}_0^{\text{r}} = \text{VQ-Encoder}\,(\mathbf{I}^{\text{r}})$
for $t = 1, \ldots, T$ **do**
  $\mathbf{x}_t^{\text{m}} \leftarrow \text{DPM-Solver++}\,(\mathbf{x}_{t-1}^{\text{m}}, t-1, \mathcal{W})$
  $\mathbf{x}_t^{\text{r}} \leftarrow \text{DPM-Solver++}\,(\mathbf{x}_{t-1}^{\text{r}}, t-1, \mathcal{W})$
end for
$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{x}_T^* \leftarrow \mathbf{x}_T^{\text{r}} \odot \mathbf{M}^{\text{user}} + \mathbf{x}_T^{\text{m}} \odot (\mathbf{1} - \mathbf{M}^{\text{user}}) + \mathbf{z} \odot (\mathbf{M}^{\text{user}} \oplus \mathbf{M}^{\text{seg}})$
// Step 2: Image Composition
for $t = T, \ldots, 1$ **do**
  $\mathbf{x}_{t-1}^{\text{m}}, \{\mathbf{A}_t^{\text{m}}\} \leftarrow \text{DPM-Solver++}\,(\mathbf{x}_t^{\text{m}}, t, \mathcal{W})$
  $\mathbf{x}_{t-1}^{\text{r}}, \{\mathbf{A}_t^{\text{r}}\} \leftarrow \text{DPM-Solver++}\,(\mathbf{x}_t^{\text{r}}, t, \mathcal{W})$
  $\{\mathbf{A}_t^{\text{cross}}\} \leftarrow \text{CrossAtten}(\mathbf{x}_t^{\text{m}}, \mathbf{x}_t^{\text{r}})$
  $\{\mathbf{A}_t^*\} \leftarrow \vartheta_{\text{compose}}\,(\{\mathbf{A}_t^{\text{m}}\}, \{\mathbf{A}_t^{\text{r}}\}, \{\mathbf{A}_t^{\text{cross}}\})$
  if $t > \text{int}(\tau_A \times T)$ **then**
    $\mathbf{x}_{t-1}^* \leftarrow \text{DPM-Solver++}\,(\mathbf{x}_t^*, t, \mathcal{E}, \{\mathbf{A}_t^*\})$
  else
    $\mathbf{x}_{t-1}^* \leftarrow \text{DPM-Solver++}\,(\mathbf{x}_t^*, t, \mathcal{E})$
  end if
  if $t > \text{int}(\tau_B \times T)$ **then**
    $\mathbf{x}_{t-1}^* \leftarrow \mathbf{x}_{t-1}^* \odot \mathbf{M}^{\text{user}} + \mathbf{x}_{t-1}^{\text{m}} \odot (\mathbf{1} - \mathbf{M}^{\text{user}})$
  end if
end for
$\mathbf{I}^* = \text{VQ-Decoder}(\mathbf{x}_0^*)$
return $\mathbf{I}^*$

sults were sampled in 50 steps using the second-order DPM-Solver++. The normal prompt for the conditional output and the output with CFG was set as *'a portrait photo'*. The CFG scale was 5. The common token value of the exceptional prompt was 7788.

In further experiments on the COCO2017 (*i.e.*, Table 2), the entire validation set with 5000 images was used. The first listed caption of each image in the annotations serves as the normal prompt. In the experiments on the ImageNet [2] (*i.e.*, Table 2), 3000 images were randomly sampled from the ImageNet validation set. *'a photo of the [class]'* was used as the normal prompt. For both datasets, the CFG scale was set at 5, and the common token value of 7788 was used in the exceptional prompt.

Image Composition. Since most baselines are trained only in the photorealism domain, where objective metrics are more effective, we conducted our quantitative comparison in this domain. However, for other domains, we relied on user study and qualitative comparisons. For quantitative comparison in the photorealism domain, we used the official implementation of Deep Image Blending (DIB)[2] [20], Blended Diffusion[3] [1], Paint by Example[4] [19], and SDEdit[5] [13]. Our framework utilizes Stable Diffusion[6] with the second-order DPM-Solver++ to solve all three ODEs in 20 steps. The first two inversion ODEs, aimed at obtaining accurate inverted noises and self-attention maps, were performed under the exceptional prompt with a common token value of 7788, while the last ODE utilized the normal prompt with a CFG scale of 2.5. The threshold values τ_A and τ_B were set at 0.4 and 0, respectively.

C. Ablation of Value Injection

We conducted an additional ablation study in which we not only injected the attention maps but also included the values information. Specifically, we multiply the attention maps with the corresponding values for both the main and reference images, and then compose and inject them. The metrics obtained on the dataset are as follows: $\text{LPIPS}_{(\text{BG})} = 0.10$, $\text{LPIPS}_{(\text{FG})} = 0.63$, $\text{CLIP}_{(\text{Image})} = 81.37$, $\text{CLIP}_{(\text{Text})} = 27.68$. These metrics are lower compared to injecting only the attention maps.

The rationale behind this is that injecting all the information might result in a more rigid generation, potentially hindering the ability to transition across visual domains due to the direct replacement of all information from the guiding images. On the other hand, by injecting self-attention maps only, we are able to preserve the semantic layouts while incorporating values derived from the inherent composition features. The visual comparison is shown in Figure 13.

D. User Study

To compare image composition baselines across various domains, we conducted a user study by recruiting 50 participants from Amazon. The participants were asked to complete 40 ranking questions, with each question comprising a foreground image, a background image with a bounding box to indicate the region of interest, and a text prompt. For each question, the participants were presented with five images generated using different methods. They were requested to rank five images from 1 to 5 (1 being the best and 5 being the worst) based on comprehensive criteria:

1. **Text Alignment:** The resulting image should match the specific style mentioned in the text prompt. For example, if the target domain is cartoon, oil painting, pencil drawing, or photorealism, the generated image should align with that style.

