DisenDreamer: Subject-Driven Text-to-Image Generation with Sample-aware Disentangled Tuning

Hong Chen, Yipeng Zhang, Xin Wang, *Member, IEEE,* Xuguang Duan, Yuwei Zhou, and Wenwu Zhu, *Fellow, IEEE*

***Abstract*—Subject-driven text-to-image generation aims to generate customized images of the given subject based on the text descriptions, which has drawn increasing attention recently. Existing methods mainly resort to finetuning a pretrained generative model, where the identity-relevant information (e.g., the boy) and the identity-irrelevant sample-specific information (e.g., the background or the pose of the boy) are entangled in the latent embedding space. However, the highly entangled latent embedding may lead to low subject identity fidelity and text prompt fidelity. To tackle the problems, we propose DisenDreamer, a sample-aware disentangled tuning framework for subject-driven text-to-image generation in this paper. Specifically, DisenDreamer finetunes the pretrained diffusion model in the denoising process. Different from previous works that utilize an entangled embedding to denoise, DisenDreamer instead utilizes a common text embedding to capture the identity-relevant information and a sample-specific visual embedding to capture the identity-irrelevant information. To disentangle the two embeddings, we further design the novel weak common denoising, weak sample-aware denoising, and the contrastive embedding auxiliary tuning objectives. Extensive experiments show that our proposed DisenDreamer framework outperforms baseline models for subject-driven text-to-image generation. Additionally, by combining the identity-relevant and the identity-irrelevant embedding, DisenDreamer demonstrates more generation flexibility and controllability.**

***Index Terms*—diffusion model, subject-driven text-to-image generation, disentangled finetuning**

I. Introduction

Training on billions of text-image pairs, large-scale text-to-image models [34], [36], [39] have recently achieved unprecedented success in generating photo-realistic images that conform to the given text descriptions. Thanks to their remarkable generation ability, a more customized generation topic, subject-driven text-to-image generation, has attracted increasing attention in the community [8], [12], [21], [24], [38], [40], [45]. Given a small set of images of a subject, e.g., *3 to 5 images of your favorite toy*, subject-driven text-to-image generation aims to generate new images of the same subject according to the text prompts, e.g., *an image of your favorite toy with green hair on the moon*. The challenge of subject-driven text-to-image generation lies in the requirement that in addition to aligning well with the text prompts, the generated images are expected to preserve the subject identity [40] as well.

Existing subject-driven text-to-image generation methods [10], [12], [24], [38] mainly rely on finetuning strategy. These methods finetune the pretrained text-to-image diffusion models [36], [39] on the small set of images, and map these images containing the subject to a special text embedding. Then, the special text can be used together with other text descriptions to generate new images of the subject. However, since the text embedding is designed to align with the small set of images, information regarding the subject will be inevitably entangled with information irrelevant to the subject identity, such as the background or the pose of the subject. This entanglement tends to impair the image generation in two ways: (i) the identity-irrelevant information hidden in the entangled embedding may dominate the generation process, resulting in the generated images being heavily dependent on the irrelevant information while ignoring the given text descriptions, e.g., DreamBooth [38] ignores the "*in the snow*" text prompts, and overfits to the input image background as shown in row 2 column 3 of Figure 1; (ii) the identity-relevant information carried in the entangled embedding can not be appropriately preserved, resulting in identity change of the subject in the generated images, e.g., DreamBooth generates a backpack with a different color from the input image in row 3 column 3 of Figure 1. Other works [8], [13], [30], [40], [45] focus on reducing the computational burden, and investigate the problem of subject-driven text-to-image generation without finetuning. They rely on additional datasets that contain many subjects to train additional modules to customize the new subject. Once the additional modules are trained, they can be used for subject-driven text-to-image generation without further finetuning. However, these methods still suffer from poor generation ability without considering the disentanglement, e.g., ELITE [45] fails to maintain the subject identity in row 2 column 1 and ignores the action "*running*" from the text prompt in row 1 column 1 of Figure 1.

The main problem of existing methods is that the learned embedding of the subject is highly entangled with sample-specific identity-irrelevant information. To tackle the problem, in this paper, we propose DisenDreamer, a sample-aware disentangled tuning framework based on pretrained diffusion models. Specifically, DisenDreamer conducts the disentangled

This work was supported by the National Key Research and Development Program of China No. 2023YFF1205001, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia. This article was recommended by Guest Editor Dr. Zhengyuan Yang. *(Corresponding authors: Xin Wang and Wenwu Zhu.)*

Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail:{h-chen20, zhang-yp22, dxg18, zhou-yw21}@mails.tsinghua.edu.cn, {xin_wang, wwzhu}@tsinghua.edu.cn). Xin Wang and Wenwu Zhu are also with Beijing National Research Center for Information Science and Technology.

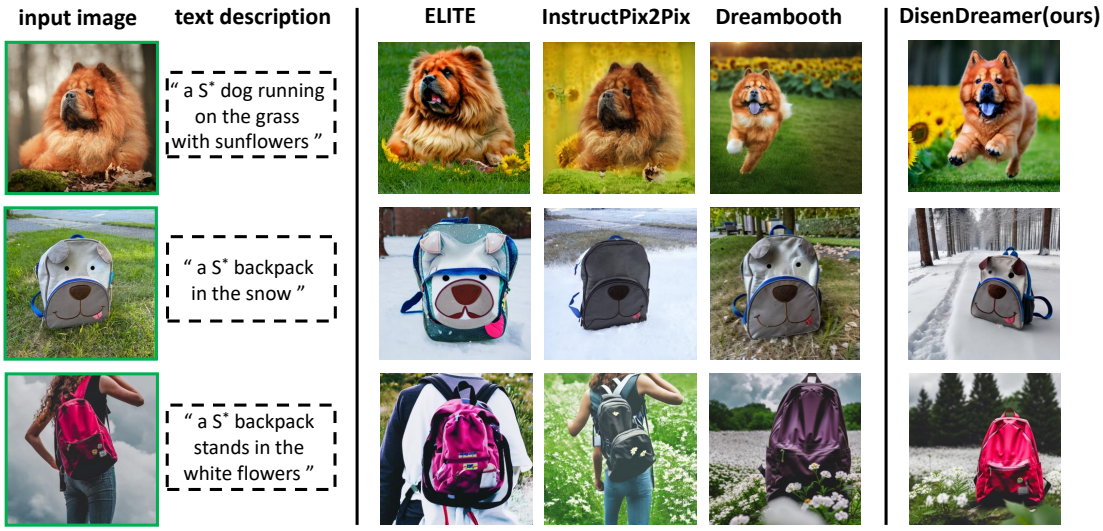


Fig. 1. Comparisons between different existing methods and our proposed DisenDreamer for subject-driven text-to-image generation. *"a S^* dog/backpack"* is a special token that represents the subject identity. The non-finetuning method ELITE and the image editing method InstructPix2Pix struggle to preserve the subject identity. The existing finetuning method DreamBooth suffers from overfitting on the input image background (row 2 column 3) and subject identity changes (row 3 column 3). Our proposed DisenDreamer can faithfully preserve the subject identity and conform to the text prompts.

tuning during the diffusion denoising process. As shown in Figure 2, different from previous works that only rely on an entangled text embedding as the condition to denoise, DisenDreamer simultaneously utilizes a textual identity-relevant embedding and a visual identity-irrelevant embedding as the condition to denoise for each image containing the subject. The identity-relevant embedding is extracted with a clip text encoder and the identity-irrelevant embedding is extracted with a clip visual encoder followed by a sample-aware adapter, where the adapter is designed to be a function of the input image, so that it can capture the image-specific information. To guarantee that the textual embedding and the visual embedding can respectively capture the identity-relevant and identity-irrelevant information, we propose three auxiliary disentangled objectives, i.e., the weak common denoising objective, the weak sample-aware denoising objective, and the contrastive embedding objective. To further enhance the tuning efficiency, parameter-efficient tuning strategies are adopted. The common identity-relevant branch and the sample-aware identity-irrelevant branch are both used during finetuning to denoise each image, but in the inference stage, only the identity-relevant embedding is utilized for subject-driven text-to-image generation, so that the generated content will be not influenced by the image-specific information such as the background. Additionally, through combining the two disentangled embeddings together, we are able to achieve more flexible and controllable image generation. Extensive experiments show that DisenDreamer can faithfully capture the identity-relevant and the identity-irrelevant information, and demonstrates superior generation ability over state-of-the-art methods in subject-driven text-to-image generation.

