ComFusion: Enhancing Personalized Generation by Instance-Scene Compositing and Fusion

Yan Hong[1], Yuxuan Duan[2], Bo Zhang[2], Haoxing Chen[1], Jun Lan[1], Huijia Zhu[1], Weiqiang Wang[1], Jianfu Zhang[3⋆]

[1] Ant Group
[2] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[3] Qing Yuan Research Institute, Shanghai Jiao Tong University
yanhong.sjtu@gmail.com, {sjtudyx2016,bo-zhang}@sjtu.edu.cn, hx.chen@hotmail.com, {yelan.lj, huijia.zhj, weiqiang.wwq}@antgroup.com, c.sis@sjtu.edu.cn

Abstract. Recent progress in personalizing text-to-image (T2I) diffusion models has demonstrated their capability to generate images based on personalized visual concepts using only a few user-provided examples. However, these models often struggle with maintaining high visual fidelity, particularly when modifying scenes according to textual descriptions. To address this challenge, we introduce ComFusion, an innovative approach that leverages pretrained models to create compositions of user-supplied subject images and predefined text scenes. ComFusion incorporates a class-scene prior preservation regularization, which combines subject class and scene-specific knowledge from pretrained models to enhance generation fidelity. Additionally, ComFusion uses coarse-generated images to ensure alignment with both the instance images and scene texts, thereby achieving a delicate balance between capturing the subject's essence and maintaining scene fidelity. Extensive evaluations of ComFusion against various baselines in T2I personalization have demonstrated its qualitative and quantitative superiority.

1 Introduction

Text-to-image (T2I) personalization aims to customize a diffusion-based T2I model with user-provided visual concepts [12, 27, 39]. This innovative approach enables the creation of new images that seamlessly integrate these concepts into diverse scenes. More formally, given a few images of a subject (no more than five), our objective is embed this subject into the model's output domain. This allows for the synthesis of the subject with a unique identifier in various scenes. The task of rendering such imaginative scenes is particularly challenging. It entails the synthesis of specific subjects (*e.g.*, objects, animals) in new contexts, ensuring their natural and seamless incorporation into the scene. Such a task demands a delicate balance between the subject's distinctive features and the new scene context. Recently, this field of T2I personalization

⋆ Corresponding author.

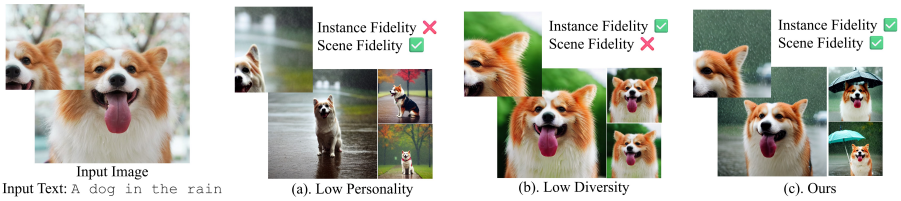

Fig. 1: Contrasting with existing methods [2,39] , which often face challenges in simultaneously preserving instance fidelity and scene fidelity, ComFusion skillfully composites the instance image with textual prompts and fuses the visual details of the subject instance with the textual variations of the scenes, yielding the creation of plausible, personalized images that exhibit a rich diversity.

has attracted significant attention from the academic community with many works [1, 2, 12, 25, 31, 41, 43, 48, 51, 53] leveraging the capabilities of advanced diffusion-based T2I models [16, 24, 34, 36, 37, 40, 50, 55, 56, 58, 59].

These approaches broadly fall into two categories: The first category [8,13,22, 27, 41] integrates additional modules with a pretrained base model. This stream enables the creation of personalized subjects without the need for finetuning during testing. However, it often struggles to maintain the subject's identity consistently across different synthesized images. In contrast, the second category [2, 12, 25, 39] focuses on finetuning the pretrained model with a select set of images, employing various regularization techniques and training strategies. This finetuning process effectively utilize the model's existing class knowledge, combined with the unique identifier of the subject, thereby allowing for the generation of diverse variations of the subject in various contexts.

Methods that finetune diffusion models for high-quality image generation face considerable challenges, primarily because existing designs [2,12,25,39] often focus on constraints for instance images and text prompts independently, without adequately considering the interplay between them. As a result, these models might lose their pre-finetuning knowledge and the ability to understand and generate a broad range of classes and scenes. This limitation becomes apparent in attempts to create images within novel scenes, leading to less-than-optimal generation or the integration of various prompts or subjects. In Fig. 1, we show an example of the personalized generation using a specific dog instance image and the text prompt "`A dog in the rain`". The images, generated by existing leading methods [2, 39] and our proposed approach. Fig. 1 (a) illustrates a lack of *instance fidelity*, where the generated images fail to preserve the subject dog's appearance, resulting in low-personality output. Fig. 1 (b) highlights examples with insufficient *scene fidelity*, failing to accurately represent the rainy scene, thus limiting the diversity of the generated images.

To enhance the instance fidelity and scene fidelity of generated images, we propose **ComFusion** (**Com**posite and **Fusion**), a novel approach designed for personalized subject generation across varied scenes. ComFusion integrates vi-

sual information from instance images with textual information from text prompts through compositing and fusion, facilitating the synthesis of new images where high-fidelity instances are seamlessly incorporated into diverse scenes. To achieve this, ComFusion is structured around two streams: a *composite stream* and a *fusion stream*. The composite stream incorporates a *class-scene prior loss* to retain the pretrained model's understanding of both the subject class and the novel scene. This approach results in the production of a wide variety of images that merge the class and scene priors with the subject instances and their respective contexts, facilitating a cohesive synthesis of subject instances within their scene contexts. The fusion stream introduces a *visual-textual matching loss* to effectively merge the subject instance's visual data with the scene's textual information. This ensures their collective depiction in the coarsely generated images, achieving a balanced representation between instance fidelity and scene fidelity. In Fig. 1, we present some impressive samples obtained by ComFusion. The images illustrate a remarkable preservation of the dog's appearance, while the scene "`in the rain`" is brought to life with vivid details such as rain spots and umbrellas. Our extensive experimental analysis across various subject instances and scenes underscores ComFusion's superior performance, both qualitatively and quantitatively, over existing approaches.

