Pacific Graphics 2024
R. Chen, T. Ritschel, and E. Whiting
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 7

CustomSketching: Sketch Concept Extraction for Sketch-based Image Synthesis and Editing

Chufeng Xiao[1,2] and Hongbo Fu[3,†]

[1] HKGAI, Hong Kong University of Science and Technology, Hong Kong, China
[2] School of Creative Media, City University of Hong Kong, Hong Kong, China
[3] Division of Arts and Machine Creativity, Hong Kong University of Science and Technology, Hong Kong, China
† Corresponding Author (hongbofu@ust.hk)

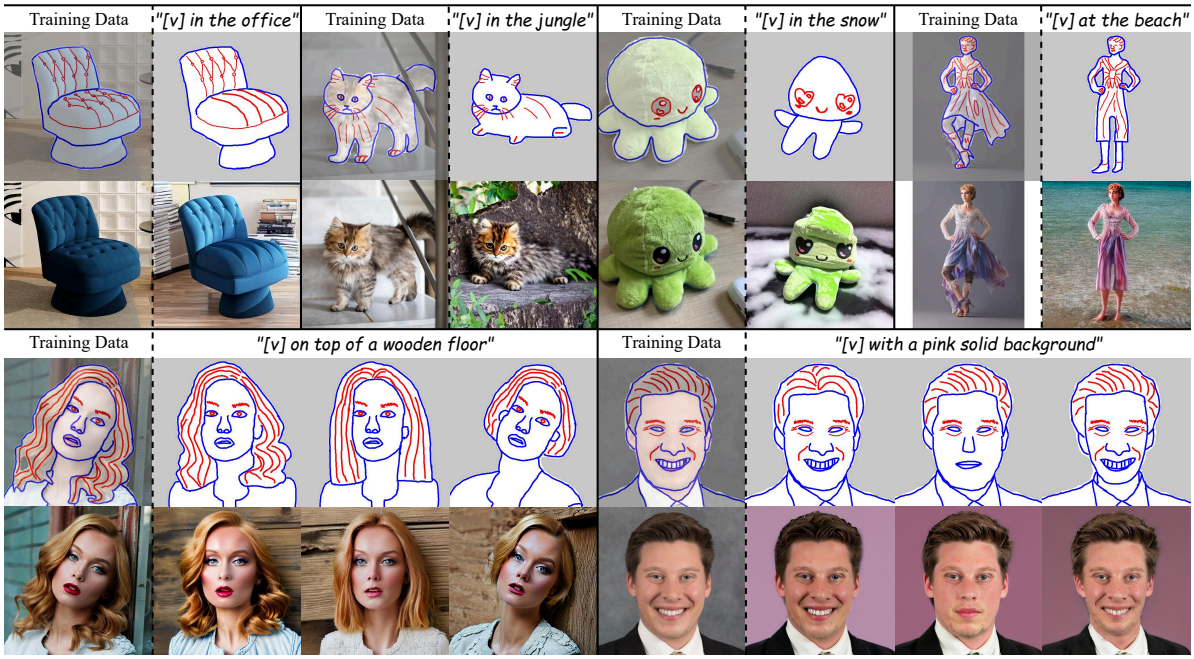


Figure 1: *Given one or several sketch-image pairs as training data, our* CustomSketching *can learn a novel sketch concept into a text token* $[v]$ *and specific sketches. We decompose a sketch into shape lines (blue strokes) and detail lines (red strokes) to reduce the ambiguity in a sketch. Users may input a text prompt and a dual-sketch to re-create or edit the concept at a fine-grained level.*

Abstract

Personalization techniques for large text-to-image (T2I) models allow users to incorporate new concepts from reference images. However, existing methods primarily rely on textual descriptions, leading to limited control over customized images and failing to support fine-grained and local editing (e.g., shape, pose, and details). In this paper, we identify sketches as an intuitive and versatile representation that can facilitate such control, e.g., contour lines capturing shape information and flow lines representing texture. This motivates us to explore a novel task of sketch concept extraction: given one or more sketch-image pairs, we aim to extract a special sketch concept that bridges the correspondence between the images and sketches, thus enabling sketch-based image synthesis and editing at a fine-grained level. To accomplish this, we introduce CustomSketching, a two-stage framework for extracting novel sketch concepts via few-shot learning. Considering that an object can often be depicted by a contour for general shapes and additional strokes for internal details, we introduce a dual-sketch *representation to reduce the inherent ambiguity in sketch depiction. We employ a shape loss and a regularization loss to balance fidelity and editability during optimization. Through extensive experiments, a user study, and several applications, we show our method is effective and superior to the adapted baselines.*

CCS Concepts

• ***Computing methodologies*** → ***Image manipulation;***

1. Introduction

The recent advent of large text-to-image (T2I) models [SCS*22, RPG*21, RBL*22] has opened up new avenues for image synthesis given text prompts. Based on such models, personalization techniques like [GAA*22, RLJ*23, KZZ*23] have been proposed to learn novel concepts on unseen reference images by fine-tuning the pre-trained models. Users can employ text prompts to create novel images containing the learned concepts in diverse contexts by leveraging the significant semantic priors of these powerful generative models.

However, the existing personalization methods fail to accurately capture the spatial features of target objects in terms of their geometry and appearance. This limitation arises due to their heavy reliance on textual descriptions during the image generation process. While some following works like [AAF*23, CHL*23b] have attempted to address this issue by incorporating explicit masks or additional spatial image features, they are still limited to providing precise controls and local editing on fine-grained object attributes (e.g., shape, pose, details) for the target concept solely through text.

To achieve fine-grained controls, sketches can serve as an intuitive and versatile handle for providing explicit guidance. T2I-Adapter [MWX*23] and ControlNet [ZRA23] have enabled the T2I models to be conditioned on sketches by incorporating an additional encoder network for sketch-based image generation. Such conditional methods perform well when an input sketch depicts the general contour of an object (e.g., the blue strokes in Figure 2 (b)). However, we observed they struggle to interpret and differentiate other types of sketches corresponding to specific local features in realistic images. As illustrated in Figure 2, these methods fail to correctly interpret detail lines for clothing folds and flow lines for hair texture (the red strokes in (b)). The primary reason behind the issue is that the sketch dataset used to train the conditional networks [MWX*23, ZRA23] is inherently ambiguous since it is generated automatically through edge detection on photo-realistic images. Consequently, directly incorporating a pre-trained sketch encoder with personalization techniques proves challenging when attempting to customize a novel concept guided by sketches.

Based on the aforementioned observation, we propose a novel task of sketch concept extraction for image synthesis and editing to tackle the issue of sketch ambiguity. The key idea is to empower users to define personalized sketches corresponding to specific local features in photo-realistic images, i.e., novel sketch-to-image mappings unseen by the pre-trained models. Users can sketch their desired concepts by first tracing upon one or more reference images and then manipulating the learned concepts by sketching, as shown in Figure 1.