[2] https://github.com/owenzlz/DeepImageBlending
[3] https://github.com/omriav/blended-latent-diffusion
[4] https://github.com/Fantasy-Studio/Paint-by-Example
[5] https://github.com/ermongroup/SDEdit
[6] https://github.com/Stability-AI/stablediffusion

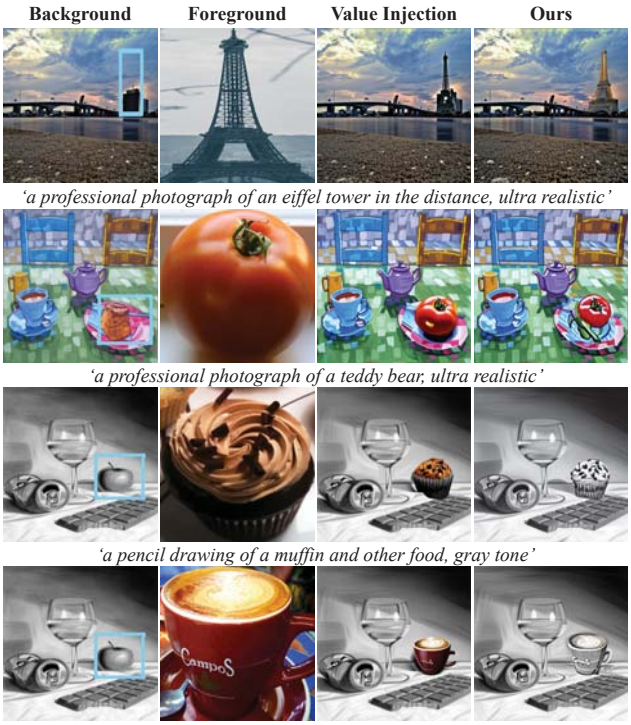
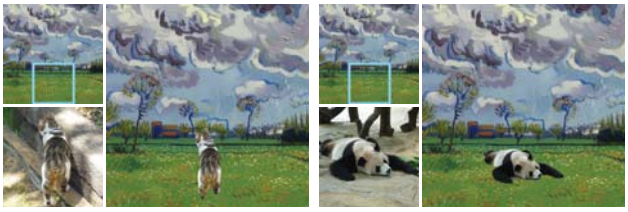


Figure 13: The visual comparison between injecting all information and our implementation. Injecting values leads to a more rigid generation, potentially impeding the smooth transition across visual domains. This impact becomes particularly evident when transferring to the sketchy domain.

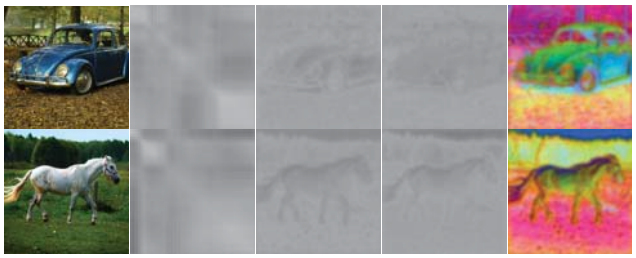
2. **Foreground Preservation:** The generated image should well-preserve the features or identity of the given object within the mask region, such that the viewers can recognize that the given and the generated objects are the same even in different domains.
3. **Background Preservation:** The background outside the mask should remain unchanged.
4. **Seamless Composition:** The resulting image should be of high quality and free from any apparent artifacts that might indicate it was generated by AI or copied and pasted.

To ensure all 40 questions are meaningful, we filtered out simple questions that, without any domain or illumination adjustment, only require copy-pasting operations to make the composition look natural despite the foreground and background being from different domains. We show examples of such cases in Figure 14. After the filtering process, we randomly sampled questions from the test benchmark. In addition to the regular ranking questions, we also included three attention-checking questions to filter out random or invalid responses. The final valid questions consisted of 20 photorealism, 7 oil painting, 7 pencil sketching, and 6 cartoon animation questions.



(a) Example 1 (b) Example 2

Figure 14: Examples of meaningless questions. The resulting images were generated by simply segmenting objects from the reference image and pasting them onto the region of interest in the background image without modifications. Despite the lack of any modification, the results appear almost seamless.



(a) Original (b) SA (c) Row (d) Column (e) PCA

Figure 15: The visualization of (a) original image; (b) self-attention (SA) maps $\in \mathbb{R}^{4096\times4096}$ of (a); (c) the averaging result of unfolding all rows $\in \mathbb{R}^{1\times4096}$ of (b) into $\mathbb{R}^{64\times64}$; (d) same operation as (c) for columns; (e) visualizing top-3 PCA components of (b).

The ranking score of the options in each question is calculated by:

$$\text{score} = \frac{1}{n} \cdot \sum_{i=1}^{5} v_i \cdot w_i \tag{12}$$

where v_i denotes the number of votes for the option to rank i, w_i indicates the weight of rank i, and n is the number of respondents. The first rank has the highest weight of 5 and the last rank has the lowest weight of 1. The resulting score reflects the overall ranking of the options, with a higher score indicating a better ranking.

E. Self-Attention Visualization

Figure 15 demonstrates how self-attention maps preserve semantic information. By unfolding the rows or columns of the self-attention maps, we can discern the underlying semantics of the image.

F. Elaboration for Toy Example

This section further analyzes the attention composition in Figure **??**. The self-attention maps of the blue region in Figure **??** (a) $\mathbf{A}^{\mathrm{r}}_{l,t} \in \mathbb{R}^{4\times4}$ are partitioned into four blocks based on the patch indices and composed into the blue regions in $\mathbf{A}^{\mathrm{m}}_{l,t} \in \mathbb{R}^{16\times16}$, as illustrated in Figure **??** (b). The dimension of $\mathbf{A}^{\mathrm{cross}}_{l,t} \in \mathbb{R}^{16\times4}$ is identical to that of the green regions, with the exception of the interactions between white patches indexed at 5, 6, 9, and 10, and blue patches with corresponding indices. Since the aim of the attention composition is to infuse contextual information from the white region into the blue region, the information from the white patches indexed at 5, 6, 9, and 10 is irrelevant and can be disregarded.

