Customized Generation Reimagined: Fidelity and Editability Harmonized

Jian Jin, Yang Shen, Zhenyong Fu*, and Jian Yang*

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology
{jinj,shenyang_98,z.fu,csjyang}@njust.edu.cn

Abstract. Customized generation aims to incorporate a novel concept into a pre-trained text-to-image model, enabling new generations of the concept in novel contexts guided by textual prompts. However, customized generation suffers from an inherent trade-off between concept fidelity and editability, i.e., between precisely modeling the concept and faithfully adhering to the prompts. Previous methods reluctantly seek a compromise and struggle to achieve both high concept fidelity and ideal prompt alignment simultaneously. In this paper, we propose a *"Divide, Conquer, then Integrate"* (DCI) framework, which performs a surgical adjustment in the early stage of denoising to liberate the fine-tuned model from the fidelity-editability trade-off at inference. The two conflicting components in the trade-off are decoupled and individually conquered by two collaborative branches, which are then selectively integrated to preserve high concept fidelity while achieving faithful prompt adherence. To obtain a better fine-tuned model, we introduce an *Image-specific Context Optimization* (ICO) strategy for model customization. ICO replaces manual prompt templates with learnable image-specific contexts, providing an adaptive and precise fine-tuning direction to promote the overall performance. Extensive experiments demonstrate the effectiveness of our method in reconciling the fidelity-editability trade-off. Code is available at https://github.com/jinjianRick/DCI_ICO.

Keywords: Model customization · Diffusion model · Trade off

1 Introduction

The recent large-scale text-to-image models, such as GLIDE [29], Imagen [38], and DALL-E 2 [31], have demenstrated astonishing image rendering ability, guided by text prompts in natural language. These models, pretrained on extensive caption-image pairs, allow users to generate visually realistic images with novel scenes and diverse styles. Although text-to-image generation has progressed impressively, implanting new visual concepts into pretrained models remains a major open challenge in this area. For example, users may wish to

* Corresponding author.

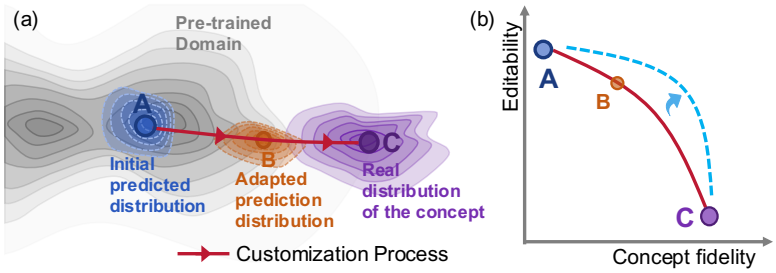

Fig. 1: Concept fidelity and editability trade-off in customized generation. (a) is an illustration of the customization process, and (b) shows the corresponding variations in concept fidelity and editability during the customization process. **(a)**: The distribution is portrayed in two dimensions. In the "pre-trained domain", deeper colors indicate higher editability. **(b)**: Concept fidelity and editability outline a trade-off curve (red line) in the customization process. Our method (blue dotted line) is designed to free the fine-tuned model from this trade-off, and provide a powerful inference-time adjustment mechanism to achieve satisfactory generations for all query prompts using a single fine-tuned model.

customize the models with their own specific photos. More personalized or some unusual visual concepts are difficult to describe merely with text prompts. In such scenarios, pretrained text-to-image models need to be adapted to incorporate these new visual concepts.

Concept customization of text-to-image models, proposed in [7, 19, 37], aims to implant a specific or personal concept into pretrained models, using a few reference images depicting the custom concept. In the customization process, the pre-trained model (and a concept token in some works) is adapted to learn the distribution of the new concept, as illustrated in Figure 1 (a), binding the concept with the concept token. Distributions that are closer to the pre-trained domain (*e.g.*, point A) tend to be more easily edited into various contexts but suffer from deficiencies in concept fidelity. In contrast, distributions closer to the customized domain (*e.g.*, point C) learn the concept well but experience a reduction in editability. Therefore, the inherent *fidelity-editability* trade-off emerges during the customization process, as shown in Figure 1 (b).

For a specific model, its performance of customized generation is confined to a fidelity-editability trade-off curve and can only be adjusted along this curve by modifying some fine-tuning parameters (*e.g.*, learning rate, fine-tuning step, *etc.*). Previous methods reluctantly seek a compromise between high concept fidelity and high editability, making it difficult to achieve both simultaneously. Besides, the generative prior varies across different concepts and query prompts, leading to variations in the optimal trade-off point, as illustrated in Figure 2. However, the raw fine-tuned models lack powerful inference-time adjustment options, making it challenging to achieve the best trade-off for all query prompts using a single fine-tuned model in practical applications. Furthermore, as shown

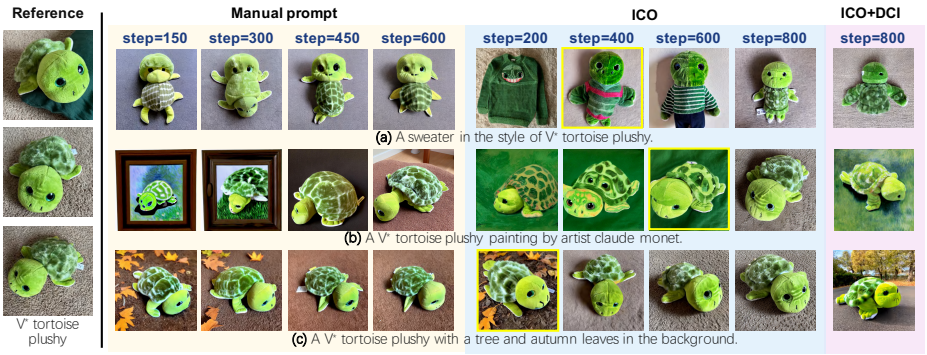

Fig. 2: Visualization of query examples varying in customization processes. We show the entire fine-tuning process (*i.e.*, until overfitting). ICO alleviates premature overfitting, resulting in a longer process. The model fine-tuned using manual prompts is prone to premature overfitting, and the generated images suffer from concept distortion. ICO can mitigate these issues and achieve a better trade-off between concept fidelity and editability. We highlight the best generation with a yellow border for each query prompt. We can find that the optimal trade-off point varies for different query prompts, and the trade-off is hard to reconcile for customized generations with low model priors (*e.g.*, sample (c)). DCI can effectively address these issues, generating satisfactory images for each prompt using one fine-tuned model. The prompt schemes of ICO used in "ICO" and "ICO+DCI" are different, which is detailed in Section 4.2.

