Supplementary for 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

Appendix

In this appendix, we provide additional material to complement our main submission. Key sections are outlined as follows: In Appendix A, we introduce related works. In Appendix B, we introduce preliminaries to duffusion models. In Appendix C, we provide the implementation details of the baseline methods. Evaluation metrics of CLIP-I, CLIP-T, and DINO used in our study are represented in Appendix D. In E, we present the specific scenes and visualize the single instance image from the training dataset. In F, we evaluate the performance of the proposed ComFusion and baseline methods in scenarios involving unseen scenes, testing the generalizability of ComFusion. In Appendix G, the performance of ComFusion, when trained with multiple instance images, against the DreamBooth baseline method, demonstrates ComFusion's effectiveness in varied training contexts. In Appendix H, we visualize additional generated images by ComFusion, further demonstrating the model's capabilities. More ablative studies are conducted in Appendix I to delve deeper into ComFusion's insights. Finally, limitation in Appendix J, discuss some failure cases in complex scenes, highlighting the current limitations and potential areas for personalized subject generation.

A Related Works

A.1 Diffuion-Based Text-to-Image Generation.

The field of Text-to-Image (T2I) generation has recently witnessed remarkable advancements [15, 22, 36, 42, 43, 52], predominantly led by pre-trained diffusion models such as Stable Diffusion [33], DALLE [32], Imagen [36] and *etc*. These

⋆ Corresponding author.

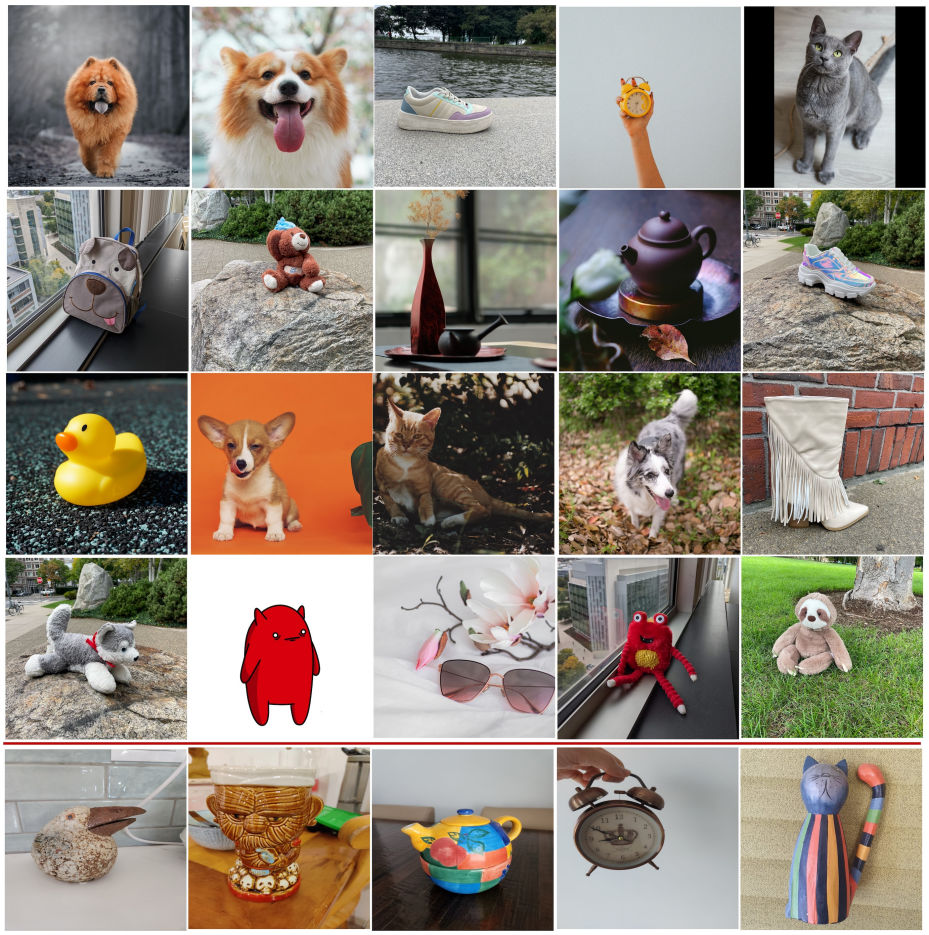

Fig. 1: A collection of 25 concept images from the DreamBooth [35] and TI [12] datasets. The images in the last row are from TI [12] dataset and others from DreamBooth [35] dataset.