G. Test Benchmark

To facilitate evaluating cross-domain image-guided composition as a unified task, we have created a comprehensive test benchmark comprising 332 samples. Each sample in the benchmark comprises a main (background) image, a reference (foreground) image, a user mask, and a text prompt. Images were collected from Open Images [8], PASCAL VOC [3], COCO [9], Unsplash[7], and Pinterest[8]. The main images comprise four visual domains: photorealism, pencil sketching, oil painting, and cartoon animation. All reference images are from the photorealism domain, as the reference requires segmentation models, which are generally more effective in this domain. The selection objective is to ensure that the main image and reference image share similar semantics, thereby guaranteeing a reasonable combination. The text prompt is manually labeled according to the semantics of the main and reference images.

The reference images comprise a wide range of object classes, including 'Car', 'Panda', 'Dog', 'Elephant', 'Fox', 'Castle', 'Buddha', 'Bird', 'Sheep', 'Fire Hydrant', 'Mailbox', 'Hamburger', 'Chicken', 'Skyscraper', 'Rocket', 'Chair', 'Cabinet', 'Bag', 'Teddy Bear', 'Mall', 'Tower', 'Building', 'Flower', 'Tortoise', 'Sparrow', 'Ostrich', 'Horse', 'Cat', 'Goose', 'Tiger', 'Eagle', 'Squirrel', 'Raccoon', 'Penguin', 'Sea Lion', 'Goat', 'Owl', 'Microwave', 'Bread', 'Cake', 'Tomato', 'Fish', 'Croissant', 'Hot Dog', 'Waffle', 'Pancake', 'Popcorn', 'Burrito', 'Muffin', 'Juice', 'Coffee', 'Paper Towel', 'Tart', 'Sandwich', 'Teapot', 'Lemon', 'Candle', 'Spoon', 'Grapefruit', 'Turkey', 'Pomegranate', 'Doughnut', 'Cantaloupe', 'Sandwich', 'Cantaloupe', and 'Turkey'. Given that most image composition baselines are trained exclusively on photorealistic images, our test benchmark contains a greater proportion of photorealism samples to enable a quantitative comparison. Specifically, the benchmark includes 237 photorealism samples, as well as 37 oil painting, 31 pencil sketching, and 27 cartoon animation samples. The benchmark will be publicly available for use in evaluating the performance of cross-domain image-guided composition methods.

[7] https://unsplash.com/

[8] https://www.pinterest.com/

H. Additional Qualitative Results

H.1. Image Reconstruction

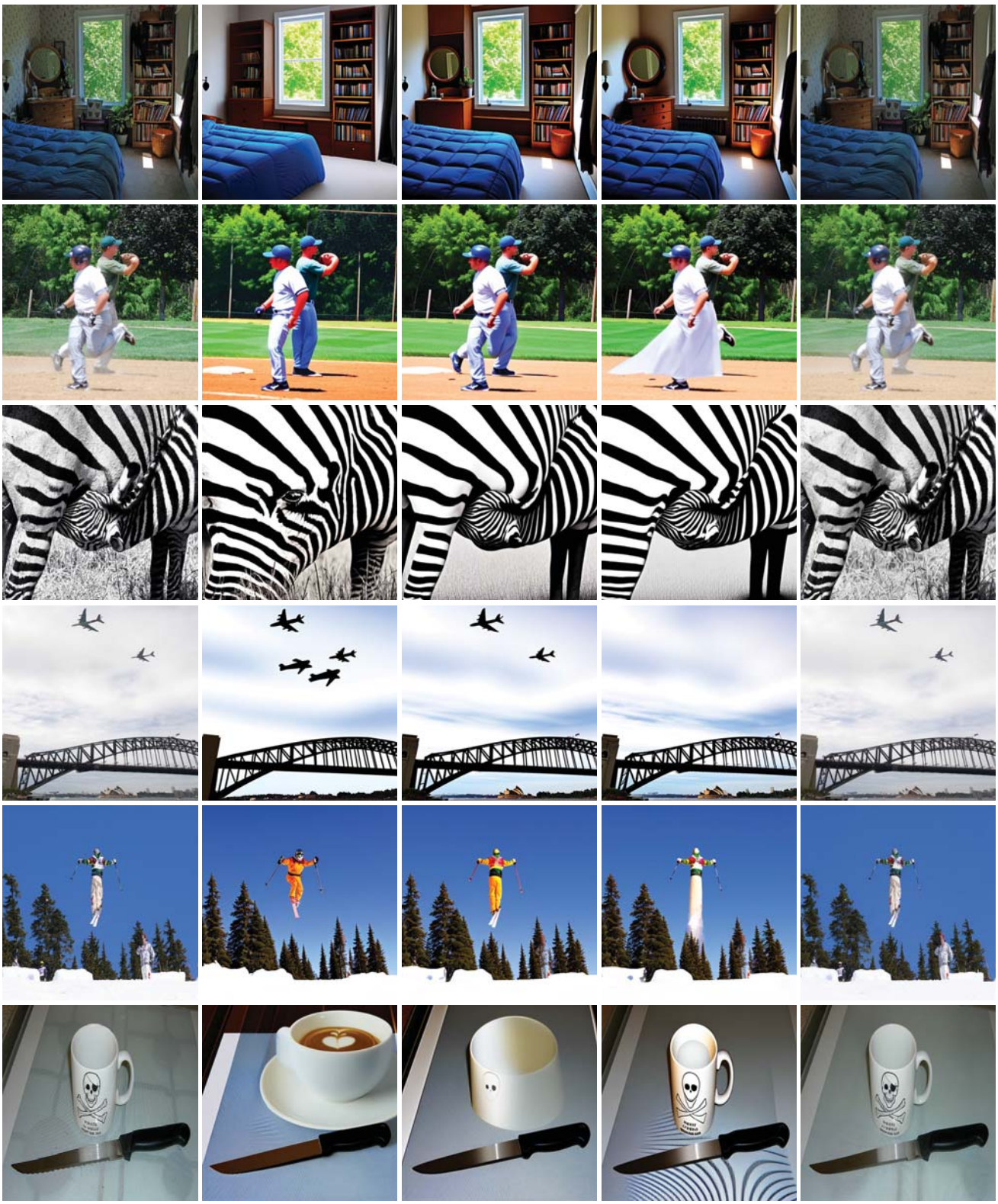
Figures 16, 17, and 18 present additional qualitative image reconstruction comparisons among different outputs of Stable Diffusion on COCO, ImageNet, and CelebA-HQ, respectively.

