Subject-to-Subject: Controllable subject guided text-to-image generation and editing in diffusion models

Saksham Jindal
Department of Electrical and Computer Engineering
University of California, San Diego
sjindal@ucsd.edu

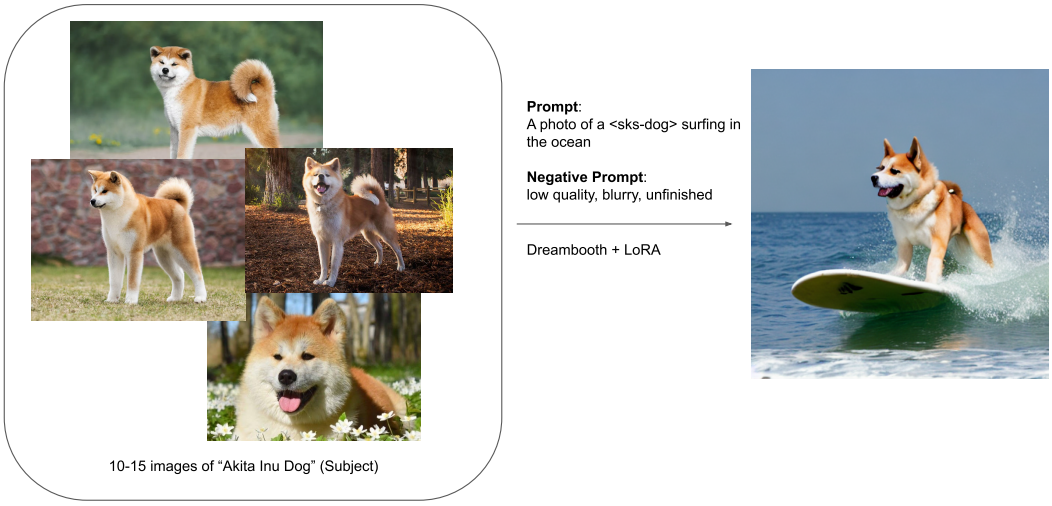

Figure 1: A schematic representation of our personalized text-to-image generation

1 Introduction

Diffusion-based models, or diffusion probabilistic models, are a family of Markov Chain models trained with variational inference. The learning goal is to reverse the process of perturbing the data with noise, i.e. diffusion, for sample generation. Denoising Diffusion Probabilistic Models (DDPMs), first introduced in a landmark paper by Ho et al. in 2020 [13], can generate images by iteratively denoising a noise-corrupted input image using a diffusion process. Since, the there have been text-to-image models such as DALLE2 [9], Imagen [7] and Stable Diffusion [22] which are popular diffusion-based models can generate diverse and high-quality images through text-based prompts.

While text-to-image models demonstrate diversity and generalization capabilities, synthesizing specific concepts from one’s personal life, like images of a pet dog, remains a complex task due to the unseen nature of these personal elements during model training. These models are trained on a vast amount of text data, but they struggle to generate novel ideas or accurately capture specific personal experiences. For instance, one might be interested in generating realistic images of your dog in various settings or scenarios, such as playing at the park or lounging on the couch. However,

the current text-to-image models may not be able to accurately recreate the intricate details and personality of the pet. The challenge lies in training these models on large-scale datasets that lack personal elements, such as the unique appearance and behavior of your pet dog.

On the other hand, there have been several diffusion-based editing methods that have made significant strides in image editing, allowing users to edit input images using simple text prompts while maintaining alignment with the target prompt and faithfulness to the original image. While earlier methods have used DDIM [25], more recent methods have incorporated techniques such as cross-attention guidance [19], plug-and-play features [27], and optimization [15]. We use the methods introduced in [11] to manipulate the attention maps of an edited image by incorporating the attention maps of the original image throughout the diffusion process. By using this approach, we can introduce learned concepts from personalized embeddings into synthetic images solely through edits made to the textual prompt. More importantly, this method allows for image-to-image translation while preserving the original content of the image. To ensure that the general content structure of the image remains unchanged after editing, we utilize cross-attention guidance. This guidance focuses on retaining the cross-attention maps of the input image, which helps guide the translation process and maintain consistency in the overall content structure.

