CLASSDIFFUSION: MORE ALIGNED PERSONALIZATION TUNING WITH EXPLICIT CLASS GUIDANCE

Jiannan Huang[1,2,4] **Jun Hao Liew**[3] **Hanshu Yan**[3] **Yuyang Yin**[1,2]
Yao Zhao[1,2] **Humphrey Shi**[4] **Yunchao Wei**[*1,2]
[1] Institute of Information Science, Beijing Jiaotong University
[2] Visual Intelligence + X International Joint Laboratory of the Ministry of Education
[3] ByteDance Inc. [4] SHI Labs@Georgia Tech

ABSTRACT

Recent text-to-image customization works have proven successful in generating images of given concepts by fine-tuning the diffusion models on a few examples. However, tunning-based methods inherently tend to overfit the concepts, resulting in failure to create the concept under multiple conditions (*e.g.*, headphone is missing when generating "a <sks> dog wearing a headphone"). Interestingly, we notice that the base model before fine-tuning exhibits the capability to compose the base concept with other elements (*e.g.*, "a dog wearing a headphone"), implying that the compositional ability only disappears after personalization tuning. We observe a semantic shift in the customized concept after fine-tuning, indicating that the personalized concept is not aligned with the original concept, and further show through theoretical analyses that this semantic shift leads to increased difficulty in sampling the joint conditional probability distribution, resulting in the loss of the compositiona ability. Inspired by this finding, we present **ClassDiffusion**, a technique that leverages a **semantic preservation loss** to explicitly regulate the concept space when learning the new concept. Although simple, this approach effectively prevents semantic drift during the fine-tuning process on the target concepts. Extensive qualitative and quantitative experiments demonstrate that the use of semantic preservation loss effectively improves the compositional abilities of fine-tuning models. Lastly, we also extend our ClassDiffusion to personalized video generation, demonstrating its flexibility.

1 INTRODUCTION

Thanks to the rapid progress in the diffusion model [31, 48, 55, 59, 63, 65, 67, 68, 72, 92, 95], the field of text-to-image generation has achieved significant progress in recent years. The leading text-to-image models [1, 20, 34, 38, 40, 66, 80] have been successful in generating high-fidelity images that align well with textual inputs. Recently, a significant part of the research [2, 4, 5, 8, 11, 28, 36, 60, 79, 88, 93, 94] has changed their focus from creating high-quality images to improving control over the generated images. Among these works, an important and widely explored research domain is subject-driven personalized generation, which aims to generate new images for a specific concept given some reference images of that concept.

Existing personalization methods [1, 9, 20, 21, 34, 38, 40, 51, 66, 78, 80, 81, 85] can generate images that closely resemble the concept by fine-tuning the base text-to-image model in a specific image set. However, all tuning-based models will inherently suffer from the over-fitting introduced by this process, which leads to weakening in the compositional ability of the model. For example, when generating "a <sks> dog wearing a headphone", though the given dog is well reconstructed, the headphone is always missing (Fig. 1). This feature affects the diversity of the generated output in practical use. A commonly accepted explanation within the community [20, 29, 66, 75] attributes this phenomenon to overfitting given a limited number of images. However, the fundamental cause of this overfit remains unexplored. In this work, our aim is to investigate the underlying causes behind the overfitting.

[*]Corresponding author.

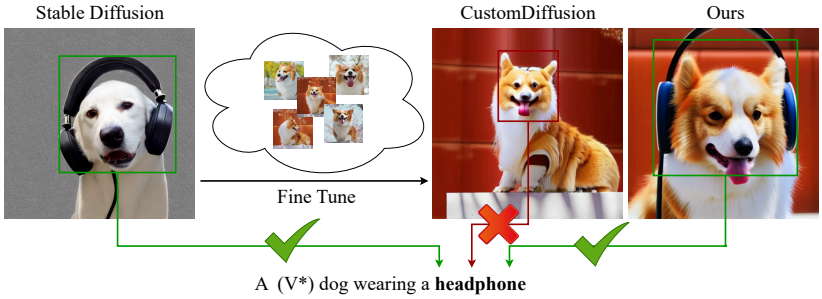


Figure 1: The base Stable Diffusion (SD) possesses the capbility to compose the concept of a dog and headphone, generating a dog wearing a headphone. However, we notice that this compositional generation capability is lost during personalization tuning. For example, when using Custom Diffusion (CD) [38], the headphone is missing despite the target corgi is generated successfully. On the other hand, our method can successfully compose the target corgi with the headphone.

Upon initial examination, it appears that the model diminishes some of its original capabilities after personalization tuning. Taking Stable Diffusion (SD) [74] as an example, from Fig. 1, we observe that the base SD model indeed has the ability to combine the concepts of a dog and a headphone. However, after fine-tuning, the model struggles to achieve compositional generation; for instance, while the target concept `<sks>` (dog) can be generated successfully, the headphone is missing. **We hypothesized that the decline in this compositional ability stems from the semantic drift of the target concept away from its superclass target during fine-tuning.** To better understand this, we conduct some empirical analysis by visualizing the CLIP text-space and cross-attention map activation area in Fig. 3a, 3b. In addition, we also perform theoretical analysis and find that the root cause lies in the semantic bias that reduces the entropy of the probability of the composed conditions, which significantly increases the difficulty to simultaneously sample the target concept combined with other elements.

Based on our experimental findings and theoretical analysis, we introduce ClassDiffusion to address the issue of weakening compositional capacity after fine-tuning. Fig. 2 shows the performance of our method. Our method uses semantic preservation loss to explicitly guide the model to restore the semantic imbalance that arises during the fine-tuning stage. In particular, it narrows the gap between the text embeddings of the target concept and its respective superclass in the textual space. Despite its simplicity, the proposed loss can successfully recover the compositional ability as shown in Fig. 1. Therefore, distinct from prevalent loss design in the community which seek to migrate the overfitting in tuning-based models, our method enhances the model's capacity of following the text prompt while maintaining the concept of a customized subject. Extensive experiments have demonstrated the effectiveness of our method in restoring the compositional generation capability of the base model. Furthermore, we explore the potential of our approach in personalized video synthesis, showcasing its ability in recovering the semantical space of the generative model. In addition, we found that the CLIP-T metric can hardly reflect the actual performance of personalized generation. Therefore, we introduce the BLIP2-T metric, a more equitable and effective evaluation metric for this particular domain. To summarize, the contributions of our work are:

- We offer a thorough examination to understand why existing tuning-based subject-driven personalized methods inherently suffer from the loss of compositional ability. This is elucidated through both experimental observations and theoretical analysis.
- We propose ClassDiffusion, a simple technique to recover the compositional capabilities lost during personalized tuning.
- Extensive experiments demonstrate that the proposed technique achieves improved personalization ability in image and video generation tasks.

2 Related Work

Text-to-Image Generation and Its Control Text-to-image generation is designed to generate high-quality, high-fidelity images that are aligned with textual prompts. This field has been under research

Figure 2: A qualitative result of two small stories produced by our model. The above showcases a bear's literary journey: from reading a book to ultimately earning a Nobel Literature Prize. The below shows the fate of a sunglasses. Finally, the bear gets the sunglasses. It shows a potential real-world application due to our model's high performance.

