Tuning-Free Image Customization with Image and Text Guidance

Pengzhi Li[1*], Qiang Nie[2,3*], Ying Chen[3], Xi Jiang[4], Kai Wu[3], Yuhuan Lin[3], Yong Liu[3], Jinlong Peng[3], Chengjie Wang[3], Feng Zheng[4,5†]

[1]Tsinghua University [2] HKUST (GZ) [3]Tencent Youtu Lab [4] Southern University of Science and Technology [5] Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory

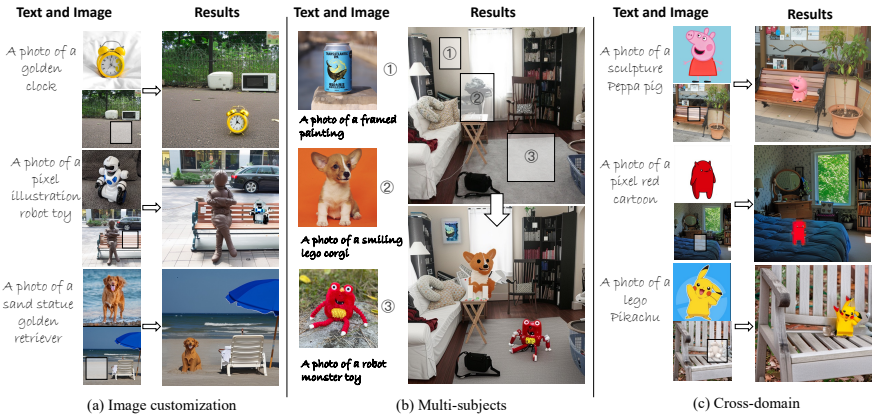

Fig. 1: Performance overview of the proposed method in image customization: (a) We enable the generation of any subject depicted in the reference image within the designated image region to be edited. Additionally, it allows for modifying the subject's attributes based on the user's prompts. (b) The versatility: our method can extend to scenarios involving multiple subjects from different reference images and multiple regions to be edited. (c) Cross-domain customization: we can transform the subject in the reference image into a different domain, such as converting it into a cartoon style.

Abstract. Despite significant advancements in image customization with diffusion models, current methods still have several limitations: 1) unintended changes in non-target areas when regenerating the entire image; 2) guidance solely by a reference image or text descriptions; and 3) time-consuming fine-tuning, which limits their practical application. In response, we introduce a tuning-free framework for simultaneous text-image-guided image customization, enabling precise editing of specific image regions within seconds. Our approach preserves the semantic features of the reference image subject while allowing modification of detailed attributes based on text descriptions. To achieve this, we propose

[†]Corresponding author, [*] Equal contribution.

an innovative attention blending strategy that blends self-attention features in the UNet decoder during the denoising process. To our knowledge, this is the first tuning-free method that concurrently utilizes text and image guidance for image customization in specific regions. Our approach outperforms previous methods in both human and quantitative evaluations, providing an efficient solution for various practical applications, such as image synthesis, design, and creative photography. Project page: https://zrealli.github.io/TIGIC.

Keywords: image editing · image customization · diffusion model

1 Introduction

Recently, with the continuous development of diffusion models [13,31,36], there have been significant advancements in customized image generation. Given different text prompts, large-scale diffusion models such as Stable Diffusion [31] have demonstrated the ability to generate high-quality images that align with specific text prompts. Based on Stable Diffusion, ControlNet [45] presents a more fancy manner to generate images according to the conditions of text description, sketch, pose, *etc*. However, methods like ControlNet re-generate the entire image according to the conditions, which always causes unintended changes in non-target regions. In some cases, people may only want to edit a certain region of the image. Moreover, existing methods [5,28] heavily rely on text description for editing, which may not always capture the desired image modifications accurately even when utilizing long sentences. If given a reference image, such a kind of information misalignment can be well-tamed. Therefore, in this paper, we are addressing the need for image customization *at specified region(s)* with concurrent guidance of *image* and *text*.

Despite the demand, previous methods have not adequately explored the potential of using both text and images to drive image generation simultaneously. For instance, Paint-by-example [42] trains a diffusion model conditioned on images, using them as templates to generate specific features in selected areas of the target image. AnyDoor [9] utilizes an ID extractor to obtain ID tokens from reference images to generate subjects with consistent features. However, replacing the text embeddings with optimized image embeddings in these methods prevents pre-trained diffusion models from retaining their text-driven generative ability, which hinders the following more detailed attribute editing on target subjects. On the other hand, text-only driven approaches like BLD use text prompts to generate new subjects in selected local areas. MasaCtrl [5] and Null-text Inversion [28] use the text prompts to edit the attributions of existing foreground or background. While text-only driven methods retain text-editing capabilities, they are unable to generate specific new subjects based on reference images with unseen attributes.