in Figure 2 (c), reconciling the trade-off becomes infeasible for customized generation with weak generative priors, such as challenging concepts (*e.g.*, those with distributions far from the pre-trained domain) or rare prompted contexts. To address these challenges, our goal is to liberate the fine-tuned model from the constraint of the fidelity-editability trade-off and achieve satisfactory generations for all query prompts using a single fine-tuned model, as shown in Figure 1 (b).

In this paper, we propose a *"Divide, Conquer, then Integrate"* (DCI) framework, which performs a surgical adjustment to overcome the fidelity-editability trade-off at inference. In DCI, the two conflicting components in the trade-off are decoupled and individually conquered by two collaborative branches, namely the concept branch and the auxiliary branch. Specifically, the concept branch is responsible for contributing the concept-relevant content with high fidelity, while the auxiliary branch is tasked with providing the concept-irrelevant content. Both branches contribute specialized contents from their entire generations, which are selectively integrated using a Dual-Branch Integration Module (DBIM) during the denoising steps to produce the final output with high concept fidelity and prompt alignment. The integration ratio between the two branches is controllable, offering an inference-time adjustment mechanism to seek the optimal trade-off for each query prompt.

The performance boundary of the proposed DCI is dependent on the overall performance of the raw fine-tuned model (*i.e.*, the red line in Figure 1 (b)). Previ-

ous works utilize manually crafted templates (*e.g.*, `"A photo of {}"`) to construct text prompts for model fine-tuning. We argue that manually designed prompts are fixed and may provide imprecise guidance for model adaptation, leading to a deterioration of the overall trade-off. Therefore, we propose an *Image-specific Context Optimization* (ICO) fine-tuning strategy for model customization, which allows us to obtain a better fine-tuned model for DCI. In ICO, we replace the manually designed prompt templates with learnable image-specific contexts for all reference images. This prompt format is more effective at disentangling irrelevant content from the concept, adaptively providing a more precise fine-tuning direction to achieve a better overall performance, as illustrated by the comparison between "Manual prompt" and "ICO" in Figure 2. Extensive experiments demonstrate the effectiveness of DCI and ICO in reconciling the fidelity-editability trade-off, particularly for generations with weak generative priors.

2 Related Work

Text-to-image Generation. The purpose of text-to-image generation is to synthesize visually realistic images based on textual descriptions. Early studies [21,34,47,55] translate natural language into images by employing generative adversarial networks [10], which are constrained within the context of small-scale input scenarios. Autoregressive models such as DALL-E [32], Cogview [4], NUWA [45], and Parti [54], reframe text-to-image generation as a sequence-to-sequence problem. Recent advancements [29, 31, 35, 38] in text-to-image generation have harnessed diffusion models (DMs) [40] as the generative backbone. These models generate images through a denoising task while incorporating text conditions during the denoising process. Such successes lay the foundations for a multitude of text-guided image synthesis works [3, 15, 17, 48, 56, 57].

Prompt Tuning. Prompting is a parameter-efficient paradigm for adapting pre-trained models to a variety of downstream tasks [25]. This line of research stems from the field of natural language processing (NLP). Previous studies in this area employ hand-crafted prompts [1, 9] or generative prompts [39], which can be inefficient and suboptimal. To alleviate this problem, recent work has focused on prompt tuning (PT) [20,24,58], which treats prompts as task-specific continuous vectors and optimizes them directly using gradients during the fine-tuning process. With the remarkable advancement of pre-trained vision-language models [22, 30] in recent years, prompt learning has also been introduced into various computer vision tasks [5, 12, 18, 23, 26, 28, 33, 43, 53, 59, 60, 62].

Customized Generation. Customized generation aims to incorporate novel concepts into pre-trained text-to-image models and generate images of these concepts in new contexts. Pioneering works implant the concept by either inverting it into a word embedding vector [7] or bounding it with a special token [37]. Subsequent works focus on developing more efficient methods [8, 13, 19], tuning-free methods [2,44,46], for customized generation. Some works focus on multi-concept customization [11, 13, 19, 27], which customize multiple new concepts and integrating them into a single image. Some studies [14, 61] aim to enhance detail

preservation in text-to-image customization. Our work, instead, explores and addresses the fidelity-editability trade-off in customized generation.