2 Related Works

2.1 Diffuion-Based Text-to-Image Generation.

The field of Text-to-Image (T2I) generation has recently witnessed remarkable advancements [15, 23, 40, 46, 47, 61], predominantly led by pre-trained diffusion models such as Stable Diffusion [37], DALLE [36], Imagen [40] and *etc*. These models are renowned for their exceptional control in producing photorealistic images that closely align with textual descriptions. This innovation has paved the way for diverse applications, including video generation [3,5,11,14,54] and 3D object creation [29,42,52,57,62]. Despite their superior capabilities in generating high-quality images, these models encounter challenges in more personalized image generation tasks , which are often difficult to precisely describe with text descriptions. This challenge has sparked interest in the rapidly evolving field of personalized T2I generation [12, 25, 31, 39, 48].

2.2 Personalized Text-to-Image Generation

Given a small set of images of the subject concept, personalized T2I generation [1, 7, 12, 13, 17, 25, 31, 36, 39–41, 44, 48, 51, 53] aims to generate new images according to the text descriptions while maintaining the identity of the subject concept. Early studies in training generative models in few-shot setting focus on alleviating mode collapse [30, 45, 49] for generative adversarial networks [9, 10, 18–20, 28, 60]. Recently, diffusion-based text-to-image models with a few images have also been explored in [2, 39]. In the stream of diffusion-based

generators, personalized T2I generation methods can be classified into two categories: The first stream involves the integration of additional modules (*e.g.*, [21, 33, 61]) with a pretrained base model. The second stream adopts a strategy of finetuning the pretrained model using a few selected images.

2.3 T2I personalization Without Finetuning.

These methods without finetuning [8, 13, 22, 27, 41] generally rely on additional modules trained on additional new datasets, such as the visual encoder in [41, 53] and the experts in [8, 27] to directly map the image of the new subject to the textual space. Specifically, [13] introduces an encoder that encodes distinctive instance information, enabling rapid integration of novel concepts from a given domain by training on a diverse range of concepts within that domain. In [41], a learnable image encoder translates input images into textual tokens, supplemented by adapter layers in the pre-trained model, thus facilitating rich visual feature representation and instant text-guided image personalization without requiring test-time finetuning. DisenBooth [7] uses weak denoising and contrastive embedding auxiliary tuning objectives for personalization. ELITE [53] introduces a method for learning both local and global maps on large-scale datasets, allowing for instant adaptation to unseen instances using a single image marked with the subject concept for personalized generation.

2.4 T2I personalization with Finetuning

Various methods employ diverse finetuning strategies to optimize different modules within pretrained models [12, 25, 39]. DreamBooth [39] and TI [12] are two popular subject-driven text-to-image generation methods based on finetuning. Both approaches map subject images to a special prompt token during finetuning. They differ in their finetuning focus: TI concentrates on prompt embedding, while DreamBooth targets the U-Net model and text-encoder. Recent finetuning-based methods [25, 51] focus on how to design training strategy to update core parameters of T2I model for subject concepts on user-provided 4-6 images. A domain-agnostic method is proposed in [1] that introduce a novel contrastive-based regularization technique. This technique aims to preserve high fidelity to the subject concept's characteristics while keeping the predicted embeddings close to editable regions of the latent space. Break-A-Scene [2] utilizes the subject concept's mask and employs a two-stage process for personalized T2I generation using a single image. However, this approach is limited in terms of the subject's diversity.

3 ComFusion

In this section, we present ComFusion, our cutting-edge technique engineered to enhance personalized subject generation across diverse scenes. Utilizing a

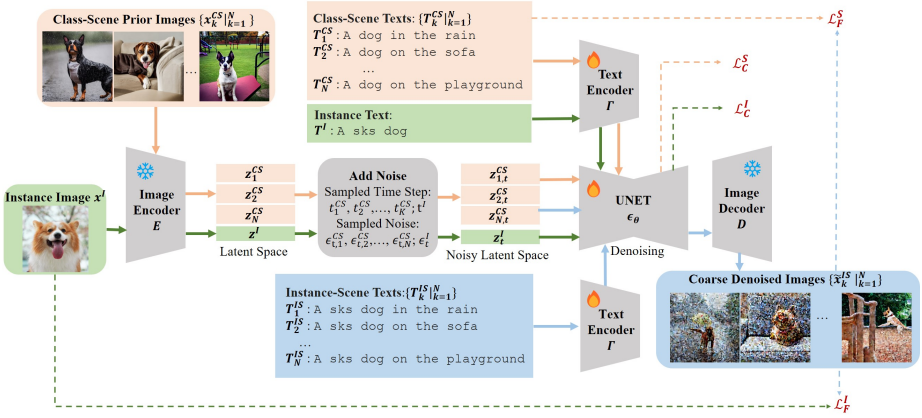

Fig. 2: The illustration of ComFusion framework. ComFusion consists of a *composite stream* (highlighted with green and orange arrows, details in Sec. 3.1 and Sec. 3.3) and a *fusion stream* (highlighted with blue arrows, details in Sec. 3.4).

constrained set of no more than five subject instance images, our goal is to create new, high-fidelity images of the subject, steered by textual prompts. These prompts facilitate a range of alterations including the subject's placement, backdrop, posture, perspective, and other context-specific transformations, without restrictions on the instance images' capture scenarios. Our main paper prioritizes the **one-shot scenario**—a setting that employs just a single instance image, marking it as **the most challenging setting** due to its minimal instance information. *While our primary emphasis is on the one-shot setting, it's crucial to recognize ComFusion's extensive adaptability in various personalized generation contexts, such as using multiple subject images, modifying properties/styles/viewpoints, etc. Due to space constraints, we direct readers to our appendix for a deeper exploration of these capabilities.* The generated images aim to be faithful (*i.e.*, accurately reflect the content) both the subject instance and the text prompts, which manifests in two key aspects - *instance fidelity*: ensuring visual congruence with the instance image and *scene fidelity*: aligning the scenes in the newly created images with the provided prompts. As shown in Fig. 2, we design a two-stream training strategy for ComFusion, consisting of a *composite stream* supervised by *instance finetune loss* and *class-scene prior loss* (denoted as $\{\mathcal{L}_C^I, \mathcal{L}_C^S\}$ in green and orange stream of Fig. 2 and demonstrated in Sec. 3.3) and a *fusion stream* supervised by *visual-textual matching loss* (denoted as $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$ in blue stream and presented in Sec. 3.4).