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 [3,5,11,14,49] and 3D object creation [27,38,47,50,53]. 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, 24, 29, 35, 44].

A.2 Personalized Text-to-Image Generation

Given a small set of images of the subject concept, personalized T2I generation [1, 6, 12, 13, 16, 24, 29, 32, 35–37, 40, 44, 46, 48] 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 [28, 41, 45] for generative adversarial networks [9, 10, 17–19, 26, 51]. Recently, diffusion-based text-to-image models with a few images have also been explored in [2,35]. 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.*, [20, 30, 52]) with a pretrained base model. The second stream adopts a strategy of finetuning the pretrained model using a few selected images.

A.3 T2I personalization Without Finetuning.

These methods without finetuning [8, 13, 21, 25, 37] generally rely on additional modules trained on additional new datasets, such as the visual encoder in [37,48] and the experts in [8, 25] 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 [37], 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 [6] uses weak denoising and contrastive embedding auxiliary tuning objectives for personalization. ELITE [48] 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.

A.4 T2I personalization with Finetuning

Various methods employ diverse finetuning strategies to optimize different modules within pretrained models [12, 24, 35]. DreamBooth [35] 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 [24, 46] 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.

B Preliminaries

Diffusion Models. Diffusion models [16, 39] can generate realistic images from a normal distribution by reversing a gradual noising process. The forward process, denoted as $q(\cdot)$, constitutes a Markov chain that incrementally transforms data from $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$ to a Gaussian distribution. A single step in the forward process is defined as $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I)$, where β_t represents a predefined variance schedule over T steps. The forward process allows for the sampling of $\boldsymbol{x}_t$ at any given timestamp t in a closed form:

$$\begin{aligned} \boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon, \\ \text{s.t.,} \quad \alpha_t &= \prod_{s=1}^{t}(1-\beta_s), \quad \epsilon \sim \mathcal{N}(0, I). \end{aligned} \tag{1}$$

The reverse process in diffusion models, which can be parameterized using deep neural networks, is defined as $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \sigma_\theta(\boldsymbol{x}_t, t)I)$. Denoising Diffusion Probabilistic Models (DDPMs) [16] have demonstrated that utilizing a noise approximation model $\epsilon_\theta(\boldsymbol{x}_t, t)$ is more effective than using $\mu_\theta(\boldsymbol{x}_t, t)$ for procedurally transforming the prior noise into data. As a result, the sampling in diffusion models is $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t\boldsymbol{\epsilon}$. Latent Diffusion Models (LDM) [33] innovate by minimizing computational demands through latent space operations, using a pretrained encoder E to embed an image into latent space, and a pretrained decoder D for image reconstruction. In LDM, the diffusion process is defined using $\boldsymbol{z} = E(\boldsymbol{x})$ instead of $\boldsymbol{x}$ itself. LDM adopts the Denoising Diffusion Implicit Models (DDIM) [40] sampling process, a neural ODE-based technique [7] enabling fast and deterministic image generation. The DDIM sampler predicts $\boldsymbol{z}_0$ directly from $\boldsymbol{z}_t$, then generates $\boldsymbol{z}_{t-1}$ through a reverse conditional distribution. By integrating the textual condition $\boldsymbol{T}$ and the text encoder $\Gamma(\cdot)$, the predicted $\boldsymbol{z}_0$ given $\boldsymbol{z}_t$ and t is:

$$h_\theta(\boldsymbol{z}_t, t, \boldsymbol{T}) = \frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t, \Gamma(\boldsymbol{T}))}{\sqrt{\alpha_t}}. \tag{2}$$

The deterministic sampling process in LDM using DDIM can be outlined as: $\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}} h_\theta(\boldsymbol{z}_t, t, \boldsymbol{T}) + \sqrt{1-\alpha_{t-1}} \epsilon_\theta(\boldsymbol{z}_t, t, \Gamma(\boldsymbol{T}))$. Once the diffusion process is complete, the image is reconstructed by the decoder D, such that $\tilde{\boldsymbol{x}} = D(\boldsymbol{z})$.

