MultiBooth: Towards Generating All Your Concepts in an Image from Text

Chenyang Zhu[1,*], **Kai Li**[2,*,†], **Yue Ma**[3], **Chunming He**[4], **Xiu Li**[1,†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Meta Platforms, Inc., USA
[3] The Hong Kong University of Science and Technology, Hong Kong
[4] Duke University, Durham, USA
{chenyangzhu.cs, li.gml.kai, mayuefighting, chunminghe19990224}@gmail.com, li.xiu@mails.tsinghua.edu.cn

Abstract

This paper introduces MultiBooth, a method that generates images from texts containing various concepts from users. Despite diffusion models bringing significant advancements for customized text-to-image generation, existing methods often struggle with multi-concept scenarios due to low concept fidelity and high inference cost. MultiBooth addresses these issues by dividing the multi-concept generation process into two phases: a single-concept learning phase and a multi-concept integration phase. During the single-concept learning phase, we employ a multi-modal image encoder and an efficient concept encoding technique to learn a concise and discriminative representation for each concept. In the multi-concept integration phase, we use bounding boxes to define the generation area for each concept within the cross-attention map. This method enables the creation of individual concepts within their specified regions, thereby facilitating the formation of multi-concept images. This strategy not only improves concept fidelity but also reduces additional inference cost. MultiBooth surpasses various baselines in both qualitative and quantitative evaluations, showcasing its superior performance and computational efficiency.

1 Introduction

The advent of diffusion models (Ramesh et al. 2022; Saharia et al. 2022; Nichol et al. 2021; He et al. 2023a, 2024c) has ignited a new wave in the text-to-image (T2I) task, leading to numerous novel methods (Hertz et al. 2022; Ye et al. 2023; Gu et al. 2023; Wang et al. 2024d,a,b). Despite their broad capabilities, users often desire to generate specific concepts such as beloved pets or personal items. These personal concepts are not captured during the training of large-scale T2I models due to their subjective nature, emphasizing the need for customized generation (Wei et al. 2023; Gal et al. 2023; Yan et al. 2023; Li et al. 2024; Zhu et al. 2024). Customized generation aims to create new variations of given concepts, including different contexts (e.g., beaches, forests) and styles (e.g., painting), based on just a few user-provided images (typically fewer than 5).

Recent customized generation methods either learn a concise token representation for each subject (Gal et al. 2022) or adopt a fine-tuning strategy to adapt the T2I model specifically for the subject (Ruiz et al. 2023). While these methods have achieved impressive results, they primarily focus on single-concept customization and struggle when users want to generate customized images for multiple subjects. This motivates the study of multi-concept customization (MCC).

Existing methods (Kumari et al. 2023) for MCC commonly employ joint training approaches. However, this strategy often leads to feature confusion. Furthermore, these methods require training distinct models for each combination of subjects and are hard to scale up as the number of subjects grows. An alternative method (Liu et al. 2023) addresses MCC by adjusting attention maps with residual token embeddings during inference. While this approach shows promise, it incurs a notable inference cost. Furthermore, the method encounters difficulties in attaining high fidelity due to the restricted learning capacity of a single residual embedding.

To address the aforementioned issues, we introduce MultiBooth, a two-phase MCC solution that accurately and efficiently generates customized multi-concept images based on user demand. MultiBooth includes a discriminative single-concept learning phase and a plug-and-play multi-concept integration phase. In the former phase, we learn each concept separately, resulting in a single-concept module for every concept. In the latter phase, we effectively combine these single-concept modules to generate multi-concept images without any extra training.

More concretely, we propose the Adaptive Concept Normalization (ACN) to enhance the representative capability of the generated customized embedding in the single-concept learning phase. We employ a trainable multi-model encoder to generate customized embeddings, followed by the ACN to adjust the L2 norm of these embeddings. Finally, by incorporating an efficient concept encoding technique, all detailed information of a new concept is extracted and stored in a single-concept module which contains a customized embedding and the efficient concept encoding parameters.

In the plug-and-play multi-concept integration phase, we further propose a regional customization module to guide the inference process, allowing the correct combination of

[*] Equal Contribution.
[†] Corresponding author.

Figure 1: MultiBooth can learn individual customization concepts through a few examples and then combine these learned concepts to create multi-concept images based on text prompts. The results indicate that our MultiBooth can effectively preserve high image fidelity and text alignment when encountering complex multi-concept generation demands, including (a) stylization, (b) different spatial relationships, and (c) contextualization.

different single-concept modules for multi-concept image generation. Specifically, we divide the attention map into different regions within the cross-attention layers of the U-Net, and each region's attention value is guided by the corresponding single-concept module and prompt. Through the proposed regional customization module, we can generate multi-concept images via any combination of single-concept modules while bringing minimal cost during inference. Fig. 1 shows some examples.

Our approach is extensively validated with various representative subjects, including pets, objects, scenes, etc. The results from both qualitative and quantitative comparisons highlight the advantages of our approach in terms of concept fidelity and prompt alignment capability. Our contributions are summarized as follows:

- We propose a novel framework named MultiBooth. It allows plug-and-play multi-concept generation after separate customization of each concept.
- The adaptive concept normalization is proposed in our MultiBooth to mitigate the problem of domain gap in the embedding space, thus learning a representative customized embedding. We also introduce the regional customization module to effectively combine multiple single-concept modules for multi-concept generation.
- Our method consistently outperforms current methods in terms of image quality, faithfulness to the intended concepts, and alignment with the text prompts.

