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

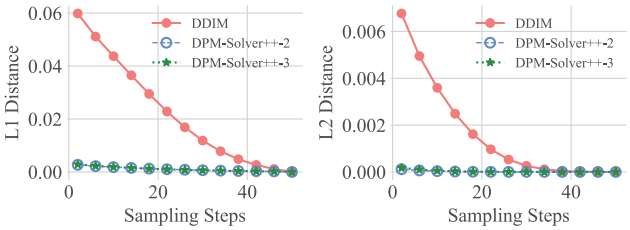


Figure 8: The comparison of the alignment of forward and backward trajectories from DDIM inversion and high-order DPM-Solver++. L1 and L2 distances were computed at each time step between the forward and backward intermediates, and then averaged over 5000 images. The curves representing the second and 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

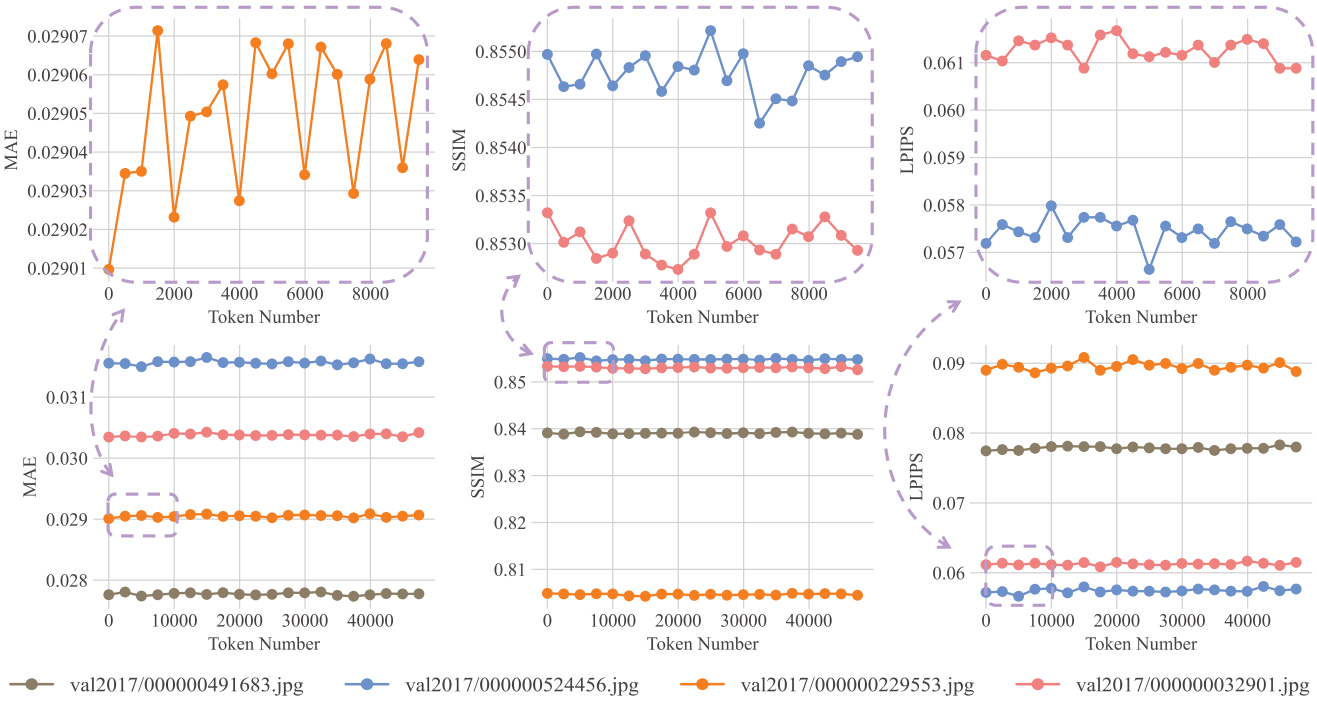


Figure 10: The analysis of the impact of the common token values in the exceptional prompt. The first row displays a magnified view of an area from the second row. For each image randomly sampled from the COCO, the exceptional prompt is applied with 100 uniformly sampled token values on Stable Diffusion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

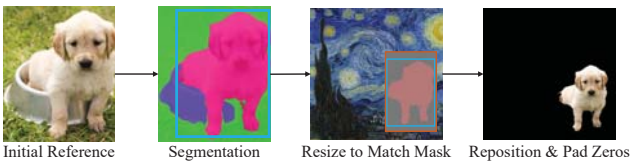


Figure 11: The preprocessing pipeline for the reference image. (1) The centralized reference image is initially processed by a pretrained segmentation model; (2) the segmented object region is then extracted, and its dimension is adjusted to match the size of the user mask; (3) the resized image is finally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