To summarize, our contributions are listed as follows,

- To the best of our knowledge, this is the first work on the investigation of disentangled finetuning for subject-driven text-to-image generation.
- We propose DisenDreamer, a sample-aware disentangled tuning framework for subject-driven text-to-image generation, which is able to learn an identity-relevant embedding through the clip text encoder and an identity-irrelevant embedding through the sample-aware clip visual encoder.
- We propose the weak common denoising, weak sample-aware denoising and the constrastive embedding, three auxiliary disentangled objectives to tune the DisenDreamer.
- Extensive experiments show that DisenDreamer has superior generation ability in subject-driven text-to-image generation over existing baseline models, and brings more generation flexibility and controllability.

II. Related Work

In this section, we discuss existing works on *Text-to-Image Generation, Text-Guided Image Editing, Subject-Driven Text-to-Image Generation* and *Disetangled Representation Learning*, which are most relevant to our work.

Text-to-Image Generation Training on large-scale datasets, the text-to-image generation models [6], [11], [22], [23], [32], [34], [35], [39], [46]–[49], [51], [52] have achieved great success recently. Among these models, diffusion-based models like Stable Diffusion [36], DALLE-2 [34], and Imagen [39] have attracted a lot of attention due to their outstanding controllability in generating photo-realistic images according to the text descriptions. Despite their superior ability, they still struggle with the more personalized generation, where we want to generate images about some specific or user-defined concepts, whose identities with detailed visual concepts are hard to be precisely described with text descriptions. This leads to the emergence of the recently popular topic, subject-driven text-to-image generation [12], [38].

Text-Guided Image Editing Text-guided image editing [2], [4], [17], [22], [25], [31] aims to edit an input image

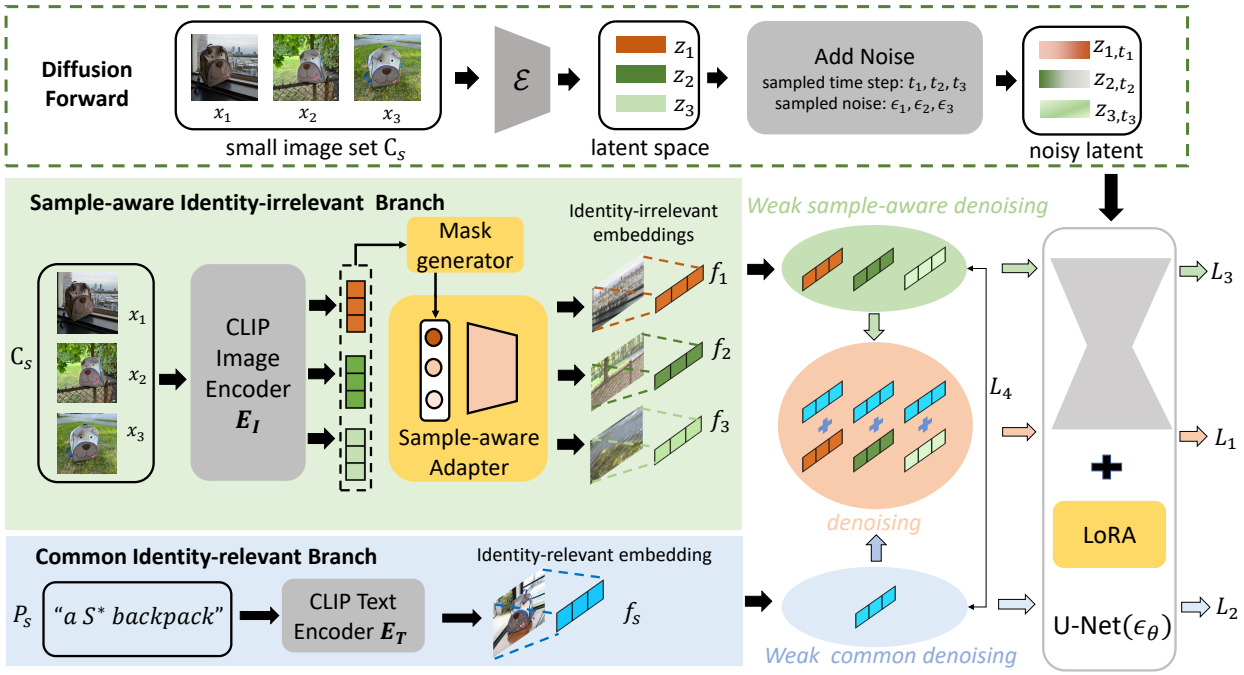


Fig. 2. DisenDreamer conducts finetuning in the denoising process, where each input image is denoised with the common textual embedding f_s shared by all the images to preserve the subject identity (the backpack), and visual embedding f_i to capture the identity-irrelevant information (e.g., the background). To make the two embeddings disentangled, the weak common denoising objective L_2, the weak sample-aware denoising objective L_3, and the contrastive embedding objective L_4 are proposed. Fine-tuned parameters include the adapter and the LoRA parameters.

according to the given textual descriptions. SDEdit [31] and Blended-Diffusion [1] blend the noisy input to the generated image in the diffusion denoising process. Prompt2Prompt [17] combines the attention map of the input image and that of the text prompt to generate the edited image. Imagic [21] utilizes a 3-step optimization-based strategy to achieve more detailed visual edits. A more recent SOTA method InstructPix2Pix [4] utilized GPT-3 [5], Stable Diffusion [36] and Prompt2Prompt [17] to generate a dataset with *(original image, text prompt, edited image)* triplets, and then train a new diffusion model which can directly conduct text-guided image editing. Despite their effectiveness, they are generally not suitable for subject-driven text-to-image generation [8], which needs to perform more complex transformations to the images, e.g., rotating the view, changing the pose, etc. Also, since these methods are not customized for the subject, their ability to preserve the subject identity is not guaranteed. Some examples of InstructPix2Pix for subject-driven text-to-image generation are presented in Figure 1, e.g., the pose of the dog is not changed to "*running*" in row 1 column 2, and the subject identity is changed in row 2 and row 3 of column 2.

Subject-Driven Text-to-Image Generation Given a small set of images of the subject, subject-driven text-to-image generation [12], [15], [24], [38] aims to generate new images according to the text descriptions while keeping the subject identity unchanged. DreamBooth [38] and TI [12] are two popular subject-driven text-to-image generation methods based on finetuning. They will both map the images of the subject into a special prompt token S^* during the finetuning process. The difference between them is that TI finetunes the prompt embedding and DreamBooth finetunes the U-Net model. Several concurrent works [8], [40], [45] propose to conduct subject-driven text-to-image generation without finetuning, which largely reduces the computational cost. They generally rely on additional modules trained on additional new datasets, like the visual encoder in [40], [45] and the experts in [8] to directly map the image of the new subject to the textual space. However, all the existing methods learn the subject embedding in an entangled manner, which will easily cause the generated image to have a changed subject identity or to be inconsistent with the text prompt.

Disentangled Representation Learning Disentangled representation learning aims to obtain representations capable of identifying and disentangling the underlying factors hidden in the observed data [3], [42], which has drawn a lot of attention in different fields, such as computer vision [9], [14], [29], recommendation [7], [28], [43], [44] and natural language processing [16], [20]. Disentangling the input data into different semantic factors can help to improve explainability and controllability [42], but [26] indicates that without supervsion and priors, unsupervised disentanglement cannot be achieved. In this paper, we utilize the priors of the pretrained Stable Diffusion and the CLIP encoders together with our proposed objectives to achieve identity-relevant and identity-irrelevant information disentanglement, which leads to more controllable subject-driven text-to-image generation.