3 Method

3.1 Preliminary of Customized Generation

The implantation of a custom concept $\mathcal{C}$ into pre-trained Diffusion Models (DMs) is accomplished through a text-to-image reconstruction task, using N image-prompt pairs $\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{P}_i)\}_{i=1}^{N}$. $\boldsymbol{X}_i$ and $\boldsymbol{P}_i$ are reference images and the corresponding textual prompts of $\mathcal{C}$. We utilize Latent Diffusion Models (LDMs) [35] as the generative backbone. Prompt $\boldsymbol{P}_i$ is firstly projected to an intermediate representation $\mathbf{c} = \tau_\theta(\boldsymbol{P}_i) \in \mathbb{R}^{M \times d_c}$ by a text encoder τ_θ. $\mathbf{c}$ is then injected into the LDMs via the cross-attention mechanism [42], serving as the conditioning for the reconstruction of image $\boldsymbol{X}_i$. The reconstruction process is regularized by the following squared error loss:

$$\mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),\mathbf{c},t}\left[w_t \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2\right] , \tag{1}$$

where $\mathbf{z}_t := \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}$ is the noised latent code at timestep t, $\mathbf{z}$ is the clean latent code of the training data, and w_t, α_t, σ_t are terms that determine the loss weight and noise schedule; $\boldsymbol{\epsilon}_\theta$ is a denoising autoencoder implemented using a conditional U-Net [36].

During inference, given the query text $\boldsymbol{P}_q(\mathcal{C})$ containing the learned concept $\mathcal{C}$, the customized DMs can generate the corresponding images conditioned on $\mathbf{c} = \tau_\theta(\boldsymbol{P}_q(\mathcal{C}))$. Concretely, an initial noise map $\mathbf{z}_T \sim \mathcal{N}(0, 1)$ is iteratively denoised from timestep $t = T$ to $t = 0$ in the latent space using either deterministic samplers [41] or stochastic sampler [16]. Then $\mathbf{z}_0$ is decoded to image space using an decoder f_D to generate the target image $\hat{\mathbf{x}} = f_D(\mathbf{z}_0)$.

3.2 Learning to Prompt for Customized Fine-tuning

In previous works [7,19,37], the textual prompts $\boldsymbol{P}$ used in model fine-tuning are constructed from manually crafted templates (*e.g.*, `"A photo of {}"`). We argue that the manually designed part in prompts may provide imprecise guidance for model adaptation, resulting in a deterioration of the overall performance.

We redesign the prompts of the reference images, and an example of the prompt generating process is illustrated in Figure 3. We introduce learnable image-specific contexts to the prompt of each reference image, completely replacing the manually designed templates that remain fixed during the fine-tuning. We refer to this fine-tuning strategy as *Image-specific Context Optimization* (ICO). Specifically, the prompts input to the text transformer are designed as follows:

$$\boldsymbol{P}_i = [\mathbf{I}]_1^i \cdots [\mathbf{I}]_M^i \, [\mathbf{T}]_1 \cdots [\mathbf{T}]_S \, [\mathbf{E}] \, [\mathbf{I}]_{M+1}^i \cdots [\mathbf{I}]_{M+L}^i , \tag{2}$$

where $\boldsymbol{P}_i$ is the prompt of image $\boldsymbol{X}_i$ and consists of three different components: image-specific context vectors $[\mathbf{I}]_m^i \, (m = 1, \ldots, M + L) \in \mathbb{R}^d$, concept-specific

context vectors $[\mathbf{T}]_s\,(s = 1, \ldots, S) \in \mathbb{R}^d$, and the concept descriptor $[\mathbf{E}] \in \mathbb{R}^{t \times d}$. The learned concept-specific context $\mathbf{T}$, in combination with the concept descriptor, is utilized to generate images of new concept at inference.

Components of Fine-tuning Prompt. *Image-specific Contexts* [**I**] are continuous vectors and are independent for each image. These vectors can be optimized end-to-end during fine-tuning, allowing them to model the image-specific content adaptively and offer more precise guidance for model adaptation. These context vectors can be positioned either at the beginning or the end of the entire prompt, serving different purposes in the learning process [60]. *Concept-specific Context* [**T**], also represented as `"V*"`, shares a learnable congruous context with all reference images, which is used to model the new concept [7,19]. *Concept Descriptor* [**E**] is the word embedding of a coarse class noun `"<noun>"` that describes the new concept [37], *e.g.*, embedding of `"cat"`. The design scheme of the fine-tuning prompt possesses flexibility in the position and length of image-specific and concept-specific contexts.

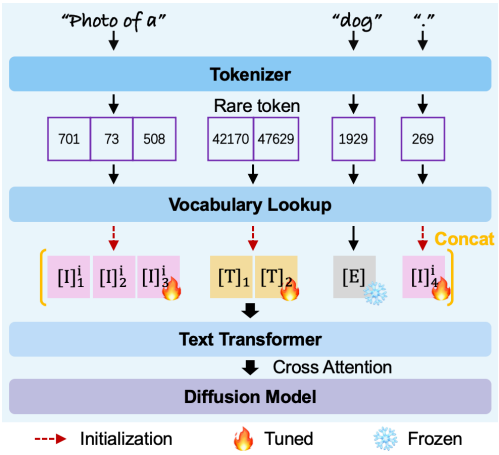

Fig. 3: Overall framework of the prompt generating process in fine-tuning.

3.3 Trade-off Reconciled at Inference

While the prompt tuning strategy improves the overall performance of the fine-tuned model, the fidelity-editability trade-off still persists. This issue becomes infeasible to reconcile for customized generation with weak model priors, such as challenging concepts (*e.g.*, those with distributions far from the pre-trained domain) or rare prompted contexts. We propose a novel framework called *"Divide, Conquer, then Integrate"* (DCI), which performs latent feature rectification to liberate the fine-tuned model from the trade-off at inference.