H.2. Image Composition

Figure 19 presents additional ablation study results. Further qualitative comparisons of image composition across various domains are exhibited in Figures 20, 21, 22, 23, 24, and 25.

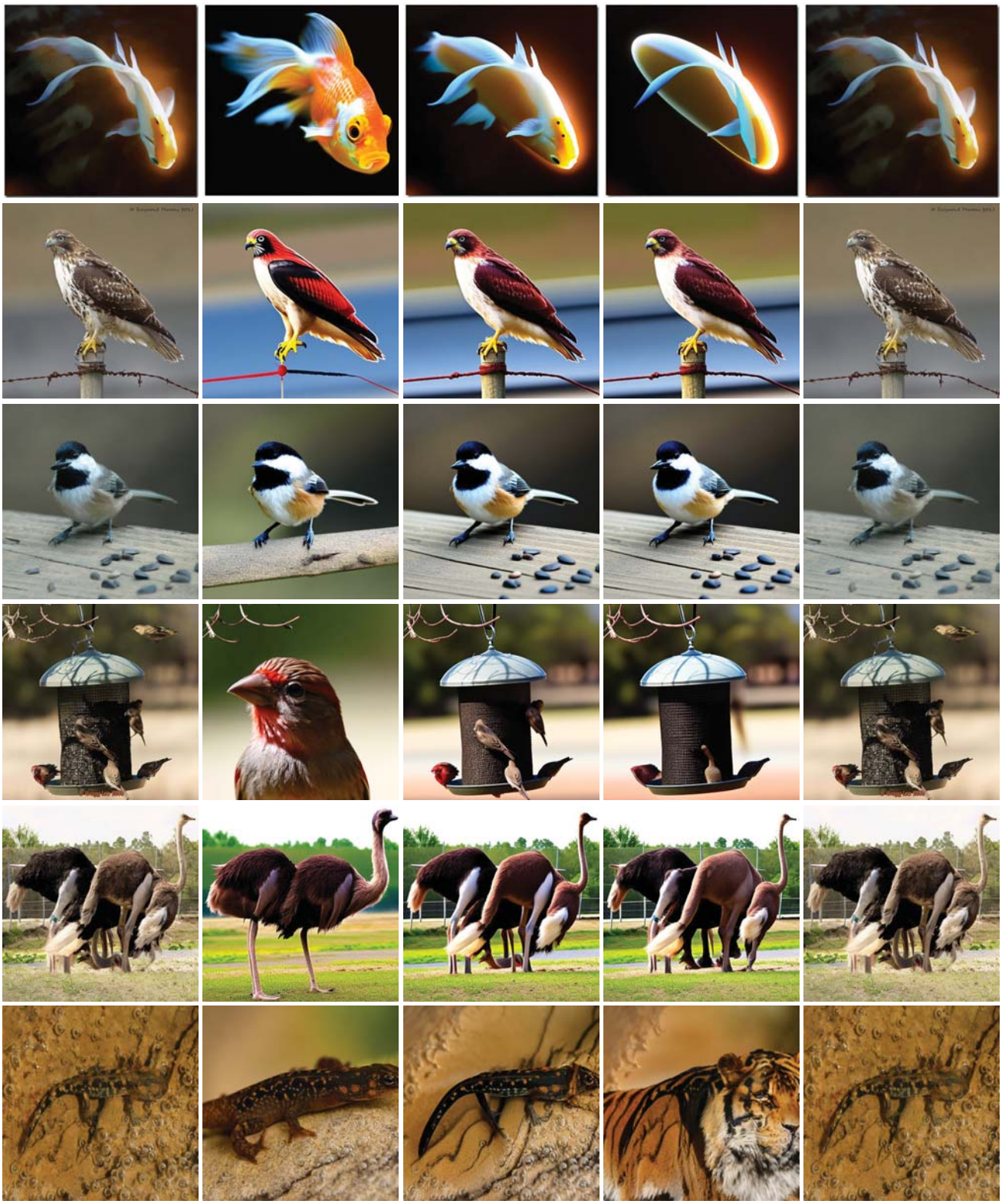
I. Societal Impacts

TF-ICON offers a means of image-guided composition that empowers individuals without professional artistic skills to create compositions. While this technology is beneficial, it can also be misused for malicious purposes, such as in cases of harassment or spreading fake news. Moreover, image composition is closely related to image generation, so it is essential to recognize that using diffusion models trained on web-scraped data, such as LAION [16], can potentially introduce biases. Specifically, LAION has been found to contain inappropriate content such as violence, hate, and pornography, as well as racial and gender stereotypes. Consequently, diffusion models trained on LAION, such as Stable Diffusion and Imagen [15], are prone to exhibit social and cultural biases. As such, using such models raises ethical concerns and should be approached with care. Finally, the capacity to compose across artistic domains has the potential to be exploited for copyright infringement purposes, as users could generate images in a similar style without the consent of the artist. Although the resulting generated artwork may be readily distinguishable from the original, future technological advances could render such infringement challenging to identify or legally prosecute. Thus, we encourage users to use this method cautiously and only for appropriate purposes.



(a) Original (b) CFG Output (c) Conditional Output (d) Unconditional Output (e) Ours

Figure 16: Comparison of image reconstruction results on the COCO using Stable Diffusion with (b) classifier-free guidance (CFG) output $\hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{E}, \varnothing)$, (c) conditional output $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{E})$, (d) unconditional output $\epsilon_\theta(\mathbf{x}_t, t, \varnothing)$, and (e) ours $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{W})$.



(a) Original (b) CFG Output (c) Conditional Output (d) Unconditional Output (e) Ours

Figure 17: Comparison of image reconstruction results on the ImageNet using Stable Diffusion with (b) classifier-free guidance (CFG) output $\hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{E}, \varnothing)$, (c) conditional output $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{E})$, (d) unconditional output $\epsilon_\theta(\mathbf{x}_t, t, \varnothing)$, and (e) ours $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{W})$.