III. Preliminaries

In this section, we will introduce the preliminaries about Stable Diffusion and subject-driven text-to-image generation, and also some notations we will use in this paper.

Stable Diffusion Models. The Stable Diffusion model is a large text-to-image model pretrained on large-scale text-image pairs $\{(P, x)\}$, where P is the text prompt of the image x. Stable Diffusion contains an autoencoder ($\mathcal{E}(\cdot)$, $\mathcal{D}(\cdot)$), a CLIP [33] text encoder $E_T(\cdot)$, and a U-Net [37] based conditional diffusion model $\epsilon_\theta(\cdot)$. Specifically, the encoder $\mathcal{E}(\cdot)$ is used to transform the input image x into the latent space $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}(\cdot)$ is used to reconstruct the input image from the latent z, $x \approx \mathcal{D}(z)$. The diffusion denoising process of Stable Diffusion is conducted in the latent space. With a randomly sampled noise $\epsilon \sim \mathcal{N}(0, I)$ and the time step t, we can get a noisy latent code $z_t = \alpha_t z + \sigma_t \epsilon$, where α_t and σ_t are the coefficients that control the noise schedule. Then the conditional diffusion model ϵ_θ will be trained with the following objective for denoising [18], [41]:

$$\min \ \mathbb{E}_{P,z,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, t, E_T(P))||_2^2]. \tag{1}$$

The goal of the conditional model $\epsilon_\theta(\cdot)$ in Eq.(1) is to predict the noise by taking the noisy latent z_t, the text conditional embedding obtained by $E_T(P)$, and the time step t as input.

Finetuning for Subject-Driven Text-to-Image Generation. Denote the small set of images of the specific subject s as $\mathbb{C}_s = \{x_i\}_{i=1}^K$, where x_i means the i^{th} image and K is the image number, usually 3 to 5. Previous works [12], [24], [38] will bind a special text token P_s, e.g., *"a S^* backpack"* in Figure 1, to the subject s, with the following finetuning objective:

$$\min \ \mathbb{E}_{z=\mathcal{E}(x),x\sim\mathbb{C}_s,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, t, E_T(P_s))||_2^2]. \tag{2}$$

This objective means that when the diffusion model $\epsilon_\theta(\cdot)$ takes the textual embedding $E_T(P_s)$ as input, it can always denoise for the images of the subject s, thus binding P_s to the subject s. Different methods use the objective in Eq. (2) to finetune different parameters, e.g., DreamBooth [38] will finetune the U-Net model $\epsilon_\theta(\cdot)$, while TI [12] finetunes the embedding of P_s in the CLIP text encoder $E_T(\cdot)$. DreamBooth finetunes more parameters than TI and achieves better subject-driven text-to-image generation ability. However, these methods bind P_s to several images $\{x_i\}$, making the textual embedding $E_T(P_s)$ inevitably entangled with information irrelevant to the subject identity, which will impair the generation results.

To tackle the problem, in this paper, we propose DisenDreamer, a sample-aware disentangled tuning framework for subject-driven text-to-image generation, whose framework is presented in Figure 2. DisenDreamer utilizes the textual embedding $E_T(P_s)$ and a visual embedding to denoise each image. Then, with our proposed three disentangled objectives, the text embedding $E_T(P_s)$ can preserve the identity-relevant information and the visual embedding can capture the identity-irrelevant information. During generation, by combining P_s with other text descriptions, DisenDreamer can generate images that conform to the text while preserving the identity. DisenDreamer can also generate images that preserve some characteristics of the input images by combining the textual identity-relevant embedding and the visual identity-irrelevant embedding, which provides a more flexible and controllable generation. Next, we will describe how DisenDreamer obtains the disentangled embeddings, how DisenDreamer finetunes the model with the designed disentangled auxiliary objectives, and then how it conducts subject-driven text-to-image generation with the finetuned model.

IV. The Proposed Method: DisenDreamer

A. DisenDreamer Model Design

DisenDreamer aims to disentangle the identity-relevant and the identity-irrelevant information in each image. Therefore, as shown in Figure 2, the common identity-relevant branch and the sample-aware identity-irrelevant branch are designed. Then, with the textual identity-relevant embedding, a better subject-driven text-to-image generation can be conducted. Next, we describe the two branches in detail.

The Common Identity-relevant Branch. This branch aims to obtain the identity-relevant embedding of the subject, which is shared by $\{x_i\}$. Like previous works [12], [38], we will map the identity of subject s to a special text token P_s. Then the identity-relevant embedding f_s can be obtained through the CLIP text encoder with P_s as its input,

$$f_s = E_T(P_s). \tag{3}$$

The Sample-aware Identity-irrelevant Branch. This branch aims to obtain the identity-irrelevant information of each image. Specifically, for each image of the subject x_i, we take the power of the pretrained CLIP image encoder E_I to obtain its image-specific visual embedding. However, since we only need the identity-irrelevant information in this embedding, we have to filter out the identity information of the subject. To achieve this, we design a sample-aware adapter for the visual feature. Specifically, for each image x_i, we extract its pretrained feature $f_i^{(p)}$ through E_I, $f_i^{(p)} = E_I(x_i)$. Considering that for two different images x_i and x_j, their identity-relevant information in f_i and f_j may be different in both dimension and value. Thus, to dynamically filter out the identity-relevant information of each sample, we design a mask generator to obtain the sample-aware mask for each image as follows,

$$M_i = \mathrm{MLP}(f_i^{(p)}), \tag{4}$$

where the mask generator is implemented by an MLP, and M_i is a vector that has the same dimension with $f_i^{(p)}$, whose value is normalized to $(0, 1)$ with the $Sigmoid(\cdot)$ funtion. M_i is a sample-aware mask for image x_i, which is expected to give smaller weights to the identity-relevant information and larger weights to the identity-irrelevant information. Additionally, since we want to use the identity-irrelevant embedding and the previous identity-relevant embedding together to denoise each image, we expect them to lie in the same space. Therefore, we add an MLP with skip connection to transform the masked identity-relevant embedding to the same space as f_s. Finally, the identity-irrelevant embedding of each image x_i is obtained as follows,

$$f_i = M_i * f_i^{(p)} + \mathrm{MLP}(M_i * f_i^{(p)}), i = 1, 2, \cdots, K, \tag{5}$$

where we will first mask the pretrained embedding $f_i^{(p)}$ to filter out the identity-relevant information, and then use the MLP with the skip connection to transform the masked embedding to the same space with f_s.

B. Tuning with Auxiliary Disentangled Objectives

With the above extracted identity-relevant and identity-irrelevant embeddings, we can conduct the finetuning with a similar denoising objective in Eq. (2) on the K images in $\mathbb{C}_s$,

$$\mathcal{L}_1 = \sum_{i=1}^{K} ||\epsilon_i - \epsilon_\theta(z_{i,t_i}, t_i, f_i + f_s)||_2^2. \tag{6}$$

As shown in Figure 2, ϵ_i is the randomly sampled noise for the i^{th} image, t_i is the randomly sampled time step for the i^{th} image, and z_{i,t_i} is the noisy latent of image x_i obtained by $z_{i,t_i} = \alpha_{t_i}\mathcal{E}(x_i) + \sigma_{t_i}\epsilon_i$ as mentioned in Sec. III. This objective means that we will use the sum of the identity-relevant embedding and the identity-irrelevant embedding as the condition to denoise each image. Since each image has an image-specific identity-irrelevant embedding, f_s does not have to restore the identity-irrelevant information. Additionally, considering that f_s is shared when denoising all the images, it will tend to capture the common information of the images, i.e., the subject identity. However, only utilizing Eq. (6) to denoise may cause several trivial solutions. One is that the visual embedding f_i captures all the information of image x_i, including the identity-relevant information and the identity-irrelevant information, while the shared embedding f_s becomes a meaningless conditional vector. The other trivial solution will be that f_i becomes a meaningless noise vector and f_s captures all the entangled information of the images, which degenerates to the situation of existing methods [12], [38]. To avoid these trivial solutions, we introduce the following three auxiliary disentangled objectives.