Furthermore, many existing methods involve time-consuming dataset processing and training phases, as seen in Paint-by-example [42] and AnyDoor [9]. Null-text Inversion [28] requires tuning the unconditional text embedding in

class-free guidance branch. Although MasaCtrl [5] is implemented without tuning, its editing ability is quite limited to the subjects that have been trained in Stable Diffusion. Time efficiency is an important consideration for the scalability of the image customization method.

To address the above issues, we propose a tuning-free framework for simultaneous text-image-driven image customization, which allows users to accurately edit the specific region(s) of an image within seconds under the guidance of reference images and text descriptions, as illustrated in Fig. 1. In the proposed framework, the content to be generated in the target region is controlled by the reference image, and the attributions of the content are edited by the text. To preserve the feature of the subject in the reference image, we create a collage by aligning the segmentation of the reference subject to the target region in the image to be edited. The collage is then inverted with DPM-Solver++ other than DDIM to obtain latent codes as the initial noise samples for the diffusion process. Next, a three-stream denoising structure is proposed for customization. The latent code with a null prompt is utilized for image reconstruction. The latent code with text prompt is used for reference subject editing. To keep the edited subject in harmony with the image to be edited, another latent code utilized for generating the edited image is attained by filling random Gaussian noise between the reference subject area and non-target area in the latent code. The customization is achieved by a self-attention blending strategy which blends the features in reconstruction and text editing streams with target image generation stream. Since the interaction among the text, the reference image and the image to be edited is limited in the target region, our method can avoid the unintended changes in non-target area and attain precise subject attribute editing.

To our knowledge, this is the first tuning-free method that concurrently utilizes text and image guidance for specific region image customization. As shown in Fig. 1, our approach exhibits significant value for applications. In summary, our work makes the following contributions:

- We propose a tuning-free image customization framework, enabling content manipulation in the given region(s) of an image according to user-provided example images and text descriptions.
- We propose a self-attention blending strategy for content customization, which addresses the issue of unintended changes in non-target area in previous image editing methods and achieves precise editing of specific attributes.
- Our method outperforms previous approaches in human and quantitative evaluations, providing an efficient solution for numerous practical applications such as image synthesis, design, and creative photography.

2 Related Work

Image Editing Guided by Text/Image. With the development of diffusion models [30,31,35], image editing and generation have grown rapidly. Most previous methods [4,7,11,14,17,22,22,27] use either text descriptions or a reference image

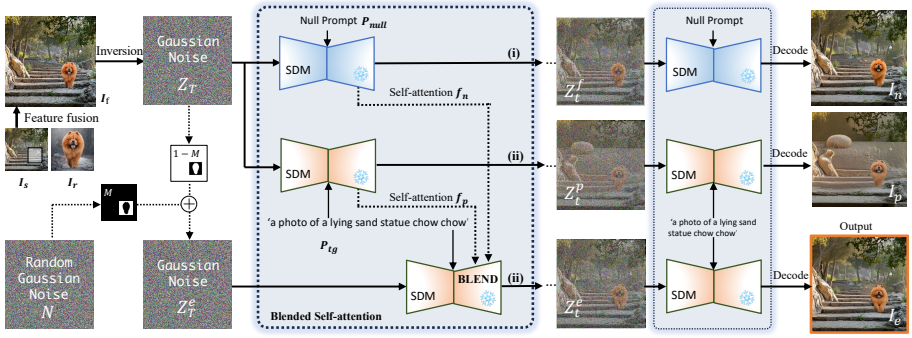

Fig. 2: The pipeline of the proposed method. Our method uses text descriptions T and the reference image I_r as guidance to customize the target region(s) of the image to be edited in a tuning-free manner. We employ blended self-attention instead of original self-attention injection throughout the denoising process, which allows us to retain (i) the generated subject features while achieving (ii) the text-driven capability for attributes modification.

to guide image editing. Some methods [17, 27, 28, 38] focus on global editing using text descriptions, with SDEdit [27] achieving editability by adding moderate noise. P2P [17] and PnP [38] employ cross-attention or self-attention mechanisms for global image editing, while Null-text Inversion [28] explores better reconstruction results during the inversion process to improve image editing. Another category of methods [2, 3, 40, 42] concentrates on local image editing. Blended Diffusion [3] and Blended Latent Diffusion [2] use mask to create a blend denoising step during editing, while DiffEdit [11] can automatically generate masks during the diffusion process to achieve local editing. In contrast to the text-based approaches, Paint-by-example [42] trains an image-conditioned diffusion model, and AnyDoor [9] uses an ID extractor for image-driven image editing. Although these works have achieved impressive results, they can only edit images based on either text or an image and learn coarse semantic information to generate low-fidelity images. Our method achieves high-fidelity image generation driven by both text and image.