C Implementation Details

C.1 Baselines

In our study, we compare ComFusion with several state-of-the-art (SOTA) methods:

- **DreamBooth** [35]: A SOTA approach that fully finetunes all layers of U-Net and text-encoder.
- **Textual-Inversion (TI)** [12]: A SOTA approach that only focuses solely on training word embeddings.
- **Custom-Diffusion (CD)** [24]: A concurrent work optimizing the cross-attention weights of the denoising model, along with a newly-added text token. Official hyperparameters are utilized.
- **Extended Textual-Inversion (XTI)** [46]: Building on TI [12], XTI inverts input images into a set of token embeddings, one per layer, demonstrating faster, more expressive, and precise results than TI.
- **ELITE** [48]: This method trains a local and global map on large-scale datasets, enabling the instant generation of new images from a single user-provided image and corresponding mask.
- **Break-A-Scene** [2]: This method employs a two-stage training strategy, optimizing token embedding, text-encoder, and U-Net under the supervision of an object mask.

It's noteworthy that existing personalization methods, such as DreamBooth [35], TI [12], CD [24], and XTI [46] typically require multiple images as input, in contrast to Break-A-Scene [2] and ELITE [48], which leverage a single image with a mask indicating the target concept. In our setting, we use a single instance image without a mask to generate new images featuring the target concept in multiple specific scenes. For our experiments, unless stated otherwise, we employ the 30-step DDIM [40] sampler with a scale of 7.5

C.2 Experimental Settings

For the methods mentioned, including DreamBooth [35], CD [24], and our proposed ComFusion, all of which utilize class-aware prior images, we generate 200 prior images to ensure a fair comparison. Besides, Break-A-Scene [2] relies on instance masks for its training process, while ELITE [48] depends on instance masks during inference. Therefore, for these methods, we obtain the concept mask of the instance image using SAM [23]. All experiments are conducted on a single A100 GPU. For all pre-trained Stable Diffusion (SD) models, we use the 1.5 checkpoint [33] for those baseline methods except for ELITE [48] without training. Here are the detailed settings for each method:

- **ComFusion**: ComFusion uses a pre-trained Stable Diffusion (SD) checkpoint 1.5 [33] and produce 200 prior images, and finetunes text-encoder Γ and denoising model ϵ_θ architectured with $U-Net$ [34] for 1200 steps, using a batch size of 1 and learning rate 1×10^{-5}. 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).
- **DreamBooth [35]**: Similar to ComFusion, DreamBooth uses instance images and 200 prior images to finetune the text-encoder and denoising model $U-Net$ based on Stable Diffusion (SD) checkpoint 1.5 [33]. The total training steps are 1200, set the batch size (*resp.,* learning rate) as 1 (*resp.,* 1×10^{-5}) .
- **TI [12]**: Based on Stable Diffusion (SD) checkpoint 1.5 [33], TI leverages instance images to learn a token embedding with a batch size of 4. The base learning rate was set to 0.005 and the model is trained with 5, 000 optimization steps.
- **CD [24]**: Following original setting in [24], CD loads a pretrained Stable Diffusion (SD) checkpoint 1.5 [33]. CD [24] learns a new token embedding and finetunes the $U-Net$ parameters with 250 steps on the combination of instance image and prior images. The batch size is set as 8 and the learning rate is 8×10^{-5}. During training, training images are randomly resized for augmentation.
- **XTI [46]**: Following original setting in [46], XTI adopts a reduced learning rate of 0.005 without scaling for optimization with batch size of 8, the model is trained for 500 steps to learn new token embeddings.
- **ELITE [48]**: ELITE is a pretrained model and can be instantly applied for generating new images with input of instance image and its mask.
- **Break-A-Scene [2]**: Following original setting in [2], we load a pretrained Stable Diffusion (SD) checkpoint 1.5 [33], the Break-A-Scene [2] adopt two-stage training strategy: in the first stage only the text embeddings is optimized with a high learning rate of 5×10^{-4}, while in the second stage, both the $U-Net$ weights and the text encoder weights are optimized with a small learning rate of 2×10^{-6}. Both stages use Adam optimizer. Each stage is trained for 400 steps.