Weak Common Denoising Objective. To avoid the first trivial solution, we proposed the weak common denoising objective. Since we expect that f_s can capture the identity-relevant information instead of becoming a meaningless vector, f_s should have the ability of denoising the common part of the images. Therefore, we add the following objective:

$$\mathcal{L}_2 = \lambda_2 \sum_{i=1}^{K} ||\epsilon_i - \epsilon_\theta(z_{i,t_i}, t_i, f_s)||_2^2. \tag{7}$$

In this objective, we expect that only with the identity-preserved embedding, the model can also denoise each image. Note that we add a hyper-parameter $\lambda_2 < 1$ for this denoising objective, because we do not need f_s to precisely denoise each image, or f_s will again contain the identity-irrelevant information. Combining this objective and the objective in Eq. (6) together, we can regard the process in Eq. (6) as a precise denoising process, and regard the process in Eq. (7) as a weak denoising process. The precise denoising process with f_s+f_i as the condition should denoise both the subject identity and some irrelevant information such as the background, while the weak denoising process with f_s as the condition only needs to denoise the common subject identity of all the images, so it requires a smaller regularization weight $\lambda_2 < 1$. We use $\lambda_2 = 0.01$ for all our experiments and find it works well.

Weak Sample-aware Denoising Objective. Similarly, to avoid the second trivial solution where the identity-irrelevant embedding becomes a meaningless noise vector and the results degenerate to an entangled f_s, we also use each f_i to denoise each image x_i as follows,

$$\mathcal{L}_3 = \lambda_3 \sum_{i=1}^{K} ||\epsilon_i - \epsilon_\theta(z_{i,t_i}, t_i, f_i)||_2^2. \tag{8}$$

λ_3 is also a hyperparameter that is smaller than 1, which means we only expected f_i can denoise the identity-irrelevant information. In our experiments, λ_3 is 0.001.

Contrastive Embedding Objective. Since we expect f_s and f_i to capture disentangled information of the image x_i, the embeddings f_s and f_i should be contrastive and their similarities are expected to be low. Therefore, we add the contrastive embedding objective as follows,

$$\mathcal{L}_4 = \lambda_4 \sum_{i=1}^{K} cos(f_s, f_i). \tag{9}$$

Minimizing the cosine similarity between f_s and f_i will make them less similar to each other, thus easier to capture the disentangled identity-relevant and identity-irrelevant information of x_i. λ_4 is a hyper-parameter which is set to 0.001 for all our experiments. Therefore, the disentangled tuning objective of DisenDreamer is the sum of the above four parts:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4. \tag{10}$$

Parameters to Finetune. In previous works, DreamBooth [38] finetunes the whole U-Net model and achieves better subject-driven text-to-image generation performance. However, DreamBooth requires high computational and memory cost during finetuning. To reduce the cost while still maintaining the generation ability, we borrow the idea of LoRA [19] to conduct parameter-efficient finetuning. Specifically, for each pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ in the U-Net $\epsilon_\theta(\cdot)$, LoRA inserts a low-rank decomposition to it, $W_0 \leftarrow W_0 + BA$, where $B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}, r \ll min(d,k)$. A is initialized as Gaussian and B is initialized as $\mathbf{0}$, and during finetuning, only B and A are learnable, while W_0 is fixed. Therefore, the parameters to finetune are largely reduced from $d \times k$ to $(d+k)\times r$. With the LoRA technique, all the parameters require finetuning for DisenDreamer containing the parameters in the previous adapter and the LoRA parameters in the U-Net, as shown in the yellow block in Figure 2.

Discussions about Disentanglement. With the above model architecture designs and auxiliary objectives, the identity-relevant and identity-irrelevant embeddings can contain different aspects of information of each image. However, only from these regularizations, the identity-relevant embedding does not necessarily contain the subject identity information. For example, for an image x_i of a backpack in the forest, with the previous regularizations, one possible solution could be the identity-relevant embedding f_s contains half of the backpack, while the rest half backpack and the forest is contained in f_i. Thanks to the priors of the pretrained models, when we use textual prompt *"a S^* backpack"* to obtain the identity-relevant embedding, the pretrained model will naturally generate an image of backpack. Therefore, with a few steps of finetuning, with this strong prior and the proposed regularization, the model parameters will be easily optimized to a point where

the textual embedding contains the identity-relevant information and the visual embedding contains the identity-irrelevant information. The disentanglement is validated in Figure 9 and Figure 10.

C. More Flexible and Controllable Generation

After the above finetuning process, DisenDreamer binds the identity of the subject s to the text prompt P_s, e.g., "*a* S^* *backpack*". When generating new images of subject s, we can combine other text descriptions with P_s to obtain the new text prompt P'_s, e.g., "*a* S^* *backpack on the beach*". Then, the CLIP text encoder will transform it to its text embedding $f'_s = E_T(P'_s)$. With f'_s as the condition, the U-Net model can denoise a randomly sampled Gaussian noise to an image that conforms to P'_s while preserving the identity of s. Moreover, if we want the generated image to inherit some characteristics of one of the input images x_i, e.g., the pose, we can obtain its visual identity-irrelevant embedding f_i through the image encoder and the finetuned adapter, and then, use $f'_s + \eta f_i$ as the condition of the U-Net model. Finally, the generated image will inherit the characteristic of the reference image x_i, and η is a hyper-parameter that can be defined by the user to decide how many characteristics can be inherited. DisenDreamer not only enables the user to control the generated image by the text but also by the preferred reference images in the small set, which is more flexible and controllable.

V. Experiments

In this section, we report the experimental results of our proposed DisenDreamer on the public dataset *DreamBench* and compare its performance with different baseline methods both quantitatively and qualitatively.

A. Experimental Settings

Dataset. We adopt the subject-driven text-to-image generation dataset *DreamBench* proposed by [38], which are downloaded from Unsplash[1]. This dataset contains 30 subjects, including unique objects like backpacks, stuffed animals, cats, etc. For each subject, there are 25 text prompts, which contain recontextualization, property modification, accessorization, etc. Totally, there are 750 unique prompts, and we follow [38] to generate 4 images for each prompt, and totally 3,000 images for robust evaluation.

Evaluation Metrics. The subject-driven text-to-image generation evaluates the generated images in two aspects: (i) The first aspect is the subject identity fidelity, i.e., whether the generated image has the same subject as the input images. To evaluate the model performance in this aspect, we adopt the DINO score proposed by [38], which is the average pairwise cosine similarity between the ViT-S/16 DINO embeddings of the generated images and the input real images. A higher DINO score means that the generated images have higher similarity to the input images, but may risk overfitting the input images. For example, generating the same image as the input image can achieve a quite high DINO score, but it does not fit the given text prompt, which is not what we expect in subject-driven text-to-image generation. (ii) The second aspect is the text prompt fidelity, i.e., whether the generated images conform to the text prompts. This aspect can be evaluated by the average cosine similarity between the text prompt and image CLIP embeddings. This metric is denoted as CLIP-T [12], [38].

Baselines. We compare our proposed method with the following baselines. TI [12] and DreamBooth [38] are finetuning methods for subject-driven text-to-image generation. InstructPix2Pix [4] is a SOTA pretrained method for text-guided text-to-image editing. We also include the pretrained Stable Diffusion (SD) [36] without finetuning as a reference model. The DINO metric of pretrained SD can be regarded as a lower bound, since it does not have prior knowledge about the subject identity. The CLIP-T metric of the pretrained SD can be regarded as the upper bound, since it will not overfit the small set of images, thus always expected to generate images that conform to the text.

Implementation Details. We implement DisenDreamer based on the pretrained Stable Diffusion 2-1 [36]. The learning rate is 1e-4 with the AdamW [27] optimizer. The finetuning process is conducted on one Tesla V100 with batch size of 1, while the finetuning iterations are $\sim$3,000. As for the LoRA rank, we use $r = 4$ for all the experiments. The MLP used in the adapter is 2-layer with ReLU as the activation function. The LoRA and adapter make the parameters to finetune about $3M$, which is small compared to the whole $865.9M$ U-Net parameters. Additionally, the special token P_s we use to obtain the identity-preserved embedding is the same as that of [38], i.e., "*a* + S^* + *class*", where S^* is a rare token and $class$ is the class of the subject. When comparing with the baselines, we only use the textual embedding f'_s mentioned in Sec. IV-C as the condition.