Training for Image Customization. Subject-driven image editing focuses on generating content consistent with subject features in scene images. Some past customization methods [14–16, 20, 21, 33, 39] required significant time and computational resources to fit new concept features. Dreambooth [33] fine-tunes diffusion models using a set of subject images, Textual Inversion employs optimized text embeddings to represent new subjects, HiPer [15] explores a set of personalized tokens to represent new subjects, and CustomDiffusion [21] captures multiple concepts simultaneously by learning new text embeddings and fine-tuning cross-attention. Break-A-Scene [1] can effectively learn multiple subject features. Some works [8, 9, 19] have explored using the large-scale datasets for pre-training to

achieve customization without fine-tuning. Although these methods can generate high-quality images, the time-consuming training phase limits their use. Training-free image generation and editing of specific subject areas remain in the exploratory stage.

Image Composition. Image composition is widely applied in various downstream tasks. A common practice in image processing is to stitch two different photos together, with many methods [12, 18, 24, 37, 41, 43, 44]focusing on image harmonization to make images more realistic. These methods can generally be divided into several categories, including object placement, image blending [37, 43], harmonization [10, 12], and shadow generation [18]. However, these methods struggle to change the original layout and content of the image, making the generation of images that conform to real human visual perception challenging. In this paper, we consider their feature preservation capabilities and leverage the powerful generation capabilities of diffusion models to drive the generation of realistic and harmonious images with consistent lighting and environmental features.

3 Method

The pipeline of our method is illustrated in Fig. 2. Given an image I to be edited and the target region(s) R that needs edition, our goal is to synthesize an image I_e that not only has the subject in the reference image(s) I_r but also satisfies the description of text T in a tuning-free manner. The text T is utilized for controlling the attributes of the customized subject in R. This is a challenging task due to the following issues: (1) maintaining consistency in the non-target region between I and I_e; (2) ensuring semantic coherence between the generated subject and the reference subject in the target region; (3) accurately controlling the attributes of the generated subject without changing the other part according to the text description; and (4) seamlessly integrating the generated subject in R with the non-target region content in I_e.

3.1 Image Guidance Injection and Inversion

To retain semantic consistency, previous methods [9, 42] often encode reference image I_r using a pre-trained visual encoder. However, such a kind of approach can hardly preserve the details of the reference subject S_r in I_r. Different from these methods, we find that the pixels of S_r contain sufficient information to keep the generated content consistent in both semantic-level and detail-level such as texture, shape and pose. Therefore, we directly utilize the pixels of S_r into the target region R for image inversion.

To achieve aforementioned goal (1) and (2), a precise inversion process is required. The mainstream DDIM [36] inversion effectively transforms an image into a latent representation which can successfully reconstruct the input image.

Fig. 3: Reconstruction results. The first row shows the initial images used for inversion, the second row represents the image reconstruction results from DDIM [36], and the third row shows our reconstruction results. DDIM's results may distort when the object's material, lighting, or additional objects in the image are artificially altered. In contrast, our method consistently generates high-quality reconstructions, a critical aspect for image editing.

However, as depicted in Fig. 3, due to differences in factors such as lighting environments between S_r and I, the collage using conventional DDIM inversion performs less favorably compared to the real image. Inspired by [26], which demonstrated that utilizing high-order ODE solvers for diffusion inversion produces superior latent representations, we employ the advanced DPM-Solver++ [25] to promote the inversion quality of the collage. As shown in Fig. 3, our approach achieves more accurate reconstruction than DDIM [36] inversion. Therefore, in this paper, we choose to apply it to the inversion process of the diffusion model. The latent code of the collage is denoted as z.

Furthermore, for the goal (4), *i.e.* the harmony between the generated subject and the non-target region in I_e, interaction between the reference subject in R and non-target area is required. To achieve this, we fill random Gaussian noise between the reference subject and non-target area in z and attain a new latent code z^e to generate the final customized image. z^e can be formulated as

$$z^e = M \odot \varepsilon + (1 - M) \odot z \tag{1}$$

where $\varepsilon \sim N(0, I)$ is the Gaussian noise. M denotes the mask to generate the region between reference subject and non-target area in the collage, as shown in Fig. 2.