To achieve sketch-based editability and identity preservation, we propose a novel personalization pipeline called *CustomSketching* for extracting sketch concepts. Note that this pipeline is a generic and few-shot method learned on one or several user-provided sketch-image pairs. It is built upon a pretrained T2I model and incorporates additional encoders to extract features from the sketch input. Since a single image may exhibit diverse local features corresponding to different types of sketches, we employ a dual-sketch representation via two sketch encoders to decouple shape and detail depiction. Our pipeline consists of two stages:

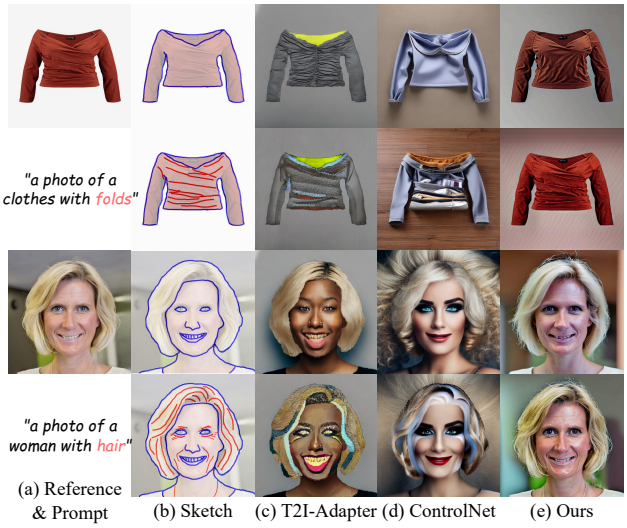


Figure 2: *Given a text prompt (a-bottom) and a sketch (b) depicting specific semantics (e.g., clothing folds and hair), T2I-adapter (c) and ControlNet (d) could not correctly interpret the out-of-domain sketch types, while our method can extract such a novel sketch concept and reconstruct the reference image (a-top). Note that the reference image is not used by T2I-adapter and ControlNet, which are a pre-trained model to sample new content but not to manipulate a given reference image, and their results are for reference only.*

in Stage I, we optimize a textual token for global semantics but freeze the weights of the sketch encoders. This stage is important to provide stable semantic priors to avoid overfitting on the limited (1-6) training samples and enable high sketch-editability for generic categories. In Stage II, we jointly fine-tune the weights of the sketch encoders and the learned token to reconstruct the reference images in terms of local appearance and geometry. To prevent overfitting, we perform data augmentation and introduce a shape loss for sketch-guided shape constraint and a regularization loss for textual prior preservation.

To the best of our knowledge, our method is the first work to extract sketch concepts using large T2I models, thus providing users with enhanced creative capabilities for editing real images. To evaluate our method, we collect a new dataset including sketch-image pairs and the edited sketches, where each sketch comprises a dual representation. Through qualitative and quantitative experiments, we demonstrate the superiority and effectiveness of *CustomSketching*, compared to the adapted baselines. Given the absence of a definitive metric to measure the performance of image editing, we conduct a user study to gather user insights and feedback. Additionally, we showcase several applications enabled by our work.

The contributions of our work can be summarized as follows. 1) We propose the novel task of sketch concept extraction. 2) We introduce a novel framework that enables a large T2I model to extract and manipulate a sketch concept via sketching, thereby improving its editability and controllability. 3) We create a new dataset for

comprehensive evaluations and demonstrate several sketch-based applications enabled by *CustomSketching*.

2. Related Work

Text-to-Image Synthesis and Editing. Text-to-image (T2I) generation has made significant strides in recent years, achieving remarkable performance. Early works [RAY*16, XZH*18, ZXL*17, ZXL*18, ZKB*21] employed RNN [CVMBB14, HS97] and GANs [GPAM*14, BDS18, KLA19] to control image generation, processing, and editing in specific scenarios, such as human faces [XYXW21], fashion [MMJ*20], and colorization [ZMG*19]. These works rely on well-prepared datasets tailored to the target scenarios, posing a bottleneck in dataset availability. To alleviate this limitation, subsequent studies [AZF*22, BAC*21, GPM*22, MTY*22, PWS*21, CBK*22] adopted CLIP [RKH*21], a large language-image representation model based on Transformer [VSP*17], to align image-text features and achieve robust performance in text-driven image manipulation tasks. Nonetheless, these approaches are still confined to limited domains, challenging their extension to other domains.

The emergence of diffusion models [HJA20, SE19, SME20, DN21, RPG*21, RBL*22] trained with large-scale image-text datasets allows for universal image generation from open-domain text, surpassing previous works based on GANs. Leveraging the power of diffusion models, several approaches have been proposed to manipulate images globally using text [BHE23, CVSC22, KZL*23, VKM*23, CWQ*23] and locally using masks [WSM*23, NDR*22, PGA*23]. For example, Mokady et al. [MHA*23] proposed an inversion method that first inverts a real image into latent representations, given which the method enables text-based image editing (e.g., changing local objects or modifying global image styles) by manipulating cross-attention maps [HMT*22]. Blended Diffusion [ALF22, AFL23] can merge an existing object into a real image. However, these approaches face challenges in modifying the fine-grained object attributes of real images due to the abstract nature of the text. Building upon Stable Diffusion [RBL*22], our method addresses this issue by incorporating sketches as an intuitive handle to manipulate real images. Inspired by [HMT*22], we introduce a shape loss that leverages cross-attention maps to provide guidance based on sketches.

Personalization Techniques. The personalization task is to produce image variations of a given concept in reference images. GAN-based methods for this task only focus on the same category (e.g., aligned faces) [RAP*21, NAH*22] or on a single image [VHZH21], and thus could not manipulate images in a new context. Most recently, diffusion-based methods based on T2I models optimize a new [GAA*22] or rare [RLJ*23] textual token to learn the novel concept and generate the concept in diverse contexts via text prompting. For fast personalization, many researchers [CHL*23a, GAA*23, JZC*23, SXLJ23, WZJ*23, CHL*23b] introduce a prior encoder with local and global mapping to save optimization time. For multi-concept personalization, Avrahami et al. [AAF*23] fine-tuned a set of new tokens and the weights of a denoising network from a single image given masks, while Kumari et al. [KZZ*23] optimized only several layers of the network based on a few images. Unlike these two methods, which need to fine-tune simultaneously the multi-concepts that are desired in generation, our method can separately extract sketch concepts for diverse targets and then work for multi-concept generation by plug-and-play without extra optimization (see Figure 11 (c)).

However, the existing personalization works do not allow precise control for novel concept generation and thus could not work for local or detailed editing (e.g., addition, removal, modification) of the learned concept. To address the issue, we introduce a new task of sketch concept extraction by optimizing sketch encoder(s) given one or more sketch-image pairs.

Sketch-based Image Synthesis and Editing. As an intuitive and versatile representation, sketch has been extensively explored to achieve fine-grained geometry control in realistic image synthesis and editing. For instance, Sangkloy et al. [SLF*17] utilized colored scribbles to depict geometry and appearance and synthesized images of various categories such as bedrooms, cars, and faces. Similarly, Chen et al. [CH18] employed freehand sketches to learn shape knowledge for diverse objects. Chen et al. [CSG*20, CLL*21] and Liu et al. [LCL*22] utilized line drawings for image synthesis, editing, and video editing of human faces. In SketchHairSalon [XYH*21], flow lines are used to represent unbraided hair, while contour lines depict braided hair. For local editing, a partial sketch has been adopted for minor image editing, e.g., FaceShop [PHS*18], Sketch2Edit [ZLP22], Draw2Edit [XGPS23]. Unlike the previous works that train dedicated networks for specific domains or limited object categories, our method is generic and few-shot, which can handle versatile sketches for image synthesis and editing using a pre-trained T2I model.

Recently, sketch-based T2I diffusion models have also been explored [WKLQ23, CCC*23, PZX*23]. Voynov et al. [VACO23] utilized sketches as a shape constraint for optimizing the latent map in a diffusion model, while T2I-Adapter [MWX*23] and ControlNet [ZRA23] are two concurrent works that train an external sketch encoder connected to a pre-trained diffusion model to enable sketch control. However, directly integrating these methods with personalization techniques may not accurately extract sketch concepts for all types of sketches (Figure 2), since the models [MWX*23, ZRA23] are biased towards training data, specifically edge maps automatically detected from images. We will establish this setup for the existing personalization methods as baselines to compare with our method, though we are the first to customize novel sketch concepts.