B. Comparison with Baselines

Comparison on DreamBench. The scores of different methods are shown in Table I and some generated result visualizations of different methods on *DreamBench* are shown in Figure 3, Figure 4, and Figure 5. Denoting the position of the first generated image of InstructPix2Pix as row 1 column 1, from the results, we can observe that:

- Pretrained SD, as the reference model, has the lowest DINO score and the highest CLIP-T score as expected. The pretrained SD is not customized for the subject, thus having the lowest image similarity, but it has the highest CLIP-T score because it does not overfit the input images.
- InstructPix2Pix has the lowest CLIP-T score, indicating it can not support the complex transformations in subject-driven text-to-image generation. On the other hand, it does not have the idea of the subject, easily making the identity of the subject changed. For example, in row 1 column 3 of Figure 4, the color of the backpack is changed to orange. In row 1 column 4 of Figure 3, the color of the can is changed to pink, and in row 1 column 2 of Figure 5, the dog is changed to blue but in fact, we only need a blue house behind the original dog. Additionally,

[1]https://unsplash.com/.

TABLE I
DINO AND CLIP-T OF DIFFERENT METHODS ON *DreamBench*. EXCEPT THE REFERENCED MODEL PRETRAINED SD, WE BOLD THE METHOD WITH THE BEST PERFORMANCE W.R.T. EACH METRIC.

	pretrained SD [36]	**InstructPix2Pix** [4]	**TI** [12]	**DreamBooth** [38]	**DisenDreamer**
DINO Score↑	0.362	0.605	0.546	**0.685**	0.680
CLIP-T Score↑	0.352	0.303	0.318	0.319	**0.329**

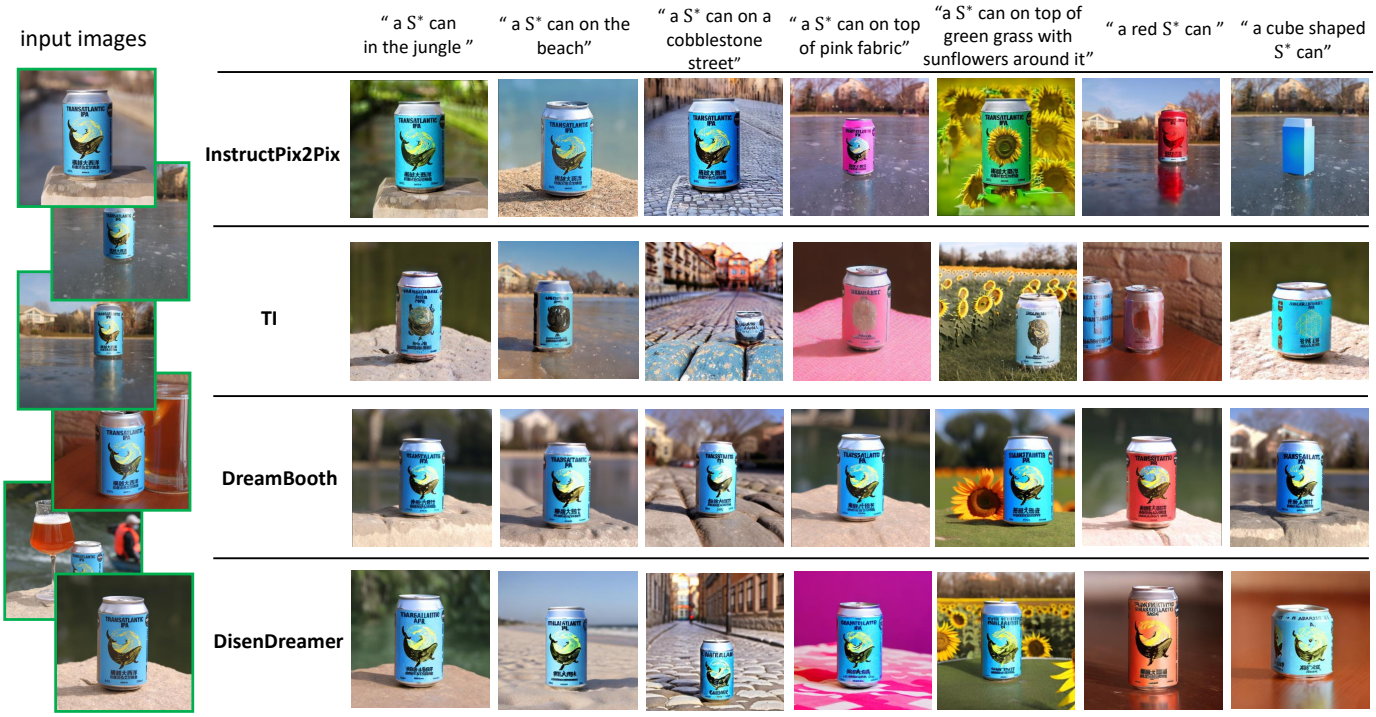


Fig. 3. Generated images of the *can* given different text prompts with different methods.

Fig. 4. Generated images of the *backpack* given different text prompts with different methods.

Fig. 5. Generated images of the *dog* given different text prompts with different methods.

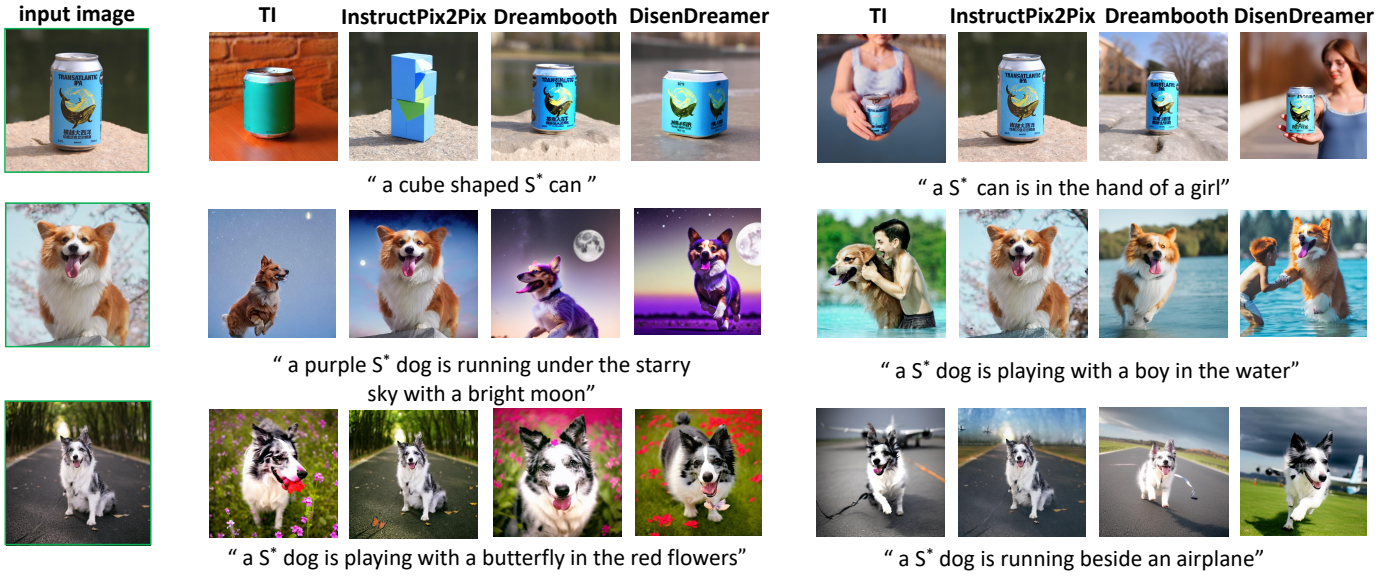


Fig. 6. Generated examples of different subjects.

the generated pictures of InstructPix2Pix always have the same pose as the input images.