Algorithm 1: Proposed Method

Input: A scene image I_s, a reference image I_r, a target prompt P_{tg}, a null prompt P_{null}, the random Gaussian noise N, the mask M.
Output: The edited image I_e corresponding to P_{tg}.

$I_f = \text{FeatureFusion}(I_s, I_r)$, $z_0^f = \text{Encoder}(I_f)$
$z_T^f \leftarrow ... \leftarrow \text{Inversion}\left(z_0^f, P_{null}\right)$
$z_T^p \leftarrow ... \leftarrow \text{Inversion}\left(z_0^f, P_{tg}\right)$
$z_T^e = \text{FusionLatent}(z_T^f, N, M)$
for $t = T, T-1, ..., 1$ **do**
  $z_{t-1}^f, f_n \leftarrow \hat{\epsilon}_\theta\left(z_t^f, t, P_{null}\right)$
  $z_{t-1}^p, f_p \leftarrow \hat{\epsilon}_\theta\left(z_t^p, t, P_{tg}\right)$
  $z_{t-1}^e, f_e \leftarrow \hat{\epsilon}_\theta\left(z_t^e, t, P_{tg}\right)$
  $z_{t-1}^e \leftarrow z_{t-1}^e \cdot M + z_{t-1}^f \cdot (1 - M)$
  $f_{BD} \leftarrow \text{BLEND}(f_e, f_p, f_n)$
  $z_{t-1}^e \leftarrow \hat{\epsilon}_\theta\left(z_t^e, P, t; f_{BD}\right),$
end for
Return $I_e = \text{Decoder}(z_0^e)$

3.2 Customization with Self-Attention Blending

To edit the attributes of generated content in target region R, the self-attention block in the U-Net [32] structure of the diffusion model provides a plug-and-play feature that can be seamlessly integrated into specific layers for content customization. In self-attention block, the intermediate features f from the previous layer $l-1$ are projected into queries Q, keys K, and values V, and the output of the self-attention block can be formulated as:

$$\mathrm{Q} = f_t^{l-1}\mathbf{W}_l^q, \quad \mathrm{K} = f_t^{l-1}\mathbf{W}_l^k, \quad V = f_t^{l-1}\mathbf{W}_l^v, \tag{2}$$

$$\boldsymbol{A}_t^l = \text{Softmax}\left(\boldsymbol{Q}_t^l \boldsymbol{K}_t^{l^T} / \sqrt{d}\right), \tag{3}$$

$$\boldsymbol{f}_t^l = \boldsymbol{A}_t^l \boldsymbol{V}_t^l \tag{4}$$

where $\boldsymbol{A}$ is the attention map. The attention map contains rich structural and content information. Manipulation in self-attention layers requires no additional optimization, allowing users to achieve image recreation within seconds effortlessly.

Specifically, as illustrated in Fig. 2, we utilize a three-stream architecture to execute the self-attention blending. Given an input image, we first obtain the latent representation z^n of the collage after feature inversion. Subsequently, at each time step t, we pass the latent code z to a denoising U-Net using the null and target text descriptions, respectively. The output feature of the self-attention block in these two streams are f_n and f_p. Similarly, f_e can be achieved

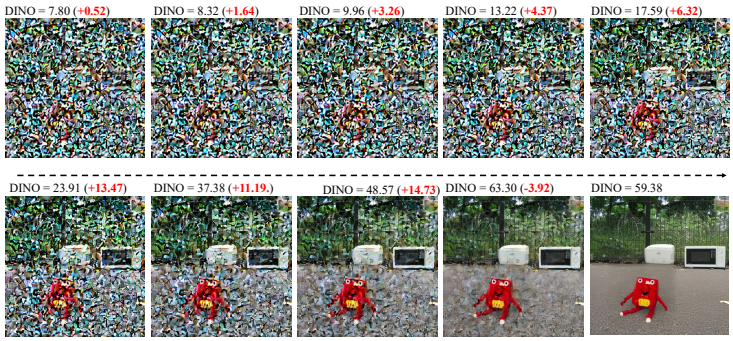

Fig. 4: Semantic information contained in different denoising steps. We observe that the layout is mainly formed in the early denoising process, while the generation of semantic information primarily begins in the latter stages. DINO [6] score can reflect the richness of semantic information. Therefore, we perform the attention enhancement at this stage.

from z^e stream. f_n from the reconstruction stream helps retain the information of non-target region. f_p from the text-driven stream provides the information for attribute editing. f_e offers diversity with additional Gaussian noise for the interaction between the generated subject and the non-target area. We then blend the self-attention, denoted as f_{BD}, using the most straightforward weighted average, as show in Eq. 5.