3. Method

Based on a pre-trained T2I diffusion model [RBL*22], our goal is to embed a new sketch concept into the model, enabling the synthesis and manipulation of diverse semantics in reference images through sketching and prompting (see Figure 1). To this end, we propose a novel few-shot framework, *CustomSketching*, which extracts a sketch concept from a small number of (one or several) reference images I and their corresponding sketches S. As illustrated in Figure 3, the framework comprises two training stages to reconstruct the reference image. During the training stage, the inputs include reference images I, the corresponding sketches S, and a templated text. Note that reference images I are only provided during training. During inference, users can flexibly control

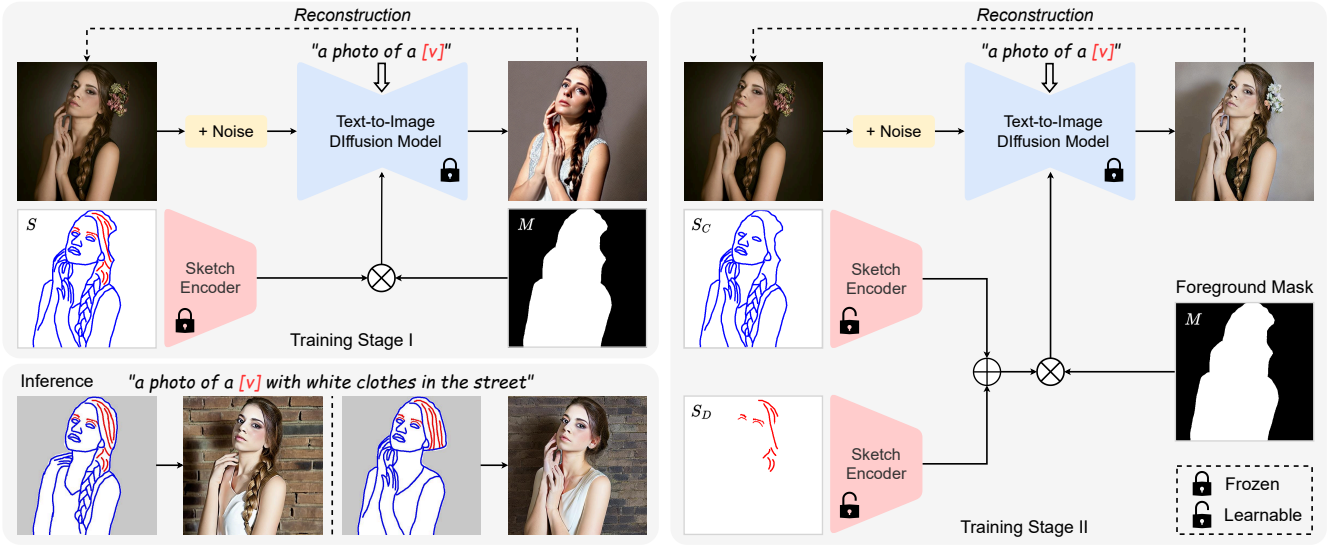


Figure 3: *The pipeline of our* CustomSketching*, which extracts novel sketch concepts for fine-grained image synthesis and editing via a two-stage framework. During training, given one or a few sketch-image pairs, Stage I only optimizes a textual embedding of a newly added token $[v]$ to represent the global semantics of the reference image(s), while Stage II jointly fine-tunes the token and two sketch encoders to reconstruct the concept in terms of local appearance and geometry. We adopt a dual-sketch representation to differentiate shape lines S_C and detail lines S_D. During inference, users may provide a text prompt and a dual-sketch to manipulate the learned concept.*

the generation of a target image that satisfies the context described by a text and faithfully reflects the input sketch in terms of geometry, without requiring reference images as input. In the following, we will describe the method in detail.

Two-stage Optimization. To leverage the robust textual prior of a large T2I model, following TI [GAA*22], we introduce a newly added textual token $[v]$ to capture global semantics while utilizing sketch representations through sketch encoder(s) to capture local features. Directly incorporating the personalization method [GAA*22] with a pre-trained encoder like [MWX*23] could not fully restore the local geometry and appearance of the target image (see the results by TI-E in Figure 5). It is because it fine-tunes merely textual embedding v for the token $[v]$. However, through joint optimization of the textual embedding and the weights of the sketch encoder(s), we encountered challenges in disentangling the global and local representations, resulting in unsatisfactory reconstruction (see Supp). To focus on learning separate features, inspired by [AAF*23], we adopt a two-stage optimization strategy. In Stage I, we optimize the textual embedding while freezing the weights of a pre-trained sketch encoder [MWX*23], establishing a pivotal initialization for the next stage. In Stage II, we jointly fine-tune the embedding and two sketch encoders to recover the target identity. Note that, in both stages, we freeze the denoising network of the pre-trained model to preserve its prior knowledge for editing.

Dual Sketch Representations. In Stage I, we fix the local features from sketches to guide the learning of the global textual embedding. To employ the prior knowledge of the sketch encoder [MWX*23], which was pre-trained on a large-scale sketch dataset, we input a binary sketch (where blue and red lines in Figure 3 are represented as black and the background as white) similar to the input used during pre-training. However, this single sketch representation inherently contains ambiguity since it combines the major contour sketch (blue lines, denoted as S_C) indicating the general shape with other minor types of sketches (red lines, denoted as S_D) capturing internal details (e.g., hair flow, clothes fold, wrinkles). This inherent ambiguity is the primary factor that biases the pre-trained sketch encoder [MWX*23, ZRA23] towards general shape, as illustrated in Figure 2. Therefore, in Stage II, optimizing the weights of a sketch encoder using the single-sketch representation would still result in ambiguous image editing (see Section 4.4).

To address this issue, we propose using a dual-sketch representation that decomposes a given sketch S into two distinct types of sketches, namely S_C and S_D as mentioned above, for Stage II. Instead of merging S_C and S_D into a single map and feeding it into a single encoder (see Supp), we employ two separate sketch encoders to extract features corresponding to each type of sketch individually. This configuration enables us to capture more distinct and recognizable features for S_C and S_D, resulting in plausible performance in decomposing shape and details, compared to the setting of the single-sketch representation. The features extracted from both types of sketches are aggregated through summation before being injected into the pre-trained T2I model.

Masked Encoder. As our focus is sketching the concept in the foreground, the sketch map S often contains significant blank areas representing the background. Therefore, fine-tuning the sketch encoder(s) on the entire map would lead to overfitting the background regions not represented in the sketch, consequently undermining the text-guided editability of the T2I model (see Figure 8). To address it, we apply a foreground mask M to remove the background features extracted from the encoder(s). The foreground mask can

either be generated automatically by filling a convex polygon following S_C, or be manually drawn by users. In summary, the sketch features are passed into the T2I model $\mathcal{F}_m$ along with a prompt p_v containing the token $[v]$ to derive the fused features $\mathcal{F}$. For Stage I, we denote it as:

$$\widehat{\mathcal{F}}^i = \mathcal{F}_e^i(S) \cdot M^i + \mathcal{F}_m^i(p_v), i \in \{1,2,3,4\}, \tag{1}$$

while for Stage II:

$$\widehat{\mathcal{F}}^i = (\mathcal{F}_c^i(S_C) + \mathcal{F}_d^i(S_D)) \cdot M^i + \mathcal{F}_m^i(p_v), \tag{2}$$

where $\mathcal{F}_e^i(S)$ is the i-th layer sketch feature extracted by the pre-trained encoder [MWX*23], while $\mathcal{F}_c^i(S_C)$ and $\mathcal{F}_d^i(S_D)$ are dual sketch representations from the fine-tuned encoders, and M_i is the resized mask fit to the feature size. We adopt four layers of the features as used in [MWX*23].

Loss Function. To optimize the sketch concept, which involves the embedding v and the weights of $\mathcal{F}_c$ and $\mathcal{F}_d$, we combine three types of losses for the text- and sketch-based problem. Firstly, we utilize a classic diffusion loss with the foreground mask M to reconstruct the target image regarding appearance and geometry. This loss encourages the optimization to concentrate on the foreground object depicted by the sketches, formulated as

$$\mathcal{L}_{rec} = \mathbb{E}_{z,t,v,\mathcal{F}_S,\epsilon}\left[\|\epsilon \cdot M - \epsilon_\theta(z_t, t, p_v, \mathcal{F}_S) \cdot M\|\right], \tag{3}$$

where $\mathcal{F}_S$ denotes the sketch features in the two stages, and ϵ_θ is the denoising network of the T2I model. At each optimization step, we randomly sample a timestep t from $[0,T]$ and add noise $\epsilon \sim \mathcal{N}(0,1)$ to the image latent z_0 to be z_t.