(a) Original (b) CFG Output (c) Conditional Output (d) Unconditional Output (e) Ours

Figure 18: Comparison of image reconstruction results on the CelebA-HQ using Stable Diffusion with (b) classifier-free guidance (CFG) output $\hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{E}, \varnothing)$, (c) conditional output $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{E})$, (d) unconditional output $\epsilon_\theta(\mathbf{x}_t, t, \varnothing)$, and (e) ours $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{W})$.

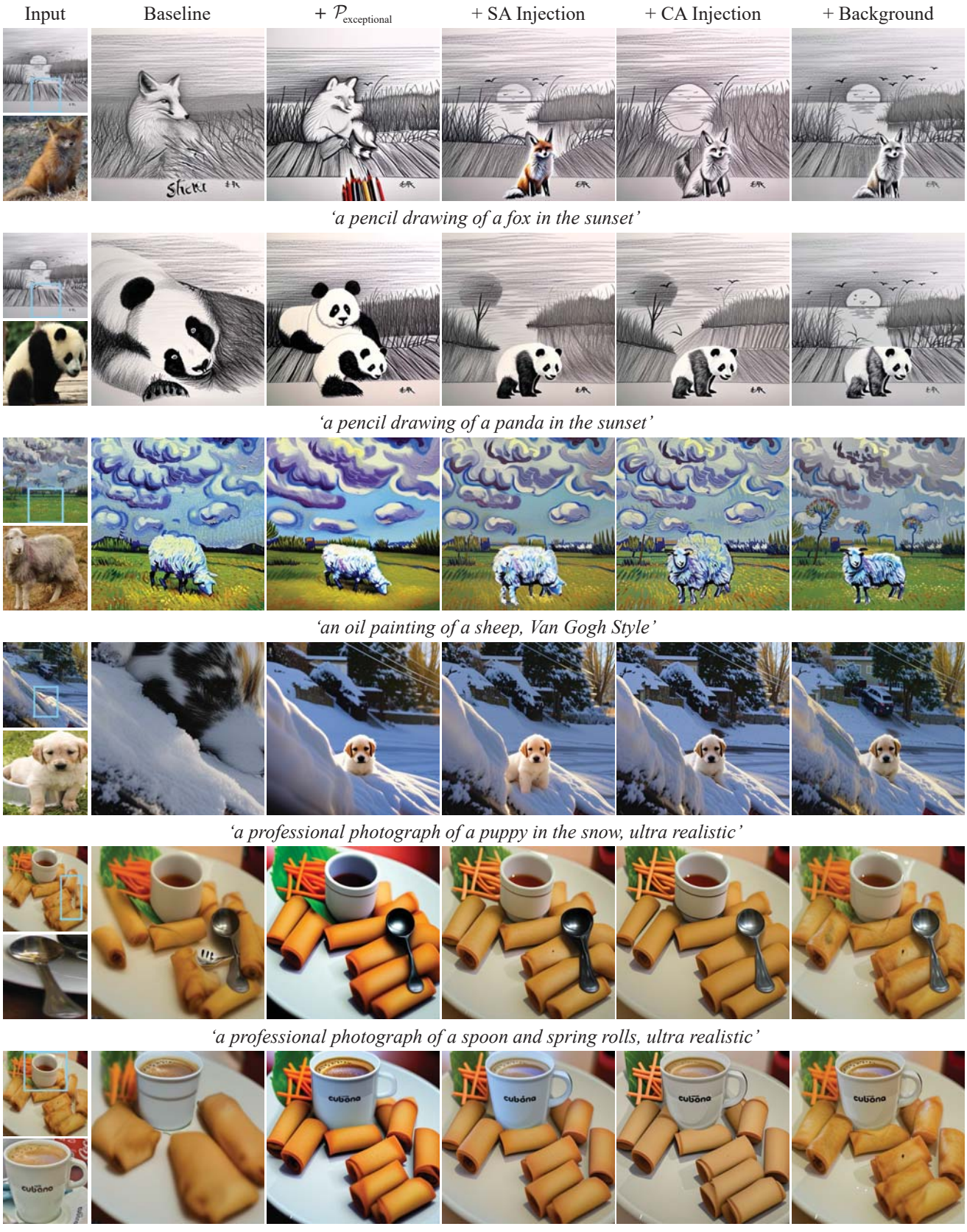


Figure 19: Ablation study of different variants of our framework. SA: self-attention. CA: cross-attention.