In the early stage of denoising ($T \sim (T - T_a)$), DCI *divides* the generation task and *conquers* it with two collaborative branches, namely the concept branch $\mathcal{B}_c$ and the auxiliary branch $\mathcal{B}_a$. $\mathcal{B}_c$ and $\mathcal{B}_a$ share the same model parameters. The overview of DCI is shown in Figure 4. Concretely, for a given prompt $\mathcal{P}$ of a custom concept `V* <noun>` (*e.g.*, $\mathcal{P}$ is `"A V* <noun> in times square"`), we first construct an auxiliary prompt $\mathcal{P}_a$ by replacing `"V* <noun>"` within $\mathcal{P}$ with an auxiliary class noun `"<auxiliary-noun>"`, while leaving the rest unaltered (*i.e.*, $\mathcal{P}_a$ is `"A <auxiliary-noun> in times square"`). $\mathcal{P}$ and $\mathcal{P}_a$ are then

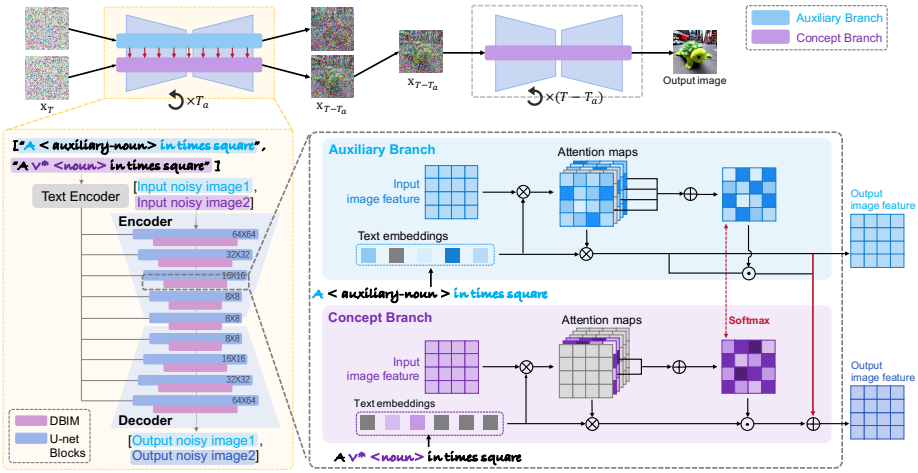

Fig. 4: An overview of the proposed *"Divide, Conquer, then Integrate"* (DCI) framework. DCI divides the customized generation task into two collaborative branches, namely the concept branch and the auxiliary branch. The concept branch takes the original query prompt as its conditioning and responsible for generating concept with high fidelity. The auxiliary branch takes an auxiliary prompt as its conditioning and contributes concept-irrelevant content. During the latent denoising steps, the contents generated by two branches are selectively integrated using a Dual-Branch Integration Module (DBIM) in each cross-attention layer.

taken as the text conditions for the concept branch $\mathcal{B}_c$ and the auxiliary branch $\mathcal{B}_a$. The auxiliary branch can regain powerful editability with the assistance of auxiliary nouns (*cf.* **Auxiliary Noun.**). The fine-tuned model used in DCI is solely required to excel at concept modeling, enabling the concept branch to generate concept-relevant content with high visual fidelity. These two branches are simultaneously forwarded within the diffusion model. Both branches contribute specialized components from their complete generations, with the concept branch providing concept-relevant content and the auxiliary branch offering concept-irrelevant content. These two branches are selectively *integrated* in the early stage of denoising using a *Dual-Branch Integration Module* (DBIM). We set $T_a = 0.15T$ and provide a quantitative analysis of the effect of T_a in the supplementary material.

Auxiliary Noun. A key observation is that the editability reduction during model customization originates from the tokens associated with the custom concept, including the entire `"V* <noun>"` and the noun `"<noun>"`, rather than from the fine-tuned model itself, as discussed in supplementary material. Therefore, `"<auxiliary-noun>"` need to be a noun distinct from `"<noun>"`. Replacing `"V* <noun>"` with `"<auxiliary-noun>"` empowers the fine-tuned model to regain the powerful editability over the query contexts, thus the content synthesized by the auxiliary branch maintains high prompt alignment. Additionally, we can select

nouns that have a high probability of occurrence within the context of the query prompt to further enhance textual editability, as explored in Section 4.5.

Dual-Branch Integration Module. To selectively integrate the generated content that each branch is responsible for, we introduce a *Dual-Branch Integration Module* (DBIM) to each cross-attention layer within the U-Net, as depicted in Figure 4. We elaborate on the design of DBIM in the following content.

We first introduce the cross-attention layer in LDMs briefly. The cross-attention layer utilizes latent image features $\mathcal{F}$ and text embeddings $\mathcal{E}$ to compute queries $Q = \ell_Q(\mathcal{F})$, keys $K = \ell_K(\mathcal{E})$, and values $V = \ell_V(\mathcal{E})$ through three projection layers: ℓ_Q, ℓ_K and ℓ_V. Then the spatial attention maps $\mathcal{M} \in \mathbb{R}^{C\times H\times W}$ are calculated as $\mathcal{M} = QK^T/\sqrt{d}$, where d is the dimension of the queries and keys, while C, H, W are the channel dimension, height, and weight of $\mathcal{M}$, respectively. Finally, the output of the cross-attention layer is obtained as $\hat{\mathcal{F}} = \mathcal{M}V$.

It is worth noting that each textual token in the conditional prompt is associated with a specific channel of the attention map $\mathcal{M}$, which determines the spatial layout and geometry of the corresponding textual semantics in the generated image [15, 57]. Inspired by this, we construct token sets containing the content that each branch is responsible for. The token set for $\mathcal{P}$ is $\mathcal{T}(\mathcal{P}) = \{$`"V*","<noun>"`$\}$, and the token set for $\mathcal{P}_a$ consists of all tokens in $\mathcal{P}_a$ except for `"<auxiliary-noun>"`, denoted as $\mathcal{T}(\mathcal{P}_a) = \mathbb{U}_{\mathcal{P}_a} \backslash \{$`"<auxiliary-noun>"`$\}$. In the i-th cross-attention layer during the denoising step t, we use the attention maps associated with the tokens in the token sets $\mathcal{T}(\mathcal{P})$ and $\mathcal{T}(\mathcal{P}_a)$ to infer the dynamic aggregation mask $\mathrm{M}^{t,i}_{\mathcal{B}_c}$ and $\mathrm{M}^{t,i}_{\mathcal{B}_a}$ for branch $\mathcal{B}_c$ and $\mathcal{B}_a$:

$$\mathrm{M}^{t,i}_{\mathcal{B}_c} = \sum_{k\in\mathcal{T}(\mathcal{P})} \mathcal{M}^{t,i}_k \,, \mathrm{M}^{t,i}_{\mathcal{B}_a} = \sum_{k\in\mathcal{T}(\mathcal{P}_a)} \mathcal{M}^{t,i}_k \,, \tag{3}$$

where $\mathcal{M}^{t,i}_k$ is the attention map related to token k.