To summarise our contribution, we use DreamBooth [24] to extend the capabilities of diffusion-based models for personalized or subject guided text-to-image generation. We perform a full finetuning and LoRA finetuning [14] of a latent diffusion model and compare them with Textual Inversion [10]. Further, we discuss extension of methods to image-to-image translation using cross attention guidance [11] and no further finetuning to introduce personalized elements into a given image while ensuring that the structure and composition of original image is preserved. The structure of this paper is as follows: the next section provides a walkthrough of existing techniques, followed by a detailed analysis of the proposed methods in the third section. Finally, we detail the experiments conducted and present our results performing a qualitative and quantitative assessment of trained models. The code of the project can be found here: *https://github.com/sakshamjindal/Subject2Subject*

2 Related Work

2.1 Diffusion models for text-based image generation

Before the success of diffusion models, the image and video generation was dominated by Generative Adversarial Networks (GANS) and VAE (VAE). However, these models lack generational capabilities beyond the domain they were trained on. Diffusion-based generative models have gained popularity due to their generalization ability and generate impressive outputs. Denoising Diffusion Probabilistic Models [13] progressively remove noise to create synthesized images. DDPMs were later enhanced by DDIMs [26], which accelerated the sampling process. Latent Diffusion Models [23] introduced multiple conditions in the latent diffusion space, resulting in realistic and high-quality text-to-image synthesis. Recently, there have been popular methods such as DALLE2 [9], Imagen [7] and Stable Diffusion [22]. These advancements in diffusion-based models have opened new avenues for creative generation tasks.

2.2 Subject-driven image generation in diffusion models

Subject-driven text-to-image generation involves generating images of a subject in different contexts based on textual prompts, given a few initial images of the subject. Textual Inversion [10] suggests representing visual concepts using a temporary text embedding and optimizing it to reconstruct subject images. DreamBooth [24] follows a similar approach of incorporating a unique identifier and finetunes all the parameters of the diffusion model and performs better than Textual Inversion in personalized image generation. Custom Diffusion [17] only fine-tune a subset of cross-attention layer parameters significantly reducing the fine-tuning time.

2.3 Text-based image editing in diffusion models

DiffusionCLIP [16] pretrained text-to-image model to convert an input image to a latent space, and then fine-tunes the the diffusion model to guide the target image to align with the text using CLIP [21]. However, a drawback is that it requires finetuning to transfer to a new domain. To avoid finetuning, LD-Edit [6] overcomes the issues of transfer to a new domain. Prompt-to-Prompt [11]

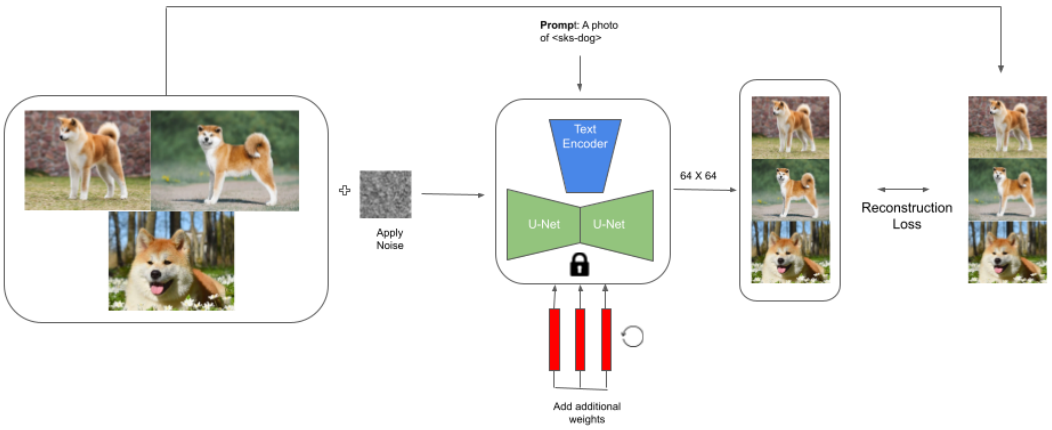

Figure 2: Schematic representation of text-guided image generation

uses cross-attention maps during diffusion to solve the problem that a simple modification of text prompt may lead to different outputs. Pix2pix-zero [20] applies cross-attention guidance to preserve structural information during motion change while allowing flexibility to change under a text prompt. Plug-and-play [28] allows for fine-grained control over the generated structure by injecting spatial features extracted from the guidance image into the generation process of the target image. In contrast to above methods, DiffEdit [8] proposes to automatically generate the mask to indicate which part to be edited and replaces the region of interest with pixels corresponding to the query text. More recently, InstructPix2Pix [5] works on the task of editing the image with human-written human instructions. Based on a large model (GPT-3) and a text-to-image model (Stable diffusion), first generates a dataset for this new task, and trains a conditional diffusion model.