2 Related Work

Layout-guided text to image generation. T2I models have benefited numerous new tasks (Ma et al. 2024b,c, 2022, 2023, 2024d; He et al. 2024b, 2023b,c; Fang et al. 2024; Zhong et al. 2024b; Tang et al. 2024, 2023a,b; Chen et al. 2024; Feng et al. 2024; Wang et al. 2024c; Zhong et al. 2024a,c). To achieve finer control that cannot be accomplished using only text prompts, many T2I methods incorporate layout as an additional input to guide the generation process. One branch of these methods (Xie et al. 2023; Phung, Ge, and Huang 2024; Chen, Laina, and Vedaldi 2024; Ma et al. 2024a) involves designing an extra loss function to update the latent variables and guide the sampling process. While these methods can achieve image generation in a single forward pass, their fidelity is inadequate when dealing with complex object interactions or attributes. The other branch of methods (Lian et al. 2023; Bar-Tal et al. 2023; Jiménez 2023) performs denoising separately for each layout and subsequently fuses the results, leading to high computational costs. Different from the aforementioned methods, our method processes all layouts simultaneously, thereby eliminating the need for additional loss functions to guide sampling. Furthermore, our method can effectively handle complex object interactions while maintaining high image fidelity and precise text alignment.

Customized text to image generation. The goal of customized text-to-image generation is to acquire knowledge of a novel concept from a limited set of examples and subsequently generate images of these concepts in diverse scenar-

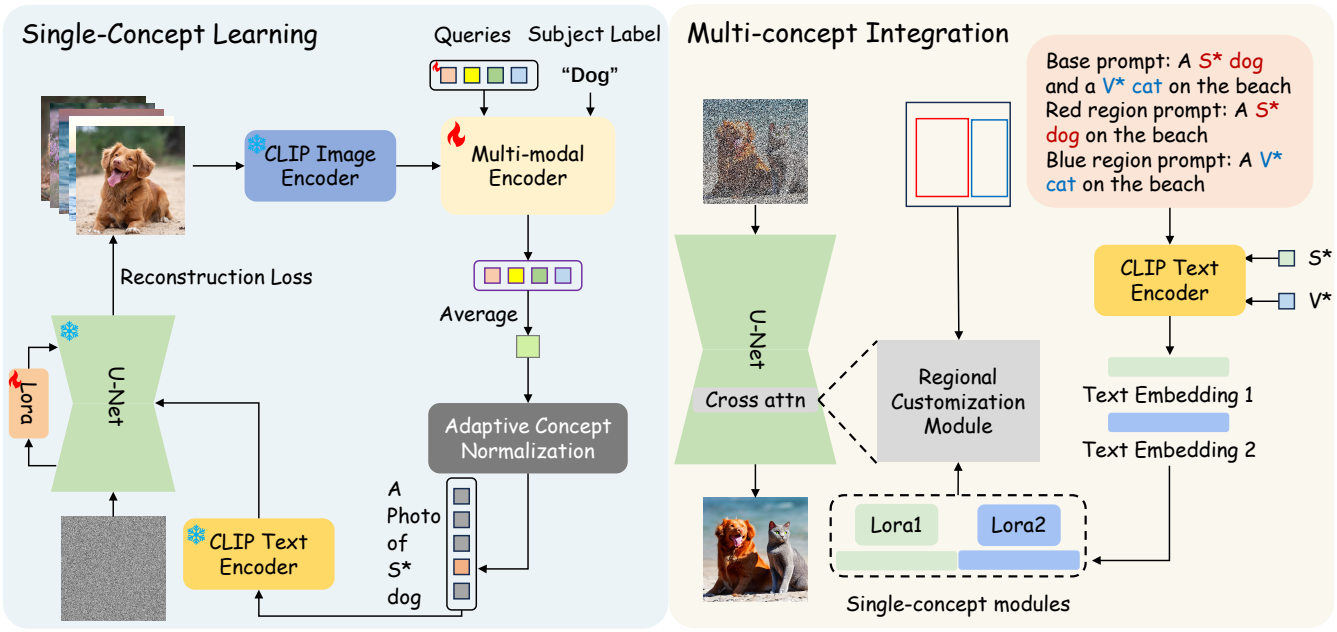

Figure 2: Overall Pipeline of MultiBooth. (a) During the single-concept learning phase, a multi-modal encoder and LoRA parameters are trained to encode every single concept. (b) During the multi-concept integration phase, we first convert S^* and V^* into text embeddings, which are then combined with the corresponding LoRA to form single-concept modules. These single-concept modules, along with the bounding boxes, are intended to serve as input for the regional customization module.

ios based on text prompts. By leveraging the aforementioned diffusion-based methodologies, it becomes possible to employ the comprehensive text-image prior to customizing the text-to-image process. The first branch of methods (Gal et al. 2022; Chen et al. 2023; Liu et al. 2023) achieves customization by creating a new embedding within the tokenizer and associating all the details of the newly introduced concept to this embedding. The second branch of methods (Wei et al. 2023; Shi et al. 2023; Gal et al. 2023) trains an adapter to generate embeddings. They need strong GPUs and large datasets for training and only support single-concept customization. To adapt to MCC, they need numerous multi-concept images and costly retraining. The third branch of methods (Ruiz et al. 2023; Kumari et al. 2023) binds the new concept to a rare token followed by a class noun. Compared to the previous two branches of methods, they often achieve the best image fidelity. However, this process is achieved by fine-tuning the entire or partial UNet. As a result, they require a larger amount of parameters to store a new concept. In this work, we utilize a multi-modal model and LoRA to discriminatively and concisely encode every single concept. Then, we introduce the regional customization module to efficiently and accurately produce multi-concept images.