However, relying solely on the masked diffusion loss may not provide sufficient constraints to ensure the faithfulness between the sketch and the generated image. For example, as shown in Figure 4 (w/o Shape Loss), unexpected elements (monster doll) or an incorrect pose (cat) without following the sketches would be produced in the generated results. Motivated by previous works [HMT*22, AAF*23, CAV*23] that leverage cross-attention maps of the T2I model to control the layout and semantics of the target, we propose a shape loss based on the cross-attention map of the token $[v]$. The shape loss $\mathcal{L}_{shape}$ comprises a foreground loss for guiding the concept shape to align with the sketch depiction via M, and a background loss for penalizing foreground pixels that violate the background region. We denote the shape loss as:

$$\mathcal{L}_{fg} = \|norm(A_\theta(z_t, v)) \cdot M - M\|, \tag{4}$$

$$\mathcal{L}_{bg} = mean(A_\theta(z_t, v) \cdot (1 - M)), \tag{5}$$

$$\mathcal{L}_{shape} = \mathcal{L}_{fg} + \mathcal{L}_{bg}, \tag{6}$$

where $A_\theta(z_t, v)$ is the cross-attention maps given the latent z_t and token $[v]$. $norm(\cdot)$ is to normalize the attention map to $[0,1]$, while $mean(\cdot)$ computes the average attention value of background pixels. As illustrated in Figure 4, the visualized attention maps indicates the shape loss can encourage the generated results to be faithful to the input sketches.

In addition, the two-stage optimization may cause the fine-tuned embedding v to increase too large so that it overfits the reference shape, thus damaging the sketch editability (see Figures 7 & 8).

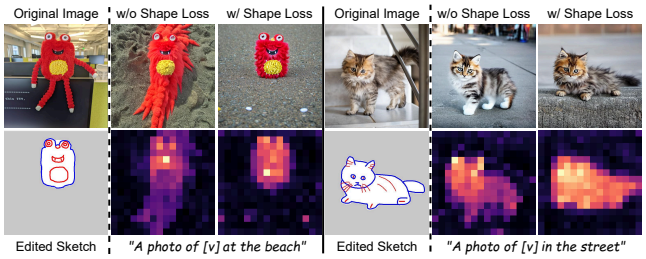


Figure 4: *Two examples comparing the results by the methods with and without the shape loss. In each group, the right part shows the generated result (Top) given the edited sketch as well as a text, and the visualized cross-attention map (Bottom) corresponding to the textual token* $[v]$.

We, therefore, introduce a regularization loss for the embedding via an $L2$ norm:

$$\mathcal{L}_{reg} = \|v\|. \tag{7}$$

In total, the loss function for the two stages is:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{shape}\mathcal{L}_{shape} + \lambda_{reg}\mathcal{L}_{reg}, \tag{8}$$

where we set the weights λ_{shape} as 0.01 and λ_{reg} as 0.001 empirically.

Implementation Details. To avoid the method overfitting limited training images, we adopt on-the-fly augmentation tricks (horizontal flip, translation, rotation) on the sketch-image pairs during optimization. Please find more implementation details in Supp.

4. Experiments

We have conducted extensive evaluations to quantitatively and qualitatively evaluate our method *CustomSketching*. We first show the comparisons between our method and the personalization baselines adapted to our proposed task. Then, we evaluate the effectiveness of our settings via an ablation study. We further conduct a perceptive user study on the edited results by the compared methods. In addition, we implement several applications based on our method to show the usefulness of the extracted sketch concepts. Please find more details, comparisons, and results in Supp and video demo.

Dataset. Before comparisons, we prepare a dataset of image-sketch pairs covering diverse categories (e.g., toys, human portraits, pets, buildings). We first collect images from the personalization works [GAA*22, KZZ*23] and the sketch-based work [XYH*21]. Next, we invite three normal users without any professional training in drawing to trace the images with separate contour lines S_C and detail lines S_D and then edit several sketches initialized with one of the traced sketches to depict a target object by changing its shape, pose, and/or details. Following the general instruction that S_C depicts a coarse shape while S_D is inside the shape, users decided S_C and S_D by themselves and drew them consistently for training and testing to personalize the sketch concept. Note that the training and testing data for a given concept are annotated by the

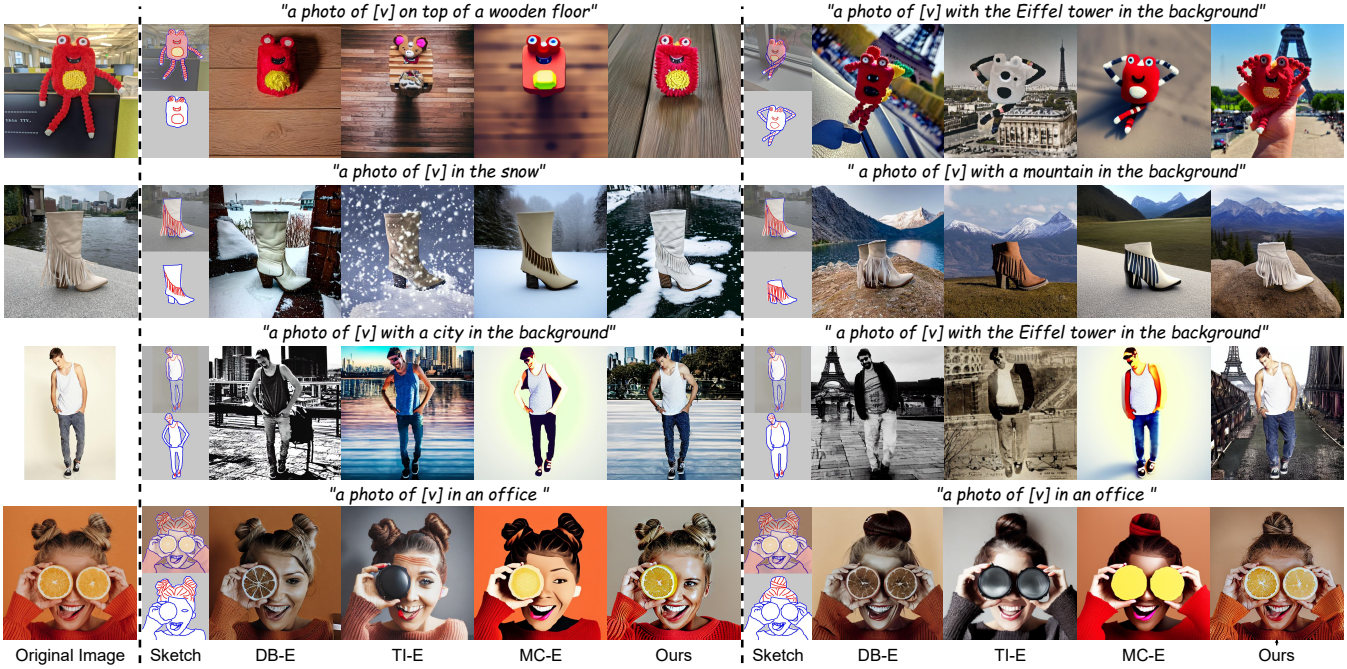


Figure 5: *Comparisons of the results generated by our method and three adapted baselines, given the same text prompt and sketch. In the sketch column, the top one is the annotated sketch corresponding to the original image for training while the bottom one is an edited sketch.*

same user to ensure consistency in concept manipulation. With the soft shape/detail instruction, the three users might have slightly different preferences for personalization. For example, one user might regard the contour of a human body as shape lines and the inside clothes as detail lines, while another user prefers the inside texture as detail lines. Since our method learns each personalized concept separately, the inter-user preference diversity is acceptable. The settings offer users a balance of freedom and controllability (see Sec. 4.3).

Finally, we obtain 35 groups of concept data. Each concept has 1-6 image-sketch pair(s) and 3-5 edited sketches. In total, the dataset contains 102 traced sketches with the corresponding images for training and 159 edited sketches without paired images. Moreover, we employ ten prompt templates for each concept, e.g., "a photo of $[v]$ at the beach", similar to [AAF*23]. Thus, the dataset includes 2,610=(102+159)×10 sketch-text pairs (see Supp) for evaluation.