- TI has a very low DINO score, which means it usually suffers severe identity change of the subject. For example, in Figure 4, the color of the backpacks generated by TI is often different from the input images. In Figure 3, in column 3 and column 4 of row 2, the cans also have different identities from the input images.
- Dreambooth has the highest DINO score, which means that the generated images are very similar to the input images. However, as observed in the generated images, this too high DINO score results from overfitting the input images. For example, in Figure 3, almost all the images of the generated cans have the same background as the last input image, making it ignore the text prompts such as *"on the beach"* in row 3 column 2, and *"on top of pink fabric"* in row 3 column 4. Similar phenomenons are also observed in Figure 5, where *"on top of a purple rug"* in row 3 column 3 and *"a purple S* dog"* in row 3 column 6 are also ignored.
- Our proposed DisenDreamer has a very high DINO score and the highest CLIP-T score, which means the subject

identity is well preserved and the generated images also conform to the text descriptions. The visualization results of DisenDreamer also show its superiority.

Besides the given prompts of *DreamBench*, we also give more complex text prompts to different subjects for subject-driven text-to-image generation. The comparison is shown in Figure 6. The results further show that DisenDreamer can better generate images with these prompts while still preserving the subject identity.

More Comprehensive Comparison with DreamBooth. Since DisenDreamer and DreamBooth show better performance than other baselines, in Table II, we provide a more comprehensive comparison between DreamBooth and our DisenDreamer, in CLIP-I, DINO, PRES, DIV and CLIP-T metrics. We have described the DINO and CLIP-T metrics before and then we will describe the other 3 metrics.

- CLIP-I is similar to DINO, which is used to measure whether the generated image contains the given subject. CLIP-I measures the average cosine similarity between the generated images and the input images using the CLIP image encoder, but its ability to identify the subject is not as good as DINO as mentioned in DreamBooth [38]. Therefore, we only use DINO in Table I.
- PRES is a prior preservation metric proposed in DreamBooth [38] to verify the effectiveness of their proposed prior preservation loss. This metric calculates the average DINO cosine similarity between the input subject images and the generated images using prompt " a + class ". A lower PRES indicates that the model will generate more different images from the input images using " a + class " as the prompt, indicating less overfitting. However, this metric does not directly relate to the performance of subject-driven text-to-image generation, because when we generate the image for the given subject, we must use the "a + S* + class" prompt. The generated images of " a + class " not overfitting the input images cannot guarantee the generated images of "a + S* + class" do not overfit the input images, "S*" may also overfit the input images.
- DIV is the average LPIPS [50] distance between the generated images of the same subject with the same prompt. A higher DIV means the generated images are more diverse.

In Table II, DreamBooth-PPL is the variant where the prior preservation loss is removed from DreamBooth. We can see that DreamBooth-PPL has the highest CLIP-I and DINO score, which means its generated images are more similar to the input images, well preserving the subject identity but overfitting the input images as mentioned in their original paper of [38]. DreamBooth has the lowest PRES because it uses additional data to directly optimize this metric to avoid overfitting the input images. Therefore, DreamBooth has a higher DIV and CLIP-T score than DreamBooth-PPL by generating more diverse images that conform to the text descriptions. This means the prior preservation loss can alleviate the overfitting problem. However, the failure generated cases presented in DreamBooth [38] show it still suffers from overfitting the identity-irrelevant input information, such as the background.

Fig. 7. Comparison with DreamBooth.

Therefore, our proposed DisenDreamer avoids this problem by disentangling the identity-relevant and identity-irrelevant information, thus achieving the best DIV and CLIP-T scores. We further provide qualitative comparisons among DreamBooth-PPL, DreamBooth and DisenDreamer in Figure 7, which further show that our proposed method can generate images that conform to the text while preserving the subject identity. DreamBooth and DreamBooth-PPL suffer overfitting the identity-irrelevant information problem.

TABLE II
COMPARISON WITH DREAMBOOTH IN CLIP-I, DINO, PRES, DIV, CLIP-T METRICS.

	CLIP-I↑	DINO↑	PRES ↓	DIV↑	CLIP-T↑
DisenDreamer	0.681	0.680	0.620	**0.668**	**0.329**
DreamBooth	0.706	0.685	**0.415**	0.654	0.319
DreamBooth-PPL	**0.712**	**0.693**	0.622	0.627	0.307

Comparison with General Text-to-image Models. To distinguish the subject-driven text-to-image methods and general text-to-image methods, we compare DisenDreamer with two general text-to-image models, i.e., Stable Diffusion [36] and Box-Diff [47], in Table III. The results show that the general text-to-image methods achieve a very low DINO score, because they are not customized for the given subject, and fail to generate images that preserve the subject identity. To further illustrate the problem, we provide the qualitative comparison results in Figure 8. The qualitative results show that both Stable Diffusion and Box-Diff fail to generate images that contain the same subject as the given images, but one goal of the subject-driven text-to-image generation is to preserve the subject identity. It is not hard to understand the phenomenon, the general methods are not customized for the given subjects, and will naturally fail in subject-driven text-to-image generation.

TABLE III
COMPARISON WITH GENERAL TEXT-TO-IMAGE METHODS.

	Stable Diffusion [36]	**Box-Diff** [47]	**DisenDreamer**
DINO	0.362	0.349	**0.680**
CLIP-T	**0.352**	0.316	0.329

Comparison on CustomConcept101. Besides the results on DreamBench, we also verify the effectiveness of our method on another benchmark for subject-driven text-to-image

Fig. 8. Comparison with general text-to-image methods.

generation, CustomConcept101 [24]. We follow [24] to use 10 subjects to evaluate different methods, which contain natural scenes, pets, and objects, and each subject is evaluated with 20 prompts. We also generate 4 images for each prompt for robust evaluation. The comparison results are shown in Table IV. The results further show that our proposed DisenDreamer achieves both the best DINO score and the CLIP-T score, which means our method can simultaneously preserve the subject identity and conform to the text description best.

TABLE IV
COMPARISON ON CUSTOMCONCEPT101.

	SD	TI	Pix2Pix	DreamBooth	DisenDreamer
DINO↑	0.353	0.567	0.593	0.603	**0.604**
CLIP-T↑	0.371	0.346	0.312	0.351	**0.359**

C. Ablation Study

Disentanglement: f_s and Identity-Relevant Information. The textual embedding f_s obtained through the special text prompt P_s aims to preserve the subject identity. We verify the relationship between f_s and the identity in Figure 9, where we only use P_s as the prompt to generate images. The results show that the shared textual embedding can faithfully capture the identity of the given subject.

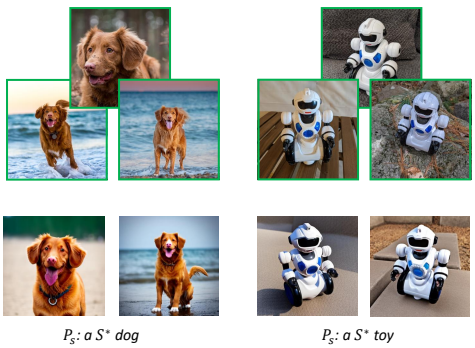


Fig. 9. Examples of using P_s to generate images with the same subject.

Disentanglement: f_i and Identity-Irrelevant Information. The visual embedding f_i aims to capture the identity-irrelevant information of image x_i. We verify this relationship by directly using f_i as the condition to generate new images. The results are shown in Figure 10. In each image pair, the left one is the input image x_i and the right one is the image generated with its visual embedding f_i. We can see that as expected, f_i indeed captures the identity-irrelevant information such as background and pose. For example, in the first row of the input *backpack* images, the f_i generates similar background with x_i while not containing any *backpack*. In the second and third rows, f_i contains not only the background information but also the pose information of the subject. Similarly, identity information about the two dogs is not included. The results show that f_i in our DisenDreamer faithfully captures the identity-irrelevant information of each image x_i.