3 Method

Given a series of images $\mathcal{S} = \{X_s\}_{s=1}^{S}$ that represent S concepts of interest, where $\{X_s\} = \{x_i\}_{i=1}^{M}$ denotes the M images belonging to the concept s which is usually very small (e.g., $M <= 5$), the goal of multi-concept customization (MCC) is to generate images that include any number of concepts from $\mathcal{S}$ in various styles, contexts, layout relationship as specified by given text prompts.

MCC faces significant challenges for two primary reasons. Firstly, learning a concept with a limited number of images is inherently difficult. Secondly, generating multiple concepts *simultaneously and coherently* within the same image while faithfully adhering to the provided text is even harder. To address these challenges, our MultiBooth initially performs high-fidelity learning of a single concept. We employ a multi-modal encoder and the adaptive concept normalization strategy to obtain text-aligned representative customized embeddings. Additionally, the efficient concept encoding technique is employed to further improve the fidelity of single-concept learning. To generate multi-concept images, we employ the regional customization module. This module serves as a guide for multiple single-concept modules and utilizes bounding boxes to indicate the positions of each generated concept.

3.1 Preliminaries

In this paper, the foundational model utilized for text-to-image generation is Stable Diffusion (Rombach et al. 2022). It takes a text prompt P as input and generates the corresponding image x. Stable Diffusion (Rombach et al. 2022) consists of three main components: an autoencoder$(\mathcal{E}(\cdot), \mathcal{D}(\cdot))$, a CLIP text encoder $\tau_\theta(\cdot)$ and a U-Net $\epsilon_\theta(\cdot)$. Typically, it is trained with the guidance of the following reconstruction loss:

$$\mathcal{L}_{rec} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t,P}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta\left(P\right)\right)\|_2^2\right], \quad (1)$$

Method	a	S*	dog	and	a	V*	cat	on	the	beach
Textual Inversion	0.35	**2.85**	-	0.34	0.35	**0.94**	-	0.34	0.34	0.37
Ours w/o ACN	0.35	**2.35**	0.37	0.34	0.35	**3.14**	0.37	0.34	0.34	0.37
Ours w/o ACN&Reg	0.35	**111.02**	0.37	0.34	0.35	**131.24**	0.37	0.34	0.34	0.37
Ours	0.35	**0.37**	0.37	0.34	0.35	**0.37**	0.37	0.34	0.34	0.37

Table 1: Quantization results of the L2 norm of each word embedding in the prompt.

where $\epsilon \sim \mathcal{N}(0,1)$ is a randomly sampled noise, t denotes the time step. The calculation of z_t is given by $z_t = \alpha_t z + \sigma_t \epsilon$, where the coefficients α_t and σ_t are provided by the noise scheduler.

Given M images $\{X_s\} = \{x_i\}_{i=1}^{M}$ of a certain concept s, previous works (Gal et al. 2022; Ruiz et al. 2023; Kumari et al. 2023) associate a unique placeholder string S^* with concept s through a specific prompt P_s like "a photo of a S^* dog", with the following finetuning objective:

$$\mathcal{L}_{bind} = \mathbb{E}_{z=\mathcal{E}(x),x\sim X_s,\epsilon,t,P_s}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta\left(P_s\right)\right)\|_2^2\right]. \tag{2}$$

Minimizing Eq. (2) can encourage the U-Net $\epsilon_\theta(\cdot)$ to accurately reconstruct the images of the concept s, effectively binding the placeholder string S^* to the concept s.

3.2 Single-Concept Learning

Multi-modal Concept Extraction. Existing customization methods (Gal et al. 2023; Wei et al. 2023) mainly utilize a single image encoder to encode the whole image into concept embeddings. However, the single image encoder may also encode unrelated objects in the images. To remedy this, we employ a multi-modal encoder that takes as input both the images and the concept name (e.g., "dog") to generate concise and discriminative customized embeddings.

Inspired by MiniGPT4 (Zhu et al. 2023) and BLIP-Diffusion (Li, Li, and Hoi 2023), we utilize the QFormer, a light-weighted multi-modal encoder, to generate the customized embeddings for each concept. As shown in the left part of Fig. 2, the QFormer encoder E has three types of inputs: visual embeddings ξ of an image, text description l of the concept of interest, and learnable query tokens $W = [w_1, \cdots, w_K]$ where K is the number of query tokens. Given an image $x_i \in X_s$ of concept s, we employ a frozen CLIP (Radford et al. 2021) image encoder to extract the visual embeddings ξ of the image. Subsequently, we set the input text l as the concept name for the image. The learnable query tokens W interact with the text description l through a self-attention layer and with the visual embedding ξ through a cross-attention layer. This interaction results in text-image aligned output tokens $O = E(\xi, l, W)$ with the same dimensions as W. Finally, we average these tokens and get initial customized embedding $v_i = \frac{1}{K} \cdot \sum_{i=1}^{K} o_i$.