D Evaluation Metrics

To assess the fidelity of both instances and scenes in the generated images, we conduct both quantitative and qualitative evaluations. Following DreamBooth [35], we use DINO score [4], and CLIP-I [31] to evaluate instance fidelity, and use CLIP-T [31] to evaluate the scene fidelity.

- **CLIP-I** [35]: Measures the average pairwise cosine similarity between CLIP [31] embeddings of generated and real images.
- **DINO** [4]: Calculates the average pairwise cosine similarity using ViT-S/16 DINO [4] embeddings of generated and real images. Unlike supervised networks, DINO does not ignore differences within the same class but rather focuses on distinct features of a subject or image, thanks to its self-supervised training objective.

– **CLIP-T** [35]: This metric evaluates the alignment between the textual prompts and the image [31] embeddings, thereby assessing the fidelity of the input scene as represented in the generated images.

E Datasets

The dataset used in this paper contains 25 concepts from DreamBooth Dataset and TI [12] dataset. The concept images from two datasets are visualized in Fig. 1. The 15 specific instance-scenes are: "`[identifier] [class noun] Scene`", the specific scenes including: "`in the rain`", "`in the river`", "`in the sky`", "`in the room`", "`in the basket`", "`in the TV`", "`in the snow`", "`on the sofa`", "`on the bed`", "`on the table`", "`on the stage`", "`on the top of mountain`", "`on the playground`", "`on the floor`", "`on the grass`".

Table 1: Adaptation to novel scenes: A quantitative comparison of instance fidelity (DINO, CLIP-I) and scene fidelity (CLIP-T) metrics.

Methods	DINO (↑)	CLIP-I (↑)	CLIP-T (↑)
Real Images	*0.795*	*0.859*	N/A
DreamBooth [35]	0.607	0.735	0.214
TI [12]	0.459	0.632	0.188
CD [24]	0.611	0.725	0.202
XTI [46]	0.431	0.602	0.185
ELITE [48]	0.415	0.607	0.241
Break-A-Scene [2]	0.618	0.749	0.261
Ours	**0.621**	**0.752**	**0.297**

Table 2: ComFusion, trained using multiple instance images, tested across specific scenes. Quantitative metrics comparison focusing on 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 (N^I=1)	0.619	0.752	0.229
Ours (N^I=1)	**0.658**	**0.814**	**0.321**
DreamBooth (N^I=3)	0.639	0.791	0.246
Ours (N^I=3)	**0.669**	**0.834**	**0.332**
DreamBooth (N^I=5)	0.629	0.761	0.261
Ours (N^I=5)	**0.661**	**0.825**	**0.348**