Metrics. We utilize prompt similarity, identity similarity, and perceptual distance as evaluation metrics. Following the prior work [AAF*23], the prompt similarity assesses the distance between a text prompt and the corresponding produced images using CLIP model [RKH*21]. For computing, the learned token $[v]$ in the prompt is replaced with its class, e.g., "a $[v]$ in the office" is modified to "a woman in the office". The identity similarity measures how the method preserves the object identity of the original image when the context by text or the structure by sketch is changed. We compute the metrics via DINO [CTM*21] features as Ruiz et al. [RLJ*23] did. Additionally, we evaluate the perceptual distance via the LPIPS metric [ZIE*18] for the reconstruction error regarding appearance and geometry between the ground truth and the generated images given the traced sketches. For identity similarity and perceptual similarity, we adopt the masked version of the results and ground truth to focus on the foreground parts depicted by sketches. Note that we evaluate prompt and identity similarity on all the sketch-text pairs while computing perceptual similarity only on the traced sketches with their paired images.

4.1. Comparison

To our knowledge, we are the first work to extract sketch concepts for image synthesis and editing. To fairly compare our method with the existing personalization techniques, we adapt two methods, TI [GAA*22] and DB [RLJ*23], to fit our proposed task by introducing a pre-trained sketch encoder [MWX*23] into their methods when training and testing. Note that we do not optimize the weights of the encoder for the two methods to keep their method intact mostly, and we thus only use a single masked encoder to preserve the pre-trained prior. The two methods receive the dual-sketch representation encoded in one map (i.e., 255 for S_C and 127 for S_D) with a mask to have the same inputs as ours. Besides the tuning-based methods, we also compare our method with a tuning-free method, MasaCtrl [CWQ*23], which can work for sketch-based editing. We directly adopt their released code that integrates the sketch encoder [MWX*23] for comparison. All the compared methods are based on SD v1.5 [RBL*22]. For simplicity, we refer to the three baselines as TI-E, DB-E, and MC-E.

Figure 5 shows a qualitative comparison between our method and the baselines. Although the editing results by the three baselines are generally faithful to the structure of the edited sketches,

Table 1: *Quantitative comparisons for diverse methods.*

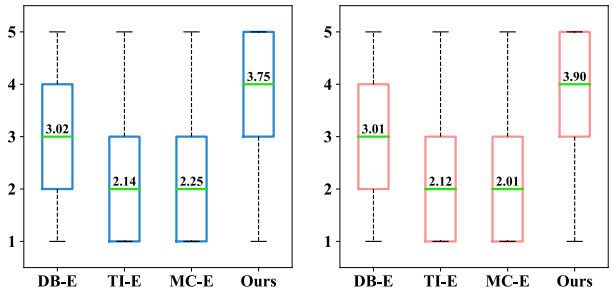
Method	Prompt ↑	Identity ↑	Perceptual ↓
DB-E	0.641	0.889	0.182
TI-E	**0.642**	0.867	0.214
MC-E	0.633	0.884	0.160
Single-sketch	0.622	0.908	0.146
w/o $\mathcal{L}_{shape}$	0.639	0.906	0.150
w/o $\mathcal{L}_{reg}$	0.618	0.909	0.142
w/o Masked $\mathcal{F}$	0.620	0.911	0.141
Ours	0.632	**0.912**	**0.134**

they could not preserve the identity or style of the objects/subjects in the original images. Specifically, DB-E can reconstruct the original images with sketches generally (see Supp), but when editing, it often loses the details depicted by the edited sketch and the correspondence between the sketch and target concept defined by the training sketch-image pairs. TI-E cannot recover the original identity in both reconstruction and editing since it merely optimizes high-level text embedding. MC-E tends to drift the result's style from the original one. It is because a) MC adopts a pre-trained sketch encoder with domain bias as discussed in Sec. 1, and thus it could not work well for novel sketch concept; b) this training-free method edits a real image by inverting it to a latent space to leverage the generative prior of a T2I model, but there is a domain gap between the generated images and real images. Our method outperforms the three baselines and maintains the original identity and the sketch-image correspondence defined in the sketch concept.

Table 1 presents the quantitative evaluation results in the three metrics. It demonstrates our method achieves the best identity preservation (identity similarity) and reconstruction quality (perceptual distance). However, our method sacrifices slightly the prompt similarity since we focus on the reconstruction of the foreground object with $\mathcal{L}_{shape}$ (see the ablation study without $\mathcal{L}_{shape}$). Such sacrifice is acceptable to trade off the concept re-creation and sketch faithfulness, as shown in Figures 5 and 6.

4.2. Perceptive User Study

To further compare our method and the three baselines, we performed a perceptive user study including two evaluations: text editability study and sketch editability study. We first prepared a subset (30 randomly picked concepts, 15 for text editability, 15 for sketch editability) of our collected dataset. For text editability, we produced the results by the three baselines and our method given a traced sketch and a prompt randomly picked from one concept. A participant was given a reference image, a prompt (e.g., "a photo of the boots in the reference image in the snow"), and the four generated results in random order. We asked the participants to rate "How the result is consistent with the prompt" on a Likert scale of 1–5 (the higher, the better). For sketch editability, we presented each participant with a reference image with the traced sketch, an edited sketch, and four results (in random order) and required them to rate "How the result is faithful to the edited sketch and consistent with the reference identity". From 40 participants, we received 600 responses for each method in each evaluation. As shown in Figure 6, the user study reflects the superiority of our method to the baselines in both evaluations.



(a) Text Editability (b) Sketch Editability

Figure 6: *Box plots of the ratings in the perceptive user study. Each value above the median line is the average rate for each method. The higher, the better.*

We also conducted one-way ANOVA tests on the rating results and found a significant difference among the four methods for text editability (F=203.21, p<0.001) and sketch editability (F=307.86, p<0.001). The further paired T-tests (with p<0.001) show our method got a significantly higher rating in the text editability term than all the other methods, i.e., DB (t=10.86), TI (t=23.96), and MC (t=25.11). In terms of sketch editability, our method also outperformed the competitive methods, i.e., DB (t=14.12), TI (t=25.76), and MC (t=26.89).

4.3. Usability Study

Note that three novice users were invited to trace the images, edit their sketches for data collection, and participate in the perceptive study. To further evaluate the applicability and generalizability of our system, we conducted a usability study by collecting more feedback from them. For each user, we randomly showed 20 groups of concepts of his/her tracing and editing, a prompt, and the generated image. Then, we asked them to fill in a questionnaire of a customized five-point System Usability Scale (SUS, 1=strongly disagree to 5=strongly agree). The questions in the SUS include: 1) Easy-to-use: This system is easy to use; 2) Quality: The results look very natural; 3) Generalizability: I can manipulate diverse objects with different edits via the system; 4) Intention Consistency: The edited images are consistent with my annotation intention of tracings.

Overall, our system was rated positive from every perspective, i.e., Easy-to-use (4.33), Quality (4.67), Generalizability (4.67), and Intention Consistency (4.33). We also asked for the users' opinions (e.g., limitations, improvement) on our system. Although they commended our system for allowing their desired editing without too much effort, one of them thought it would work for daily use if they could obtain the results in real-time after tracing and editing.