After obtaining the customized embedding v_i of concept s, we introduce a placeholder string S^* to represent the concept s, with v_i representing the word embedding of S^*. Through this placeholder string S^*, we can easily activate the customized word embedding v_i to reconstruct the input concept image x_i with prompts like "a photo of a S^* dog".

Adaptive Concept Normalization. We have observed a domain gap between our customized embedding v_i and other word embeddings in the prompt. As shown in Tab. 1, the L2 norm of our customized embedding is considerably larger than that of other word embeddings in the prompt. Notably, these word embeddings, belonging to the same order of magnitude, are predefined within the embedding space of the CLIP text encoder $\tau_\theta(\cdot)$. This significant difference in quantity weakens the model's ability of multi-concept generation. To remedy this, we further apply the Adaptive Concept Normalization (ACN) strategy to the customized embedding v_i, adjusting its L2 norm to obtain the final customized embedding $\hat{v}_i$.

Our ACN consists of two steps. The first step is L2 normalization, adjusting the L2 norm of the customized embedding v_i to 1. The second step is adaptive scaling, which brings the L2 norm of v_i to a comparable magnitude as other word embeddings in the prompt. Specifically, let $c_l \in \mathbb{R}^d$ represent the word embedding corresponding to the subject name of v_i (e.g., the word embedding of "dog"), where d is the dimension of embeddings. The adaptive concept normalization $\hat{v}_i = v_i \cdot \frac{\|c_l\|_2}{\|v_i\|_2}$. As shown in Tab. 1, this operation effectively addresses the problem of domain gap in the embedding space.

Efficient Concept Encoding. To further improve the concept fidelity during single-concept learning and avoid language drift caused by finetuning the U-Net, we incorporate the LoRA technique (Hu et al. 2021; He et al. 2024a) for efficient concept encoding. Specifically, we incorporate a low-rank decomposition to the key and value weight matrices of attention layers within the U-Net $\epsilon_\theta(\cdot)$. Each pre-trained weight matrix $W_{init} \in \mathbb{R}^{d\times k}$ of the U-Net $\epsilon_\theta(\cdot)$ is utilized in the forward computation as follows:

$$h = W_{init}x + \Delta W x = W_{init}x + BAx, \tag{3}$$

where $A \in \mathbb{R}^{r\times k}$, $B \in \mathbb{R}^{d\times r}$ are trainable parameters of efficient concept encoding, and the rank $r \ll \min(d,k)$. During training, the pre-trained weight matrix W_{init} stays constant without receiving gradient updates. We also use a regularization term to lower the L2 norm of v_i before ACN. Without this term, the L2 norm of v_i can grow large as shown in Tab. 1. Scaling v_i with ACN could greatly alter its magnitude, causing information loss. As a result, the whole single-concept learning framework can be trained as follows:

$$\mathcal{L} = \mathbb{E}_{z=\mathcal{E}(x),x\sim X_s,\epsilon,t,P_s}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta\left(P_s\right)\right)\|_2^2\right] + \lambda\|v_i\|_2^2, \tag{4}$$

where λ denotes a balancing hyperparameter and is consistently set to 0.01 across all experiments.

So far, we can learn a new concept efficiently and store its information in a dedicated single-concept module. This module contains a customized embedding along with the corresponding LoRA parameters. The extra parameter for a new concept is less than 7MB, which is significantly lower compared to 3.3GB in DreamBooth (Ruiz et al. 2023) and 72MB in Custom Diffusion (Kumari et al. 2023). Furthermore, the single-concept module is plug-and-play for multi-concept generation, as users can combine any single-concept module through the Regional Customization Module to perform multi-concept generation.