3.1 Finetuning Text-to-Image Diffusion Models.

ComFusion finetunes specific pretrained diffusion models, *e.g.*, Stable Diffusion [37], which consists of an auto-encoder (encoder E and a decoder D), a

text-encoder Γ, and a denoising model ϵ_θ architectured with U-Net [38]. The auto-encoder maps an image $\boldsymbol{x}$ into low-dimensional latent $\boldsymbol{z}$ with encoder E and recover the original image $\tilde{x}$ with decoder D after the denoising process. The denoising model ϵ_θ is trained on the latent space to produce denoised latent based on the textual condition source from $\Gamma(\boldsymbol{T})$, where $\boldsymbol{T}$ is the user-provided prompt providing the information (*e.g.*, subject classes, instance attributes, scenes) of the generated images and Γ denotes the pretrained CLIP text encoder [35]. Given a single subject *instance image* $\boldsymbol{x}^I$ from a *subject class*, the instance image is captioned with *instance text* $\boldsymbol{T}^I =$ "`a [identifier] [class noun]`" (*e.g.*, "`a sks dog`"), where "`[class noun]`" is a coarse class (*e.g.*, "`dog`") provided by user and "`[identifier]`" is an unique identifier for the subject concept (*e.g.*, "`sks`"). Given a single instance image $\boldsymbol{x}^I$, the pretrained models will be finetuned with *instance finetune loss*:

$$\mathcal{L}_C^I = \mathbb{E}_{\boldsymbol{z}\sim\{\boldsymbol{z}^I\},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{z}_t, t, \Gamma(\boldsymbol{T}^I))\|_2^2\right], \tag{1}$$

where $t \sim \mathcal{N}(0,1)$ is the time step, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ is the unscaled Gaussian noise, $\boldsymbol{z}_t$ are the noisy latent at time t, and $\{\boldsymbol{z}^I\}$ is the latent of instance images $\{\boldsymbol{x}^I\}$ processed by encoder E. This finetuning mechanism integrates a new (unique identifier, subject) air into the model's knowledge base, capitalizing on the model's inherent class understanding while embedding the subject's unique identifier. This strategy enhances the model's ability to generate novel subject variations in varied contexts using existing visual priors.

3.2 Class Prior Loss.

Finetuning diffusion models introduces the challenge of *language drift* [26,32], a phenomenon where models diverge from understanding the essence of language syntax and semantics, concentrating excessively on task-specific details. In our context, this results in the finetuned model losing prior knowledge, including the ability to recognize various classes and scenes integral to pretrained models, thereby diminishing *scene fidelity*. To mitigate this issue, existing methods employ a specific *prior loss* to regularize the model. Typically, this loss function involves using designated *prior texts* $\{\boldsymbol{T}_i^P|_{i=1}^N\}$, input into the pretrained model to generate *prior images* $\{\boldsymbol{x}_i^P|_{i=1}^N\}$ based on prior texts, where N is the number of prior text-image pairs. his loss function maintains the model's dedication to its pretrained knowledge base, essential for preserving foundational knowledge during few-shot finetuning. An exemplary approach is the class prior loss proposed by DreamBooth [39]. It leverages coarse class to form *class text* $\boldsymbol{T}^C =$ "`a [class noun]`" (*e.g.*, "`a dog`"), which is fed into the pretrained model to produce *class prior images* $\{\boldsymbol{x}_i^C|_{i=1}^N\}$ depicting the coarse class without requiring the preservation of specific subject instances. Given class prior images $\{\boldsymbol{x}_i^C|_{i=1}^N\}$, the pretrained models will be finetuned with:

$$\mathcal{L}_{dream} = \mathbb{E}_{\boldsymbol{z}\sim\{\boldsymbol{z}^C\},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{z}_t, t, \Gamma(\boldsymbol{T}^C))\|_2^2\right], \tag{2}$$

where $\{\boldsymbol{z}^C\}$ represents latent of class prior images $\{\boldsymbol{x}^C\}$ and the other terms are defined similar to Eq. (1). This class-specific prior-preservation loss supervises

the model to reconstruct class prior images, balancing this objective with the instance finetune loss. It facilitate the generation of varied instances within the subject's class.

3.3 Composition: Class-Scene Prior Loss

The class prior loss in Eq. (2) leverages semantic priors related to classes, embedding them into the model's framework to enable the creation of diverse instances within a given subject class. Ideally, models fine-tuned with both instance finetune loss and class prior loss, drawing on the vast knowledge base of large-scale T2I pretrained models capable of rendering any scene, should maintain high levels of both instance and scene fidelity. However, this approach primarily addresses language drift related to the subject class but may overlook drifts in text prompts describing the scenes of the generated images, leading to *catastrophic neglecting* [6]. In large-scale pretrained T2I models like Stable Diffusion [37], models trained on a vast array of image-text pairs demonstrate proficiency in generating novel images based on combinations of random texts. Nonetheless, the neglecting phenomenon remains an issue in certain scenarios, where some prompts or subjects are not adequately generated or integrated by these large-scale models [6]. In contrast, few-shot learning paradigms, relying on a limited set of image-text pairs, often yield less optimal responses to complex subject instances and scene texts than their large-scale counterparts, potentially exacerbating the neglecting issue. This can lead to loss of scene fidelity or instance fidelity for personalized generation, as illustrated in Fig. 1.