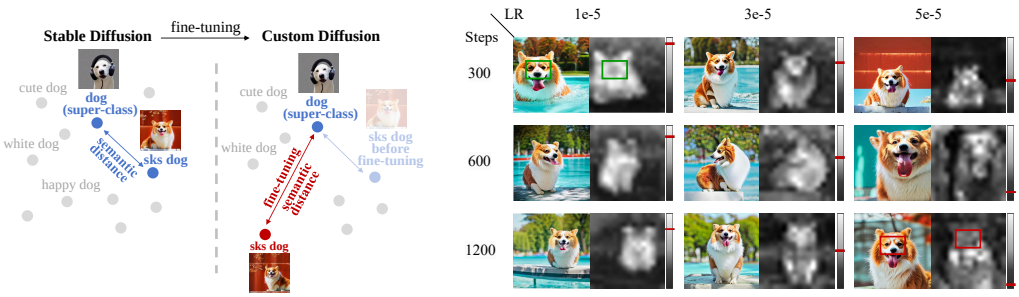
for an extended period. Recently, the field of Text-to-image generation has made significant progress with extensive research in Generative Adversarial Network (GAN) [23, 24, 53, 98], Variational Autoencoder (VAE) [6, 13, 37, 76], and Diffusion models [31, 48, 55, 59, 63, 65, 67, 68, 72, 95]. The diffusion models achieve a new state-of-the-art (SOTA) in unconditional image generation. Numerous works [12, 33, 43, 46, 55, 61–63, 67, 91, 96, 99] have been done to make the image generated by diffusion models more aligned with the textual prompts. Among them, Stable Diffusion [65] is a widely recognized model in the field, utilizes a cross-attention mechanism to integrate textual conditions into the image generation process and employs the Latent Diffusion Model, which maps the image to latent space [65]. Our research is based on the Stable Diffusion framework due to its adaptability and wide use in the community. Furthermore, different methods exist for controlling generative models. The primary categories for controlling generative models typically include Text-guided [7, 18, 22, 60, 64], Image-guided [1, 20, 34, 38, 40, 66, 80], Additional Sparse conditions [2, 4, 5, 8, 11, 28, 36, 49, 60, 79, 88, 93, 94], Brain-guided [3, 10, 19, 47, 54, 57, 74], Sound-Guided [60, 89], and some universe control [42, 49, 90]. Text-guided control utilizes textual descriptions to directly influence the outcome, guiding the model based on specific verbal instructions. Our method focused on the text-guided controllable generative model.

Subject-Driven Personalized Generation Subject-driven personalized generation is focused on creating images based on reference images. Recent works [1, 9, 20, 21, 34, 38, 40, 51, 58, 66, 73, 78, 80, 81, 85, 87] have explored techniques for producing striking resemblance images in multiple ways. One of the primary ways is to fine-tune the base text-to-image models. Furthermore, there has been a significant effort in research [14, 29, 32, 35, 38, 39, 71, 80, 84] aimed at integrating various concepts in personalization. While striking resemblance images are produced, fine-tuning the base model on a small set of images leads to overfitting, resulting in unexpected issues. One prevalent issue discussed in prior research is the decrease in diversity. Recent studies have proposed various methods to address these issues. For instance, DreamBooth [66] introduced the Class-Specific Priori Loss, which effectively addresses diversity reduction by recovering the class's prior knowledge. However, it does not effectively maintain the ability of the model to follow the text prompt. Another common issue is the inability to generate images under multiple conditions. Some Recent Research [29, 30, 40, 50, 75] proposed some methods to migrate this appearance. However, the underlying reasons for this phenomenon have not been thoroughly investigated. Our research endeavors to explore these reasons and develop solutions to overcome this challenge.

3 METHOD

3.1 PRELIMINARY

Text-to-Image Diffusion Model Stable Diffusion [65] is widely used in image generation task. For any input image, Stable Diffusion first transforms it into a latent representation x using the encoder ε of a variant auto-encoder [37]. For any input image, Stable Diffusion first transforms it into a



(a) Comparison of distances in CLIP text space. (b) Visualization of the cross-attention map activation area.

Figure 3: (a) Each dot represents the position of a phrase combining an adjective and "dog" in the CLIP text-space. After fine-tuning, customized concepts move further away from the the distribution of super-class. (b) Visualization results of cross-attention map activation maps corresponding to the dog token. The bar chart on the right shows the average activation level in the dog area. Experiments show that the activation strengths of the corresponding classes decrease with the increase of the learning rate and the total number of training steps. These demonstrate that the customized concepts likely no longer belong to the super-class, resulting in a loss of super-class semantic information, such as wearing a headphone.

latent representation x using the encoder ε of a variant auto-encoder [37]. The diffusion process then operates on x by incrementally introducing noise, resulting in a fixed-length Markov chain represented as $x_1, x_2, ..., x_T$, where T is the chain's length. Stable Diffusion uses a UNet architecture to learn the reverse of this diffusion process, predicting a denoised version of the latent input x_t at each timestep t from 1 to T. In the context of text-to-image generation, the text prompts' conditioning information y is encoded into an intermediate representation $\tau_\theta(y) = c$, where τ_θ is a pre-trained CLIP [61] text encoder. The primary objective in training this text-to-image diffusion model involves optimizing this transformation and prediction process, and it can be expressed as:

$$\mathcal{L}_{recon} = \mathbb{E}_{x,y,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, \tau_\theta(y))\|_2^2\right] \tag{1}$$

where ϵ and ϵ_θ represent the noise samples from the standard Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ and predicted noise residual, respectively.

Subject-driven Diffusion Model Although text-to-image models have achieved remarkable performance, their controllability is limited. To personalize the generated outputs, DreamBooth[66] fine-tunes the diffusion U-Net to fit several target concept images. Custom Diffusion[38] introduces a new modifier token V* in front of the category name and optimizes only the key and value matrices in the cross-attention layers, thereby improving efficiency.

3.2 Experimental Analysis

We begin by observing simple experimental test cases to realize that the loss of compositional ability after personalization tuning is a common phenomenon. We then analyze the underlying logic through visualizations of the CLIP text-space and cross-attention strength map. Finally, we conduct a theoretical analysis to support our hypothesis.

Simple experimental test cases. As shown in Fig. 1, we observe a loss of compositional ability after fine-tuning. We then conduct additional test cases and find that the headphone concept is not the only one affected; other concepts also experience a similar loss. Furthermore, this situation occurs in both the dog case and other classes.

Semantic drift in CLIP text-space. To elucidate the reasons, we project text into the CLIP text space and use two dimensions for simplified visualization in Fig. 3a. Each dot represents phrases composing the super-class ("dog" in this case) with different adjectives (*e.g.*, "a cute dog", "a white dog" *etc.*). Before fine-tuning, the customized concepts (<sks> dog) have no special meaning and are at a similar distance from the super-class as other words. After fine-tuning, we observe a significant increase in the distance of the target concept from its corresponding super-class. This indicates that

the semantics of the personalized concepts change during fine-tuning. In short, the model increasingly fails to recognize that personalized concepts belong to the dog category. This shift may lead to an inability to access the knowledge associated with the super-class(like wearing a headphone). More details can be found in Appendix D, E.

Reduction of cross-attention activation strength. We further investigate the model by visualizing the cross-attention layers in Fig. 3b. The attention maps indicate the activation area of super-class words in cross-attention layers. It shows that while the "dog" token activates the relevant region in the image, its activation level is notably lower than that of the pre-trained model. Furthermore, its activation level decreases with the increase in epochs and the learning rate. These findings align with our observations in the CLIP text space and provide support for hypothesis that the customized concepts are increasingly not recognized as part of the super-class during fine-tuning.

Next, we theoretically analyze why the semantic drift of personalized phrases results in the weakening of the compositional ability from the respective reduction in the entropy of composable conditional probability.