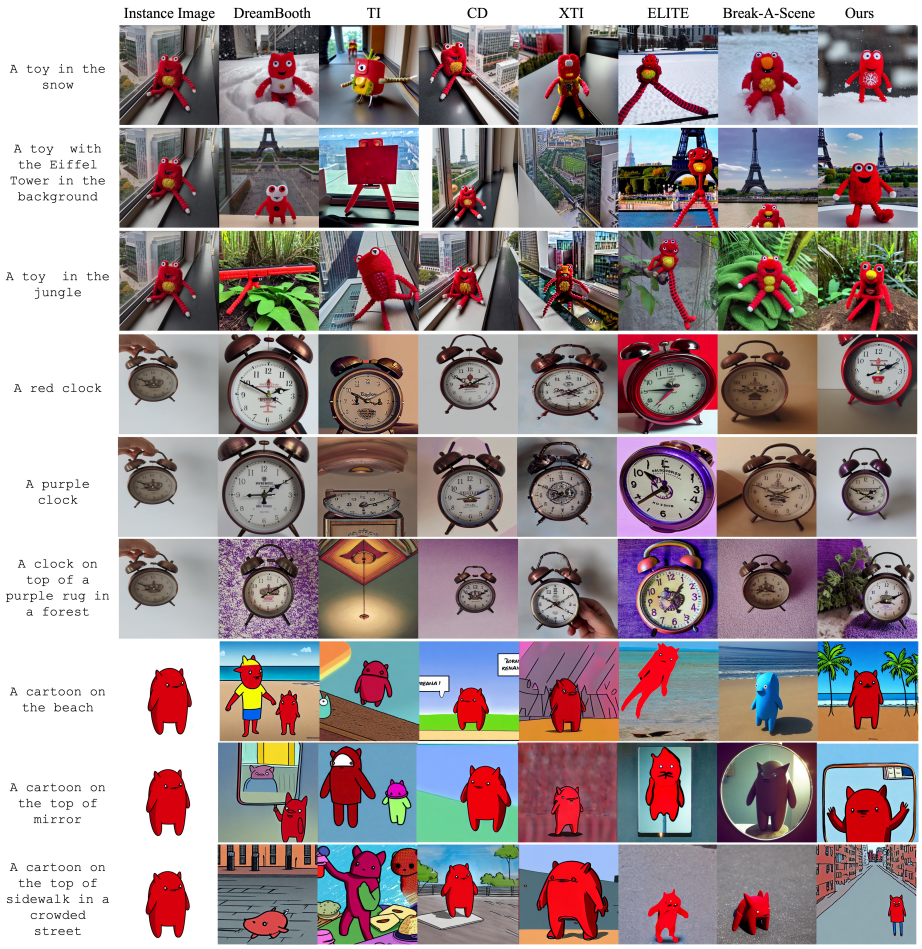

Fig. 2: Images in unseen scenes generated by baseline methods and our proposed ComFusion trained from a single instance image.

Table 3: ComFusion trained on multiple instance images, and testing in unseen scenes. Quantitative metrics comparison of 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 (N^I=1)	0.607	0.735	0.214
Ours (N^I=1)	**0.618**	**0.749**	**0.297**
DreamBooth (N^I=3)	0.613	0.752	0.219
Ours (N^I=3)	**0.640**	**0.767**	**0.301**
DreamBooth (N^I=5)	0.611	0.748	0.231
Ours (N^I=5)	**0.622**	**0.753**	**0.306**

Fig. 3: Images generated by DreamBooth [35] and our proposed ComFusion in specific scenes (the left four columns) and unseen scenes (the right four columns).

F Generalization to Unseen Scenes

To assess ComFusion's capability in generating images for unseen scenes, we follow DreamBooth [35] by using 25 diverse prompts, which include 20 recontextualization prompts and 5 property modification prompts. For each prompt, ComFusion and the baseline methods are employed to sample 10 images. The instance fidelity and scene fidelity of these images are then evaluated using CLIP-I, CLIP-T, and DINO metrics. Quantitative comparison results are reported in Tab. 1, while qualitative outcomes are illustrated in Fig. 2. The results from Tab. 1 indicate that ComFusion achieves the highest scene fidelity scores, maintaining instance fidelity comparable to Break-A-Scene [2]. This performance can be attributed to the integration of class-scene prior images in the training process, which supplements the model with additional textual information. This enhancement aids the model in generalizing to unseen scenes and mitigates the risk of overfitting to the specific prompt structure "`a [identifier] [class noun]`". However, the combination of instances with unseen scenes, which is not encountered during training, may result in a slightly lower instance fidelity score.

G Multiple Instance Images

In the main paper, we utilize a single instance image to train ComFusion, aiming to assess its few-shot learning ability. To further evaluate the impact of the number of instance images $\boldsymbol{x}^I$ on ComFusion's performance, we conduct additional experiments. In these tests, we maintain a constant number of class-scene images at $N = 200$ while varying the number of instance images N^I. Specifically, we explore scenarios where N^I is set to either 3 or 5, allowing us to observe how changes in the number of instance images influence the effectiveness of our model in few-shot learning contexts.