Subsequently, we first perform softmax $\mathbf{S}_s$ in the spatial dimension to sharpen the spatial distribution of each mask, and then apply softmax $\mathbf{S}_c$ across two branches at each pixel to regularize the overall aggregation strength. The final aggregation masks $\hat{\mathrm{M}}^{t,i}$ are defined as follows:

$$\left(\hat{\mathrm{M}}^{t,i}_{\mathcal{B}_c}, \hat{\mathrm{M}}^{t,i}_{\mathcal{B}_a}\right) = \mathbf{S}_c \left(\mathbf{S}_s \left(\mathrm{M}^{t,i}_{\mathcal{B}_c}\right), \lambda \mathbf{S}_s \left(\mathrm{M}^{t,i}_{\mathcal{B}_a}\right)\right) \,, \tag{4}$$

where λ is a factor that can control the injection strength of concept-irrelevant information at inference time.

$\hat{\mathrm{M}}^{t,i}$ dynamically determines the spatially varying influences of each generative branch, allowing for the adaptive suppression or enhancement of the relative expression strength at different spatial regions. The output of DBIM is a linear combination of the two generative branches that weighted by $\hat{M}^{t,i}$:

$$\hat{\mathcal{F}}^{t,i} = \hat{\mathcal{F}}^{t,i}_{\mathcal{B}_c} \odot \hat{\mathrm{M}}^{t,i}_{\mathcal{B}_c} + \hat{\mathcal{F}}^{t,i}_{\mathcal{B}_a} \odot \hat{\mathrm{M}}^{t,i}_{\mathcal{B}_a} \,, \tag{5}$$

where $\hat{\mathcal{F}}^{t,i}_{\mathcal{B}_c}$ and $\hat{\mathcal{F}}^{t,i}_{\mathcal{B}_a}$ are the raw outputs of the concept branch and auxiliary branch at the i-th cross-attention layer in denoising step t, and $\odot$ denotes pixel-wise multiplication. $\hat{\mathcal{F}}^{t,i}$ is used to update the concept branch output $\hat{\mathcal{F}}^{t,i}_{\mathcal{B}_c}$.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on 16 datasets from previous works [7,19,37], each of which contains several reference images of a unique subject. These subjects cover a diverse range of categories, including pets, toys, plushies, *etc.* Each subject is evaluated using different types of query prompts, such as recontextualization, artistic variation, and attribute modification.

Evaluation Metrics. Concept reconstruction capacity and editability are the two primary metrics for evaluating customized generation [7, 19, 37]. The visual alignment of the generated images with the target concept reflects the model's reconstruction capacity, which is quantified using the average pairwise cosine similarity between the CLIP [30] embeddings of the generated and reference images. Editability is represented by the text alignment of the generated images with the query prompts. We measure this metric using the average CLIP-space cosine similarity between the generated images and their textual prompts, with the exclusion of the task-specific context `"V*"`. We also calculate the DINO-based image similarity [37] and present it in the supplementary material.

Baselines. We compare our method with Textual Inversion [7], DreamBooth [37], and Custom Diffusion [19]. Textual Inversion [7] inverts the new concept into a special token. DreamBooth [37] fine-tunes the entire diffusion model to bind the new concept with a fixed identifier. Custom Diffusion [19] introduces a new token for each novel concept and jointly optimizes the token embedding along with a subset of parameters in cross-attention layers. The implementation details of our method and baselines are provided in the supplementary material.

4.2 Comparisons

Comparisons of ICO with Baselines. We evaluate the proposed ICO in this section. We utilize the following prompt scheme to strike a balance between concept fidelity and editability: $M = 4$, $L = 2$ and $S = 2$ (cf. ablation study). Since it is challenging to determine the optimal trade-off point during the fine-tuning process for different methods, we evaluate all methods at various fine-tuning steps (until convergence or overfitting) for a more comprehensive comparison. The quantitative evaluation results are depicted in Figure 6. We mark the different evaluation points. We can observe that the fine-tuning process of each method delineates a fidelity-editability trade-off curve, and ICO achieves the best performance in balancing concept reconstruction and prompt adherence. We present sample generations of ICO and the other three baselines in Figure 5. Comparing to baselines, ICO better captures both the high-level semantics and the visual details of the target concept, while also demonstrating higher text-alignment for various types of prompts.