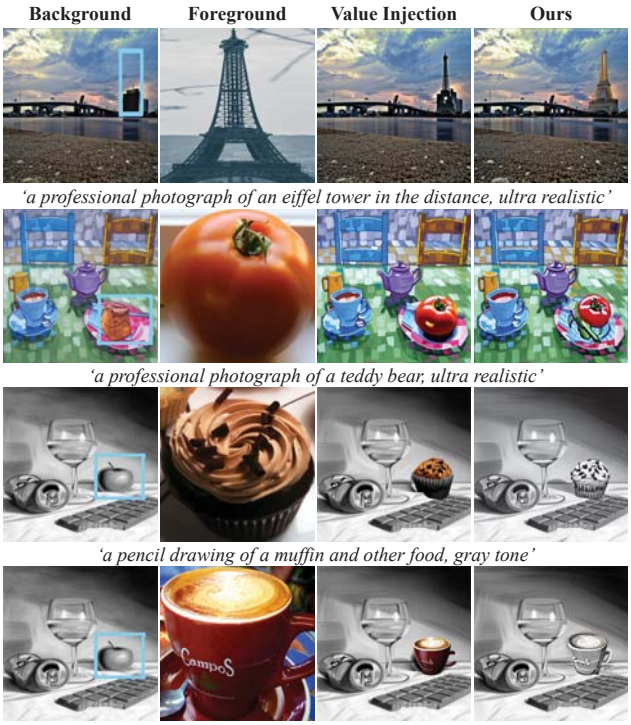
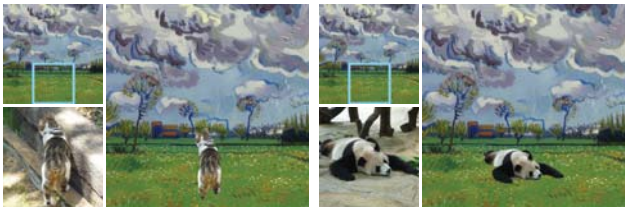


Figure 13: The visual comparison between injecting all information and our implementation. Injecting values leads to a more rigid generation, potentially impeding the smooth transition across visual domains. This impact becomes particularly evident when transferring to the sketchy domain.

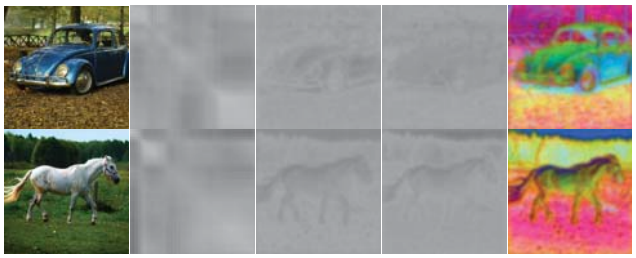
2. **Foreground Preservation:** The generated image should well-preserve the features or identity of the given object within the mask region, such that the viewers can recognize that the given and the generated objects are the same even in different domains.
3. **Background Preservation:** The background 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

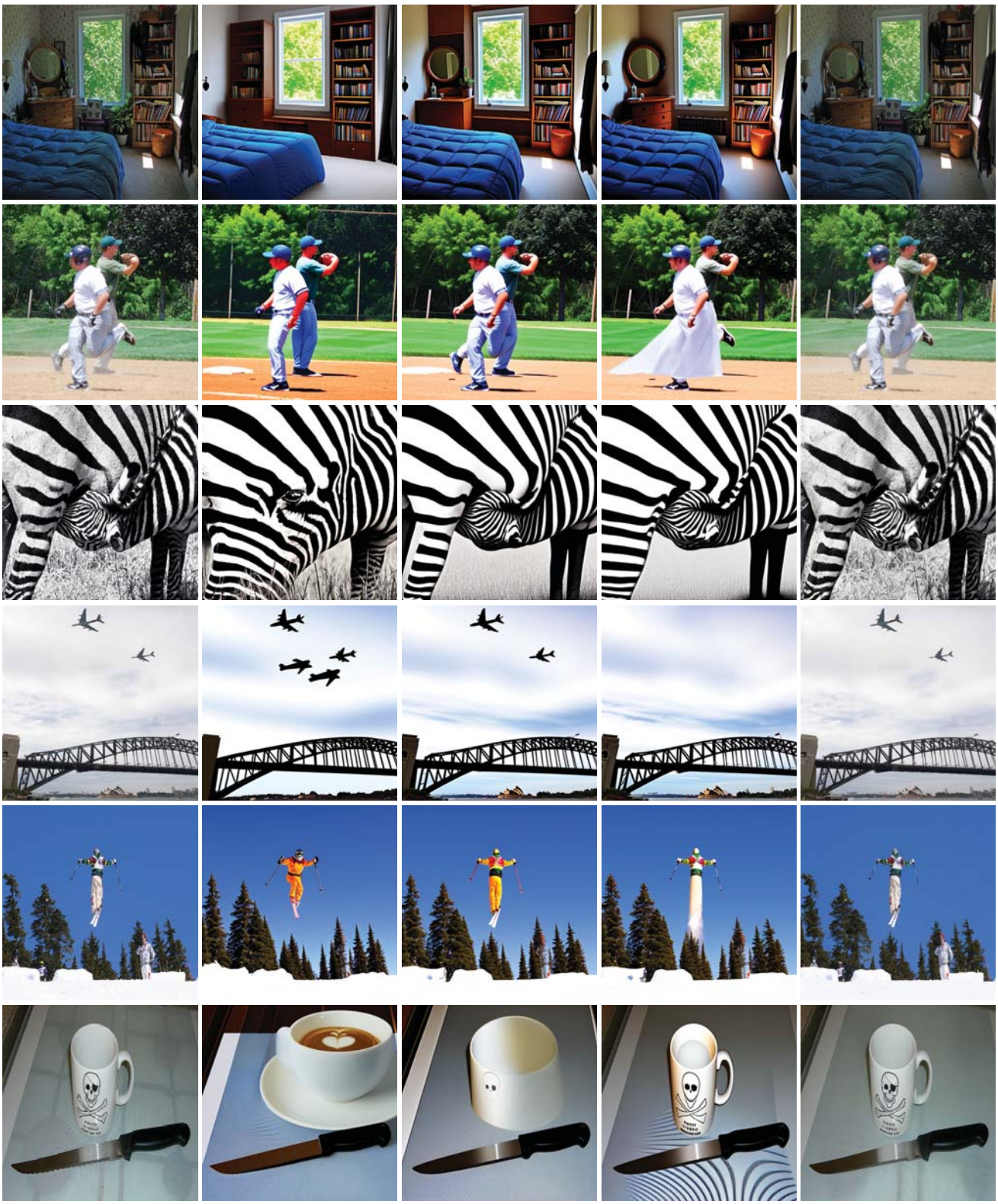
Figures 16, 17, and 18 present additional qualitative image reconstruction comparisons among different outputs of Stable Diffusion on COCO, ImageNet, and CelebA-HQ, respectively.