G.1 Generalization on Specific Scenes

In accordance with the experimental setting described in Sec. 4.1 in main paper, we generated 10 images for each of the 25 subjects across each of the 15 scenes, resulting in a total of 3750 images for evaluation. Additionally, we calculated the CLIP-I, CLIP-T, and DINO metrics to assess both instance fidelity and scene fidelity, as detailed in Tab. 2. From the table, it is evident that the proposed ComFusion model surpasses DreamBooth in performance when the number of instance images increases. A notable trend observed is the enhancement in scene fidelity, as indicated by the CLIP-T score, with the increase in the number of instance images. However, this trend is not mirrored in the instance fidelity metrics (CLIP-I and DINO), where the scores for "DreamBooth ($N^I = 3$)" (*resp.*, "ComFusion ($N^I = 3$)") are higher than for "DreamBooth ($N^I = 5$)" (*resp.*, "ComFusion ($N^I = 5$)"). We hypothesize that the use of multiple instance images can reduce overfitting to a specific instance image and introduce greater diversity to the target concept. This hypothesis is supported by the visual evidence in

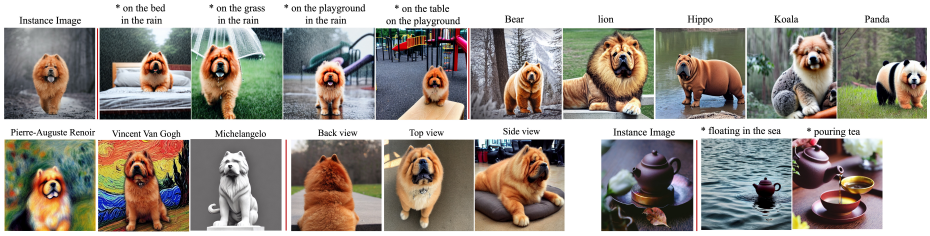

Fig. 4: Results of more challenging examples generated from ComFusion.

Fig. 3, which shows a rich variety of bird poses and shapes when models are trained on either 3 or 5 instance images. This alleviation of overfitting, thanks to multiple instance images, also helps the pretrained model retain prior knowledge, thus achieving higher scene fidelity.

G.2 Generalization on Unseen Scenes

Expanding on the 25 unseen scenes described in F, we generated 10 images for each of these scenes and assessed instance fidelity and scene fidelity using the CLIP-I, DINO, and CLIP-T metrics. A comparison between Tab. 2 and Tab. 3 reveals that the overall performance in unseen scenes is not as promising as in specific scenes. Analyzing Fig. 3 and Tab. 3, we observe a trend consistent with the findings in specific scenes. Specifically, ComFusion surpasses DreamBooth in performance when an equal number of instance images are used. When we increase the number of instance images from 1 to 5, there is a noticeable improvement in scene fidelity as evaluated by the CLIP-T metric. The best results for instance fidelity are achieved when the model is trained on 3 instance images.

H More Visualization Results

H.1 More Challenging Examples

To showcase the adaptability of ComFusion, we present a range of challenging examples in Fig. 4, including alterations in viewpoints, a variety of artistic styles (*e.g.*, Van Gogh, Michelangelo), dynamic scenes (*e.g.*, a teapot pouring tea), and property transformations (*e.g.*, transforming a dog into a panda). These examples underscore ComFusion's proficiency in navigating complex and diverse contexts, all derived from a singular instance image.

H.2 More Visualization Comparison to SOTA Methods

In this section, we present more visualization comparisons as shown in Fig. 5 and Fig. 6. Observing Fig. 5, it's evident that images generated by ComFusion