Comparisons of DCI with Baselines. In this section, we evaluate the effectiveness of DCI. DCI is an inference-time framework that requires combining with a specific fine-tuning strategy, which in our experiment is ICO. To empower the concept branch with high reconstruction capacity, the models used for DCI

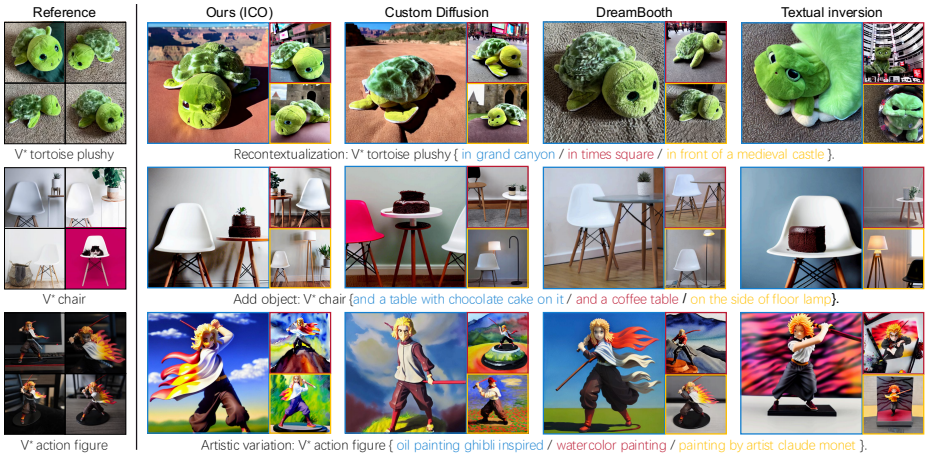

Fig. 5: Visual comparsion of ICO with baselines. We illustrate the effectiveness of ICO by comparing it to three baselines. Reference images are shown on the left. *First row*: customizing various scenes for the target concept. Our method can better reconstruct the target concept while maintaining high text alignment. *Second row* : adding new objects. *Third row* : customizing artistic styles for the target concept. While other methods suffer from concept distortion, our method can generate the target concept with higher concept fidelity.

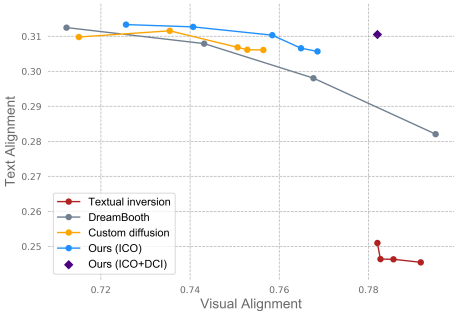

Fig. 6: Quantitative comparisons. The quantitative evaluation results of our method, Textual Inversion, DreamBooth, and Custom Diffusion. We mark the different evaluation points.

	Metric	Textual Inversion	DreamBooth	Custom Diffusion
ICO	Text Alignment	**84.38** %	**58.33** %	**55.21** %
	Visual Alignment	**71.88** %	**53.13** %	**57.29** %
ICO +DCI	Text Alignment	**89.58** %	**66.67** %	**63.54** %
	Visual Alignment	**76.04** %	**60.42** %	**68.75** %

Table 1: User study. In each paired comparison, our method is preferred ($\geq$ 50%) over the baseline methods in terms of both visual alignment and text alignment. The combination of DCI and ICO demonstrates a overwhelming user preference over other methods.

are fine-tuned using the following prompt scheme: $M = 1$, $L = 1$, and $S = 3$ (cf. ablation study). Since DCI incurs additional denoising steps T_a at inference, we reduce the total denoising steps (*i.e.*, T_a+T) of DCI to match other single-branch baselines' steps for fair comparison. We depicted the quantitative evaluation results of DCI in Figure 6, comparing with the other three baselines and the raw model fine-tuned using ICO. We can observe that DCI overcomes the inherent

Fig. 7: Visual comparisons of DCI with baselines. Reference images are shown on the left. We show sample generations for three types of query prompts: Recontextualization (*first row*), Property modification (*second row*), and Artistic variations & Add object (*third row*). In the context of customized generation with weak model priors, the outputs of the raw fine-tuned models either suffer from severe subject distortions or overfit to the given subject. DCI can significantly alleviate this fidelity-editability trade-off. The images generated by DCI can preserve subject details while adhering faithfully to the query prompt.

fidelity-editability trade-off in the fine-tuned model at inference time. It achieves both higher concept fidelity and text alignment compared to the other raw fine-tuned models. We show sample generations of our method and the other baselines in Figure 7. The depicted generations of baselines are selected from all evaluation steps. As for DCI, images are generated from one fine-tuned model for each concept. As we can see, the raw fine-tuned models struggle to reconcile the trade-off, particularly for generations with weak model priors. These models either suffer from severe concept distortions or overfit to the given concept throughout the entire customization process. DCI can overcome these inherent limitations of fine-tuned models and generate satisfactory images. The target concepts are successfully rendered in the query contexts with high visual fidelity by DCI, while still maintaining a high variability and diversity in poses and articulations. We also compare our method to ED-LoRA [11] in the supplementary material.

User Study. We further evaluate the proposed method through a user study. We conduct paired tests comparing the two versions of our method, namely "ICO" and "ICO+DCI", with DreamBooth, Textual Inversion, and Custom Diffusion. In each comparison, users are presented with reference images $\mathcal{I}$, the query prompt $\mathcal{P}$, and two corresponding generations from two compared method (ours and baseline). Two questions are designed to compare the visual alignment and text alignment of the two images. For the visual alignment comparison, users are asked to select the better image by answering the question-"which image contains the subject that is more similar to the provided reference images.". In the text alignment comparison, users are asked to select the better image

with the question-"which image is more consistent with the textual description of the query prompt.". A total of 576 comparison results are collected for the evaluation, and the aggregated result is shown in Table 1. As observed, our method, particularly the "ICO+DCI" version, demonstrates a dominant user preference over other methods. This further validates the effectiveness of the proposed method in enhancing the overall performance of customized generation.

4.3 Ablation Studies

Prompt Schemes. We explore the impact of prompt design schemes on concept customization using a subset of subjects, focusing on three aspects: suffix length M, prefix length L, and concept-specific context length S. We fine-tune Stable Diffusion [35] using different prompt schemes, comparing them to baselines that are fine-tuned using manually fixed prompt templates with $S = 1$, $S = 2$, and $S = 3$. As shown in Figure 8, our method and baselines outline fidelity-editability trade-off curves, with our method achieving a better overall performance. This validates our hypothesis that replacing manual prompt templates with learnable contexts can improve model customization. Besides, the prompt scheme impacts model convergence. In general, $M + L$ (the total length of image-specific contexts) influences the overall trade-off (distance to the origin). S affects model's convergence point on trade-off curve. An increase in S enhances the model's concept reconstruction capacity but diminishes its editability.