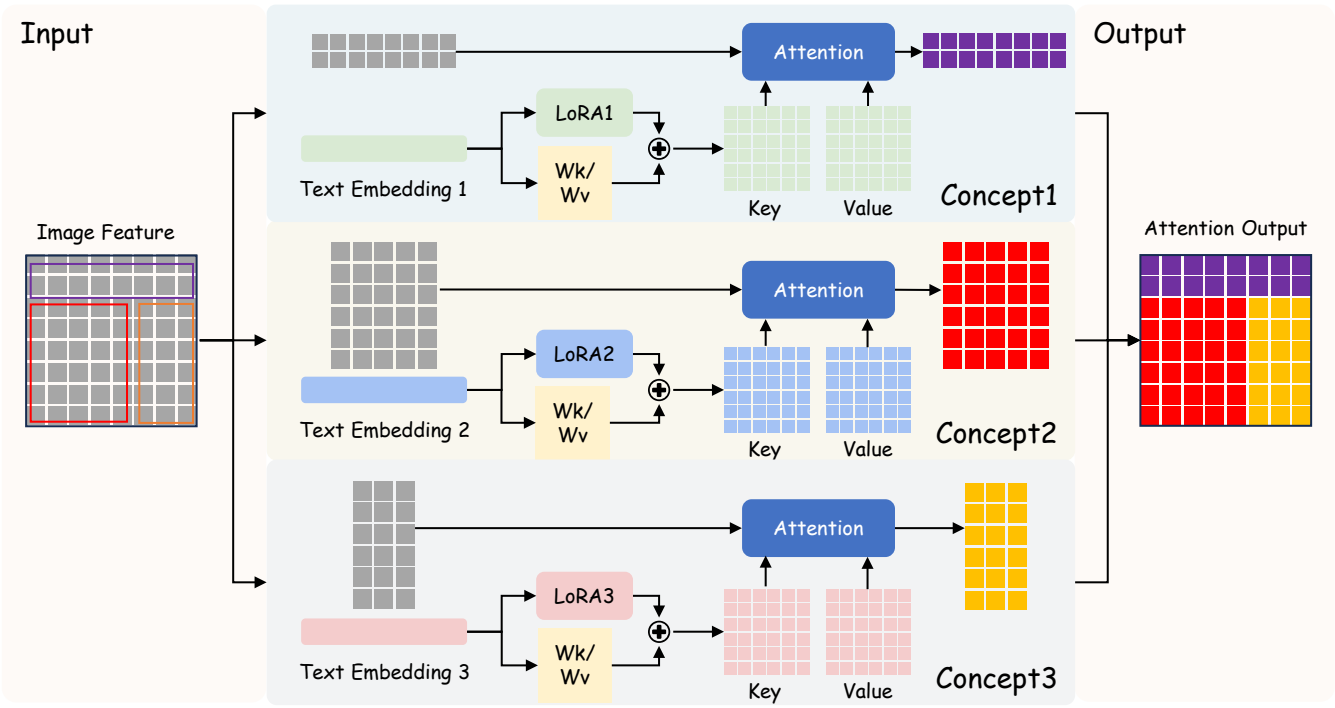

Figure 3: Regional Customization Module. We initially divide the image feature into several regions via bounding boxes to acquire the query Q for each concept. Subsequently, we combine the single-concept module with W_k and W_v to derive the corresponding key K and value V. After that, we perform the attention operation on the obtained Q, K, and V to get a partial attention output. The above procedure is applied to each concept simultaneously, forming the final attention output.

3.3 Multi-Concept Integration

Regional Customization Module. To integrate multiple single-concept modules for multi-concept generation, we propose the Regional Customization Module (RCM) in cross-attention layers. The key insight of our RCM is to generate each concept within the specified region and allow different concepts to interact accurately in overlapping regions.

As shown in the right part of Fig. 2, given a base prompt p_{base} describing the desired generated results, we can obtain the bounding boxes $B = \{b_i\}_{i=1}^{S}$ and the corresponding region prompts $P_r = \{p_i\}_{i=1}^{S}$ for each concept either through user-defined methods or automated processes (see Section 4.3). The region prompt guides the concept generation within each specific region, while the base prompt ensures interaction among concepts across different regions. As a result, the text embeddings $C = \{c_i\}_{i=1}^{S}$ for each region can be acquired through the combination of the region prompt and the base prompt:

$$c_i = \tau_\theta(p_i) + \tau_\theta(p_{base}), i = 1, 2, \cdots, S, \tag{5}$$

where $c_i \in \mathbb{R}^{k \times d}$, k is the the maximum length of input words and $\tau_\theta(\cdot)$ is the CLIP text encoder.

Then, we integrate the text guidance from text embeddings and the concept information in LoRA into each region *simultaneously* within the cross-attention layers. As shown in Fig. 3, the image feature $F \in \mathbb{R}^{h \times w}$ is the input of RCM. For the i^{th} concept, the image feature F is cropped using the bounding box $b_i \in \mathbb{R}^{h_i \times w_i}$, resulting in the partial image feature $f_i \in \mathbb{R}^{h_i \times w_i}$. With f_i, we can obtain the query vector Q_i through $Q_i = W_q \cdot f_i$. Next, we derive the key and value vector K_i and V_i using the text embedding c_i and corresponding LoRA parameters $\{A_{ij}, B_{ij}\}_{i=1}^{S}$ through:

$$K_i = W_k \cdot c_i + B_{i1}A_{i1} \cdot c_i, \tag{6}$$

$$V_i = W_v \cdot c_i + B_{i2}A_{i2} \cdot c_i, \tag{7}$$

where $A_{ij} \in \mathbb{R}^{r \times k}$ and $B_{ij} \in \mathbb{R}^{d \times r}$, $j = 1$ and $j = 2$ indicating the low-rank decomposition of W_k and W_v respectively. In order to derive the text-aligned image feature with concept information, we then apply the attention operation to the query, key, and value vectors:

$$\text{Attn}\left(Q_i, K_i, V_i\right) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d'}}\right) V_i, \tag{8}$$

where d' represents the output dimension of key and query features. The image feature $\hat{f}_i = \text{Attn}\left(Q_i, K_i, V_i\right) \in \mathbb{R}^{h_i \times w_i}$ contains both the text guidance and concept information through the attention mechanism and retains its original dimensions. For overlapping regions, we use a weighted average strategy to ensure the generation of each concept:

$$\hat{f} = \frac{1}{\eta} \cdot \sum_{i=1}^{\eta} w_i \cdot \hat{f}_i, \ \sum_{i=1}^{\eta} w_i = 1, \ \bigcup_{i=1}^{\eta} b_i \neq \varnothing, \tag{9}$$

where η is the number of overlapping concepts, $\hat{f}$ is the output feature of the overlapping region, w_i is the average