$$f_{BD} = \begin{cases} \alpha f_e + \beta f_p + \gamma f_n & \text{if } t \in (\tau_a T, \tau_b T) \\ \frac{1}{2}(f_e + f_p) & \text{if } t > \tau_b T \\ f_e & otherwise \end{cases} \tag{5}$$

where $\alpha + \beta + \gamma = 1$. This core operation maintains semantic information consistency while enabling text guided capabilities. Finally, we inject f_{BD} during the specific denoising steps of z_t^e:

$$z_{t-1}^e = \hat{\epsilon}_\theta \left(z_t^e, P, t; f_{BD} \right), \tag{6}$$

where $\hat{\epsilon}_\theta$ represents the modified denoising step with f_{BD}. Therefore, in time step $t-1$, the self-attention block is calculated using f_{BD}.

$$\mathrm{Q}_{t-1} = f_{BD}\mathbf{W}^q, \quad \mathrm{K}_{t-1} = f_{BD}\mathbf{W}^k, \quad V_{t-1} = f_{BD}\mathbf{W}^v, \tag{7}$$

Blended Enhancement While the proposed blending self-attention manipulation simply but effectively integrates target semantic information into the image structure, it can still result in inaccurate edition in output image and the unintended content. We believe this issue stems from the overfitting of global semantic information. Given that there's ample semantic information in the late diffusion denoising stages, further injection of blended self-attention may lead to

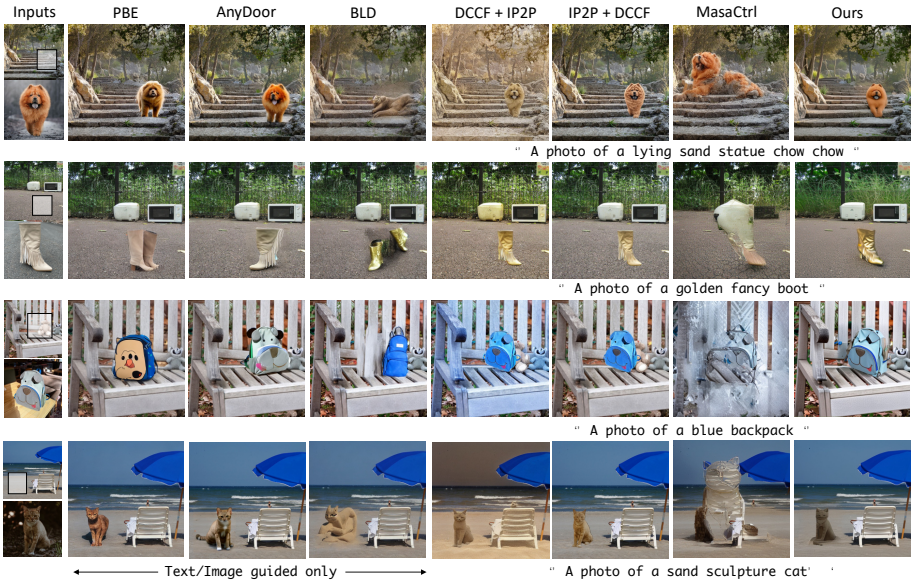

Fig. 5: Qualitative comparison with existing state-of-the-art methods. PBE [42] and AnyDoor [9] are methods guided only by images, while BLD [2] uses text as the only guidance. To evaluate the efficiency of our method, we set up an additional group of two-step methods, including first using image stitching and harmonization followed by text guided image editing (DCCF [41] + IP2P [4], MasaCtrl [5]) and another method involving editing first and then harmonizing (IP2P [4] + DCCF [41]). These methods can only focus on text or image, global or local editing. Our method outperforms all these methods and overcomes their limitations, achieving text and image guided local editing and generation.

more artifacts. To address this, we establish a threshold to determine when to cease injecting blended self-attention during the denoising steps.

We draw inspiration from [47] that the diffusion model denoising U-Net generates images in the order of "layout → content → material/style. Specifically, only at a specific time step (τ_a, τ_b), we pass the latent representations f_t^p and f_t^n to the denoising U-Net of z_T^e, as shown in the first row of eq. 5. When time step t larger than $\tau_b T$, we only blend the f_t^p with f_t^e to inject more information from the text-guided stream for better attributes modification in the target region, as shown in second row of eq. 5. For early stage before $\tau_a T$, we don't apply any blending to avoid affecting the layout generation. This enhancement strategy effectively corrects the inaccuracy of semantic information in target region, improving the overall quality of text-image-guided editing in the final output.