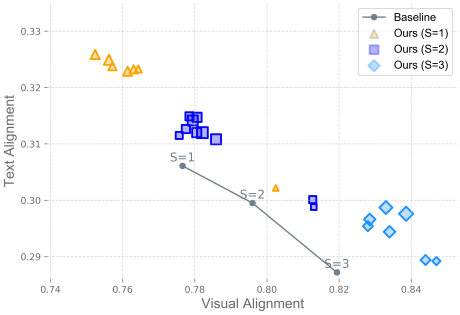

Fig. 8: Investigations of ICO's prompt scheme. S is the length of concept-specific contexts. A larger data point size indicates a larger $M + L$ (the length of image-specific contexts).

Injection Strength λ. In DCI, the injection ratio λ in Eqn. 4 controls the injection strength of the concept-irrelevant content, which offers users with flexibility for inference-time adjustment. We explore the influence of different values of λ on customized generation and provide visual examples in Figure 9. Specifically, the sample generations closely resemble reference images when λ is small. As λ gradually increases, more concept-irrelevant content is injected into the concept branch, leading to a significant improvement in text alignment and a slight decrease in concept fidelity. When λ becomes too large, concept fidelity will sharply decline. This inference-time adjustment mechanism allows users to obtain the optimal generation for each individual query prompt. In practical applications, λ is typically set to $-2 \sim 2$ based on user demand.

Ablation of Key Components in DCI. We ablate the following three components in DCI: 1) We use masks in Eqn. 5 to selectively aggregate two branches.

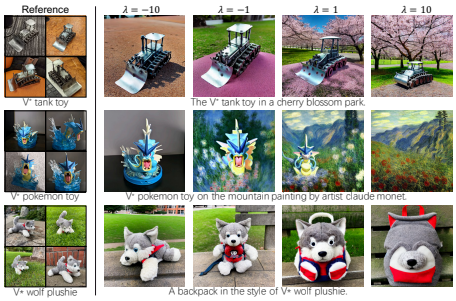

Fig. 9: Comparisons of generations with different injection strength λ. The reference images are shown on the left. The values of λ used in DCI are indicated at the top of each column.

Fig. 10: Visual comparisons of different variants. The reference images are shown on the left. Each column verifies the necessity of one component, which is indicated at the top.

In the *"w/o mask"* scenario, we verify the necessity of masks by ablating it and directly adding two branches. 2) We use entire prompts as the conditioning for each branch in DCI. In the *"w/o Entire Prompt"* scenario, we generate samples by simply using the content that each branch is responsible for as its prompt. For example, for the query prompt `"V* shoes carved from ice"`, we use `"V* shoes"` and `"carved from ice"` as the prompts for the concept branch and the auxiliary branch. 3) We aggregate two branches in each cross-attention layer within the U-Net. In the *"w/o Deep Agg."* scenario, we aggregate two branches by only adding the output of the last U-Net cross-attention layer. We show sample generations in Figure 10. As we can see, without masks, the influence scope of each branch extends to the entire image, resulting in integration confusion and subject distortion. The use of the entire prompt is essential for determining and aligning the layout of the images in two branches. When the prompt is separated, the two branches fail to target the correct regions they are responsible for and may generate artifacts. Finally, the injection capacity of the auxiliary branch will significantly degrade without deep integration.

4.4 Visualization of Aggregation Masks

The aggregation masks of the two branches vary spatially to enable adaptive integration. To better explain our method, we visualize aggregation masks of two samples in Figure 11. Since the masks are normalized across the two branches, we only display the masks of the auxiliary branch (*i.e.*, $\hat{\mathrm{M}}_{\mathcal{B}_a}^{t,i}$ in Eqn. 5) at various scales. The brighter pixels represent higher aggregation weights of the auxiliary branch. As observed, different content (*e.g.*, objects or style) appears to be injected at varying scales, thereby avoiding spatial confusion. Objects are injected at small scales, while style is injected at larger scales. Objects are injected at small scales, while style is injected at larger scales. Therefore, DCI is capable of performing prompts such as style transfer and property modification, where the functional region overlaps with the concept region.

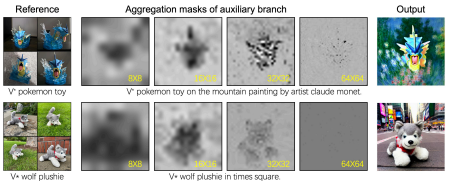

Fig. 11: Visualization of aggregation masks. We visualize aggregation masks of the auxiliary branch at various scales (indicated in each image). The brighter pixels represent higher aggregation weights.

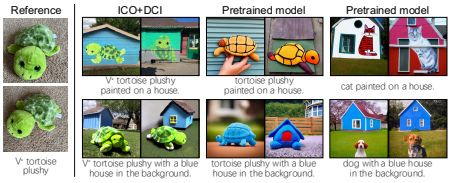

Fig. 12: Editability enhancement by noun replacement. By selecting appropriate auxiliary nouns, DCI can successfully render the given subject in contexts where the pretrined models struggle.