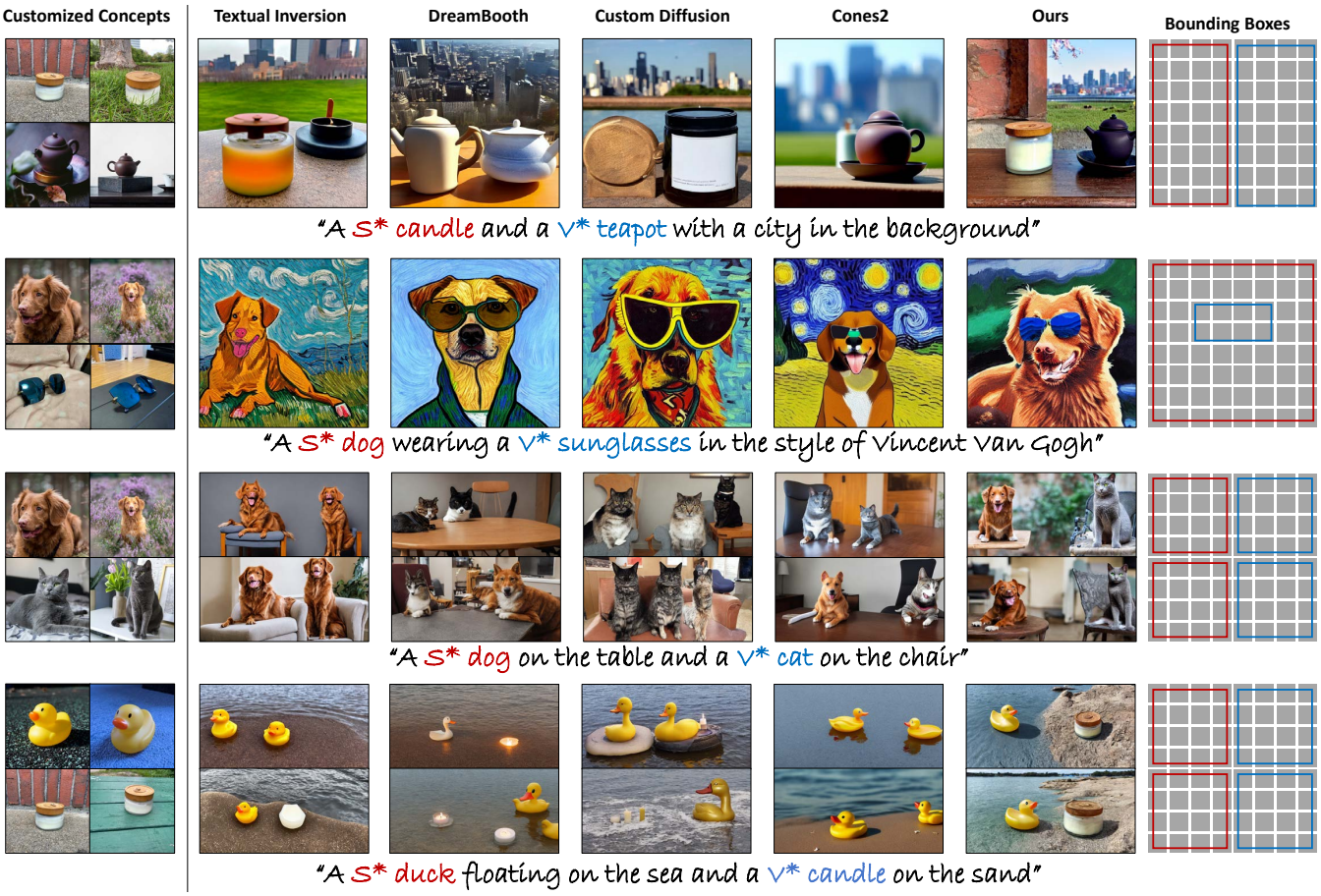

Figure 4: Qualitative comparisons. Our method outperforms all the compared methods in image fidelity and prompt alignment.

weight of the i^{th} concept. The setting of w_i is further discussed in *Suppl*.

Compared to (Kumari et al. 2023; Liu et al. 2023), our RCM offers more flexible and precise customization that cannot be achieved solely through text prompts. Once the single-concept modules are obtained, RCM can combine multiple single-concept modules in a plug-and-play manner to perform multi-concept generation without retraining. With bounding boxes indicating the regions of the generated concepts, RCM can generate each concept according to different region prompts (see Section 4.1) and handle complex object interactions under the guidance of the base prompt (see Section 4.3). Moreover, despite the superior multi-concept customization performance achieved by our RCM, it incurs minimal cost during inference. This is because the RCM generates all the customized concepts *simultaneously*, rather than *sequentially*, which is further discussed in Section 4.3. We also provide a thorough comparison between our RCM and other layout T2I methods (Lian et al. 2023; Xie et al. 2023), detailed in Section 4.2.