To determine the appropriate threshold range for parameters τ_a and τ_b, we conduct an analysis of the generated images and observe that altering semantic information at early or late stages can deteriorate the final results. As shown

Algorithm 2: BLEND

Input: f_e, f_p, f_n.
if $t > \tau_a T$ and $t < \tau_b T$
then $f_{BD} \leftarrow \alpha f_e + \beta f_p + \gamma f_n$
else if $t > \tau_b T$ **then** $f_{BD} \leftarrow \frac{1}{2}(f_e + f_p)$ **else** $f_{BD} \leftarrow f_e$
Return f_{BD}

in Fig. 4, for our denoising process, semantic information (quantitatively represented by DINO score) appears infrequently between $(0,\ 0.5 \times T)$, and rapidly increases starting at $(0.5 \times T,\ 0.8 \times T)$. In the next section, we conduct detailed experimental analysis on the two thresholds. The whole framework of our algorithm is illustrated in Alg. 1. The BLEND function is illustrated in Alg. 2.

4 Experiments

4.1 Experimental Setup

Benchmarks Since there is no existing dataset to evaluate specified region customization with both text and image inputs, we collect a dataset comprising 3000 images for quantitative evaluation. The sample images in this dataset span 30 categories from DreamBooth [34], including 21 objects and nine living subjects. We select ten representative scenes from the COCO dataset [23], covering indoor and outdoor environments, and provided corresponding bounding box information. Subsequently, we generate ten sets of text for attribute modification and applied them to scenes across all categories. We provide more details in the supplementary material.

Evaluation metrics Our approach employs a dual-driven mechanism using both images and text, necessitating evaluation from both textual and visual perspectives. Our metrics is same as Dreambooth [34], prioritize subject fidelity. This involves ensuring that the generated images maintain consistency with the reference subject's features. To achieve this, we utilize CLIP-I and DINO [6] metrics to calculate the similarity of subject features within the edited regions. The second metric assesses the consistency between the edited regions and textual descriptions. Furthermore, we employ the CLIP-T metric to measure the cosine similarity between text prompts and CLIP [29] embeddings. Additionally, we conduct user studies to comprehensively evaluate the feasibility of our approach.

4.2 Comparison with Previous Works

Single-driven method As shown in Fig. 5, we present the generated results of the currently outstanding single-driven methods. The Paint-by-example [42] utilizes example images as references to generate images in the selected regions of the scene image, matching the features of the example image. However, it can only

Table 1: Quantitative comparison of different methods. We report three scores: DINO [6] score, CLIP-I, and CLIP-T, which are used to comprehensively evaluate the similarity of subject features and the matching degree of text descriptions. Our method achieves the best scores on all three metrics.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Paint-by-example [42]	15.26	66.94	20.62
Stable Diffusion Inpainting [31]	14.43	62.17	21.04
Blended Latent Diffusion [2]	17.17	63.21	21.14
Ours	**51.18**	**78.08**	**26.86**

Table 2: Quantitative comparison of different methods. Left: Text-driven capabilities. Right: single-driven comparisons.

Method	CLIP-T ↑
Paint-by-example [42]	20.62
Anydoor [9]	21.22
Ours	**26.86**

Method	DINO ↑	CLIP-I ↑
Paint-by-example [42]	15.26	66.94
Anydoor [9]	59.19	78.92
Ours	**62.88**	**80.28**

learn coarse semantic information and fails to capture the features that best reflect the details of the subject. Based on text descriptions, Blended Latent Diffusion [2] can generate corresponding images in the target area. We input as detailed text descriptions as possible, but due to the limited training samples of diffusion models, they can only generate rough object categories. In contrast, our method produces subject features highly consistent with both the example image and text descriptions. We provide a quantitative evaluation of these methods in Tab. 1. We also include quantitative comparisons with Anydoor in Tab. 2. It's important to highlight that our method excels in retaining text-driven capabilities, while Anydoor falls short in this aspect, relying solely on reference images. Furthermore, Anydoor requires the collection of 14 datasets and involves significant training time compared to our effective tuning-free method.

Two-steps method Due to the absence of a one-step, dual-driven generation method capable of simultaneously incorporating both text and images, we adopt a two-step generation strategy for comparison. The first strategy involves segmenting the subject and seamlessly integrating it into the scene image. This is achieved using the harmonization algorithm DCCF [41] and the text-driven image editing algorithm IP2P [4], resulting in the outcomes (DCCF [41] + IP2P [4], DCCF [41] + MasaCtrl [5]). The second strategy entails editing the example image first and then proceeding with integration and harmonization (IP2P [4] + DCCF [41]). These methods require intricate procedures and face challenges in achieving harmonious interactions between the background and subject. In contrast, our approach accomplishes the task in a single operation and generates higher-quality images.