3.3 Theoretical Analysis

Drawing on the insights of [45], a trained diffusion model can be seen as implicitly defining an Energy-Based Model (EBM) [17]. This perspective allows us to build on prior research in composing EBMs and adapting them for use in diffusion models. Building on the work of [16] in the context of generating images with multiple attributes and Bayes' theorem, the conditional probability can be decomposed as:

$$p(x|c_{class}, c_1, c_2, \ldots, c_i) \propto p(x, c_{class}, c_1, c_2, ..., c_3) = p(c_{class}|x)p(x) \prod_{i \in T} p(c_i|x) \tag{2}$$

$$= p(c_{class}|x)p(x) \prod_{i \in T} \frac{p(c_i)p(x|c_i)}{p(x)} \tag{3}$$

where T is all the set of conditions in prompts except for the class, $p(c_i)$ represents the probability of occurrence of condition c_i in the training dataset and can be regarded as a constant for large-scale pre-training models. $p(c_i|x)$ represents an implicit classifier, denoting the probability of categorizing a concept as c_{class}. Specifically, $p(c_{class}|x)$ represents a specific implicit classifier for the super-category. Thus, we have:

$$p(x|c_{class}, c_1, c_2, \ldots, c_i) \propto p(c_{class}|x)p(x) \prod_{i \in T} \frac{p(c_i)p(x|c_i)}{p(x)} \tag{4}$$

Denoted $p(x) \prod_{i \in T} \frac{p(c_i)p(x|c_i)}{p(x)}$ as $d(x)$, $p(c_{class}|x)$ as $q(x)$, and $p(x|c_1, c_2, \ldots, c_i)$ as $a(x)$. The entropy of a is calculated as:

$$\mathrm{H}(a) = -\sum_x q(x)d(x)[\log(q(x)) + \log(d(x))] \tag{5}$$

After fine-tuning, the components of $d(x)$ change only slightly and can be treated as unchanged, and the implicit classifier $p_\theta(c_{class}|x)$ changes to $p_{\theta'}(c_{class}|x)$, Thus the difference in entropy before and after can be expressed as:

$$\begin{aligned} \triangle \mathrm{H} = &\sum_x q_\theta(x)d(x) \left[\log(q_\theta(x)) + \log(d(x))\right] \\ &- \sum_x q_{\theta'}(x)d(x) \left[\log(q_{\theta'}(x)) + \log(d(x))\right] \end{aligned} \tag{6}$$

$$\begin{aligned} = d(x) \sum_x \{&[q_\theta(x) \log q_\theta(x) - q_{\theta'}(x) \log q_{\theta'}(x)] \\ &+ \log d(x) \left[q_\theta(x) - q_{\theta'}(x)\right]\} \end{aligned} \tag{7}$$

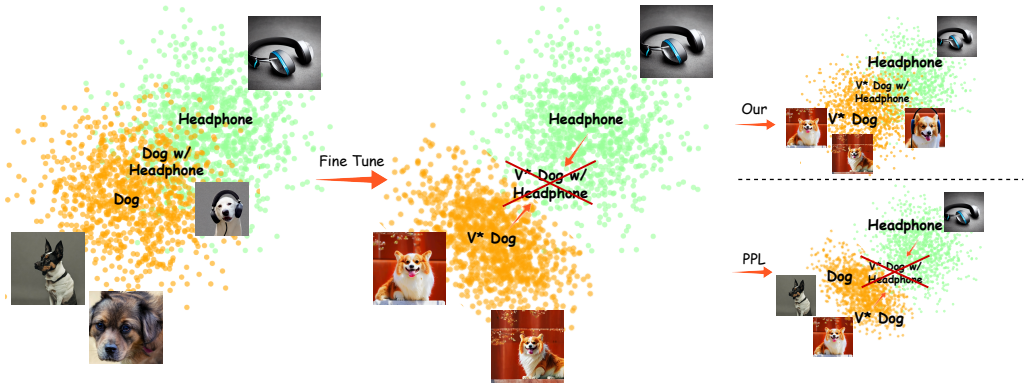


Figure 4: The orange and green point sets represent the distributions of dogs and headphones, respectively, and their overlapping regions represent their joint probability distributions. During the tuning process, the conditional distribution of dogs and headphones shrinks, which gradually increases the difficulty of sampling. Unlike the Prior Preservation Loss (PPL) in DreamBooth [66], which aims to maintain class diversity, our proposed Semantic Preservation Loss (SPL) focuses on recovering the semantic space of the customized concept. This approach enables our method to synthesize images that are more consistent with the text prompt.

Based on our observations in Fig. 3a, 3b we can show that $q_\theta(x) > q_{\theta'}(x)$, combining the properties of probability theory and the monotonically decreasing nature of $x \log x$ at (0,1), we have:

$$q_\theta(x) \log q_\theta(x) - q_{\theta'}(x) \log q_{\theta'}(x) < 0; \log d(x) < 0 \tag{8}$$

Thus, we have:

$$\triangle \mathrm{H}(a) < 0 \tag{9}$$

As a result, it is more difficult to sample from our demanded conditional distributions under $c_{\text{class}}, c_1, ..., c_i$ conditions than before the fine-tuning, leading to the phenomenon that the combining ability is weakened after the fine-tuning. We will discuss the theoretical reasons here in more detail in Appendix B. Fig. 4 illustrates the changes in the distribution during this process that lead to a weakening of the compositional generation capability.

The diminished combinatorial capacity is due to the increased difficulty in sampling from joint conditional probabilities caused by shifts in the semantics of customisation concepts. Therefore, in order to reduce the difficulty of sampling in joint conditional probability distributions, we need to recover the original semantic space of the text, i.e. to recover the semantic distance between custom concepts and superclasses in the semantic space that has been distanced by the fine-tuning process. So, in the next section, we present our proposed semantic preservation loss to mitigate the semantic drift that occurred during fine-tuning.

3.4 Semantic Preservation Loss

As analyzed above, the key challenge lies in preserving semantic information during fine-tuning to reduce the difficulty of sampling from the joint conditional distribution. To address this, we propose a novel loss function aimed at constraining semantic variation throughout the fine-tuning process.

Specifically, considering that there are N special tokens, each representing a customized concept. During the process, the embeddings associated with these tokens are fine-tuned to align with the target concepts. Our loss function is designed to minimize the semantic distance between the phrase containing the special token (*e.g.*, a photo of a V* dog) and the phrase containing only the class word (*e.g.*, a photo of a dog). By labeling the training prompt as P_{tp} (*e.g.*, a photo of a V* dog), and the class prompt as P_{cp} (*e.g.*, a photo of a dog), we use a text encoder to get their embeddings E_{tp} and E_{cp} in Stable Diffusion's semantic space. The semantic preservation loss (SPL) is calculated by the sum of the cosine distance of their text embeddings. Formally speaking, our proposed SPL can be

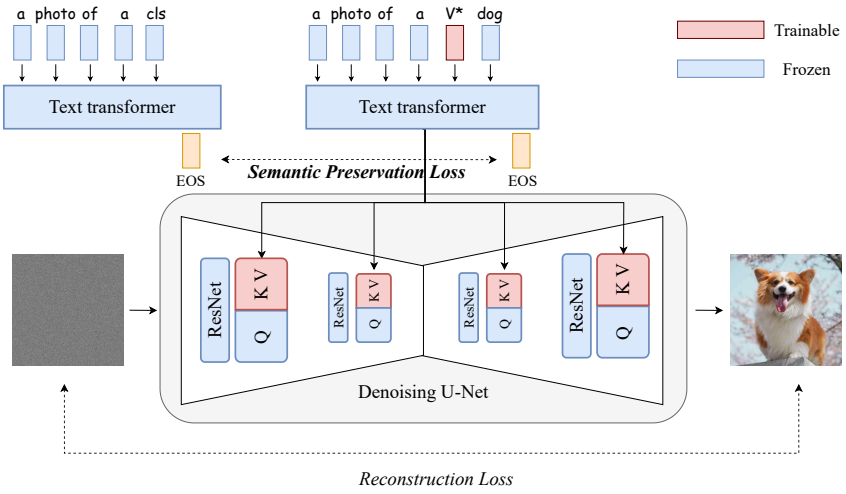


Figure 5: The framework of ClassDiffusion. The personalization fine-tuning strategy is based on Custom Diffusion [39], which primarily fine-tunes the K and V parameters in the transformer block. Our **semantic preservation loss (SPL)** is calculated by measuring the cosine distance between text features extracted from the same text transformer (using EOS tokens as text features following CLIP) for phrases with personalized tokens and phrases with only super-class.