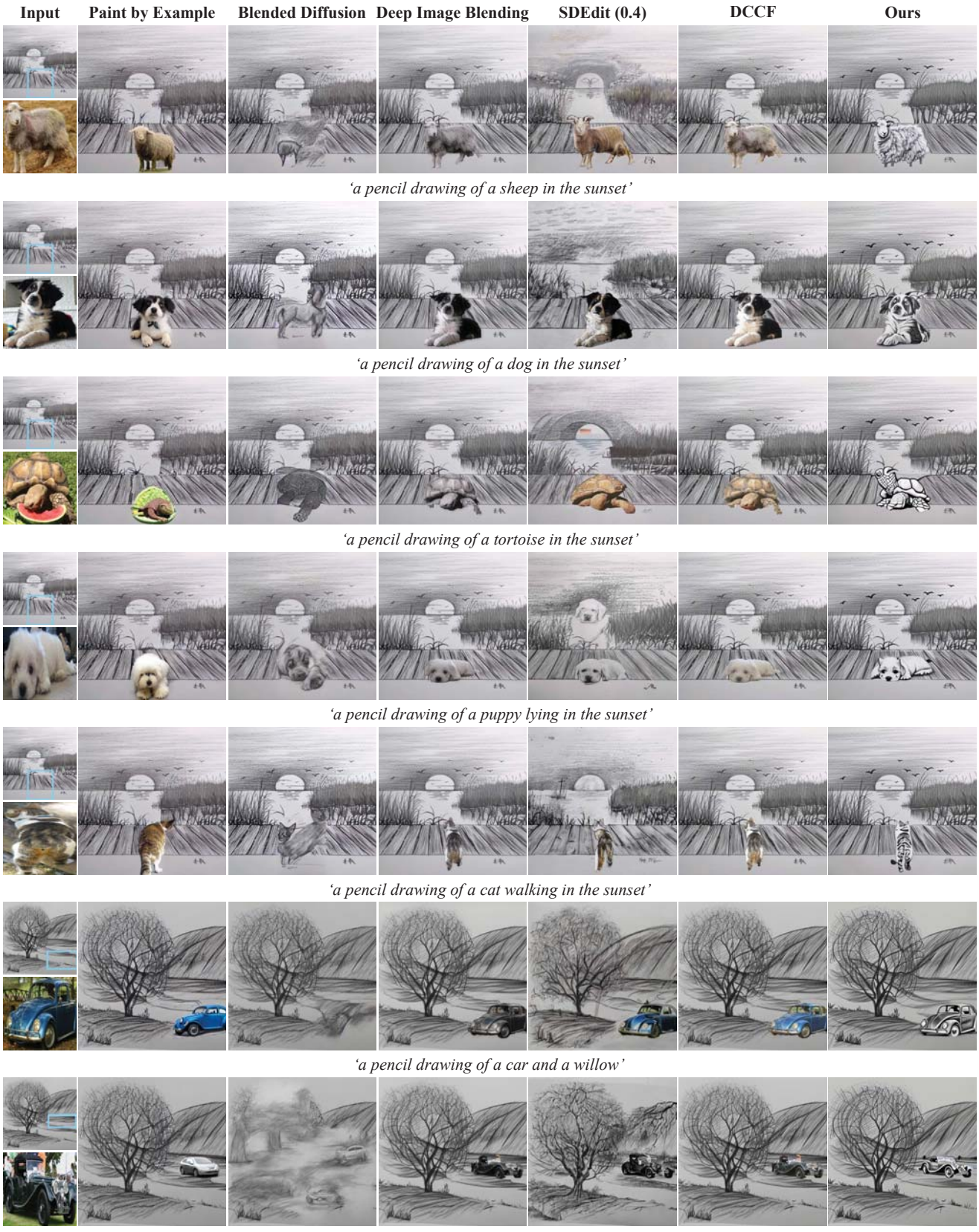


Figure 20: Qualitative comparison with SOTA baselines in image composition for the pencil sketching domain.

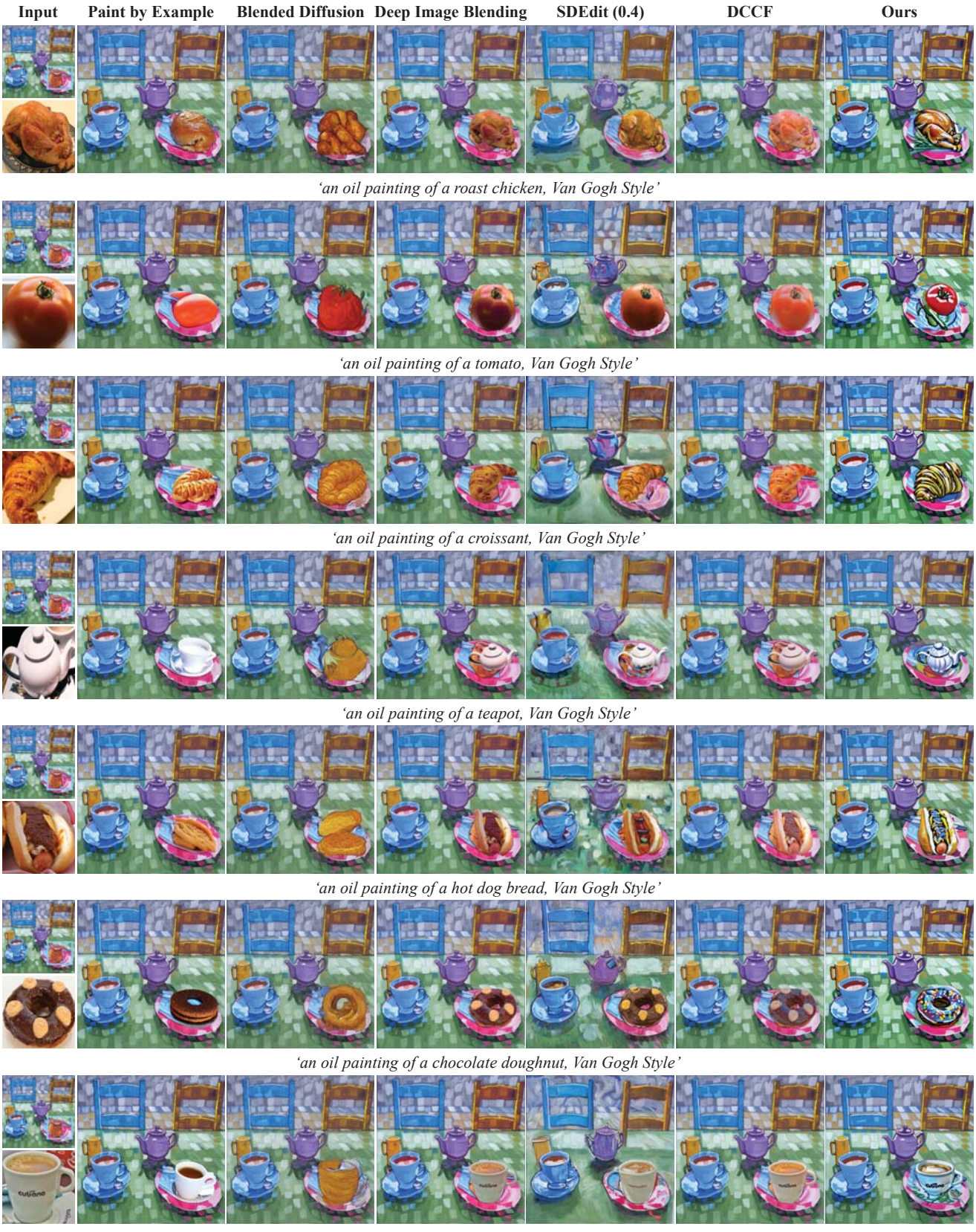


Figure 21: Qualitative comparison with SOTA baselines in image composition for the oil painting domain.