Table 3: Quantitative comparison of reconstruction results. We reconstruct a higher quality image compared to DDIM [36].

Method	MAE ↓	LPIPS ↓	SSIM ↑
DDIM	0.128	0.472	0.697
Ours	**0.041**	**0.106**	**0.806**

Table 4: Quantitative ablation studies on the core components of our method. The values of α and β represent the different order of attention blending processes, resulting in different weights.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Baseline	32.30	73.88	25.93
+ Blended Self-attention	50.43	77.73	26.16
+ Enhancement($\alpha = 1/4$, $\beta = 1/4$)	50.71	77.93	26.44
+ Enhancement($\alpha = 1/3$, $\beta = 1/3$)	47.90	77.75	26.35
+ Enhancement($\alpha = 1/4$, $\beta = 1/2$)	**51.18**	**78.08**	**26.86**
Blended $\rightarrow (0, T)$	47.32	75.96	24.97
Enhancement $\rightarrow (0, T)$	39.08	74.37	24.85
Enhancement $\rightarrow (0, T/2)$	50.34	77.80	26.31

User Study We conduct an user study to compare our work in detail with previous methods. We select six methods for evaluation, including Paint-by-example [42], Belended Latent Diffusion [2], DCCF [41] + IP2P [4], IP2P [4] + DCCF [41], and DCCF [41] + Masactrl [5]. For each method, we generate 20 groups from 120 different images. Each set of images included additional scene images, text descriptions, and example images. Clear rules are established for evaluating fidelity, quality, and text alignment, with scores ranging from 1 to 5, representing the worst to the best. Fidelity is designed to evaluate the similarity of image features, quality to judge the harmony of images, and text alignment to evaluate whether the generated subjects within the region matched the text descriptions. We collected a total of 2513 valid answers. As shown in Tab. 5, our method achieves outstanding scores on all metrics.

4.3 Ablation Studies

We conduct an extensive ablation study to validate the effectiveness of our designed core components.

Inversion. Firstly, we verify the effectiveness of the DPM-Solver++ [25] solver in inversion inharmonious lighting images after feature fusion. In comparison to the DDIM [36] method, our approach demonstrates superior performance. We employ 300 sets of scene images with different subjects for validation, as illustrated in Fig. 3 and Tab. 3, DDIM [36] exhibits varying degrees of image distortion due to significant differences between the pasted subject and the original image. Experimental results show that our method outperforms DDIM [36] significantly in metrics such as MAE, LPIPS [46], and SSIM, designed to assess image generation quality.

Table 5: Result of user study. Our method achieves the highest scores.

Method	Fidelity ↑	Quality ↑	Text alignment ↑
Paint-by-example [42]	3.46	3.49	2.88
Blended Latent Diffusion [2]	2.93	3.55	3.88
DCCF [41] + IP2P [4]	3.63	3.55	3.88
IP2P [4] + DCCF [41]	3.91	3.31	4.11
MasaCtrl [5]	2.33	2.34	2.52
Ours	**4.02**	**3.93**	**4.28**

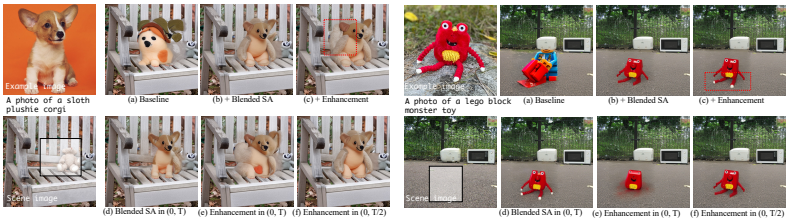

Fig. 6: Ablation studies of each components. (a) is the baseline with only inversion performed. (b) represents the blended self-attention method, while (c) adds the enhancement strategy. (d), (e) and (f) show the results with different threshold values.

Blended self-attention. As shown in Tab. 4, we conduct comprehensive ablation experiments on the blended self-attention method to validate its effectiveness. (a) Baseline: No attention injected, only inversion performed. (b) Integration of blended attention. (c) Integration of blended enhancement strategy. (d) Setting the threshold of blended attention to 1. (e) Setting the threshold of enhancement strategy to (0, 1). (f) Setting the threshold of the enhancement strategy to (0, 0.5). In Fig. 6, we also provide visual results of generated images under different settings. These results strongly support the effectiveness of our blended self-attention method. By applying blended self-attention and enhancement strategies, our approach achieves a significant improvement.