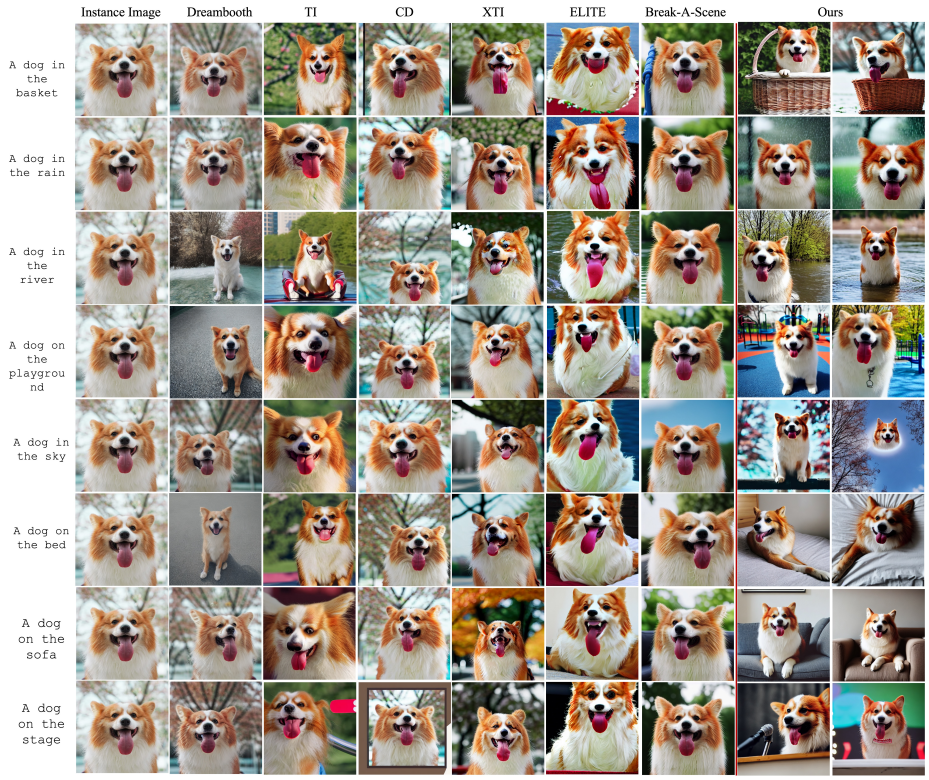

Fig. 5: Images generated by DreamBooth [35], TI [12],CD [24],XTI [46],ELITE [48], Break-A-Scene [2], and our proposed ComFusion in multiple specific scenes from a single instance image.

Fig. 6: Images generated by DreamBooth [35], TI [12],CD [24],XTI [46],ELITE [48], Break-A-Scene [2], and our proposed ComFusion in multiple specific scenes from a single instance image.

not only exhibit high instance accuracy but also align well with the input prompt in terms of the background scene. Break-A-Scene [2] and CD [24] demonstrate strong instance fidelity, yet they lack diversity in the background and do not adequately respond to the input prompts. DreamBooth [35] tends to either replicate the instance image closely or generate scene-specific images with compromised instance fidelity. Both TI [12] and XTI [46] consistently struggle to accurately depict specific scenes described in the input prompts. ELITE [48], not being trained with instance images, falls short in instance fidelity compared to the other baseline methods.

I More Ablative Studies

Table 4: Ablation analysis of the different number of scenes.

Methods	DINO (↑)	CLIP-I (↑)	CLIP-T (↑)
Real Images	*0.795*	*0.859*	N/A
DreamBooth (Unseen)	0.607	0.735	0.214
Ours ($L = 5$, Unseen)	0.613	0.737	0.258
Ours ($L = 10$, Unseen)	0.617	0.741	0.274
Ours ($L = 15$, Unseen)	0.618	0.749	0.297
Ours (Full-Fledged)	0.658	0.814	0.321

Table 5: Quantitative results of only instance/scene generation.

Methods	DINO (↑)	CLIP-I (↑)	CLIP-T (↑)
Real Images	*0.795*	*0.859*	N/A
Ours (Only Instance)	0.724	0.872	0.369
Ours (Only Scene)	N/A	N/A	0.378
DreamBooth (Only Instance)	0.638	0.841	0.359
DreamBooth (Only Scene)	N/A	N/A	0.287
Ours (Full-Fledged)	0.658	0.814	0.321

I.1 Impact of Scene Counts

In our ablation study, we assessed the influence of varying scene counts in our regularization set (*i.e.*, 5, 10 scenes) on performance. As demonstrated in Tab. 4, results for unseen scenes reveal that performance marginally declines with a reduced number of scenes. Notably, ComFusion maintains superior performance over DreamBooth, irrespective of the scene count variations.