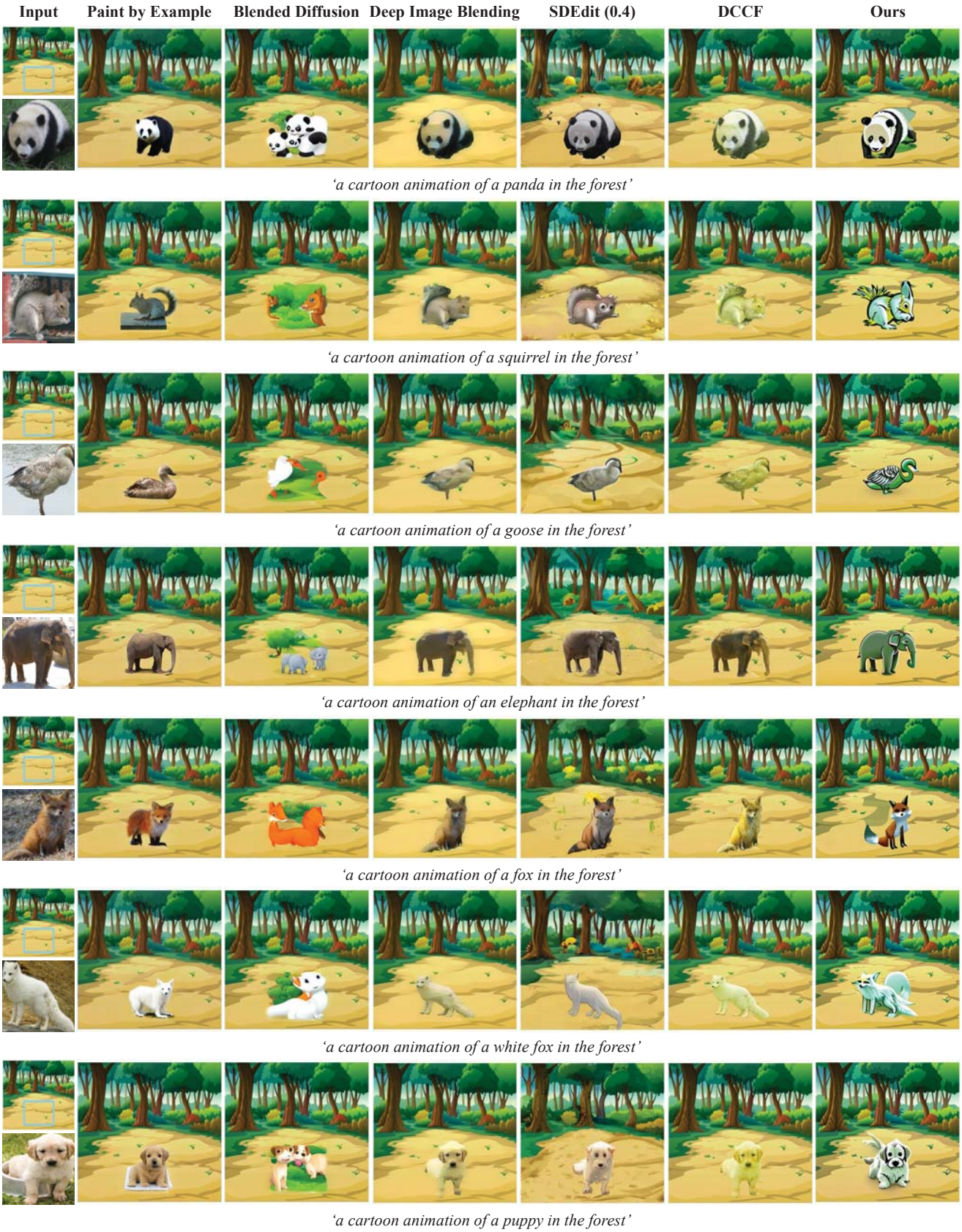


Figure 22: Qualitative comparison with SOTA baselines in image composition for the cartoon animation domain.

Figure 23: Qualitative comparison with SOTA baselines in image composition for the photorealism domain.