4 Experiment

Implementation details. All of our experiments are based on Stable Diffusion v1.5 and are conducted on one RTX3090. We set the rank of LoRA to be 16. During training, we randomly select text prompts P_s from the CLIP ImageNet templates (Radford et al. 2021) following the Textual Inversion (Gal et al. 2022). During training, we optimize for 900 steps with a learning rate of 8×10^{-5}. During inference, we sample for 100 steps with the guidance scale $\omega = 7.5$. More detailed settings can be found in the *Suppl*.

Datasets. Following Custom Diffusion (Kumari et al. 2023), we conduct experiments on twelve subjects selected from the DreamBooth dataset (Ruiz et al. 2023) and CustomConcept101 (Kumari et al. 2023). They cover a wide range of categories including two scene categories, two pets, and eight objects.

4.1 Comparative Study

We conduct comparisons between our method and four existing methods: Textual Inversion (TI) (Gal et al. 2022), DreamBooth (DB) (Ruiz et al. 2023), Custom Diffusion (CD) (Kumari et al. 2023), and Cones2 (Liu et al. 2023).

Qualitative comparison. As shown in Fig. 4, TI and DB are limited to generating a single concept, whereas CD and Cones2 can produce multiple concepts but struggle with maintaining high fidelity. In contrast, our method excels in multi-concept generation, achieving both high image fidelity and prompt alignment, even in challenging long-format scenarios (third and fourth rows).

Method	Single-Concept			Multi-concept			Training Time	Inference Time
	CLIP-I	Seg CLIP-I	CLIP-T	CLIP-I	Seg CLIP-I	CLIP-T		
TI(2022)	0.738	0.721	0.752	0.666	0.660	0.736	23min	**7.50s**
DB(2023)	**0.769**	0.736	0.775	0.637	0.652	**0.828**	10min	**7.35s**
Custom(2023)	0.654	0.661	**0.813**	0.624	0.637	0.812	**4min**	7.53s
Cones2(2023)	0.768	**0.747**	0.758	**0.670**	**0.685**	0.816	26min	21.41s
Ours	**0.783**	**0.761**	**0.780**	**0.714**	**0.713**	**0.838**	**6min**	8.29s

Table 2: Quantitative comparisons. The best and second best results are in red and blue, respectively.

Quantitative comparison. We assess all the methods using three evaluation metrics: CLIP-I, Seg CLIP-I, and CLIP-T. (1) CLIP-I measures the average cosine similarity between the CLIP (Radford et al. 2021) embeddings of the generated images and the source images. (2) Seg CLIP-I is similar to CLIP-I, but all the subjects in source images are segmented. (3) CLIP-T calculates the average cosine similarity between the embeddings of prompt and image. As presented in Tab. 2, our method demonstrates superior image alignment and comparable text alignment in the single-concept setting. In the multi-concept setting, our method outperforms all the compared methods in the three selected metrics. Moreover, with excellent image fidelity and prompt alignment ability, our method does not incur significant training and inference costs.

Figure 5: Qualitative ablation results.

4.2 Ablation Study

Regional Customization Module (RCM). We first verify the effectiveness of RCM by simply removing it. As shown in Fig. 5 and Tab. 3, without RCM, the features of the candle and teapot have fused to some extent. To further validate the effectiveness of RCM, we retain our single concept learning (SCL) and replace our RCM with other layout T2I methods. We select two representative methods: LLM-grounded Diffusion (LG) (Lian et al. 2023) and BoxDiff (Xie et al. 2023), with the bounding boxes used displayed on the left. On the one hand, LG (Lian et al. 2023) denoises each concept within the bounding boxes sequentially and then integrates them at the latent level, resulting in concept fusion in the overlapping regions. On the other hand, BoxDiff (Xie et al. 2023) employs the cross-attention map to construct a loss function for updating the latent variables. Although it can generate two concepts simultaneously, it suffers from low image fidelity. Furthermore, neither of these methods can handle complex object interactions according to the given text prompt. In contrast, our method allows different single-concept modules to target specific regions at the cross-attention level, thereby generating multiple concepts simultaneously. By using a base prompt to guide complex object interactions across various regions, we can produce images with both high image fidelity and precise text alignment.

Method	CLIP-I	Seg CLIP-I	CLIP-T
w/o Region	0.691	0.707	0.710
w/o QFormer	0.691	0.694	0.823
w/o ACN	0.694	0.695	0.826
Ours	**0.713**	**0.712**	**0.838**

Table 3: Quantitative ablation results.

QFormer and Adaptive Concept Normalization (ACN). We also demonstrate the effectiveness of the QFormer and the ACN by removing them either. As shown in Fig. 5 and Tab. 3, without QFormer or ACN, the fidelity of our method has decreased. In contrast, our full method can faithfully perform multi-concept generation.

4.3 Discussions

Inference time $\times N$ for N concepts? We also analyze the inference time of our method with the increasing number of concepts. As shown in Tab. 4, the inference time of our method increases only slightly as the number of concepts grows. This is because increasing concepts only leads to additional cross-attention computation in our RCM; other operations, like self-attention, residual addition, etc. remain the same as generating a single concept.

	2 Concepts	3 Concepts	4 Concepts
Inference Time	8.29s	10.07s	10.53s

Table 4: Inference Time with more concepts.

5 Conclusion

We introduce MultiBooth, a novel and efficient framework for multi-concept customization (MCC). Compared with existing MCC methods, our MultiBooth allows plug-and-play multi-concept generation with high image fidelity while bringing minimal cost during training and inference. By conducting qualitative and quantitative experiments, we demonstrate our superiority over state-of-the-art methods within diverse customization scenarios. We believe that our approach provides a novel insight for the community.