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.

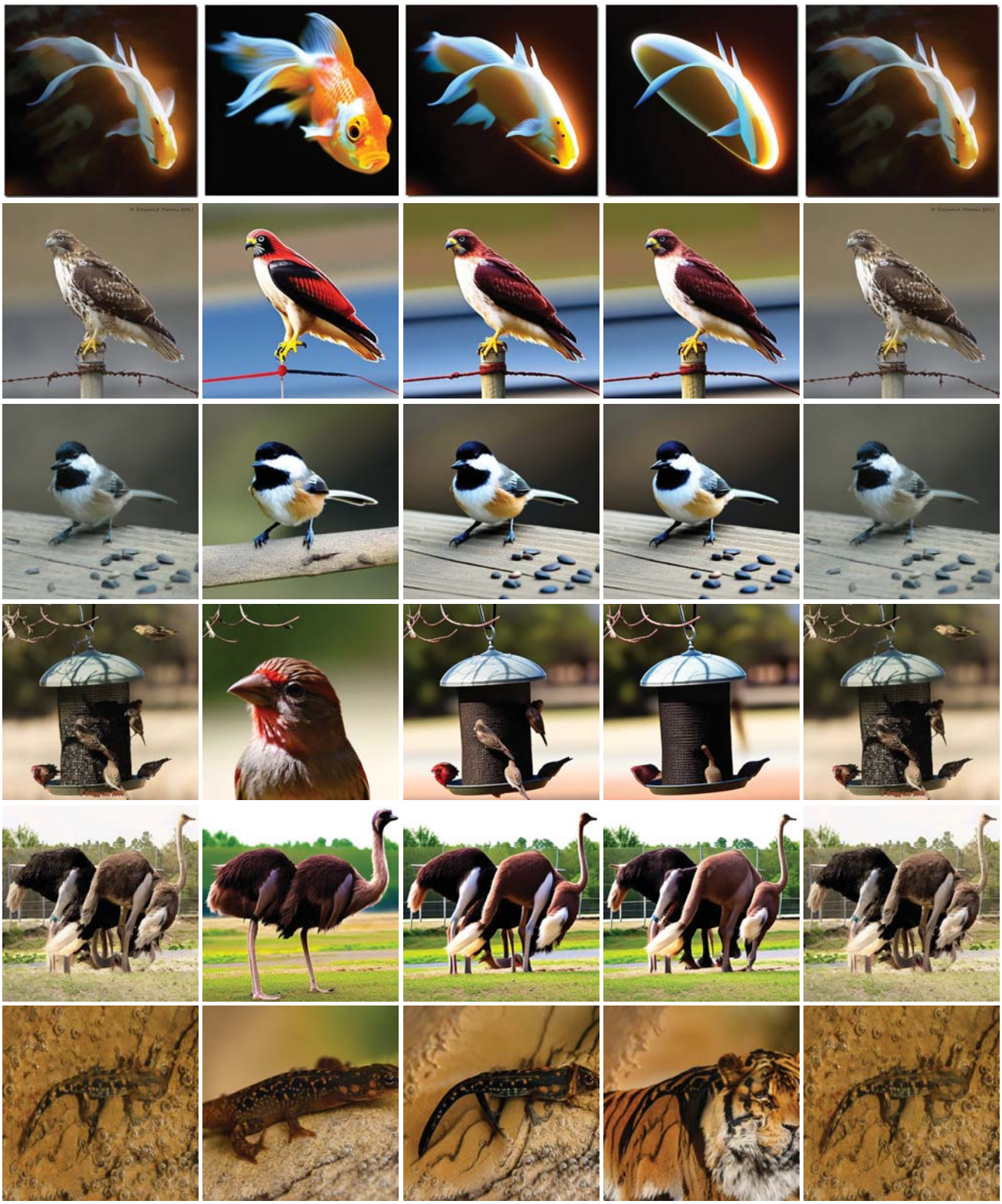
I. Societal Impacts

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 $\epsilon_\theta(\mathbf{x}_t, t, \mathcal{W})$.

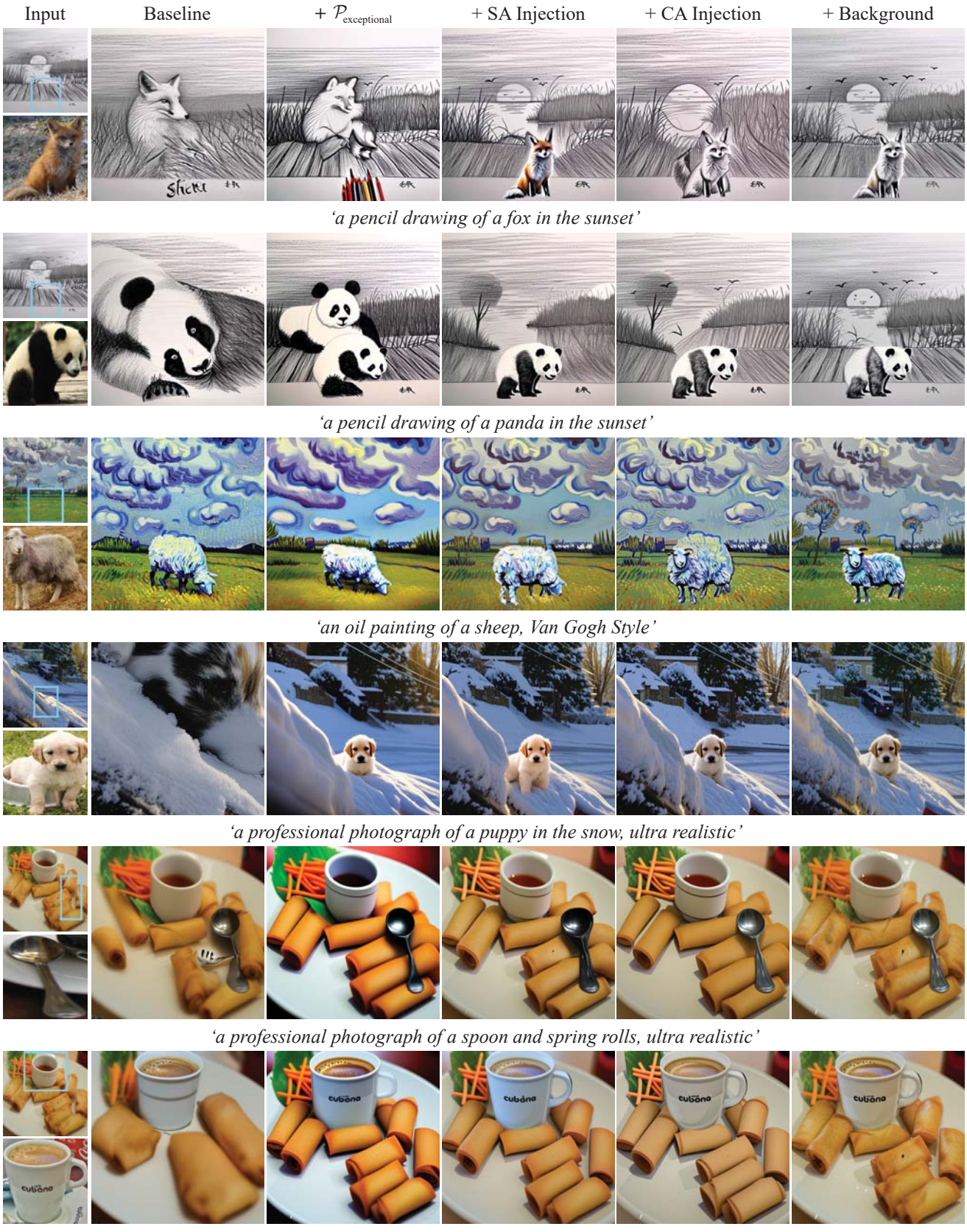


Figure 19: Ablation study of different variants of our framework. SA: self-attention. CA: cross-attention.