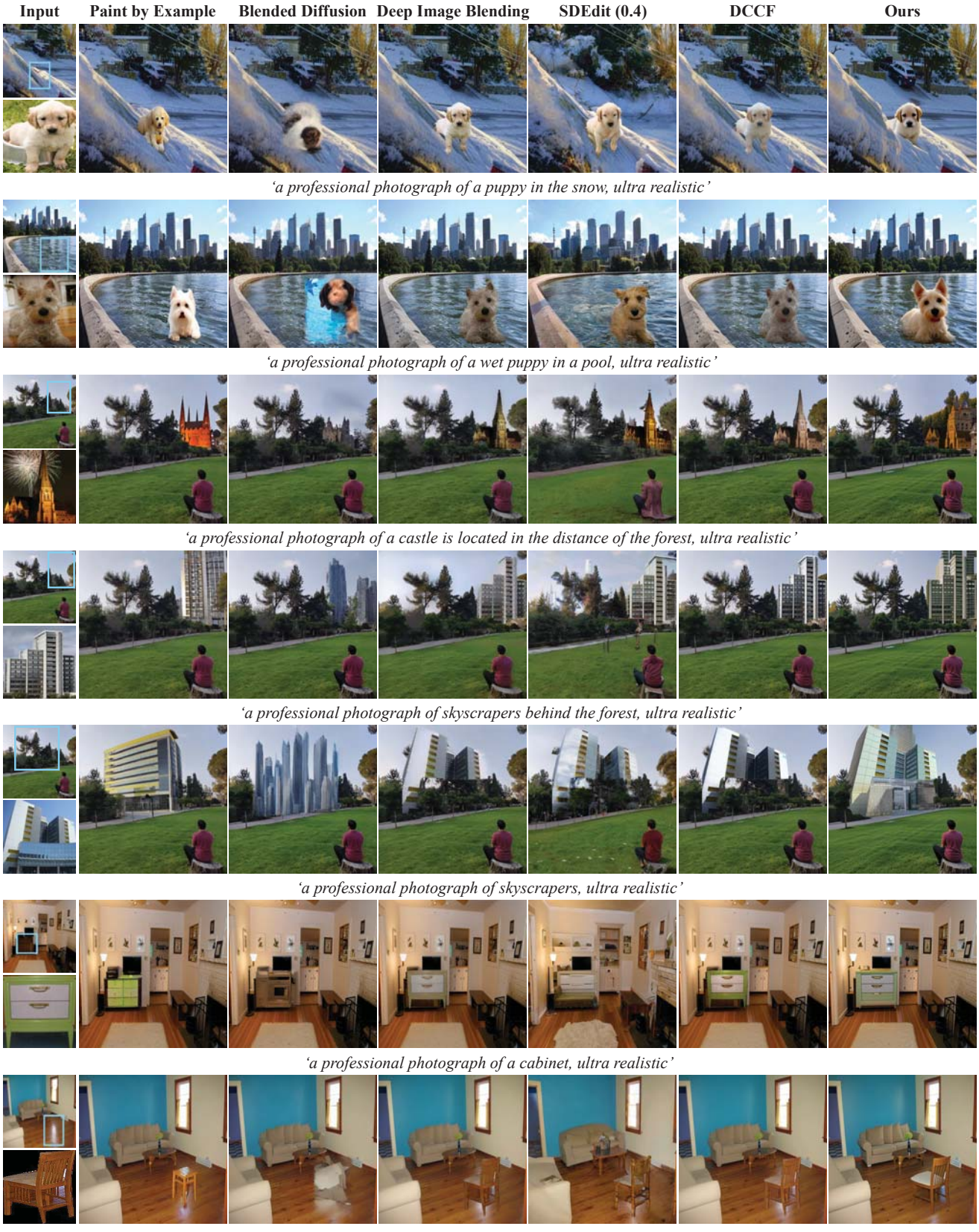


Figure 24: Qualitative comparison with SOTA baselines in image composition for the photorealism domain.

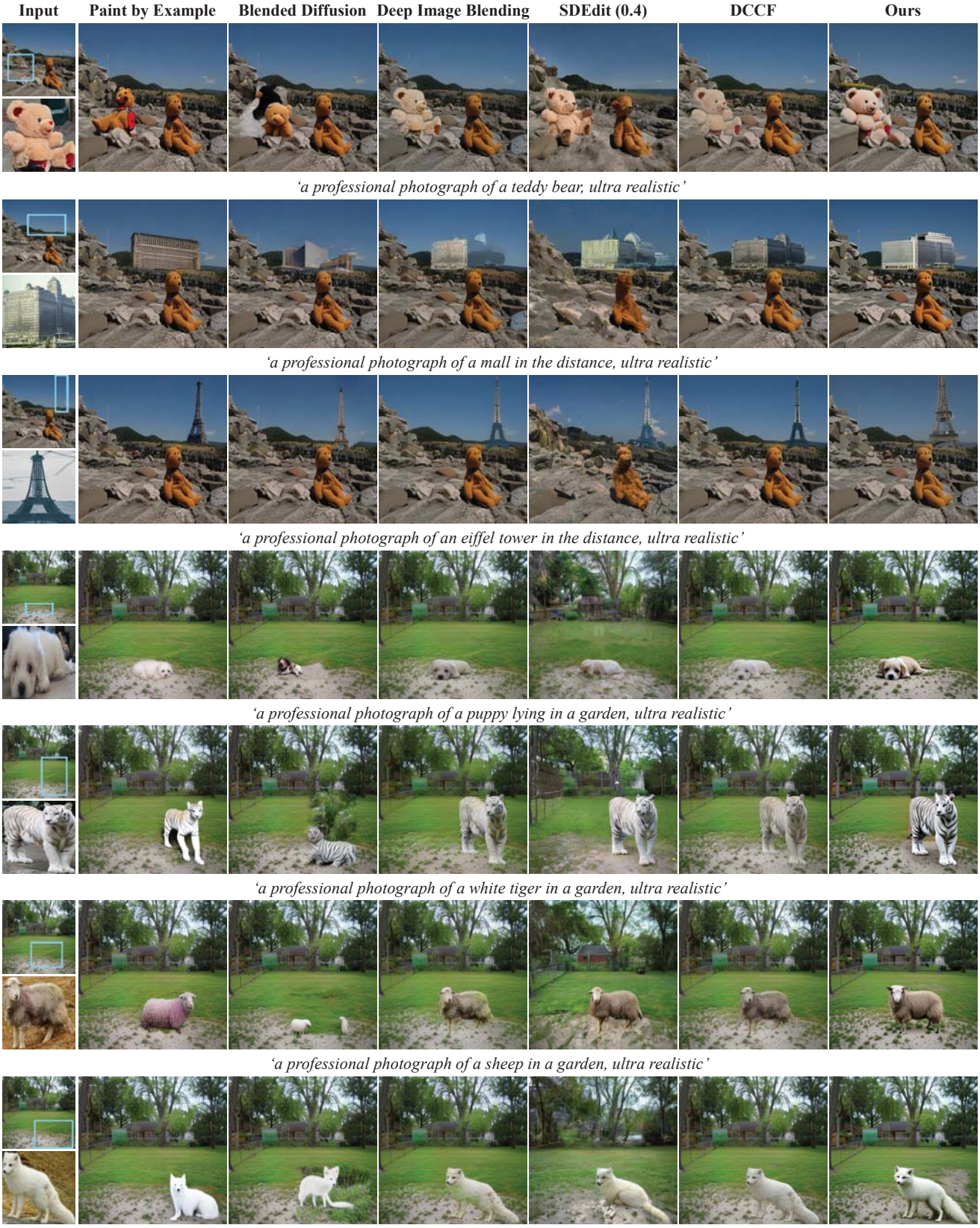


Figure 25: Qualitative comparison with SOTA baselines in image composition for the photorealism domain.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 4

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 6

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[6] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 1

[7] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[8] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 6

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 6

[10] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 1

[11] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 1

[12] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 1

[13] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 4

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 6

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 6

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[18] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 3

[19] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 4

[20] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 231–240, 2020. 4

[21] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 2