Class-Scene Prior Loss. Given our objective to *composite* new subject instance images within various specific scenes, we updated the prior-preservation loss in Eq. (2) to *class-scene prior loss*. This update is specifically designed to maintain the pretrained model's knowledge of both class and scene, thereby significantly enhancing *scene fidelity*. By integrating class-scene prior loss with instance finetune loss, ComFusion effectively preserves and leverages the extensive understanding of class and scene inherent to the pretrained model. To elaborate, we initially generate a descriptive *class-scene text set* $\{\boldsymbol{T}^{CS}\}$. This set *composites* the subject class information "`[class noun]`" (*e.g.*, "`dog`") and scene information "`[scene]`" (*e.g.*, "`in the rain`"), resulting in class-scene texts "`a [class noun] [scene]`" (*e.g.*, "`a dog in the rain`"). The *class-scene prior images* $\{(\boldsymbol{x}^{CS}\}$ is generated by pretrained model with the corresponding $\{\boldsymbol{T}^{CS}\}$. Subsequently, these richly detailed class-scene image-text pairs $(\boldsymbol{x}_k^{CS}, \boldsymbol{T}_k^{CS})$ combined with instance image-text pairs $(\boldsymbol{x}^I, \boldsymbol{T}^I)$ are fed into ComFusion to finetune the diffusion model. In a manner akin to Eq. (2), the trainable parameters of ComFusion are optimized by class-scene prior loss:

$$\mathcal{L}_C^S = \mathbb{E}_{(\boldsymbol{z},\boldsymbol{T})\sim\{(\boldsymbol{z}^{CS},\boldsymbol{T}^{CS})\},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{z}, t, \Gamma(\boldsymbol{T}))\|_2^2\right], \tag{3}$$

where $\{(\boldsymbol{z}^{CS}, \boldsymbol{T}^{CS})\}$ is the set of latent-text pairs corresponding to $\{(\boldsymbol{x}^{CS}, \boldsymbol{T}^{CS})\}$. Similar to $\mathcal{L}_{dream}$ in Eq. (2), this class-scene prior loss will be trained together with instance finetune loss $\mathcal{L}_C^I$ in Eq. (1) with a λ_C controls for the relative

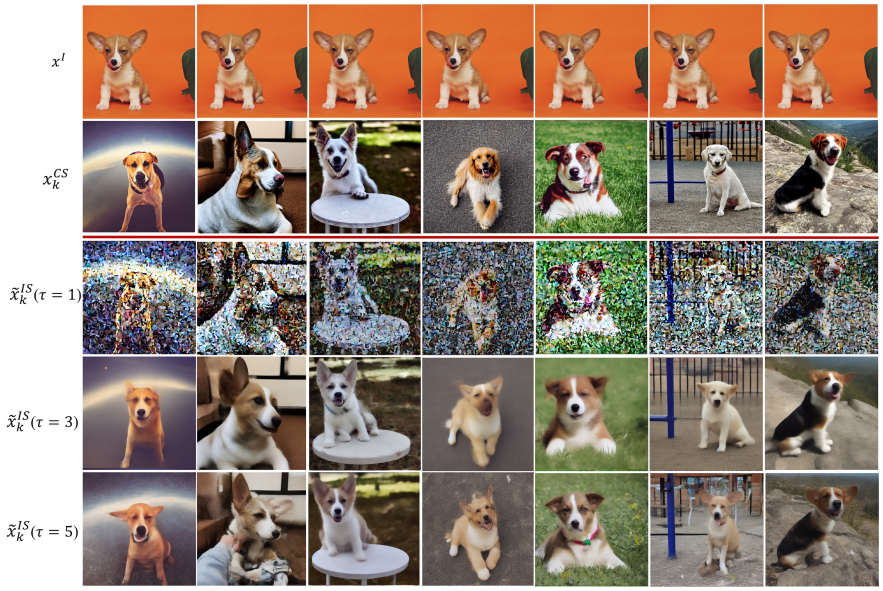

Fig. 3: The coarse generated results $\tilde{\boldsymbol{x}}_k^{IS}$ under the supervision of visual-textual fusion loss $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$ with denoising steps τ selected from $\{1, 3, 5\}$. The instance within these coarse-generated images closely resembles the instance image $\boldsymbol{x}^I$, while harmoniously aligning with the specific scene depicted in prior images $\boldsymbol{x}_k^{CS}$.

weight of this term. $\mathcal{L}_C^I$ and $\mathcal{L}_C^S$ formulate the objective of the *composite stream* in ComFusion. Different from $\mathcal{L}_{dream}$, the class-scene prior text $\boldsymbol{T}^{CS}$ in $\mathcal{L}_C$ articulates a comprehensive delineation of subject class information and meticulous scene descriptors as specified by $\boldsymbol{T}^{CS}$. This loss formulation adeptly tackles the language drift issue related to class and scene knowledge in the finetuned model. It facilitates the generation of a varied collection of images that capture the essence of both class and scene priors from the pretrained model while integrating these elements with subject instances and their specific contexts. Such integration enables ComFusion to produce images that accurately depict subject instances within their respective scenes.

3.4 Fusion: Visual-Textual Matching Loss

ComFusion integrates a *visual-textual matching loss* to *fuse*the visual characteristics of subject instances with the textual information of scenes, thereby amplifying the fidelity of both instances and scenes. This integration ensures a cohesive and precise portrayal of subject instances within their contextual scenes. The key idea behind the visual-textual matching loss is to generate coarse images that encapsulate both the subject instance and scene texts. This loss mechanism ensures that these initial images blend the distinct visual attributes of the in-

stances with the descriptive intricacies of the scenes. Consequently, it aligns the generated images closely with both the visual appearance of the subject instance and the descriptive elements of the scene texts.

Coarse Generated Images. Specifically, for a class-scene prior image $\boldsymbol{x}^{CS}$ annotated with detailed class-scene text $\boldsymbol{T}^{CS}$, according to the standard forward process of DDPMs [17], we add a random and appropriate level of noise $\boldsymbol{\epsilon}_t$ with timestep t to obtain noisy latent $\boldsymbol{z}_t^{CS}$. This is designed to infuse new information while preserving the structural features of class-scene prior images. We then create *instance-scene text* $\boldsymbol{T}^{IS}$ by replacing the "`[class noun]`" with "`[identifier] [class noun]`" in class-scene text (*e.g.*, "`a sks dog in the rain`"). This modified text is processed by the text-encoder Γ to obtain conditional textual information, which is then used to iteratively denoise the noisy latent $\boldsymbol{z}_t^{CS}$ to denoised latent $\tilde{\boldsymbol{z}}^{IS}$. Utilizing the accelerated generation mechanism in DDIMs, we set τ as the number of steps required to denoise $\boldsymbol{z}_t^{CS}$ into $\tilde{\boldsymbol{z}}^{IS} = f_\theta^\tau(\boldsymbol{z}_t^{CS}, t, \boldsymbol{T}^{IS}, \tau)$, where $f_\theta(\cdot)$ is the coarse denoising function. Generally, $f_\theta(\cdot)$ significantly reduces the computational effort required for denoising from t steps to τ steps, producing a coarse denoised latent. The denoised latent $\tilde{\boldsymbol{z}}^{IS}$ is then decoded using decoder D, resulting in a *coarse denoised image* $\tilde{\boldsymbol{x}}^{IS} = D(\tilde{\boldsymbol{z}}^{IS})$. This image, guided by the instance-scene text, fuses both the subject and the scene's features. Fig. 3 illustrates examples of coarse denoised images under different settings of τ.