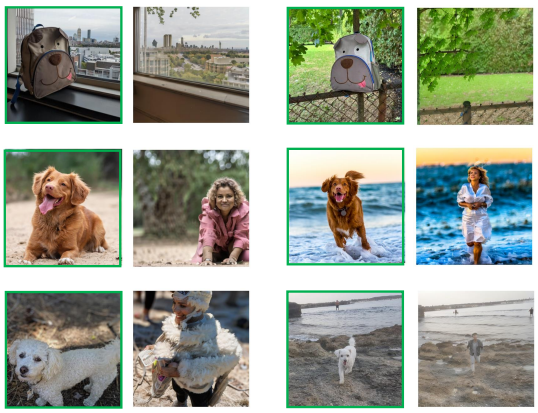


Fig. 10. Examples of using f_i to generate images irrelevant to the subject identity.

The disentanglement explains why our DisenDreamer outperforms current methods. The shared textual embedding only contains the information about the subject identity, making generating new background, pose, or property easier and resulting in better generation results.

More Flexible and Controllable Generation with f_i. From Figure 3 to 6, we only combine the identity-relevant information P_s with other text prompts to obtain f_s' as the condition. As aforementioned, if we want to inherit some characteristics of the input image x_i, we can add the visual embedding f_i to f_s' with a user-defined weight η, i.e., the condition is $f_s' + \eta f_i$. In Figure 11, we provided examples of two subjects. For each subject, with the same text prompt *"a S* dog on the Great Wall"* to obtain f_s', we select 2 different images of the *dog* to obtain f_i, and the weight η is linearly increased. The results show that with larger η, the generated image will be more similar to the reference image. With a relatively small η, the generated image can

TABLE V
EFFECTIVENESS OF THE PROPOSED COMPONENTS.

	w/o wcd	w/o wsad	w/o ce	w/o wcd&wsad	w/o mask	w/o skip	DisenDreamer
DINO↑	0.663	0.674	0.670	0.662	0.656	**0.678**	**0.678**
CLIP-T↑	0.331	0.323	0.327	0.331	0.328	0.322	**0.334**

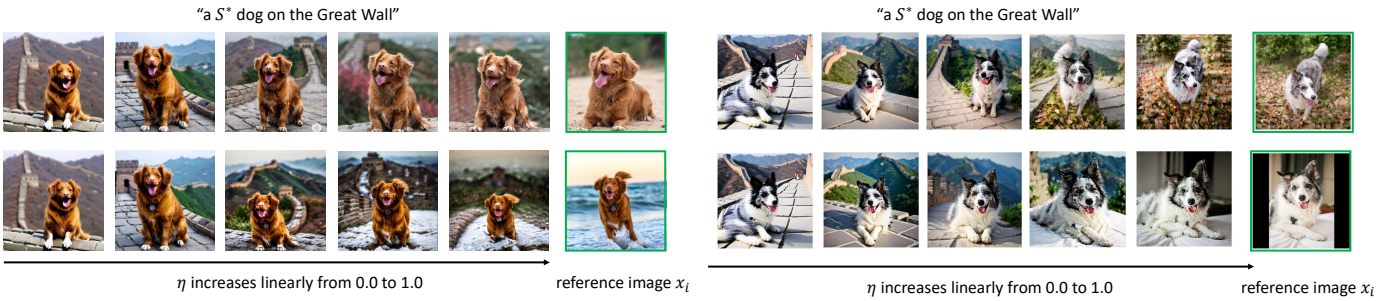


Fig. 11. Generating images with different reference images with different η.

simultaneously conform to the text and inherit some reference image characteristics, which gives a more flexible and controllable generation. However, as η becomes large, the text prompt will be ignored and the generated image will be the same as the reference image. This phenomenon means the identity-irrelevant part will impair the subject-driven text-to-image generation, which inspires us to disentangle the tuning process.

Effectiveness of the Proposed Components. We validate the effectiveness of each of our proposed components on randomly sampled 10 subjects of DreamBench. The results are reported in Table V, where we respectively remove the weak common denoising objective(w/o wcd), remove the weak sample-aware denoising objective(w/o wsad), remove the contrastive embedding objective(w/o ce), remove both the weak common denoising and weak sample-aware objectives(w/o wcd&wsad), remove the sample-aware mask and remove the skip connection(w/o skip). The results show that:

- Without the weak common denoising objective(wcd), the DINO score will significantly decrease, which means the weak common denoising objective is important to preserving the subject identity.
- Without the weak sample-aware denoising objective(wsad) to make the identity-irrelevant branch learn the identity-irrelevant information, the identity-relevant branch will risk in overfitting the identity-irrelevant information, causing the model to have a low CLIP-T score.
- The contrastive embedding objective(ce) boosts disentanglement and improves both the DINO and the CLIP-T score.
- Without the sample-aware mask to filter out the identity-relevant information, the sample-aware identity-irrelevant branch will also contain identity information. This will cause the common identity-relevant branch to contain less identity information, thus w/o mask suffers a low DINO score.
- Without the skip connection(skip), it will be hard for the identity-irrelevant branch to learn the identity-irrelevant information. This will cause all the information to be entangled in the common identity-relevant branch, thus the model will have a low CLIP-T score.

To further demonstrate the effectiveness of the components, we provide the qualitative results in Figure 12, which are consistent with the quantitative results.

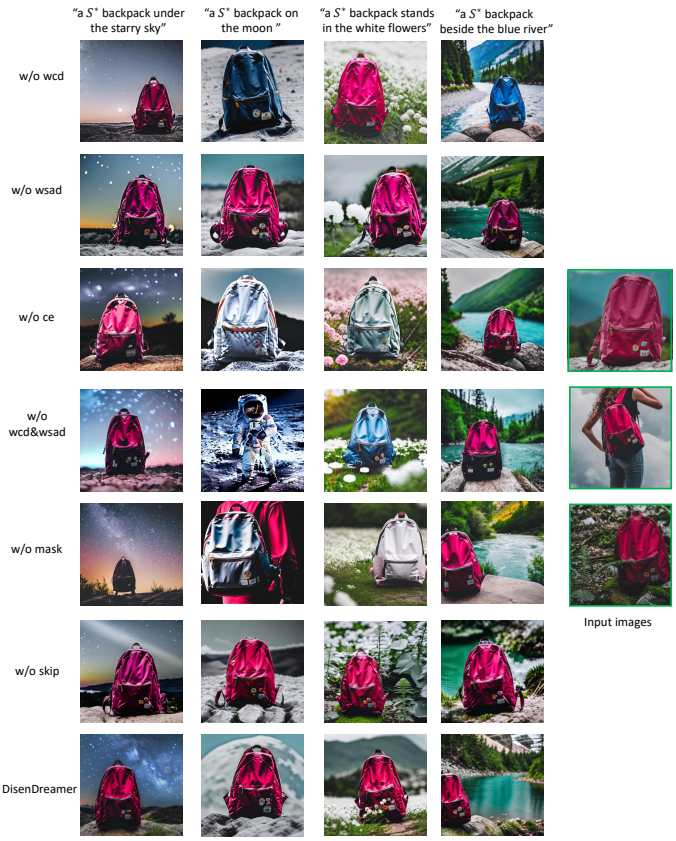


Fig. 12. The effectiveness of the proposed components.

Sample-aware Mask to Filter Out the Identity Information. In Figure 13, we visualize the disentangling ability of DisenDreamer with and without the sample-aware mask. We can see that with the sample-aware mask, the images generated by f_i and P_s are well disentangled, and the identity information is filtered out from f_i. In contrast, without the sample-aware mask, the image generated by f_i will contain the identity information(the backpack), which will make P_s has less information about the backpack, resulting in the subject generated by P_s less similar to the input image and a lower DINO score as shown in w/o mask Table V.

Hyper-parameter Experiments. In our experiments, we fix $\lambda_2, \lambda_3, \lambda_4$ to $0.01, 0.001, 0.001$ respectively. We conduct ablation studies on these hyper-parameters on the "berry bowl" subject by changing the hyper-parameters in $\{0.0, 0.001, 0.01, 0.1\}$, and the results are shown in Table VI.