expressed by the following equation:

$$\mathcal{L}_{sp} = \sum^{N}\sum^{B}\sum^{L} D_c\left(E_{SC}, E_C\right) \tag{10}$$

where B represents the batch size, L denotes the hidden dimensions of the text encoder, and D_c implies the cosine distance. We can represent the final training objective as:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda\mathcal{L}_{sp} \tag{11}$$

The overview of our proposed model is shown in Fig. 5

4 Experiments

4.1 Experiment Details

Implementation details Our method is built on Stable Diffusion V1.5, with a learning rate 10^{-6}, and batch size 2 for fine-tuning. We used 500 optimization steps for a single concept and 800 for multiple concepts, respectively. During inference, the guidance scale is set to 6.0 and the inference steps are set to 100. The semantical preservation loss weight is set to 1.0 during all experiments. All experiments are conducted on 2×RTX4090 GPUs. Our method uses $\sim$ 6 min for the generation of single concepts and $\sim$ 11 min for the generation of multiple concepts.

To better preserve the semantic space, we compute SPL between text embeddings embedded in the semantic space of the Stable Diffusion model. Therefore, we utilize the CLIP [61] text encoder from Stable Diffusion v1.5 [63], specifically clip-vit-large-patch14 [47], to extract the text embeddings of phrases. Following common practice, we use the End of Sequence (EOS) token to represent the semantics of embeddings.

Baselines We compare our method with state-of-the-art (SOTA) competitors, including DreamBooth [66], Textual Inversion [20], Custom Diffusion [38], NeTI [1], SVDiff [29]. For DreamBooth, CustomDiffusion, and Textual Inversion, we used the diffusers [77] version of the implementation. For NeTI, we use its official implementation. Given that SVDiff does not have an official open-source repository. For SVDiff, we use the implementation of [69]. All training parameters follow the recommendations of the official paper. To ensure fairness of comparison, all these baselines are built on Stable Diffusion V1.5.

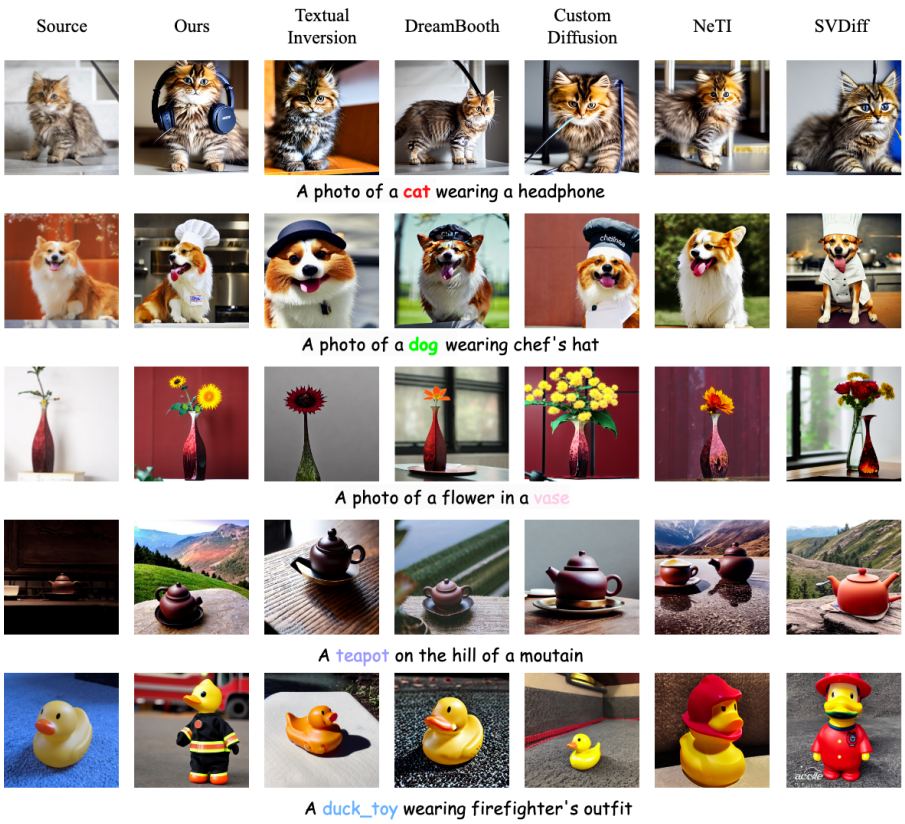


Figure 6: Qualitative comparison between our method and baselines with single given concept. Our method generates images that align with the prompts, surpassing all baselines.

Datasets Following previous work [29, 66, 75], we conduct quantitative experiments on DreamBooth Dataset [66]. It contains 30 objects including both live objects and non-live objects. In addition, we used images from the Textual Inversion Dataset [20] and CustomConcept101 [38] in qualitative experiments.

Evaluation metrics We assess our approach using three metrics: CLIP-I, CLIP-T, and DINO-I. CLIP-I calculates the visual similarity between the produced images and the target concept images by utilizing CLIP [61] visual features. CLIP-T evaluates the similarity between text prompts and images. If one baseline contains the special token S*, it will be replaced with a prior class word. In the case of DINO-I, we evaluate the cosine similarity between the ViT-S/16 DINO [56] embeddings of the generated images and the concept images. Further, we note the impact of CLIP's outdated performance on the fairness of the evaluation. Therefore, we introduce the BLIP2-T Score, which calculates the similarity between text features extracted from BLIP2's Q-former and image features extracted from Vision Encoder as a score. This metric is designed by calculating the similarity between image and text embeddings extracted by the BLIP2 model. Our approach involved utilizing the Transformer [82] implementation and the fine-tuned weights of BLIP2-IMT on CoCo [44], with ViT/L [15]. This new metric aims to offer a more equitable and efficient evaluation measure for future studies in this field. Empirical findings from various studies [25, 40, 41, 70, 83, 86] indicate that BLIP2 outperforms CLIP significantly in the assessment of text-image alignment.

4.2 Qualitative & Quantitative Experiments

Qualitative Experiments We compare our method with DreamBooth [66], Textual Inversion [20], NeTI [1], SVDiff [29], and Custom Diffusion [38] on challenging prompts. The results depicted in Fig. 6 demonstrate the outcomes obtained from these prompts. Fig. 2 shows a story of a given dog and sunglasses. The experimental findings indicate a substantial superiority of our approach over other techniques regarding alignment with text prompts, without any decline in similarity to the specified concept. More qualitative results are shown in Appendix. J. Also, we conduct

Figure 7: Qualitative comparison between our method and Custom Diffusion(CD) in multiple concepts. Our method has better text alignment than custom diffusion.

qualitative experiments with multiple concepts on the combinations <cat>, <sunglasses>; <bear>,<barn>; <dog>,<backpack>; Fig. 7 shows the results of the experiments. The experiments show that our method can be aligned with prompts better than custom diffusion in multi-concept generation.

Quantitative Experiments Following the previous work, we used 20 concepts for quantitative experiments. For Single Concept text similarity metrics (CLIP-T, BLIP2-T), we followed the 25 prompts used in [66], sampling 20 images per prompt. The results of the experiment are shown in Tab. 1. Experimental results show that our method obtains new SOTA on each text similarity metric, indicating that we have good compositional generation capability.

Table 1: Quantitative Results on all Metrics and Results of the User Study. The last two columns display the win rates of our method compared to other approaches in the user study, evaluated in terms of text similarity and image similarity. A win rate exceeding 50% (highlight by ✓) indicates that our method outperforms the compared methods on the corresponding metric, as judged from a human perspective.