$f_\theta(\cdot)$ is specifically designed to generate coarse denoised images through a τ steps iteration process. It is intended for recursive application, executed τ times. Each iteration of $f_\theta(\cdot)$ progressively reduces the noise in $\boldsymbol{z}_t$, finally resulting in a coarse denoised latent. For a single-step iteration ($\tau = 1$), the function simplifies to $f_\theta(\boldsymbol{z}_t, t, \boldsymbol{T}, \tau) = h_\theta(\boldsymbol{z}_t, t, \boldsymbol{T})$. For multiple iterations ($\tau > 1$), the function recursively applies as follows:

$$\begin{aligned} f_\theta(\boldsymbol{z}_t, t, \boldsymbol{T}, \tau) = f_\theta(\sqrt{\alpha_{r(\tau,t)}} h_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{T}\right) + \\ \sqrt{1-\alpha_{r(\tau,t)}} \epsilon_\theta\left(\boldsymbol{z}_t, t, \Gamma(\boldsymbol{T})\right), t, \boldsymbol{T}, \tau - 1), \end{aligned} \tag{4}$$

where $r(\tau, t) = \lceil \frac{v \times t}{\tau} \rceil$. In the implementation, we carefully select the timestep t from $[\lceil 0.2T \rceil, \lceil 0.8T \rceil]$. This choice aims to guarantee that the coarse denoised image aptly merges details from both the subject instance and the scene description. If t is too close to 1, the influence of the instance-scene text $\boldsymbol{T}^{IS}$ becomes limited, resulting in $\tilde{\boldsymbol{x}}^{IS}$ lacking sufficient visual cues of the subject instance. If t approaches T, the effect of the class-scene latent $\boldsymbol{z}^{CS}$ diminishes due to excessive noise, causing a loss of scene information in $\tilde{\boldsymbol{x}}^{IS}$.

Visual-Textual Matching. To ensure that the instance image $\tilde{\boldsymbol{x}}^{IS}$ closely mirrors the textural structure (*resp.,* visual appearance) of class-scene prior image $\boldsymbol{x}^{CS}$ (*resp.,* instance image $\boldsymbol{x}^I$), we employ a dual visual-textual fusion loss:

$$\mathcal{L}_F^I = \mathbb{E}_{\boldsymbol{x} \sim \{\tilde{\boldsymbol{x}}_k^{IS}\}}\left[-f_{\mathbf{V}}(\boldsymbol{x}, \boldsymbol{x}^I)\right], \quad \mathcal{L}_F^S = \mathbb{E}_{(\boldsymbol{x}', \boldsymbol{T}) \sim \{(\tilde{\boldsymbol{x}}_k^{IS}, \boldsymbol{T}_k^{CS})\}}\left[-f_{\mathbf{T}}(\boldsymbol{x}', \boldsymbol{T})\right], \tag{5}$$

where the first (*resp.,* second) term is represented by $\mathcal{L}_F^I$ (*resp.,* $\mathcal{L}_F^S$) in Fig. 2. $f_{\mathbf{V}}(\tilde{\boldsymbol{x}}_k^{IS}, \boldsymbol{x}^I)$ is used to calculate the cosine similarity between DINO embedding:

$\mathbf{DINO}(\tilde{\boldsymbol{x}}_k^{IS})$ and $\mathbf{DINO}(\boldsymbol{x}^I)$ with pretrained ViT-S/16 DINO [4]. DINO [4] is a self-supervised pretrained transformer that excels in visual information extraction from images. By employing self-supervised learning techniques, DINO adeptly identifies and encodes complex visual patterns, making it particularly effective for image analysis tasks. Consequently, a higher similarity score from $f_{\mathbf{V}}(\cdot)$ indicates that the generated $\tilde{\boldsymbol{x}}_k^G$ closely resembles the instance image $\boldsymbol{x}^I$ in visual terms. $f_{\mathbf{T}}(\tilde{\boldsymbol{x}}_k^{IS}, \boldsymbol{T}_k^{CS})$ aims at calculating the cosine similarity of visual embedding of generated image $\mathbf{CLIP}(\tilde{\boldsymbol{x}}_k^{IS})$ and textual embedding of class-scene prior text $\mathbf{CLIP}(\boldsymbol{T}_k^{CS})$. CLIP [35] is a prototype in image-text cross-modal pretraining model, aligns visual and textual information to enhance model comprehension of and correlation between image contents and their textual descriptions. Its capability to connect visual and textual domains is crucial for tasks requiring an in-depth understanding of both, ensuring accurate and effective image-text alignments. A higher similarity from $f_{\mathbf{T}}(\cdot)$ suggests that the model is inclined to produce images that encapsulate specific scene details accurately. By applying both visual loss $\mathcal{L}_F^I$ and textual loss $\mathcal{L}_F^S$ on the same generated image $\tilde{\boldsymbol{x}}_k^{IS}$, ComFusion mitigates the catastrophic neglecting problem and achieves a better balance between instance fidelity and scene fidelity. As a result, it yields a more harmonious and precise depiction that effectively captures the core characteristics of instance fidelity and scene fidelity.