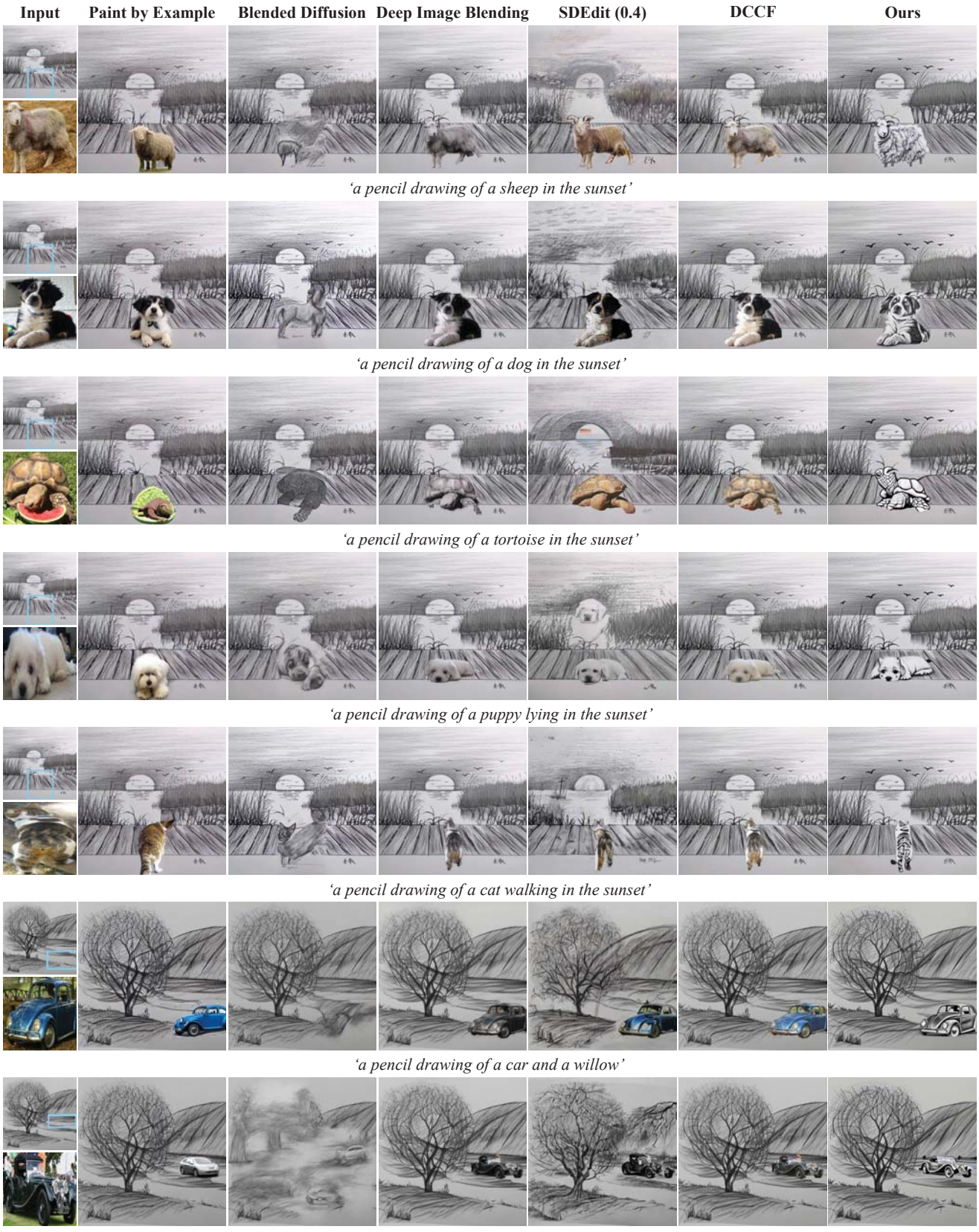


Figure 20: Qualitative comparison with SOTA baselines in image composition for the pencil sketching domain.