	Method	CLIP-T↑	CLIP-I↑	DINO-I↑	BLIP2-T ↑	TIFA↑	User-T Win Rate ↑	User-I Win Rate ↑
Single Concept	DreamBooth [66]	0.249	**0.855**	**0.700**	0.295	0.559	95.4% ✓	42.1%
	Textual Inversion [20]	0.242	0.825	0.631	0.308	0.505	95.1% ✓	75.0% ✓
	Custom Diffusion [38]	0.286	0.837	0.693	0.416	0.746	79.1% ✓	40.0%
	NeTI [1]	0.290	0.838	0.648	0.329	0.607	78.8% ✓	70.0% ✓
	SVDiff [29]	0.293	0.834	0.606	0.418	0.835	56.6% ✓	95.8% ✓
	Our	**0.300**	0.828	0.673	**0.460**	**0.843**	-	-
Multiple Concepts	Custom Diffusion [38]	0.282	0.813	**0.636**	0.380	-	-	
	Our	**0.320**	**0.821**	0.604	**0.477**	-	-	

4.3 Additional Experiments

User Study We also performed user study to validate the effectiveness of our method. We used the same set of images generated in the Section 4.2 for user study, details of which are available in Appendix. I. The results of the user study are located in Tab. 1. The numbers in the table indicate at what percentage our method is considered by humans to be superior to the compared methods. The result of the user study shows that our method outperforms all methods in text similarity (> 50%).

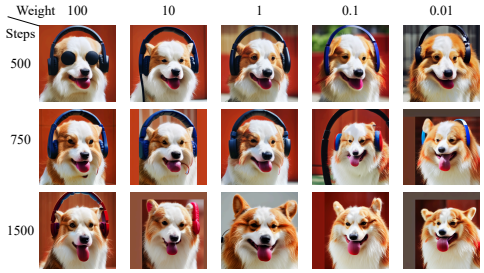


Figure 8: Generation results for the prompt "a photo of a dog wearing a headphone" with different step counts and SPL weights. All results are generated using the same random seed.

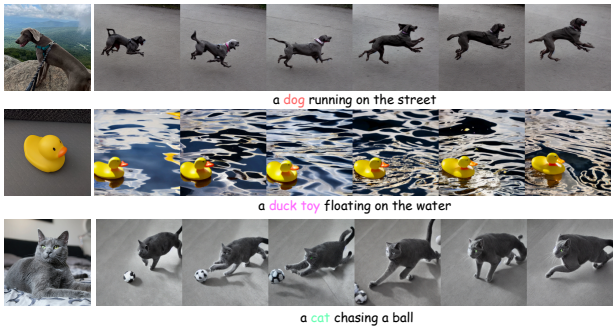


Figure 9: Result of generated videos, showing good textual alignment and similarity of given concepts.

Although our method does not outperform all methods in image similarity, given the high CLIP-I scores, our method still produces images that are highly consistent with the given concept.

Personalized Video Generation We investigate the implementation of our method in personalized video generation. We utilized AnimateDiff V2 [26, 27] for video generation, configuring parameters to a resolution of 512×512, a guidance scale of 7.5, and 25 inference steps. The outcomes of the video generation process are illustrated in Fig. 9. Utilizing AnimateDiff, our technique produces videos that exhibit strong textual and conceptual coherence without the need for additional training. This demonstrates that our approach, which aligns personalized phrases with superclass-centric semantics, can generate engaging videos with dynamic generation capabilities stemming from pre-training, along with the ability to transition across corresponding domains.

Abalation Experiments We studied the influence of different weights of semantic preservation loss (SPL). The results show that higher SPL weights preserve combining ability better. At s SPL weight of 100, all steps successfully depict "wearing a headphone." However, at lower weights of 0.1 and 0.01, the headphone details diminish by 750 steps and disappear by 1500 steps. On the other hand, lower SPL weights restore specific concept features more effectively.

5 Conclusions

In this work, we highlight the problem of weakened compositional ability due to individualized fine-tuning and provide an analysis of the causes of this problem from experimental observations and information-theoretic perspectives. We discovered that this weakening effect is primarily attributed to the semantic shift of the customized concepts throughout the fine-tuning process. As the model undergoes fine-tuning, the representations of these concepts gradually drift away from their original meanings, leading to a misalignment with the intended semantics. This semantic drift complicates the model's ability to accurately sample from the joint conditional distribution, ultimately hindering its performance in generating or understanding the intended outcomes based on the fine-tuned concepts. We then introduce a new approach, termed ClassDiffusion, which mitigates the weakening of compositional ability by restoring the original semantic space. Finally, we present comprehensive experimental results showcasing the efficacy of ClassDiffusion and the fresh perspectives it offers on interconnected fields.

6 ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China under Grant No. 2022YFC3310200, the National Natural Science Foundation of China (No.92470203, U23A20314), Beijing Natural Science Foundation (No. L242022), and the Fundamental Research Funds for the Central Universities (No. 2024XKRC082)

REFERENCES

[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023.

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023.

[3] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[5] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[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] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[8] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*, 2023.

[9] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[10] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023.

[11] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

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

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.

An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020.

[17] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[19] Honghao Fu, Zhiqi Shen, Jing Jih Chin, and Hao Wang. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction. *arXiv preprint arXiv:2312.14871*, 2023.

[20] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] 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 e-prints*, pages arXiv–2302, 2023.

[22] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023.

[23] Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2019.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[25] Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. Tiam – a metric for evaluating alignment in text-to-image generation, 2024.

[26] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023.

[27] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

[28] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[29] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.

[30] Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Williams, George J Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J Zico Kolter. Automated black-box prompt engineering for personalized text-to-image generation. *arXiv preprint arXiv:2403.19103*, 2024.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[32] Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-imagen: Image generation with multi-modal instruction, 2024.

[33] Yihan Hu, Yiheng Lin, Wei Wang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Diffusion for natural image matting. In *European Conference on Computer Vision*, pages 181–199. Springer, 2025.

[34] Miao Hua, Jiawei Liu, Fei Ding, Wei Liu, Jie Wu, and Qian He. Dreamtuner: Single image is enough for subject-driven generation. *arXiv preprint arXiv:2312.13691*, 2023.

[35] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models, 2024.

[36] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023.

[37] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[38] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[39] Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models, 2024.

[40] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[42] Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion, 2024.

[43] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[45] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[46] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023.

[47] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5899–5908, 2023.

[48] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[49] Yuanhuiyi Lyu, Xu Zheng, and Lin Wang. Image anything: Towards reasoning-coherent and training-free multi-modal image generation, 2024.

[50] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023.

[51] 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.

[52] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[53] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[54] Pengyu Ni and Yifeng Zhang. Natural image reconstruction from fmri based on self-supervised representation learning and latent diffusion model. In *Proceedings of the 15th International Conference on Digital Image Processing*, pages 1–9, 2023.

[55] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[57] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

[58] Jae Wan Park, Sang Hyun Park, Jun Young Koh, Junha Lee, and Min Song. Cat: Contrastive adapter training for personalized image generation, 2024.

[59] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[60] Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23085–23096, 2023.

[61] 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.

[62] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[63] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[64] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[65] 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 (CVPR)*, pages 10684–10695, June 2022.

[66] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[67] 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.

[68] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023.

[69] Mingkai Shing. Svdiff: Stochastic video diffusion for conditional video generation. https://github.com/mkshing/svdiff-pytorch, 2023.

[70] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[71] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[72] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[73] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation, 2024.

[74] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[75] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[76] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

[77] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[78] Anton Voronov, Mikhail Khoroshikh, Artem Babenko, and Max Ryabinin. Is this loss informative? faster text-to-image customization by tracking objective dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[80] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

[81] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

[82] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[83] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[84] You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, and Jintao Li. U-vap: User-specified visual appearance personalization via decoupled self augmentation. *arXiv preprint arXiv:2403.20231*, 2024.