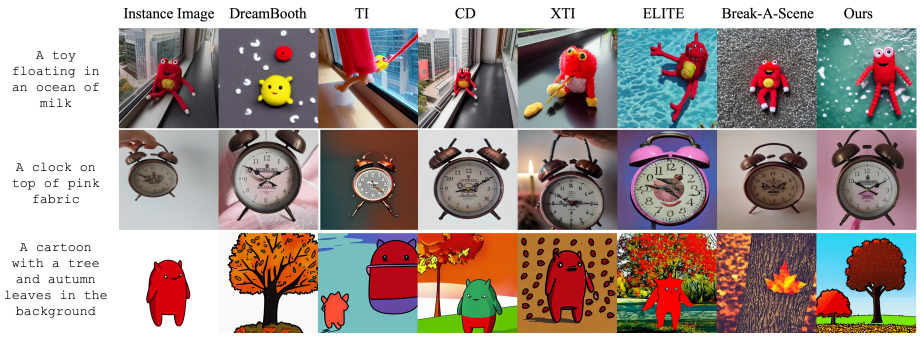

Fig. 7: Failure cases in unseen scenes.

Table 6: 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

I.2 Trade-offs between training time, computational resources, and the quality of the generated images.

We have compared the training resource consumption , training duration in Tab. 6, and the final image generation quality across different settings in Tab. 6 and in visual comparison results in main paper. τ does increase memory usage during finetuning. As shown in the memory usage and training time in Tab. 6, our method requires more memory and time than DreamBooth due to the forward pass of τ times. However, the quality of images generated by ComFusion surpasses those of DreamBooth. Importantly, the finetuning time remains reasonable (597.7s vs. 491.8s). Furthermore, *ComFusion does not introduce additional time during inference* (10.0G memory usage). Also, *there is no extra memory consumption for more scene prompts*, because these prompts are encoded by the text encoder into embeddings of the same size, and the finetuning steps remain unchanged. These comparisons provide a detailed analysis of the training efficiency and the rationale behind our model settings.

I.3 Special Cases with Only Instance or Scene

ComFusion is adaptable to special cases involving solely instances or scenes. Specific evaluations were carried out for sets comprising exclusively scenes or instances. The results of these assessments, as detailed in Tab. 5, demonstrate ComFusion's superior performance over DreamBooth in these distinct scenarios.

J Limitations

We visualize some failure cases in Fig. 7, highlighting areas where both the baseline methods and ComFusion encounter challenges. The first row of Fig. 7 demonstrates that both baseline methods and ComFusion struggle with understanding and rendering creative scenes, such as "`in an ocean of milk`". The second row shows that when it comes to descriptions of material properties (*e.g.,* fabric), the methods exhibit limited capability in integrating instance concepts with such specific prompts. This suggests a gap in accurately representing detailed material textures and properties. The third row highlights the challenge with long-term prompts that describe composite semantics, like a scene with a tree and autumn leaves in the background. Both baseline and proposed methods find it difficult to coherently integrate the target concept from the instance image with the background scene, often neglecting the target concept. These limitations point to areas where further research and development could enhance the model's understanding and rendering capabilities, particularly in contexts involving creative, material, or composite semantic descriptions.

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) 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) 3, 4, 5, 6, 7, 10, 12, 13, 14
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) 6
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. 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
7. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018) 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) 3
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) 2, 3, 5, 6, 7, 12, 13, 14
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) 3
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) 1
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3, 4

17. Hong, Y., Niu, L., Zhang, J., Zhang, L.: Matchinggan: Matching-based few-shot image generation. In: ICME (2020) 3
18. Hong, Y., Niu, L., Zhang, J., Zhang, L.: DeltaGAN: Towards diverse few-shot image generation with sample-specific delta. ECCV (2022) 3
19. 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
20. 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) 3
21. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023) 3
22. 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) 1
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 5
24. 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) 3, 4, 5, 6, 7, 12, 13, 14
25. 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) 3
26. 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
27. 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
28. 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
29. 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) 3
30. 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) 3
31. 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, 7
32. 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) 1, 3
33. 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) 1, 4, 5, 6
34. 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
35. 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) 2, 3, 5, 6, 7, 9, 10, 12, 13, 14
36. 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) 1, 3
37. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023) 3
38. 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
39. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 4
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 3, 4, 5
41. 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
42. 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) 1
43. 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) 1
44. 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) 3
45. 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
46. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023) 3, 4, 5, 6, 7, 12, 13, 14
47. 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
48. 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) 3, 5, 6, 7, 12, 13, 14

49. 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
50. 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
51. 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
52. 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) 1, 3
53. 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