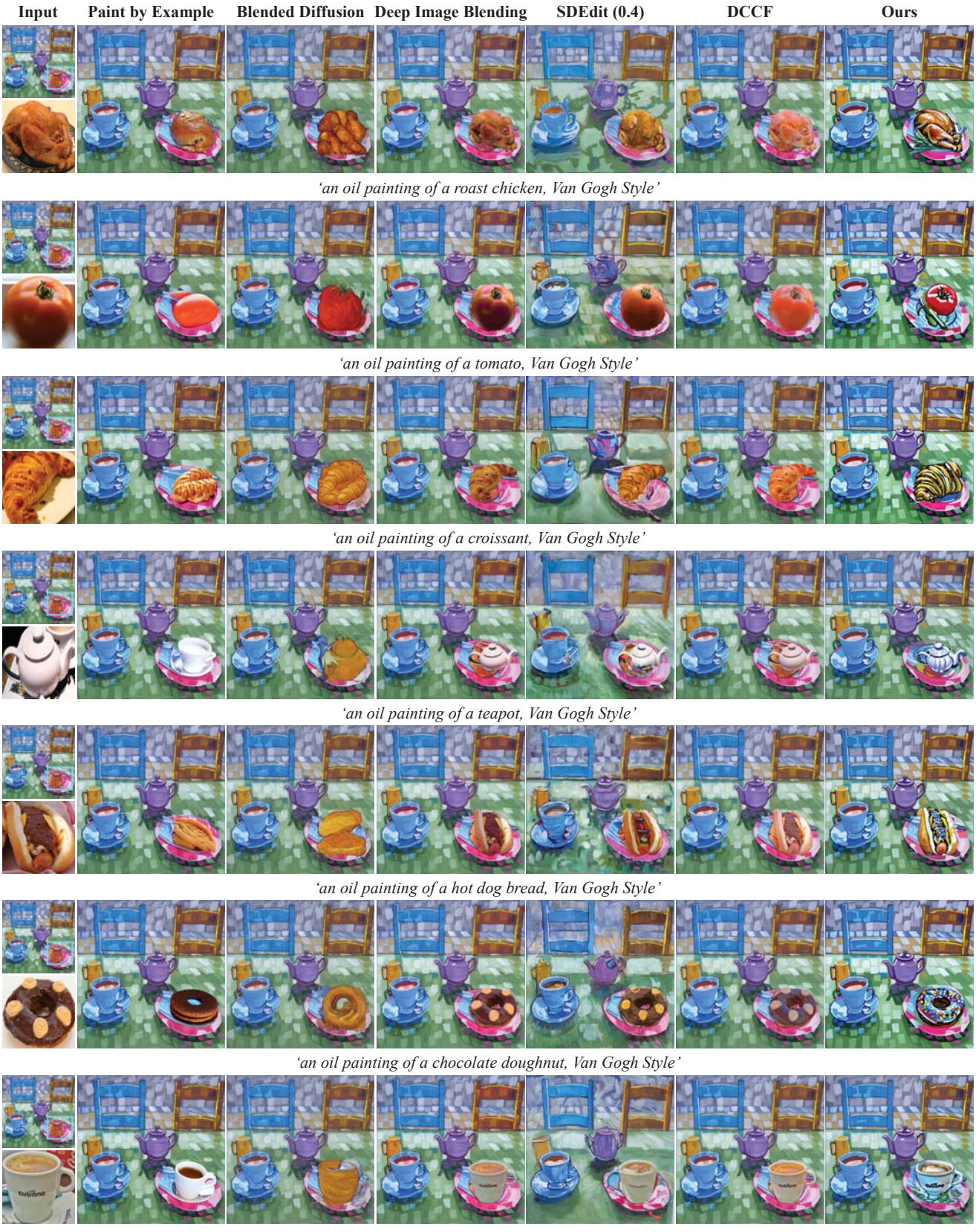


Figure 21: Qualitative comparison with SOTA baselines in image composition for the oil painting domain.