[85] Chendong Xiang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. A closer look at parameter-efficient tuning in diffusion models. *arXiv preprint arXiv:2303.18181*, 2023.

[86] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

[87] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.

[88] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023.

[89] Yue Yang, Kaipeng Zhang, Yuying Ge, Wenqi Shao, Zeyue Xue, Yu Qiao, and Ping Luo. Align, adapt and inject: Sound-guided unified image generation. *arXiv preprint arXiv:2306.11504*, 2023.

[90] Yuyang Yin, Dejia Xu, Chuangchuang Tan, Ping Liu, Yao Zhao, and Yunchao Wei. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8145–8156, 2023.

[91] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

[92] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency, 2024.

[93] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

[94] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[95] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[96] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.

[97] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):244:1–244:14, 2023.

[98] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[99] Yiming Zhong, Xiaolin Zhang, Yao Zhao, and Yunchao Wei. Dreamlcm: Towards high quality text-to-3d generation via latent consistency model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1731–1740, 2024.

A VISUALIZATION AFTER FINE TUNNING

In section 3, we present visualization results for the text space and the cross-attention layer of other methods, highlighting the semantic bias that emerges in the text space, leading to a decline in compositional ability. To further affirm the effectiveness of our method and our hypothesis, we also visualize the textual feature space and the cross-attention layer with our method in this section. The visualization results are depicted in Fig. 10. The ranking of the distance is 36 out of 71, as opposed to 26 out of 71 before fine-tuning and 67 out of 71 for other methods. A comparison with the visualizations in Fig. 3b and Fig. 3a reveals that our model effectively addresses the semantic drift in the text space.

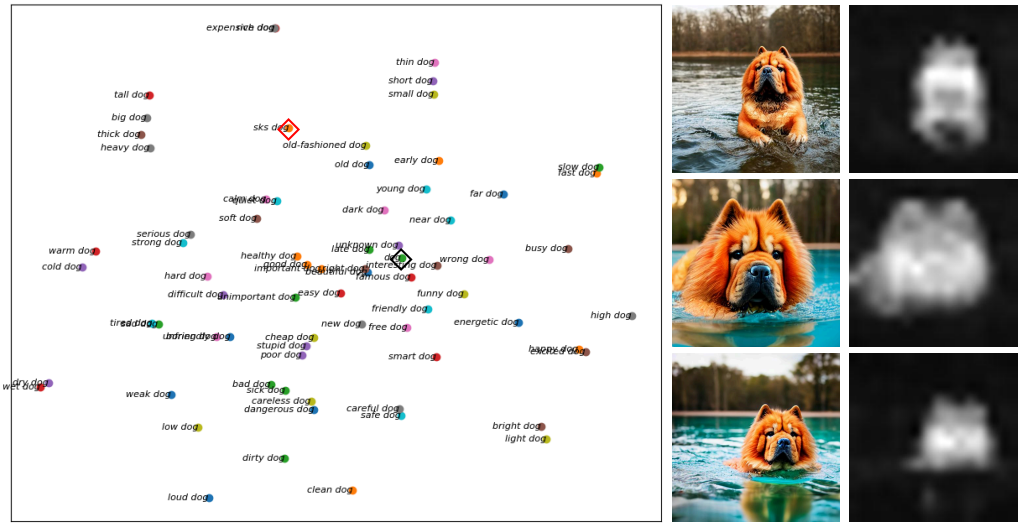


Figure 10: Visualization results after fine-tuning of our approach.

B ENTROPY REDUCTION DURING THE FINE-TUNING

In section 3.3, we provide a solid theoretical analysis for the weakening of compositional ability due to the semantic drift from the perspective of information theory and probability distributions. In this section, we will discuss it in a more detailed way.

In the field of subject-driven personalization generation, two manifest phenomena are caused by overfitting: weakening of diversity in classes of given concepts and weakening of compositional ability. In addition to the calculations mentioned in the main body of the text, the entropy of combined conditional probability can also be calculated as conditional entropy:

$$\mathrm{H}(X|c_1, c_2, \ldots, c_i) \tag{1}$$

Make the given concept into a series of specific conditions: $c_{s1}, \ldots, c_{si}$, each condition describes one of the features of the given concept. The entropy after fine-tuning will be:

$$\mathrm{H}(X|c_1, c_2, \ldots, c_n, c_{s1}, \ldots, c_{si}) = \mathrm{H}(X|c_1 \ldots, c_n) - I(X|c_1 \ldots, c_n; c_{s1}, \ldots, c_{si}) \tag{2}$$

Where I represent the mutual information, According to [52], we have:

$$I(X|c_1 \ldots, c_n; c_{s1}, \ldots, c_{si}) \geq 0 \tag{3}$$

$$\mathrm{H}(X|c_1, c_2, \ldots, c_n, c_{s1}, \ldots, c_{si}) < \mathrm{H}(X|c_1, c_2, \ldots, c_i) \tag{4}$$

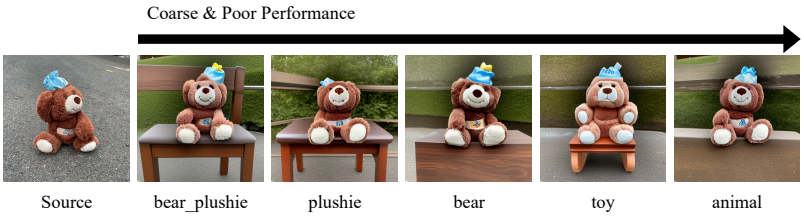


Figure 11: Visualization of generating "a <center word> sitting on the chair.". This <center word> is used in both prompt and class token. Experiments show that a fine-grained center word will benefit our proposed method.

However, the entropy reduction here is different from the entropy reduction in the main text. The entropy reduction here leads to a reduction in the diversity of the generated images. Specifically, when the cue word is "a photo of a dog", the image generated is closer to the given concept than to the diversity of dogs.

C Fine-grained Experiments

At the core of our approach, we want the semantics of personalized phrases to be closer to the category-centered words. In this section, we explore the effect of different category-centered words on the results for the same given concept. Fig. 11 shows the training results using different category-centered words. The results show that the use of different center words leads to significant differences in performance, and a fine-grained center word benefits our method. Also, it indicates the importance of recovering the semantical space of the customized concept.

D Prompt used in the visualization of CLIP text sample space

In this section, we provide a realistic visualization of the schematic in Fig. 3a, and discuss the prompt to generate the 70 phrases that include adjectives and super-categories "dog", and the whole 71 adjectives.

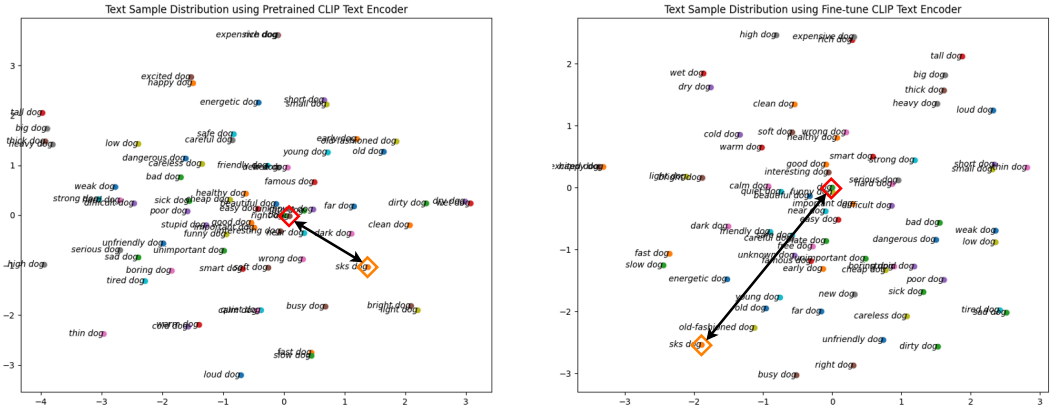
The realistic visualization of the schematic is:

The prompt we use is:

Please help me generate some adjectives that can describe an attribute of a dog in a photo.