Fig. 13. Visualization for how the sample-aware mask helps to filter out the identity information.

Fig. 14. DisenDreamer generated examples on some anime subjects.

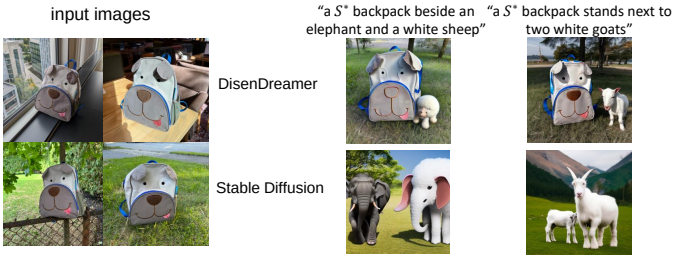


Fig. 15. Failure generation cases.

We can see that as the λ_2 (the weight of the weak common denoising objective) increases, the subject will be more similar to the input images, thus having a higher DINO score, but may risk overfitting the input images and a low CLIP-T score. Increasing λ_3 (the weight of the weak sample-aware objective) can make the identity-irrelevant branch contain identity-irrelevant information, which can prevent the common identity-relevant branch overfitting the identity-irrelevant information, thus increasing the CLIP-T score. However, too large λ_3 may make the identity-irrelevant branch also contain identity-relevant information, resulting the identity-relevant branch containing less identity information, thus having a low DINO score. From our empirical observation, setting λ_2 to 0.01, λ_3 and λ_4 to $0.001 - 0.01$ works well for most cases.

TABLE VI
HYPER-PARAMETER EXPERIMENTS ON $\lambda_1, \lambda_2, \lambda_3$.

λ_2	0.000	0.001	0.010	0.100
DINO	0.756	0.762	0.803	0.866
CLIP-T	0.273	0.272	0.270	0.224
λ_3	0.000	0.001	0.010	0.100
DINO	0.790	0.803	0.797	0.717
CLIP-T	0.244	0.270	0.272	0.275
λ_4	0.000	0.001	0.010	0.100
DINO	0.791	0.803	0.802	0.798
CLIP-T	0.232	0.270	0.269	0.258

D. Generation with Anime Subjects

In previous examples, we finetune the subjects on *DreamBench*. We also use DisenDreamer to finetune on some anime characters, and the results are shown in Figure 14. The results show that our DisenDreamer works well for these anime subjects.

VI. LIMITATIONS

The limitations of DisenDreamer lie in the following two aspects. The first one is that since our DisenDreamer is a finetuning method on pretrained Stable Diffusion, it inherits the limitations of the pretrained Stable Diffusion, e.g., compositional abilities as shown in Figure 15. When generating multiple subjects simultaneously, both Stable Diffusion and our method will miss some subjects, e.g., our method misses the elephant in the first image and one goat in the second image, and Stable Diffusion misses the backpack in both images and generates a white elephant instead of a white sheep in the first image. The second limitation is that since our proposed method utilizes the model prior and does not require any additional supervision, it can only disentangle the subject identity-relevant and identity-irrelevant information. How to conduct more fine-grained disentanglement in the identity-irrelevant information, e.g., the pose, the background and the image style, to achieve a more flexible and controllable generation is worth exploring in the future.

VII. CONCLUSION

In this paper, we propose DisenDreamer for subject-driven text-to-image generation. Different from existing methods which learn an entangled embedding for the subject, DisenDreamer will use an identity-relevant embedding and several sample-specific identity-irrelevant embeddings for all the images in the finetuning process. During generation, with the identity-relevant embedding, DisenDreamer can generate images that simultaneously preserve the subject identity and conform to the text descriptions. Additionally, DisenDreamer shows superior subject-driven text-to-image generation ability and can serve as a more flexible and controllable framework.

REFERENCES

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[7] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. Curriculum disentangled recommendation with noisy multi-feedback. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 26924–26936, 2021.

[8] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023.

[9] Xin Deng, Enpeng Liu, Shengxi Li, Yiping Duan, and Mai Xu. Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. *IEEE Transactions on Image Processing*, 32:1078–1091, 2023.

[10] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022.

[11] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023.

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[13] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023.

[14] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1294–1305, 2018.

[15] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023.

[16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *International Conference on Learning Representations*, 2022.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[19] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

[20] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 424–434. Association for Computational Linguistics, 2019.

[21] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[23] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023.

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

[25] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[26] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Reproducibility in Machine Learning, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

[28] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5712–5723, 2019.

[29] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 99–108. Computer Vision Foundation / IEEE Computer Society, 2018.

[30] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023.

[31] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations.

[32] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015.

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[40] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.
[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*.
[42] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *CoRR*, abs/2211.11695, 2022.
[43] Xin Wang, Hong Chen, and Wenwu Zhu. Multimodal disentangled representation for recommendation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.
[44] Xin Wang, Zirui Pan, Yuwei Zhou, Hong Chen, Chendi Ge, and Wenwu Zhu. Curriculum co-disentangled representation learning across multiple environments for social recommendation. In *International conference on machine learning*. PMLR, 2023.
[45] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023.
[46] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022.
[47] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. *arXiv preprint arXiv:2307.10816*, 2023.
[48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.
[49] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.
[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
[51] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. *arXiv preprint arXiv:2303.13126*, 2023.
[52] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10157–10166, 2023.

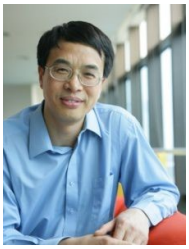
Hong Chen received B.E. from the Department of Electronic Engineering, Tsinghua University, Beijing, China in 2020. He is currently a PH.D. candidate in the Department of Computer Science and Technology of Tsinghua University. His main research interests include auxiliary learning and multi-modal learning. He has published several papers in top conferences and journals including NeurIPS, ICML, IEEE TPAMI, etc.

Yipeng Zhang is currently a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. He got his B.E. degree at the Department of Computer Science and Technology, Tsinghua University. His research interests include machine learning, disentangled representation learning and auxiliary learning.

Xin Wang is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications in multimedia big data analysis. He has published over 150 high-quality research papers in top journals and conferences including IEEE TPAMI, IEEE TKDE, ACM TOIS, ICML, NeurIPS, ACM KDD, ACM Web Conference, ACM SIGIR and ACM Multimedia etc., winning three best paper awards. He is the recipient of 2020 ACM China Rising Star Award, 2022 IEEE TCMC Rising Star Award and 2023 DAMO Academy Young Fellow.

Xuguang Duan is a graduate student at the Department of Computer Science and Technology, Tsinghua University. He got his B.E degree at the Department of Electronic Engineering, Tsinghua University. His research interests include machine learning, neural-symbolic systems, video understanding. He has published some research paper in top conferences and journals include NeurIPS, ICML,TPAMI, ACM Multimedia, etc.

Yuwei Zhou is currently a Ph.D. student at the Department of Computer Science and Technology, Tsinghua University. His email is zhou-yw21@mails.tsinghua.edu.cn. He received his B.E. degree from the Department of Computer Science and Technology, Tsinghua University. His main research interests include machine learning, curriculum learning and multimodal learning.



Wenwu Zhu is currently a Professor in the Department of Computer Science and Technology at Tsinghua University, the Vice Dean of National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs New Jersey as Member of Technical Staff during 1996-1999. He received his Ph.D. degree from New York University in 1996.

His research interests include graph machine learning, curriculum learning, data-driven multimedia, big data. He has published over 400 referred papers, and is inventor of over 80 patents. He received ten Best Paper Awards, including ACM Multimedia 2012 and IEEE Transactions on Circuits and Systems for Video Technology in 2001 and 2019.

He serves as the EiC for IEEE Transactions on Circuits and Systems for Video Technology, the EiC for IEEE Transactions on Multimedia (2017-2019) and the Chair of the steering committee for IEEE Transactions on Multimedia (2020-2022). He serves as General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019. He is an AAAS Fellow, IEEE Fellow, ACM Fellow, SPIE Fellow, and a member of Academia Europaea.