Acknowledgments

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404.

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*.

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5343–5353.

Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-Your-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation. *arXiv preprint arXiv:2409.01055*.

Fang, C.; He, C.; Xiao, F.; Zhang, Y.; Tang, L.; Zhang, Y.; Li, K.; and Li, X. 2024. Real-world Image Dehazing with Coherence-based Pseudo Labeling and Cooperative Unfolding Network. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2024. Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–13.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2023. Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. *arXiv preprint arXiv:2305.18292*.

He, C.; Fang, C.; Zhang, Y.; Li, K.; Tang, L.; You, C.; Xiao, F.; Guo, Z.; and Li, X. 2023a. Reti-Diff: Illumination Degradation Image Restoration with Retinex-based Latent Diffusion Model.

He, C.; Li, K.; Xu, G.; Yan, J.; Tang, L.; Zhang, Y.; Wang, Y.; and Li, X. 2023b. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *IEEE Transactions on Neural Networks and Learning Systems*.

He, C.; Li, K.; Xu, G.; Zhang, Y.; Hu, R.; Guo, Z.; and Li, X. 2023c. Degradation-resistant unfolding network for heterogeneous image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12611–12621.

He, C.; Li, K.; Zhang, Y.; Xu, G.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2024a. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36.

He, C.; Li, K.; Zhang, Y.; Zhang, Y.; You, C.; Guo, Z.; Li, X.; Danelljan, M.; and Yu, F. 2024b. Strategic Preys Make Acute Predators: Enhancing Camouflaged Object Detectors by Generating Camouflaged Objects. In *The Twelfth International Conference on Learning Representations*.

He, C.; Shen, Y.; Fang, C.; Xiao, F.; Tang, L.; Zhang, Y.; Zuo, W.; Guo, Z.; and Li, X. 2024c. Diffusion Models in Low-Level Vision: A Survey. *arXiv preprint arXiv:2406.11138*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiménez, Á. B. 2023. Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, D.; Li, J.; and Hoi, S. C. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8640–8650.

Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

Liu, Z.; Zhang, Y.; Shen, Y.; Zheng, K.; Zhu, K.; Feng, R.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023. Cones 2: Customizable Image Synthesis with Multiple Subjects. *arXiv preprint arXiv:2305.19327*.

Ma, W.-D. K.; Lahiri, A.; Lewis, J. P.; Leung, T.; and Kleijn, W. B. 2024a. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4098–4106.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024b. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024c. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024d. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022. Visual Knowledge Graph for Human Action Reasoning in Videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Phung, Q.; Ge, S.; and Huang, J.-B. 2024. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7932–7942.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.

Tang, L.; Li, K.; He, C.; Zhang, Y.; and Li, X. 2023a. Consistency regularization for generalizable source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4323–4333.

Tang, L.; Li, K.; He, C.; Zhang, Y.; and Li, X. 2023b. Source-free domain adaptive fundus image segmentation with class-balanced mean teacher. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 684–694. Springer.

Tang, L.; Tian, Z.; Li, K.; He, C.; Zhou, H.; Zhao, H.; Li, X.; and Jia, J. 2024. Mind the Interference: Retaining Pre-trained Knowledge in Parameter Efficient Continual Learning of Vision-Language Models. *arXiv preprint arXiv:2407.05342*.

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024a. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*.

Wang, J.; Pu, J.; Qi, Z.; Guo, J.; Ma, Y.; Huang, N.; Chen, Y.; Li, X.; and Shan, Y. 2024b. Taming Rectified Flow for Inversion and Editing. *arXiv preprint arXiv:2411.04746*.

Wang, J.; Pu, Y.; Han, Y.; Guo, J.; Wang, Y.; Li, X.; and Huang, G. 2024c. GRA: Detecting Oriented Objects through Group-wise Rotating and Attention. *arXiv preprint arXiv:2403.11127*.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024d. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv preprint arXiv:2401.07519*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Yan, Y.; Zhang, C.; Wang, R.; Zhou, Y.; Zhang, G.; Cheng, P.; Yu, G.; and Fu, B. 2023. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhong, X.; Chen, B.; Fang, H.; Gu, X.; Xia, S.-T.; and Yang, E.-H. 2024a. Going Beyond Feature Similarity: Effective Dataset distillation based on Class-aware Conditional Mutual Information. arXiv:2412.09945.

Zhong, X.; Fang, H.; Chen, B.; Gu, X.; Dai, T.; Qiu, M.; and Xia, S.-T. 2024b. Hierarchical Features Matter: A Deep Exploration of GAN Priors for Improved Dataset Distillation. *arXiv preprint arXiv:2406.05704*.

Zhong, X.; Sun, S.; Gu, X.; Xu, Z.; Wang, Y.; Wu, J.; and Chen, B. 2024c. Efficient Dataset Distillation via Diffusion-Driven Patch Selection for Improved Generalization. arXiv:2412.09959.

Zhu, C.; Li, K.; Ma, Y.; Tang, L.; Fang, C.; Chen, C.; Chen, Q.; and Li, X. 2024. InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences. *arXiv preprint arXiv:2412.01197*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.