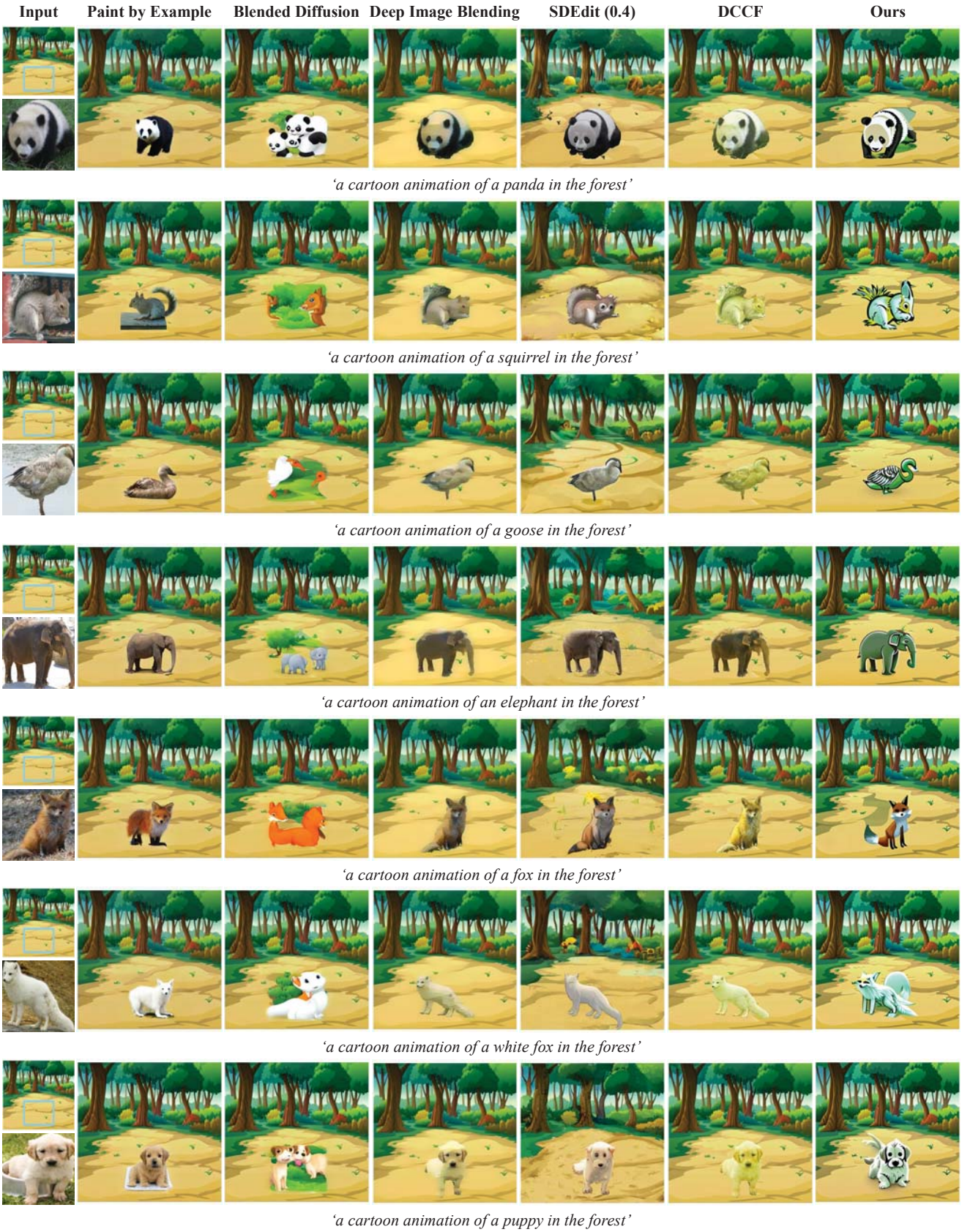


Figure 22: Qualitative comparison with SOTA baselines in image composition for the cartoon animation domain.