4.5 Advances and Limitations

A common noun that just differs from the original noun works well as the auxiliary noun in most cases. We preset common nouns for automatic auxiliary prompt construction. Furthermore, by selecting appropriate auxiliary nouns (*i.e.*, nouns that have a high probability of occurrence within the given context), DCI can overcome some limitations exhibited by pretrained models and achieve higher editability on the given concept. As exemplified in Figure 12, `"cat"` has a higher model prior than `"tortoise plushy"` in the context `"painted on a house"`. By selecting `"cat"` as the auxiliary noun, DCI successfully renders `"V* tortoise plushy"` in such a context, which even pretrained models struggle with. Here, `"cat"` serves as an example to represent a large group of common nouns. However, for some counterfactual or rarely occurred contexts, such as `"swimming on the moon"`, pretrained models have low model priors for all available nouns. In such cases, DCI fails to generate corresponding images faithfully.

5 Conclusion

In this paper, we explore and address the inherent fidelity-editability trade-off in customized generation. We propose a *"Divide, Conquer, then Integrate"* (DCI) framework, along with an *Image-specific Context Optimization* (ICO) fine-tuning strategy. ICO rectifies the customization process to obtain superior fine-tuned models, while DCI performs surgical adjustment in the early stage of denoising to liberate the fine-tuned models from the constraint of the fidelity-editability trade-off at inference. The synergy of ICO and DCI effectively reconciles the fidelity-editability trade-off, particularly in generations with low model priors, generating customized images with high concept fidelity while faithfully adhering to the query prompt. Furthermore, our method is capable of rendering concepts even in novel contexts where pre-trained models face challenges. Our future work will explore more controllable customized generation conditioned on additional information such as layout [6, 52] and depth [49–51].

Acknowledgements

The authors would like to thank the area chair and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62361166670, 62276132 and 61876085).

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS. pp. 1877–1901 (2020)
2. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024)
3. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: ICLR (2023)
4. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. In: NeurIPS. pp. 19822–19835 (2021)
5. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: CVPR. pp. 14084–14093 (2022)
6. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. In: NeurIPS. pp. 18225–18250 (2024)
7. 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: ICLR (2023)
8. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Designing an encoder for fast personalization of text-to-image models. In: Siggraph (2023)
9. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: ACL. pp. 3816–3830 (2021)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
11. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In: NeurIPS. pp. 15890–15902 (2024)
12. Guo, J., Wang, C., Wu, Y., Zhang, E., Wang, K., Xu, X., Shi, H., Huang, G., Song, S.: Zero-shot generative model adaptation via image-specific prompt learning. In: CVPR. pp. 11494–11503 (2023)
13. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: ICCV. pp. 7323–7334 (2023)
14. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Plug-and-play visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023)

15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: ICLR (2023)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020)
17. Huang, Z., Chan, K.C., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: CVPR. pp. 6080–6090 (2023)
18. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: ECCV. pp. 105–124 (2022)
19. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR. pp. 1931–1941 (2023)
20. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP. pp. 3045–3059 (2021)
21. Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. In: NeurIPS. pp. 2065–2075 (2019)
22. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022)
23. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR. pp. 10965–10975 (2022)
24. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL. pp. 4582–4597 (2021)
25. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
26. Liu, Y., Lu, Y., Liu, H., An, Y., Xu, Z., Yao, Z., Zhang, B., Xiong, Z., Gui, C.: Hierarchical prompt learning for multi-task learning. In: CVPR. pp. 10888–10898 (2023)
27. Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones: Concept neurons in diffusion models for customized generation. arXiv preprint arXiv:2303.05125 (2023)
28. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: CVPR. pp. 5206–5215 (2022)
29. Nichol, A., 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: ICML. pp. 16784–16804 (2022)
30. 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: ICML. pp. 8748–8763 (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
32. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831 (2021)
33. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: DenseCLIP: Language-guided dense prediction with context-aware prompting. In: CVPR. pp. 18082–18091 (2022)
34. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML. pp. 1060–1069 (2016)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

36. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
37. 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: CVPR. pp. 22500–22510 (2023)
38. 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. In: NeurIPS. pp. 36479–36494 (2022)
39. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: EMNLP. pp. 4222–4235 (2020)
40. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265 (2015)
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
43. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: CVPR. pp. 139–149 (2022)
44. 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. In: ICCV. pp. 15943–15953 (2023)
45. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In: ECCV. pp. 720–736 (2022)
46. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)
47. 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: CVPR. pp. 1316–1324 (2018)
48. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: CVPR. pp. 14256–14266 (2023)
49. Yan, Z., Li, X., Wang, K., Zhang, Z., Li, J., Yang, J.: Multi-modal masked pre-training for monocular panoramic depth completion. In: ECCV. pp. 378–395. Springer (2022)
50. Yan, Z., Lin, Y., Wang, K., Zheng, Y., Wang, Y., Zhang, Z., Li, J., Yang, J.: Tri-perspective view decomposition for geometry-aware depth completion. In: CVPR. pp. 4874–4884 (2024)
51. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J., Yang, J.: Rignet: Repetitive image guided network for depth completion. In: ECCV. pp. 214–230. Springer (2022)
52. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: CVPR. pp. 14246–14255 (2023)
53. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR. pp. 6757–6767 (2023)
54. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. TMLR (2022)

55. 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: ICCV. pp. 5907–5915 (2017)
56. Zhang, Z., Han, L., Ghosh, A., Metaxas, D.N., Ren, J.: Sine: Single image editing with text-to-image diffusion models. In: CVPR. pp. 6027–6037 (2023)
57. Zhao, J., Zheng, H., Wang, C., Lan, L., Yang, W.: Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. In: ICCV. pp. 22592–22602 (2023)
58. Zhong, Z., Friedman, D., Chen, D.: Factual probing is [mask]: Learning vs. learning to recall. In: NAACL. pp. 5017–5033 (2021)
59. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
60. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)
61. Zhou, Y., Zhang, R., Sun, T., Xu, J.: Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. arXiv preprint arXiv:2305.13579 (2023)
62. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. arXiv preprint arXiv:2205.14865 (2022)