3.5 Overall Objectives and Inference Process

ComFusion's objective function integrates the instance finetune loss in Eq. (1) is combined with the class-scene prior loss in Eq. (3) and visual-textual fusion loss in Eq. (5):

$$\mathcal{L}_{total} = \mathcal{L}_C^I + \lambda_C^S \mathcal{L}_C^S + \lambda_F^I \mathcal{L}_F^I + \lambda_F^S \mathcal{L}_F^S, \tag{6}$$

where λ_C^S, λ_F^I, and λ_F^S represent the respective weights of $\mathcal{L}_C^S$, $\mathcal{L}_F^I$, and $\mathcal{L}_F^S$. $\mathcal{L}_{total}$ is employed to finetune the trainable parameters of text-encoder Γ and U-Net ϵ_θ based on pretrained Stable Diffusion [37]. During this process, the parameters of the auto-encoders remain fixed. In the inference phase, ComFusion follows the standard T2I inference protocol: generating a random latent, followed by denoising this latent using the prompt "`a [identifier] [class noun] [scene]`" with the U-Net. Finally, the denoised latent is decoded to produce new images.

4 Experiments

In this section, we compare ComFusion to both state-of-the-art and concurrent work baselines and provide comprehensive quantitative and qualitative comparisons. Then, we study in greater depth the properties of ComFusion by ablation studies.

4.1 Experimental Settings and Details

Implementation Details. All methods were applied using a pre-trained Stable Diffusion (SD) checkpoint 1.5 [37]. We trained ComFusion and DreamBooth for

Table 1: Comparative analysis of quantitative metrics for instance fidelity (DINO, CLIP-I) and scene fidelity (CLIP-T).

Methods	DINO (↑)	CLIP-I (↑)	CLIP-T (↑)
Real Images	*0.795*	*0.859*	N/A
DreamBooth [39]	0.619	0.752	0.229
TI [12]	0.465	0.634	0.185
CD [25]	0.615	0.724	0.205
XTI [51]	0.435	0.601	0.198
ELITE [53]	0.405	0.615	0.249
Break-A-Scene [2]	0.632	0.771	0.294
Ours	**0.658**	**0.814**	**0.321**

Table 2: User preference for instance fidelity and scene fidelity across various methods.

Methods	Instance fidelity (↑)	Scene fidelity (↑)
DreamBooth [39]	3.1%	3.8%
TI [12]	0.3%	0.0%
CD [25]	6.2%	1.0%
XTI [51]	0.3%	1.8%
ELITE [53]	0.0%	11.1%
Break-A-Scene [2]	34.5%	20.2%
Ours	**55.6%**	**62.1%**

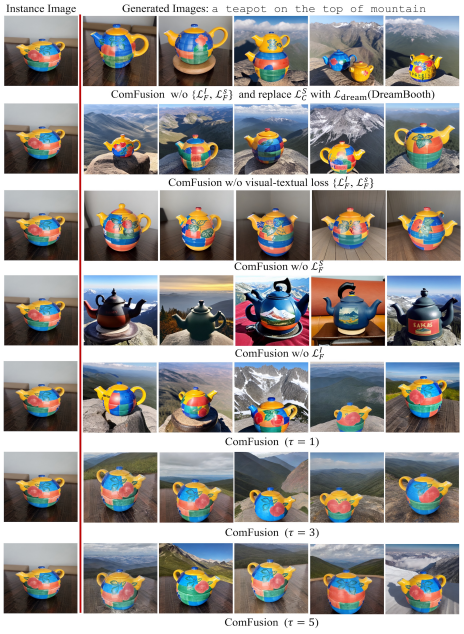

Fig. 4: Visual ablative results.

1200 steps, using a batch size of 1 and learning rate 1×10^{-5}. The number of prior images N is set as 200 for fair comparison. During training, the hyper parameters λ_C^S (*resp.*, λ_F^S, λ_F^I,τ) is set as 1 (*resp.*, 0.01, 0.01,3). All experiments are conducted with 1 $A100$ GPU. Detailed implementation information for all baselines is provided in appendix.

Datasets. To evaluate the effectiveness of our proposed methods among different datasets, we use a combined dataset of the TI [12] dataset of 5 concepts, and the dataset from DreamBooth [39] with 20 concepts. For both datasets, each concept selects one original image as an instance image. We perform experiments on 25 subject datasets spanning a variety of categories and varying training samples. We evaluate all the methods with 15 distinct scenes. Also, we use $\boldsymbol{T}^{CS}$ = "`a [class noun] Scene`" with the same scene prompts to sample prior images with 15 scenes for ComFusion. Detailed information about the subject datasets and scene prompts is available in the appendix. Experiments involving more than one instance image, other specific scenes, and scenarios without specific scenes for all methods are also documented in appendix.

Baselines. We compare our ComFusion described in Section Sec. 3 with DreamBooth [39], Textual-Inversion(TI) [12], Custom-Diffusion (CD) [25], Extended Textual-Inversion (XTI) [51], ELITE [53], and Break-A-Scene [2]. Details of these baseline methods are reported in appendix.

Evaluation Metrics. Following DreamBooth [39], for each method, we generated 10 images for each of 25 instances and each of 15 scenes, totaling 3750

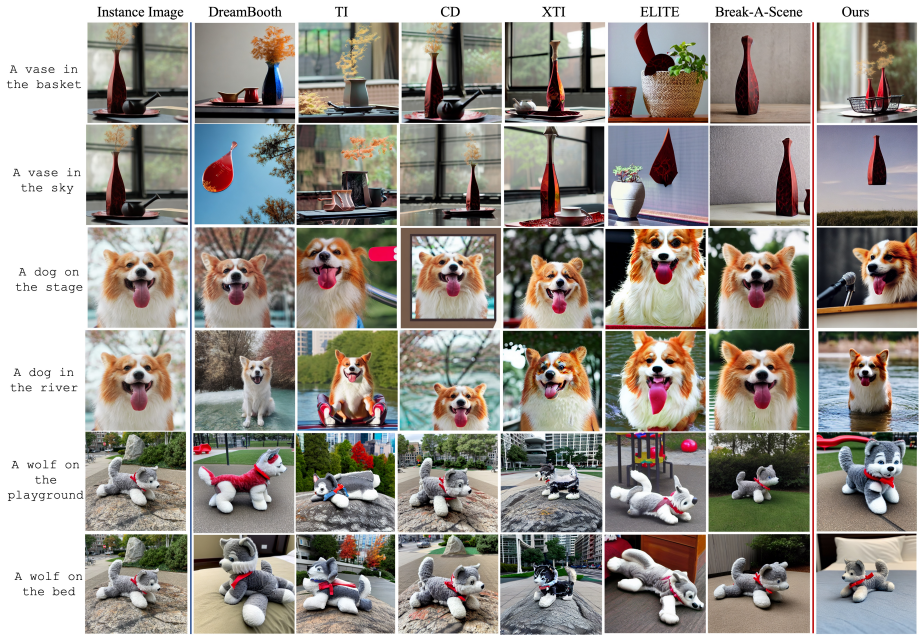

Fig. 5: Comparative display of images generated from a single instance image in various specific scenes by DreamBooth [39], TI [12],CD [25],XTI [51],ELITE [53], Break-A-Scene [2], and our proposed ComFusion.