5 Application

Creative photography Post-production in photography is crucial, and inserting or editing objects in photos has excellent potential. Relying on traditional image processing software like Photoshop requires significant time to adjust the attributes of inserted objects to make them harmonious with the environment. Fig. 1 (a) and (b) show that our method can achieve text and example image dual-driven post-production within seconds.

Graphic design As shown in Fig. 7, we demonstrate that our method can provide strong support for many creative design applications, such as interior design, poster design, and more.

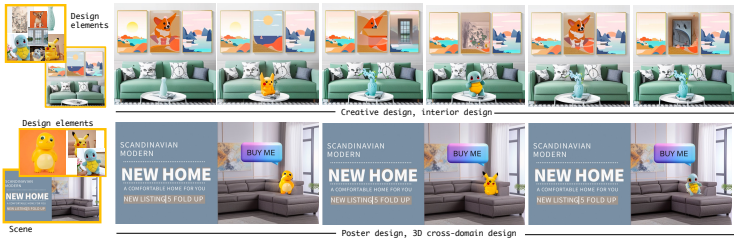

Fig. 7: Some creative applications. As shown in the first row, given an indoor scene and a collection of materials, our method can edit the interior decorations and furnishings using reference subjects from the material library. Our method can also be applied to cross-domain graphic design creations, as shown in the second column, where cartoon characters are generated directly in real-world scenes.

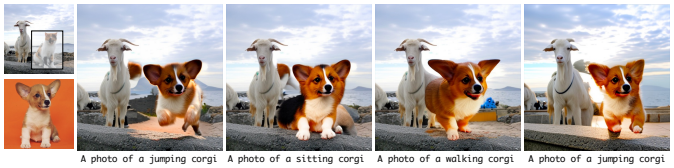

Fig. 8: Non-rigid and perspective editing may sometimes loss the subject features.

6 Limitations

Our method employs a simple yet effective hybrid strategy that maintains the subject's characteristics while possessing text-driven capabilities. However, due to self-attention blending mechanism without tuning, as other tuning-free methods, generating images from multiple perspectives is still challenging. As shown in Fig. 8, editing non-rigid motion can also result in losing subject features. This issue has long troubled the field and urgently needs to be addressed.

7 Conclusion

We introduce a novel tuning-free framework for image customization that effectively leverages both text prompts and reference images. Our innovative blended self-attention strategy ensures precise editing while enabling us to maintain generated subject features and simultaneously achieve text-driven capability. As a pioneering approach in this domain, it demonstrates superior performance in evaluations and provides an efficient, versatile solution for a wide range of practical applications.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant NO. 62122035)

References

1. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:2305.16311 (2023)
2. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)
3. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR. pp. 18208–18218 (2022)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. arXiv preprint arXiv:2301.13826 (2023)
8. Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. arXiv preprint arXiv:2304.00186 (2023)
9. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
10. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8394–8403 (2020)
11. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
12. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing **29**, 4759–4771 (2020)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
15. Han, I., Yang, S., Kwon, T., Ye, J.C.: Highly personalized text embedding for image manipulation by stable diffusion. arXiv preprint arXiv:2303.08767 (2023)
16. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. arXiv preprint arXiv:2303.11305 (2023)
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
18. Hong, Y., Niu, L., Zhang, J.: Shadow generation for composite image in real-world scenes. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 914–922 (2022)

19. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023)
20. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276 (2022)
21. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
22. Li, P., Huang, Q., Ding, Y., Li, Z.: Layerdiffusion: Layered controlled image editing with diffusion models. arXiv preprint arXiv:2305.18676 (2023)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
24. Liu, D., Long, C., Zhang, H., Yu, H., Dong, X., Xiao, C.: 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. pp. 8139–8148 (2020)
25. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022)
26. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023)
27. Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
28. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
33. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
37. Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J.M., Chari, V.: Learning to generate synthetic data via compositing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 461–470 (2019)
38. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. arXiv preprint arXiv:2211.12572 (2022)
39. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
40. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)
41. Xue, B., Ran, S., Chen, Q., Jia, R., Zhao, B., Tang, X.: Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In: European Conference on Computer Vision. pp. 300–316. Springer (2022)
42. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
43. Zhang, L., Wen, T., Min, J., Wang, J., Han, D., Shi, J.: Learning object placement by inpainting for compositional data augmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 566–581. Springer (2020)
44. Zhang, L., Wen, T., Shi, J.: Deep image blending. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 231–240 (2020)
45. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
47. Zhang, Y., Dong, W., Tang, F., Huang, N., Huang, H., Ma, C., Lee, T.Y., Deussen, O., Xu, C.: Prospect: Expanded conditioning for the personalization of attribute-aware image generation. arXiv preprint arXiv:2305.16225 (2023)