4.4. Ablation Study

We ablated one of the key settings of our method to validate their effectiveness, including 1) w/ single-sketch representation; 2) w/o

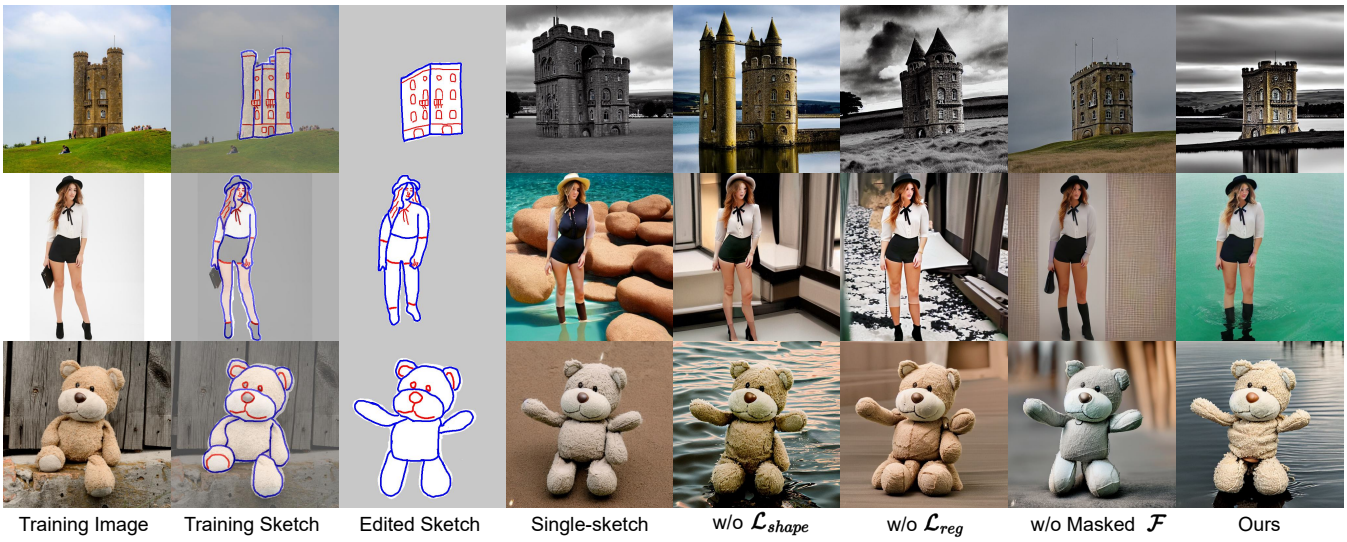


Figure 7: *Comparisons of our results and those by the ablated variants, given the text prompt "A photo of* $[v]$ *floating on top of water".*

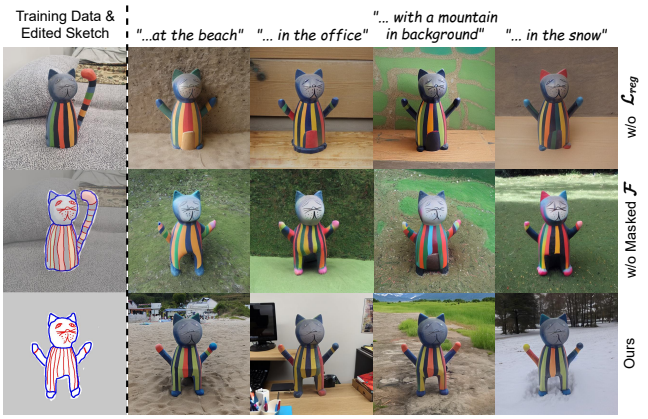


Figure 8: *Comparisons of the results by ours and the ablated variants using one edited sketch and diverse prompts indicating different contexts. The prefix of the prompt is "A photo of* $[v]$ *...".*

shape loss $\mathcal{L}_{shape}$; 3) w/o regularization loss $\mathcal{L}_{reg}$; 4) w/o masked encoder $\mathcal{F}$. As shown in Figure 7, using the single-sketch representation could not provide sufficient constraints on shapes (e.g., the castle and bear toy) and details (e.g., the woman's clothes), damaging the identity preservation. Removing $\mathcal{L}_{shape}$ would produce redundant parts and weaken the concept reconstruction. Without $\mathcal{L}_{reg}$, the method would overfit to the original shape and worsen the sketch editability (see Figures 7 & 8). Additionally, removing either $\mathcal{L}_{reg}$ or the masked $\mathcal{F}$ would affect a lot the text editability for background, shown as Figure 8. It is because $\mathcal{L}_{reg}$ can prevent the global embedding from enlarging significantly to outweigh the background token, while the masked $\mathcal{F}$ can filter out the local background features from the empty region of the sketch. The quantitative results in Table 1 further confirm the above conclusions.

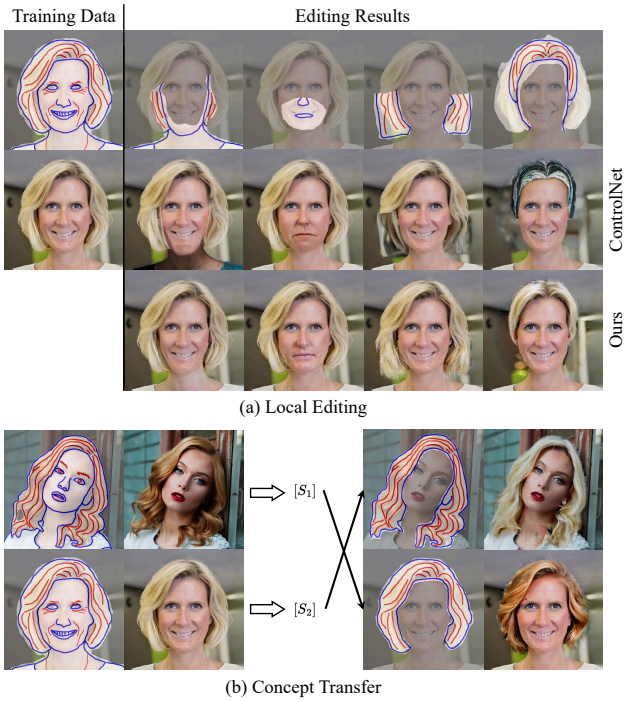


Figure 9: *Two applications, (a) local editing and (b) concept transfer, enabled by our* CustomSketching. *The text prompt for the results is "A photo of a* $[v]$*".*

4.5. Applications

We implemented four applications enabled by our method: local editing, concept transfer, multi-concept generation, and text-based style variation. We showcase the applications in Figure 9-12 to

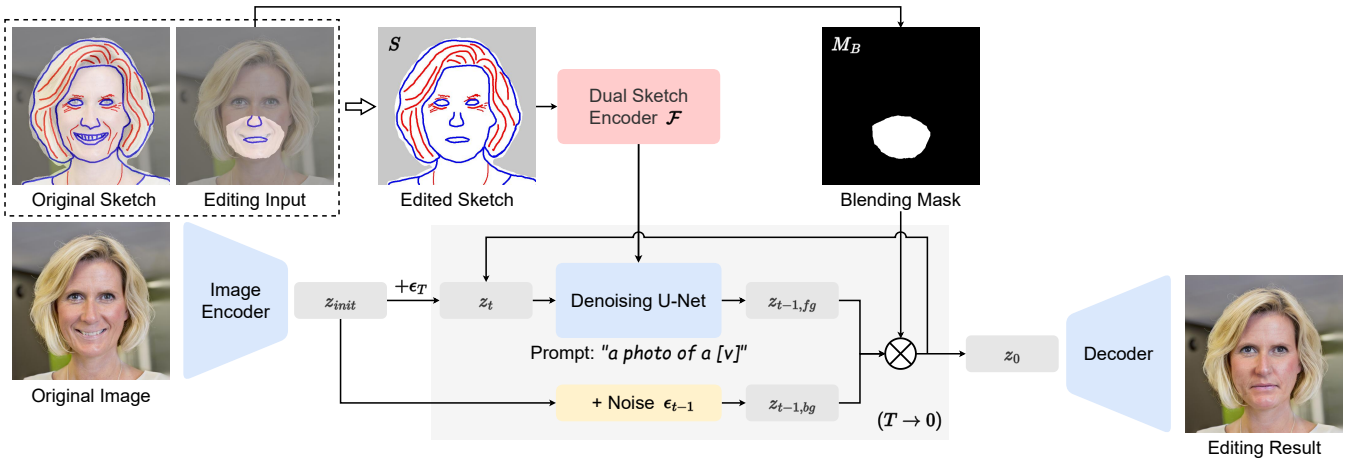


Figure 10: *The pipeline of local editing enabled by our method. Incorporating [AFL23], our method can be applied to local image editing, which allows users to edit a local region of a given real image via sketching while keeping the unedited region intact.*

demonstrate the effectiveness and versatility of *CustomSketching*. Please refer to Supp for the implementation details and more results for each application.