images for evaluation of robustness and generalization abilities of each method. Following DreamBooth [39] and CD [25], we evaluate those methods on two dimensions including instance fidelity and scene fidelity. CLIP-I [35] and DINO score [4] were used to evaluate instance fidelity by measuring the similarity between generated images and instance images, and the alignment between textual scene with generated images are measured by CLIP-T [35]. Detailed descriptions of these measurement metrics are provided in appendix.

4.2 Comparisons with Baselines

Quality Assessments of Generated Images. We perform the quantitative evaluation on the instance fidelity using DINO score and CLIP-I score, and scene fidelity with CLIP-T score. In Tab. 1, "Real Images" represents a measure of the similarity between a given single image and the remaining real images belonging to the same subject as the given image, providing the upper bound of fidelity of the subject. Comparisons in Tab. 1 indicate that our ComFusion achieves the highest scores for DINO, CLIP-I, and CLIP-T, indicating that it can generate high-fidelity images with higher instance fidelity and scene fidelity than baseline methods.

Human Perceptual Study. Further, following DreamBooth [39], we conduct a user study to evaluate the instance fidelity and scene fidelity of generated images. In detail, based on generated 3750 images per method including 6 baseline methods and our method, we present results generated from different methods in random order and we ask 12 users to choose. (1) Instance fidelity: determining which result better preserves the identity of the instance image, and (2) Scene fidelity: evaluating which result achieves better alignment between the given textual scene and the generated image. We collect 90,000 votes from 12 users ($12 \times 3750 \times 2$) for instance fidelity and scene fidelity, and show the percentage of votes for each method in Tab. 2. The comparison results demonstrate that the generated results obtained by our method are preferred more often than those of other methods.

Qualitative Evaluations. To evaluate the superiority of our ComFusion in balancing the accuracy of subjects and the consistency of multiple specific scenes, we visualize comparison results in Fig. 5. We can see that images generated by TI [12], CD [25], and XTI [51] are similar to input instance image in terms of structure, those methods fail to make response to the specific scene in given testing prompts. ELITE [53] may generate distorted images in unexpected scenes. Images generated by Break-A-Scene [2] maintain instance fidelity while may fail to composite subject instance in specific scenes. In contrast, our ComFusion can generate images of higher instance fidelity and scene fidelity. This is attributed to the class-scene prior loss can introduce specific scene information during the process of learning subject instance and visual-textual matching loss can enhance the fusion between visual instance image and textual scene context.

4.3 Ablation Studies

Effect of Class-Scene Prior Loss. Compared with prior preservation loss (Eq. (2)) proposed in DreamBooth [39], our class-scene prior loss $\mathcal{L}_C^S$ (Eq. (3)) utilizes detailed texts for prior images to incorporate multiple specific scenes. During training, this loss explicitly enforces the model to retain prior scene knowledge while incorporating new information from instance images within these scenes. From the visual comparison between ComFusion("w/o visual-textual loss $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$") and DreamBooth [39] in Fig. 4, and quantitative results in Tab. 3, we can see that class-scene prior loss $\mathcal{L}_C^S$ significantly improves the CLIP-T score while achieves comparable CLIP-I(*resp.,* DINO) score, which indicates that it can effectively improve scene fidelity without undermining the instance fidelity.

Effect of Visual-Textual Matching Loss. We further conduct ablation to evaluate the effect of the proposed visual-textual loss $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$ in Eq. (5). To evaluate the effect of each item in visual-textual loss, we alternatively remove $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$ (*resp.*,$\mathcal{L}_F^I$, $\mathcal{L}_F^S$) from total loss function in Eq. (6), and report visual results in Fig. 4 and quantitative results in Tab. 3. The comparison results indicate that $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$ are well-design to balance the instance fidelity and scene fidelity, removing them degrades the instance fidelity in generated images in 2nd row in Fig. 4. To further study effect of each item in visual-textual loss, removing

Table 3: Ablation analysis of individual loss components and alternative designs on seen/unseen scene.

Methods	DINO (↑)	CLIP-I (↑)	CLIP-T (↑)	Time(/s)	Memory(G)
Real Images	*0.795*	*0.859*	N/A	N/A	N/A
DreamBooth	0.619/0.607	0.752/0.735	0.229/0.214	491.8	20.2
DreamBooth(w/ L_F^I and L_F^S)	0.643/0.616	0.786/0.744	0.232/0.221	597.7	44.5
Ours (w/o $\{\mathcal{L}_F^I, \mathcal{L}_F^S\}$)	0.627/0.597	0.771/0.717	0.301/0.276	491.8	20.2
Ours (w/o $\mathcal{L}_F^I$)	0.586/0.561	0.697/0.629	0.342/0.321	597.3	44.5
Ours (w/o $\mathcal{L}_F^S$)	0.716/0.685	0.828/0.769	0.189/0.161	597.1	44.5
Ours ($\lambda_C^S = 10$)	0.619/0.584	0.768/0.707	0.334/0.311	597.7	44.5
Ours ($\lambda_C^S = 0.1$)	0.669/0.641	0.826/0.768	0.296/0.267	597.7	44.5
Ours ($\lambda_F^S = 0.1$)	0.529/0.495	0.638/0.578	0.351/0.328	597.7	44.5
Ours ($\lambda_F^S = 0.001$)	0.732/0.698	0.851/0.794	0.272/0.247	597.7	44.5
Ours ($\lambda_F^I = 0.1$)	0.715/0.684	0.842/0.789	0.302/0.252	597.7	44.5
Ours ($\lambda_F^I = 0.001$)	0.546/0.509	0.672/0.621	0.341/0.321	597.7	44.5
Ours ($\tau = 1$)	0.641/0.608	0.806/0.744	0.334/0.301	537.9	30.9
Ours ($\tau = 5$)	0.698/0.651	0.825/0.763	0.309/0.274	623.1	60.1
Ours($\tau = 3$)	0.658/0.621	0.814/0.752	0.321/0.297	597.7	44.5

$\mathcal{L}_F^S$ leads to degrading the score of DINO and CLIP-I reflecting by poor instance fidelity in 4th row in Fig. 4, while lower scene fidelity in 3rd row in Fig. 4 is caused by removing $\mathcal{L}_F^C$.