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

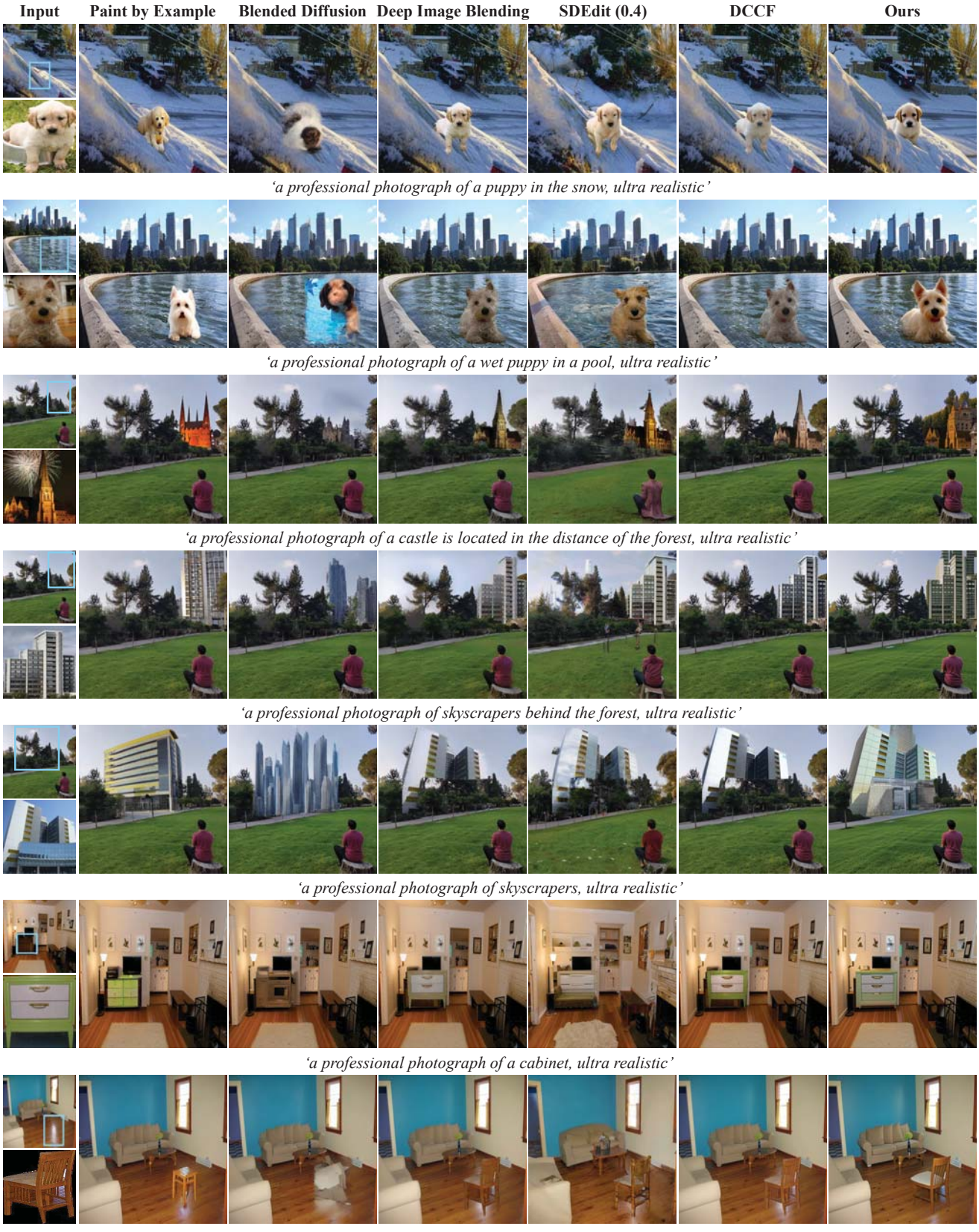


Figure 24: Qualitative comparison with SOTA baselines in image composition for the photorealism domain.

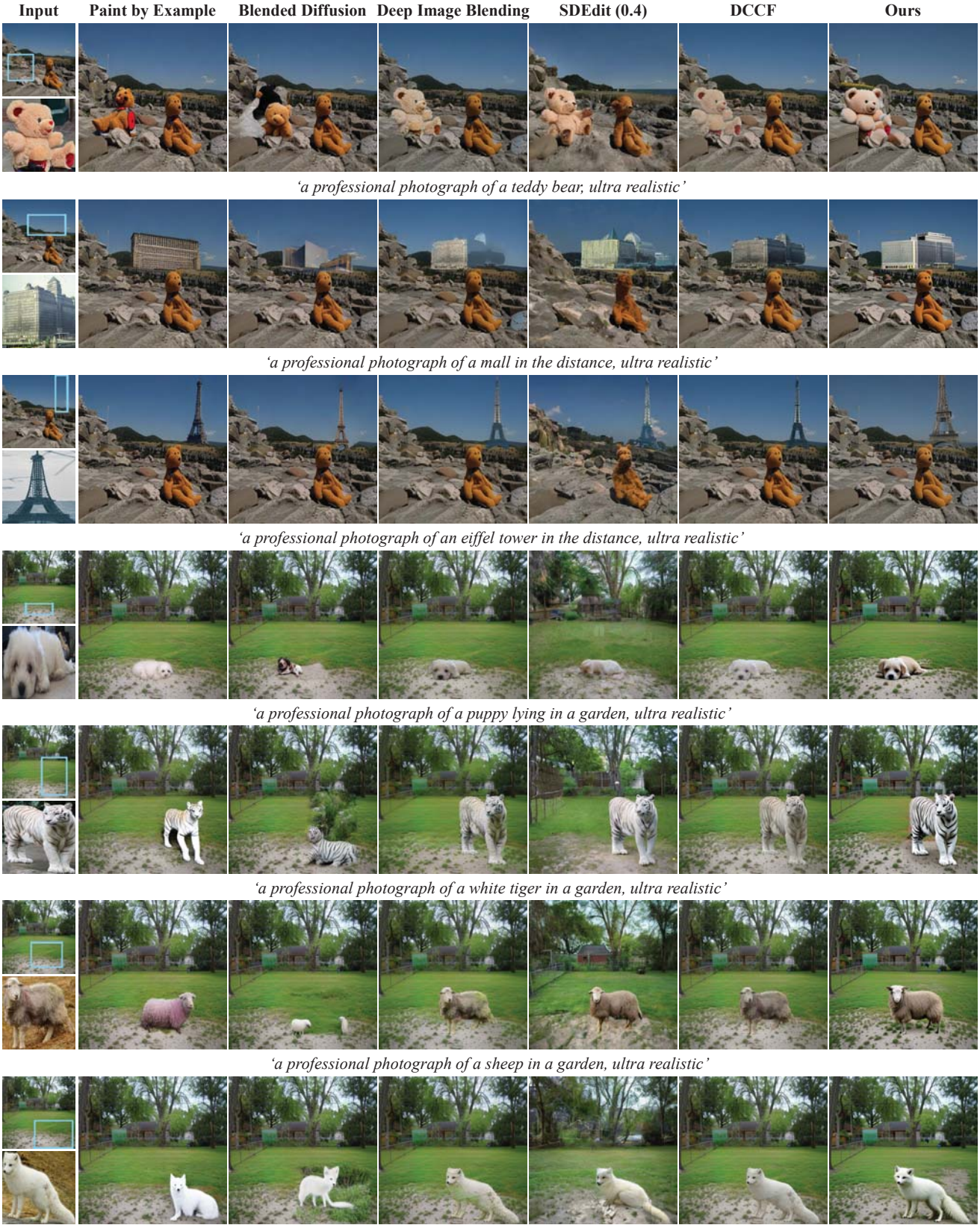


Figure 25: Qualitative comparison with SOTA baselines in image composition for the photorealism domain.

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