The adjectives we use are:

beautiful	*happy*	*sad*	*tall*	*short*
bright	*dark*	*big*	*small*	*young*
old	*fast*	*slow*	*warm*	*cold*
soft	*hard*	*heavy*	*light*	*strong*
weak	*good*	*bad*	*rich*	*poor*
thick	*thin*	*expensive*	*cheap*	*quiet*
loud	*clean*	*dirty*	*smart*	*stupid*
interesting	*boring*	*new*	*old-fashioned*	*safe*
dangerous	*healthy*	*sick*	*easy*	*difficult*
right	*wrong*	*high*	*low*	*near*
far	*early*	*late*	*wet*	*dry*
busy	*free*	*careful*	*careless*	*friendly*
unfriendly	*important*	*unimportant*	*famous*	*unknown*
excited	*calm*	*serious*	*funny*	*tired*
energetic				



Rank:26/71 Rank:67/71

Figure 12: Visualization of the CLIP sample space. Using ChatGPT, we created 70 phrases containing adjectives related to the superclass of dogs. Subsequently, text features derived from these phrases were processed through the CLIP text encoder, downscaled, and their distance from the central point of the superclass (representing a dog image) was calculated. The comparison between the pre-trained model (illustrated in the left figure) and the fine-tuned model (depicted in the right figure) indicates that in the pre-trained model, phrases with special tokens ranked 26 out of 71, while in the fine-tuned model, they ranked 67 out of 71. Moreover, it is evident that phrases containing special tokens are situated further away from the central point of the superclass.

E 2D Text Space Distance Calculation

In this section, we discuss how we visualized CLIP's text sample space in Fig. 3a. First, we collect 70 phrases that combine adjectives with words representing the class of given concepts (e.g., "a happy dog" or "a cool dog"). Using the CLIP text encoder, we extract text embeddings for these phrases. To visualize the semantic space and intuitively track modifications within it, we use t-SNE to reduce these high-dimensional embedding vectors to 2D. An overview of this result are shown in Fig.3(a), meanwhile we show the real experimental results in Fig.12. Formally, we use the adjectives generated which are described in Section. D as the initialize set S, and use the following pseudocode to get a 2D point set T:

Algorithm 1 Algorithm to Convert Character Set to 2D Point Set

Input: Initial character set S
Output: 2D point set T, Distance set Dis
$E \leftarrow$ CLIP text encoder encoding(S) ▷ Encode the character set to an encoding set
$T \leftarrow$ TSNE(E) ▷ Dimensionality reduction of the encoding set to a 2D point set
$Dis \leftarrow \{\|T_i - T_{class}\| \mid i = 1, 2, \ldots, |T|\}$ ▷ Calculate the 2D distance to T_{class} for each point in T
return T, Dis

F Multi-concepts Experiments

In this section, we demonstrate the ability of ClassDiffusion to generate multiple concepts, specifically 3 concepts in one model. Fig 13 shows the result of this experiment. The experiments demonstrate that ClassDiffusion generates high-quality results when combining multiple concepts, validating the effectiveness of our proposed methods.

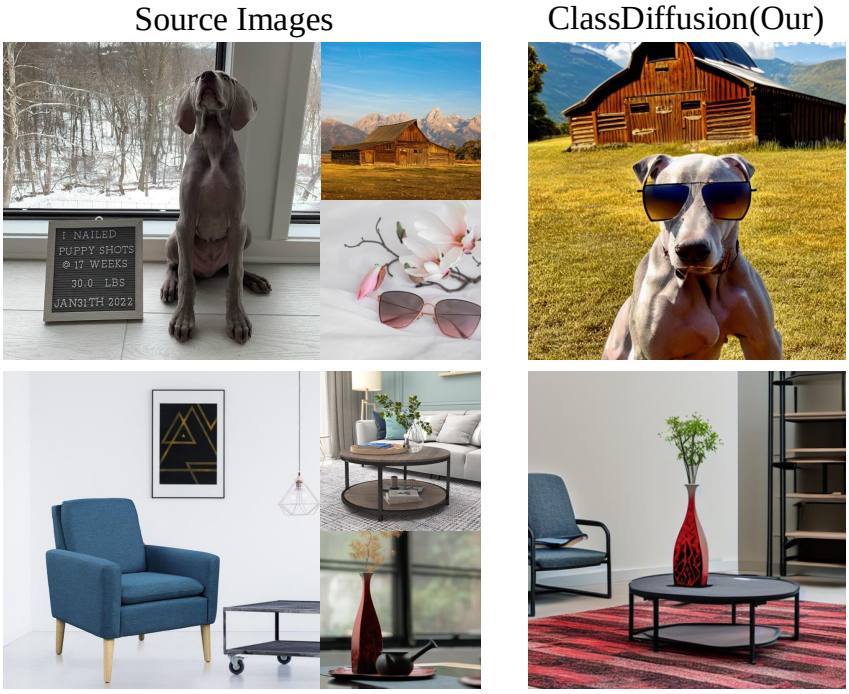


Figure 13: Qualitative result of generating three concepts.

G More baseline comparison

To further evaluate the performance of our proposed method, we further introduce a new baseline Prospect [97] which aims at solving similar problems we observe. The result of the experiment is shown in Tab. 2. The quantitative results below show that our model achieves superior performance.

Models	CLIP-T↑	CLIP-I↑	DINO-I↑
Prospect	0.294	0.815	0.588
ClassDiffusion(Ours)	**0.300**	**0.828**	**0.673**

Table 2: Quantitative results comparing with Prospect

H Quantitative ablation of SPL weight

In this section, we conduct a quantitative ablation experiment on the choose of SPL weight. Tab. 3 shows the result of the experiment result. This result shows that CLIP-T becomes higher (increase in the ability to follow prompts) and DINO-I decreases(decrease in the ability to customize concepts) with increasing SPL weight, which is consistent with our expectations. Meanwhile, we find that SPL is a loss function that is insensitive to weight. Combined with the qualitative observation in Fig. 8, we prefer to choose 1 as the SPL weight.

I Details of User Study

We offer users a comprehensive user study guide that includes user selection criteria which is shown in Fig. 14. Additionally, to maintain fairness, we positioned our method alongside the baseline method randomly to prevent users from showing bias towards either method. Our percentages in Tab. 1 are obtained by calculating the number that chose our model better as a percentage of the overall number that made a preference.