Effect of Hyperparameters. To assess the impact of coarse denoising timesteps τ in the fusion stream, we experimented with varying τ values from $\{1, 3, 5\}$ to train ComFusion. From Fig. 4 and Tab. 3, our observations indicate that a larger τ value tends to better preserve instance fidelity but at the expense of reduced scene fidelity. We selected $\tau = 3$ as the default setting for ComFusion, considering time cost and a balance between instance fidelity and scene fidelity Furthermore, to investigate the impact of λ_C^S, λ_F^I, and λ_F^S, we tune them to see the influence on performance, we adjusted these hyperparameters and analyzed the outcomes in-depth, as detailed in Tab. 3. This analysis justified our chosen settings of $\lambda_C^S = 1$, $\lambda_F^S = 0.01$, and $\lambda_F^I = 0.02$, ensuring an optimal balance for model performance.

5 Conclusions

We present ComFusion, a novel approach designed to facilitate personalized subject generation within multiple specific scenes from a single image. ComFusion introduces a class-scene prior loss to composite knowledge of subject class and specific scenes from pretrained models. Moreover, a visual-textual matching loss to further improve the fusion of visual object feature and textual scene feature. Extensive quantitative and qualitative experiments demonstrate the effectiveness of ComFusion.

References

1. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., Bermano, A.H.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06925 (2023) 2, 3, 4
2. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:2305.16311 (2023) 2, 3, 4, 11, 12, 13
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023) 3
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 10, 12
5. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23040–23050 (2023) 3
6. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023) 7
7. Chen, H., Zhang, Y., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374 (2023) 3, 4
8. Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. arXiv preprint arXiv:2304.00186 (2023) 2, 4
9. Clouâtre, L., Demers, M.: Figr: Few-shot image generation with reptile. arXiv preprint arXiv:1901.02199 (2019) 3
10. Ding, G., Han, X., Wang, S., Wu, S., Jin, X., Tu, D., Huang, Q.: Attribute group editing for reliable few-shot image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11194–11203 (2022) 3
11. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023) 3
12. 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: The Eleventh International Conference on Learning Representations (2022) 1, 2, 3, 4, 11, 12, 13
13. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023) 2, 3, 4
14. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023) 3

15. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022) 3
16. 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. Advances in Neural Information Processing Systems **36** (2024) 2
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3, 9
18. Hong, Y., Niu, L., Zhang, J., Zhang, L.: Matchinggan: Matching-based few-shot image generation. In: ICME (2020) 3
19. Hong, Y., Niu, L., Zhang, J., Zhang, L.: DeltaGAN: Towards diverse few-shot image generation with sample-specific delta. ECCV (2022) 3
20. Hong, Y., Niu, L., Zhang, J., Zhao, W., Fu, C., Zhang, L.: F2gan: Fusing-and-filling gan for few-shot image generation. In: ACM MM (2020) 3
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 4
22. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023) 2, 4
23. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023) 3
24. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023) 2
25. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023) 2, 3, 4, 11, 12, 13
26. Lee, J., Cho, K., Kiela, D.: Countering language drift via visual grounding. arXiv preprint arXiv:1909.04499 (2019) 6
27. Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv preprint arXiv:2305.14720 (2023) 1, 2, 4
28. Li, L., Zhang, Y., Wang, S.: The euclidean space is evil: Hyperbolic attribute editing for few-shot image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22714–22724 (2023) 3
29. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023) 3
30. Liu, K., Tang, W., Zhou, F., Qiu, G.: Spectral regularization for combating mode collapse in gans. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6382–6390 (2019) 3
31. 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) 2, 3

32. Lu, Y., Singhal, S., Strub, F., Courville, A., Pietquin, O.: Countering language drift with seeded iterated learning. In: International Conference on Machine Learning. pp. 6437–6447. PMLR (2020) 6
33. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023) 4
34. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36** (2024) 2
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 6, 10, 12
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 2, 3
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3, 5, 7, 10
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 6
39. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 1, 2, 3, 4, 6, 11, 12, 13
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 2, 3
41. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023) 2, 3, 4
42. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 3
43. Smith, J.S., Hsu, Y.C., Zhang, L., Hua, T., Kira, Z., Shen, Y., Jin, H.: Continual diffusion: Continual customization of text-to-image diffusion with c-lora. arXiv preprint arXiv:2304.06027 (2023) 2
44. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 3
45. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. Advances in neural information processing systems **30** (2017) 3
46. Tao, M., Bao, B.K., Tang, H., Xu, C.: Galip: Generative adversarial clips for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14214–14223 (2023) 3
47. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: Df-gan: A simple and effective baseline for text-to-image synthesis. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 16515–16525 (2022) 3
48. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 2, 3
49. Thanh-Tung, H., Tran, T.: Catastrophic forgetting and mode collapse in gans. In: 2020 international joint conference on neural networks (ijcnn). pp. 1–10. IEEE (2020) 3
50. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 2
51. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023) 2, 3, 4, 11, 12, 13
52. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023) 3
53. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023) 2, 3, 4, 11, 12, 13
54. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023) 3
55. Wu, Q., Liu, Y., Zhao, H., Bui, T., Lin, Z., Zhang, Y., Chang, S.: Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7766–7776 (2023) 2
56. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023) 2
57. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20908–20918 (2023) 3
58. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7754–7765 (2023) 2
59. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. Advances in Neural Information Processing Systems **36** (2024) 2
60. Yang, M., Wang, Z., Chi, Z., Feng, W.: Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In: European Conference on Computer Vision. pp. 1–17. Springer (2022) 3
61. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 3, 4
62. Zhuang, J., Wang, C., Lin, L., Liu, L., Li, G.: Dreameditor: Text-driven 3d scene editing with neural fields. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023) 3