Local Editing. Our *CustomSketching* can be easily applied for local editing on the original images, including modification, addition, and removal. Users can edit the training sketch and provide a mask for the region they want to manipulate for fine-grained local editing. To keep the unedited region intact, we incorporate our method with an off-the-shelf local editing method by Avrahami et al. Figure 10 shows the pipeline of such an application. After extracting a novel concept $[S]$=$\{[v],\mathcal{F}\}$ given reference sketch-image pair(s), users can provide a blending mask M_B and a part sketch inside the mask to indicate an editing input. Then, our method blends the local sketch with the original sketch to be an edited sketch S fed into the learned dual-encoder $\mathcal{F}$. Given the extracted sketch feature and a prompt *"a photo of a* $[v]$*"*, the denoising U-Net produces a foreground latent, which is blended with the background latent inverted from the original image via M_B, to achieve the final editing result. The two latents are blended during all the inference time steps (T=50). We also compare our method with ControlNet [ZRA23] incorporated with [AFL23] for image local editing. As illustrated in Figure 9 (a), our method performs much better in preserving the identity and appearance of the original image than ControlNet [ZRA23], thanks to our two-stage optimization.

Concept Transfer. Given different concepts separately learned from the corresponding sketch-image pairs, our method can transfer between the concepts ($[S_i]$=$\{[v_i],\mathcal{F}_i\}$) with similar semantics via sketches. Figure 9 (b) shows an example of hairstyle transfer. Note that we also resort to [AFL23] for local transfer.

Multi-concept Generation. For multi-concept generation, prior works [KZZ*23, AAF*23] need to fine-tune the model on all the concepts desired in generation jointly. Unlike these works, which optimize the entire denoising network, we only optimize $[v]$ and $\mathcal{F}$ for one concept. This lightweight setting enables our method to achieve plug-and-play multi-concept generation by separately learning each concept and then combining them freely without extra optimization. Figure 11 presents two cases of the combinations among three extracted sketch concepts ($[S_1],[S_2],[S_3]$).

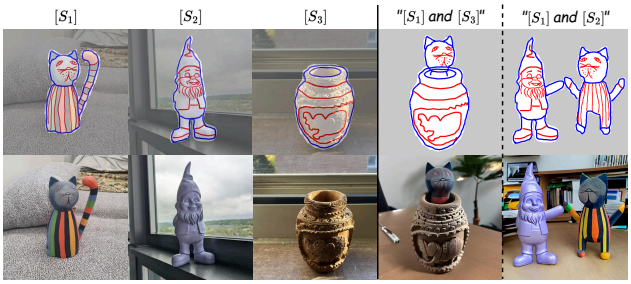


Figure 11: *The results of multi-concept generation enabled by our* CustomSketching. *Our method supports separately learning each concept and then combining them together without further optimization. The prompt is "A photo of* $\{[S_0],...,[S_i]\}$ *in an office".*

Text-based Style Variation. Our method decouples global semantics and local features of a reference image to a textual token $[v]$ and a sketch encoder $\mathcal{F}$. Thus, our method can be used to produce diverse style variations of the target object while preserving its geometry (shape and details), as shown in Figure 12. To this end, our method takes as input the sketch (regarded as an intermediate representation of object geometry) and a style prompt without $[v]$ (e.g., "a crayon drawing") to control the target style. We compare our method with PnP [TGBD23], a text-based image-to-image translation method in two cases, i.e., inputting the image without and with masking out the background, given a style prompt. Thanks to the extraction of a novel sketch concept, our method can better disentangle the geometry (depicted by a sketch) and style (depicted by a text), thus offering more user controllability and flexibility via sketching.

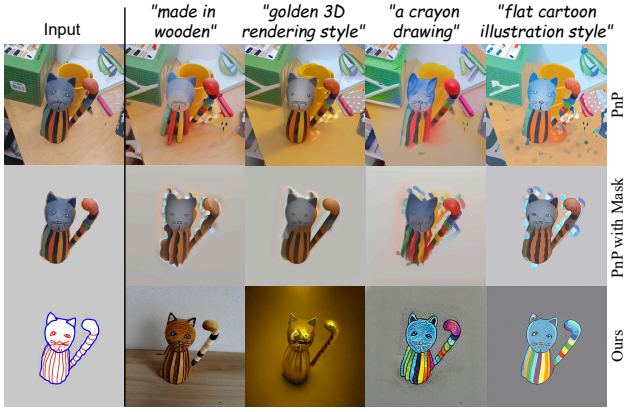


Figure 12: *Comparison of style variation enabled by our* CustomSketching *and PnP [TGBD23]. Our method can successfully decouple global semantic and local features of the reference image. Thus, it can vary the global style of a given object by directly inputting the style prompt like "made in wooden" without the textual token* [v].

4.6. Limitations

While our method improves the controllability and flexibility of the personalization task, it has several limitations. First, inherited from latent diffusion models, our method processes images in a low-resolution latent space (64×64). It thus struggles to control an object's tiny shape and details by sketching thin strokes. As shown in Figure 13, the car's details could not be changed following the edited sketch. Another limitation is the learning efficiency. Currently, our method requires almost 30-min optimization to learn one concept in few-shot, which limits practical applications. In the future, we may use fast personalization techniques [GAA*23, JZC*23] and extend our method to be zero-shot to address this issue, i.e., directly inputting pairs of reference images and sketches into an extra encoder to fast embed novel sketch concepts into the pre-trained models without optimization. Additionally, we asked novice users to trace and edit sketches for each concept to collect the evaluation dataset (Sec. 4). Note that tracing a sketch over an image costed 30s-2min, while editing a sketch took less than 1min. Although the users confirmed that the annotation did not require too much effort, the requirement of tracing and editing sketches for users is not negligible. To reduce users' load for tracing images, we might initialize a sketch via edge-detection methods for further annotation.

5. Conclusion

We proposed *CustomSketching*, a novel approach to extract sketch concepts for sketch-based image synthesis and editing based on a large T2I model. This method decouples reference image(s) into global semantics in a textual token and local features in two sketch encoders. We presented a dual-sketch representation to differentiate the shape and details of one concept. In this way, our method empowers users with high controllability in local and fine-grained image editing. Extensive experiments and several applications have shown the effectiveness and superiority of our proposed method to the alternative solutions. We will release the dataset and code to the research community.

Figure 13: *One failure case of our method. Our method could not change the car's tiny details by sketching thin strokes.*

6. Acknowledgments

We thank the anonymous reviewers for their constructive comments and user study participants for their time and effort. This project is partially supported by the Hong Kong Generative AI Research and Development Center (HKGAI) and a Start-up Project (Project No. R9913) from the Hong Kong University of Science and Technology.

References

[AAF*23] 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, 3, 4, 5, 6, 9

[AFL23] AVRAHAMI O., FRIED O., LISCHINSKI D.: Blended latent diffusion. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–11. 3, 9

[ALF22] AVRAHAMI O., LISCHINSKI D., FRIED O.: Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18208–18218. 3

[AZF*22] ABDAL R., ZHU P., FEMIANI J., MITRA N., WONKA P.: Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings* (2022), pp. 1–9. 3

[BAC*21] BAU D., ANDONIAN A., CUI A., PARK Y., JAHANIAN A., OLIVA A., TORRALBA A.: Paint by word. *arXiv preprint arXiv:2103.10951* (2021). 3

[BDS18] BROCK A., DONAHUE J., SIMONYAN K.: Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations* (2018). 3

[BHE23] 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* (2023), pp. 18392–18402. 3

[CAV*23] CHEFER H., ALALUF Y., VINKER Y., WOLF L., COHEN-OR D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–10. 5

[CBK*22] CROWSON K., BIDERMAN S., KORNIS D., STANDER D., HALLAHAN E., CASTRICATO L., RAFF E.: Vqgan-clip: Open domain image generation and editing with natural language guidance. Springer, pp. 88–105. 3

[CCC*23] CHENG S.-I., CHEN Y.-J., CHIU W.-C., TSENG H.-Y., LEE H.-Y.: Adaptively-realistic image generation from stroke and sketch with diffusion model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), pp. 4054–4062. 3

[CH18] CHEN W., HAYS J.: Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 9416–9425. 3