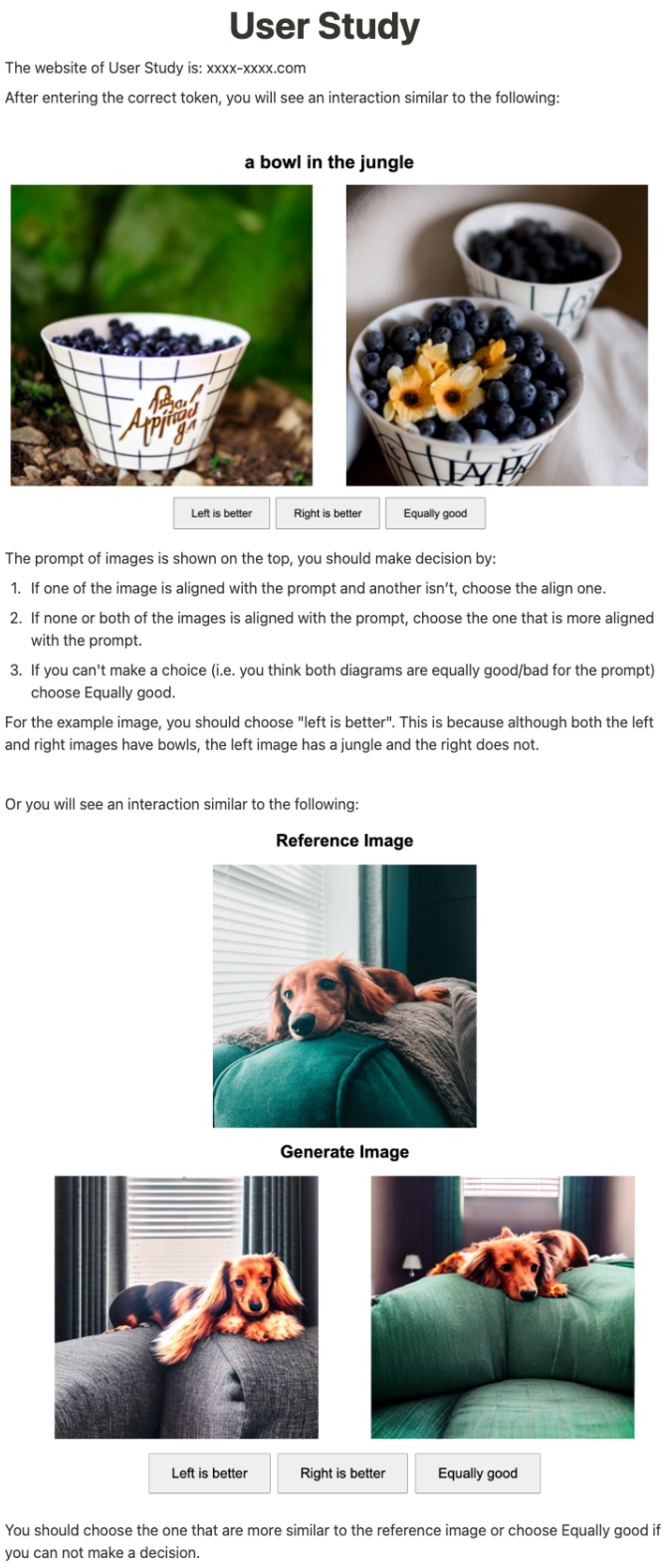


Figure 14: User Study Guide, which describes the user selection criteria and provides an example for reference.

SPL weight	CLIP-T ↑	DINO-I ↑
0.01	0.299	0.677
0.1	0.300	0.674
1	0.300	0.673
10	0.300	0.665
100	0.301	0.661

Table 3: Performance metrics for different SPL weights.

J More Qualitative Result

In Fig. 6, one generated image is provided for each prompt. Fig. 15 presents more images generated from the same prompts, thereby reinforcing the efficacy of our approach.

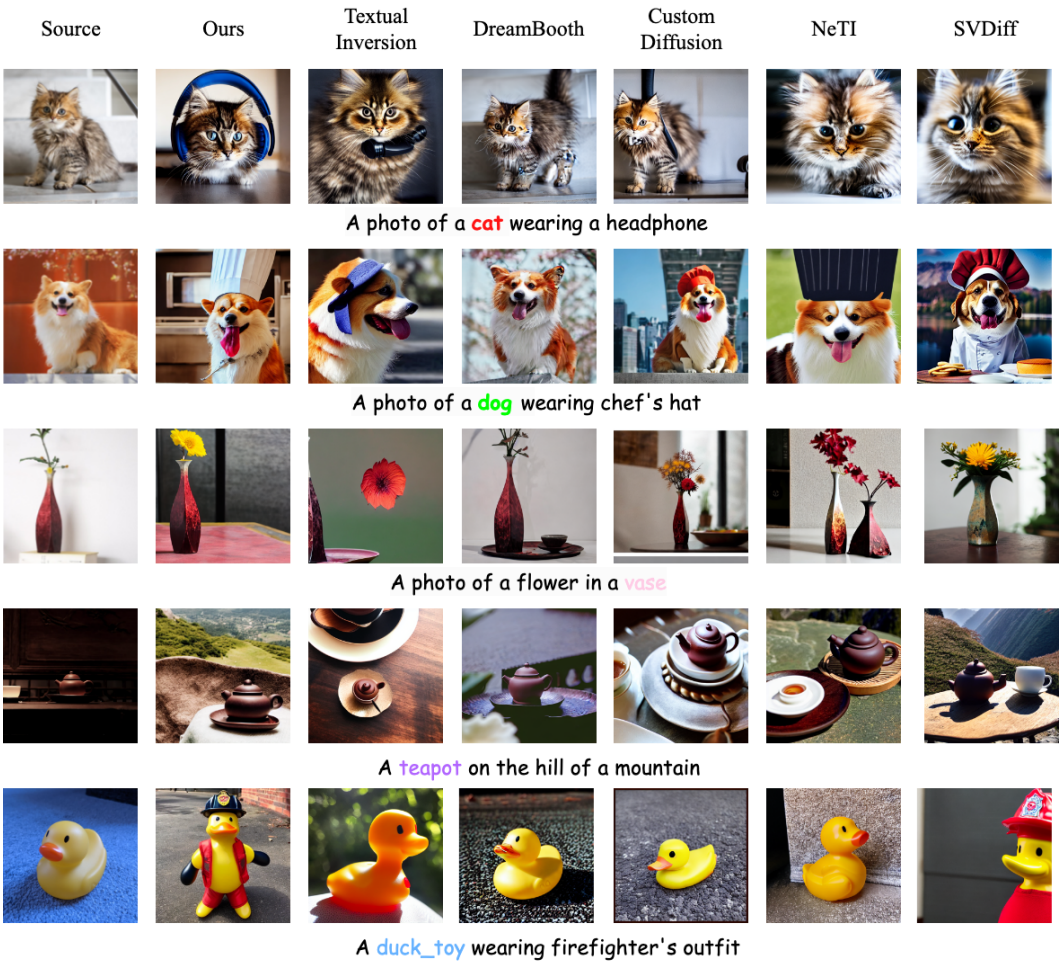


Figure 15: Qualitative result of same prompts the main text.

K Limitation

In this section, we discuss the limitations of our work. Our work are able to generate images that are aligned with the given prompt while keeping the features of the given concept. However, there are two major limitations in our work:

- Considering that reconstruction of the human face is fine-grained, and the phrase "a photo of a human" or "a photo of a human face" can not include extensive information about humans. Whether our work can transfer to human-driven personalized generation remains explored.
- For objects that have a combination of categories, choosing an appropriate center word requires some experimentation.

L SOCIAL IMPACT

The advancements in text-to-image customization through fine-tuning diffusion models, as evidenced by our work on ClassDiffusion, have significant social implications. By enhancing the compositional capabilities of these models, our approach can contribute to a variety of fields, including digital content creation, and education. In the realm of digital content creation, ClassDiffusion enables artists, designers, and marketers to generate more precise and complex images based on textual descriptions. This improvement reduces the time and effort required to produce customized visual content, fostering creativity and innovation. It also allows for the seamless incorporation of personalized elements into digital artworks, advertising materials, and user-generated content, thereby enhancing user engagement and satisfaction. ClassDiffusion can also be a powerful tool in educational settings. Educators can use this technology to create illustrative materials that are tailored to specific learning objectives. For instance, teachers could generate images that accurately depict historical events, scientific concepts, or literary scenes, making learning more interactive and engaging for students. Furthermore, this technology can aid in the development of educational content for diverse learning needs, including materials for students with disabilities.

While the advancements in text-to-image generation hold promise, it is essential to address the ethical considerations associated with their use. Ensuring that these models are free from biases and do not perpetuate harmful stereotypes is crucial. Our work on ClassDiffusion includes measures to mitigate semantic drift, which helps maintain the integrity and accuracy of generated content. Continuous evaluation and updates are necessary to uphold these standards and ensure the technology benefits society as a whole.