[CHL*23a] 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. 3

[CHL*23b] 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). 2, 3

[CLL*21] CHEN S.-Y., LIU F.-L., LAI Y.-K., ROSIN P. L., LI C., FU H., GAO L.: Deepfaceediting: deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–15. 3

[CSG*20] CHEN S.-Y., SU W., GAO L., XIA S., FU H.: Deepfacedrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 72–1. 3

[CTM*21] 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* (2021), pp. 9650–9660. 6

[CVMBB14] CHO K., VAN MERRIËNBOER B., BAHDANAU D., BENGIO Y.: On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014). 3

[CVSC22] COUAIRON G., VERBEEK J., SCHWENK H., CORD M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations* (2022). 3

[CWQ*23] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (October 2023), pp. 22560–22570. 3, 6

[DN21] DHARIWAL P., NICHOL A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems 34* (2021), 8780–8794. 3

[GAA*22] 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. In *International Conference on Learning Representations* (2022). 2, 3, 4, 5, 6

[GAA*23] GAL R., ARAR M., ATZMON Y., BERMANO A. H., CHECHIK G., COHEN-OR D.: Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–13. 3, 10

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. *Advances in neural information processing systems 27* (2014). 3

[GPM*22] GAL R., PATASHNIK O., MARON H., BERMANO A. H., CHECHIK G., COHEN-OR D.: Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG) 41*, 4 (2022), 1–13. 3

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems 33* (2020), 6840–6851. 3

[HMT*22] HERTZ A., MOKADY R., TENENBAUM J., ABERMAN K., PRITCH Y., COHEN-OR D.: Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations* (2022). 3, 5

[HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780. 3

[JZC*23] 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). 3, 10

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4401–4410. 3

[KZL*23] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 6007–6017. 3

[KZZ*23] 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* (2023), pp. 1931–1941. 2, 3, 5, 9

[LCL*22] LIU F.-L., CHEN S.-Y., LAI Y., LI C., JIANG Y.-R., FU H., GAO L.: Deepfacevideoediting: sketch-based deep editing of face videos. *ACM Transactions on Graphics (TOG) 41*, 4 (2022), 167. 3

[MHA*23] 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* (2023), pp. 6038–6047. 3

[MMJ*20] MEN Y., MAO Y., JIANG Y., MA W.-Y., LIAN Z.: Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 5084–5093. 3

[MTY*22] MOKADY R., TOV O., YAROM M., LANG O., MOSSERI I., DEKEL T., COHEN-OR D., IRANI M.: Self-distilled stylegan: Towards generation from internet photos. In *ACM SIGGRAPH 2022 Conference Proceedings* (2022), pp. 1–9. 3

[MWX*23] MOU C., WANG X., XIE L., ZHANG J., QI Z., SHAN Y., QIE X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023). 2, 3, 4, 5, 6

[NAH*22] NITZAN Y., ABERMAN K., HE Q., LIBA O., YAROM M., GANDELSMAN Y., MOSSERI I., PRITCH Y., COHEN-OR D.: Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG) 41*, 6 (2022), 1–10. 3

[NDR*22] NICHOL A. Q., DHARIWAL P., RAMESH A., SHYAM P., MISHKIN P., MCGREW B., SUTSKEVER I., CHEN M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning* (2022), PMLR, pp. 16784–16804. 3

[PGA*23] PATASHNIK O., GARIBI D., AZURI I., AVERBUCH-ELOR H., COHEN-OR D.: Localizing object-level shape variations with text-to-image diffusion models. 3

[PHS*18] PORTENIER T., HU Q., SZABÓ A., BIGDELI S. A., FAVARO P., ZWICKER M.: Faceshop: deep sketch-based face image editing. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 1–13. 3

[PWS*21] PATASHNIK O., WU Z., SHECHTMAN E., COHEN-OR D., LISCHINSKI D.: Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 2085–2094. 3

[PZX*23] PENG Y., ZHAO C., XIE H., FUKUSATO T., MIYATA K.: Difffacesketch: High-fidelity face image synthesis with sketch-guided latent diffusion model. *arXiv preprint arXiv:2302.06908* (2023). 3

[RAP*21] RICHARDSON E., ALALUF Y., PATASHNIK O., NITZAN Y., AZAR Y., SHAPIRO S., COHEN-OR D.: Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2287–2296. 3

[RAY*16] REED S., AKATA Z., YAN X., LOGESWARAN L., SCHIELE B., LEE H.: Generative adversarial text to image synthesis. In *International conference on machine learning* (2016), PMLR, pp. 1060–1069. 3

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P.,

OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695. 2, 3, 6

[RKH*21] 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* (2021), PMLR, pp. 8748–8763. 3, 6

[RLJ*23] 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* (2023), pp. 22500–22510. 2, 3, 6

[RPG*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-shot text-to-image generation. In *International Conference on Machine Learning* (2021), PMLR, pp. 8821–8831. 2, 3

[SCS*22] 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* (2022), 36479–36494. 2

[SE19] SONG Y., ERMON S.: Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems 32* (2019). 3

[SLF*17] SANGKLOY P., LU J., FANG C., YU F., HAYS J.: Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 5400–5409. 3

[SME20] SONG J., MENG C., ERMON S.: Denoising diffusion implicit models. In *International Conference on Learning Representations* (2020). 3

[SXLJ23] SHI J., XIONG W., LIN Z., JUNG H. J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411* (2023). 3

[TGBD23] TUMANYAN N., GEYER M., BAGON S., DEKEL T.: Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1921–1930. 9, 10

[VACO23] VOYNOV A., ABERMAN K., COHEN-OR D.: Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023), pp. 1–11. 3

[VHZH21] VINKER Y., HORWITZ E., ZABARI N., HOSHEN Y.: Image shape manipulation from a single augmented training sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 13769–13778. 3

[VKM*23] VALEVSKI D., KALMAN M., MOLAD E., SEGALIS E., MATIAS Y., LEVIATHAN Y.: Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–10. 3

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *Advances in Neural Information Processing Systems 30* (2017). 3

[WKLQ23] WANG Q., KONG D., LIN F., QI Y.: Diffsketching: Sketch control image synthesis with diffusion models. 3

[WSM*23] WANG S., SAHARIA C., MONTGOMERY C., PONT-TUSET J., NOY S., PELLEGRINI S., ONOE Y., LASZLO S., FLEET D. J., SORICUT R., ET AL.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18359–18369. 3

[WZJ*23] 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. 3

[XGPS23] XU Y., GUO R., PAGNUCCO M., SONG Y.: Draw2edit: Mask-free sketch-guided image manipulation. In *Proceedings of the 31st ACM International Conference on Multimedia* (2023), pp. 7205–7215. 3

[XYH*21] XIAO C., YU D., HAN X., ZHENG Y., FU H.: Sketchhairsalon: deep sketch-based hair image synthesis. *ACM Transactions on Graphics (TOG) 40*, 6 (2021), 1–16. 3, 5

[XYXW21] XIA W., YANG Y., XUE J.-H., WU B.: Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2256–2265. 3

[XZH*18] XU T., ZHANG P., HUANG Q., ZHANG H., GAN Z., HUANG X., HE X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 1316–1324. 3

[ZIE*18] 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/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 586–595. 6

[ZKB*21] ZHANG H., KOH J. Y., BALDRIDGE J., LEE H., YANG Y.: Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 833–842. 3

[ZLP22] ZENG Y., LIN Z., PATEL V. M.: Sketchedit: Mask-free local image manipulation with partial sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5951–5961. 3

[ZMG*19] ZOU C., MO H., GAO C., DU R., FU H.: Language-based colorization of scene sketches. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–16. 3

[ZRA23] 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* (2023), pp. 3836–3847. 2, 3, 4, 9

[ZXL*17] ZHANG H., XU T., LI H., ZHANG S., WANG X., HUANG X., METAXAS D. N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2017), pp. 5907–5915. 3

[ZXL*18] ZHANG H., XU T., LI H., ZHANG S., WANG X., HUANG X., METAXAS D. N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence 41*, 8 (2018), 1947–1962. 3