3 Method

3.1 Background

Denoising Diffusion Probablistic Models: Diffusion models are probabilistic generative models that generate an image by progressively removing noise from an initial Gaussian noise image, $x_T \sim N(0, I)$. These models rely on two complementary random processes: the forward process, where Gaussian noise is progressively added to a clean image x_0, and the backward process, where a cleaner image is obtained at each step by applying a neural network $\epsilon_\theta(x_t, t)$ that predicts added noise z. The noise schedule α_t determines the strength of the noise added at each step.

Recently, diffusion models have been extended to text-guided image generation, where a random noise vector z_t and textual condition P are mapped to an output image z_0 that corresponds to the given conditioning prompt. The network $\epsilon_\theta(z_t, t, C)$ is trained to predict artificial noise, where $C = \psi(P)$ is the embedding of the text condition and noise is added to the sampled data z_0 according to timestamp t. The objective is to minimize the mean squared error between the predicted noise and the true noise over all timestamps t and noise samples ϵ.

Denoising Diffusion Implicit Models: The process of inversion involves finding a noise map that can reconstruct the original latent code after it has been noised. In the DDPM method, this involves a stochastic forward noising process, followed by de-noising with a stochastic reverse process, which does not always result in an accurate reconstruction. Instead, the DDIM reverse process is adopted, which is deterministic. This involves adding noise to the latent code at each timestep, and then using a UNet-based denoiser to predict the added noise based on the timestep and encoded text features. The process is gradually repeated until a final noised latent code is obtained, which is assigned as the inverted code.

Latent Diffusion Models: Latent Diffusion Models (LDMs) are a variant of Denoising Diffusion Probabilistic Models (DDPMs) specifically designed for generating images from text. In LDMs, we utilize Stable Diffusion (SD) as the underlying latent diffusion model, which leverages the CLIP text encoder to generate embeddings from text prompts. SD employs a mostly pretrained autoencoder,

consisting of an encoder E and a corresponding decoder D, to extract latent codes from images and reconstruct the original images, respectively. The encoder maps images x from the set I to latent codes $z = E(x)$, and the decoder maps latent codes back to images $\hat{x} = D(E(x))$ such that $\hat{x}$ approximates x. SD adopts a diffusion model in the latent space of the autoencoder, enabling the incorporation of text conditions during the diffusion process. The diffusion process can be formulated as an iterative denoising operation that predicts the noise at each timestep. This is accomplished through the following loss function:

$$L_{SD} = \mathbb{E}_{z\sim E(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\pi(y))\|_2^2\right]$$

where t is the timestep, z_t is the latent code at timestep t, c_π is the text encoder that maps text prompts y into text embeddings, ϵ is the noise sampled from a Gaussian distribution, and ϵ_θ is the denoising network (i.e., U-Net) that predicts the noise. Training SD is flexible, such that we can jointly learn c_π and ϵ_θ or exclusively learn ϵ_θ with a frozen pretrained text encoder.

3.2 Subject guided image generation

We use DreamBooth [24] for personalised text-to-image generation. Give only few images (5 to 10) captured images of a subject in different poses, lightning conditions, resolutions and without textual descriptions or captions, our objective is to generate realistic images of subjects in different scenarios. We achieve this by introducing a new concept into the model's understanding of our subject by using a unique identifier for our subject. DreamBooth recommends using a rare-token identifier that has a very weak understanding in language and diffusion model. Based on few blogs online and experience of people shared online, we label all the given input images with "a photo of <sks> [class noun]" where <sks> is the unique identifier and [class noun] (for instance, dog or cat) is a class description of the the subject.

Class-specific prior preservation Loss: DreamBooth [24] recommends using class-specific prior preservation loss while finetuning the diffusion model. This is because there have been two major problems encounters in lanaguage models: language drift and reduced output diversity.

Language drift refers to the loss of syntactic and semantic knowledge in a language model that is pre-trained on a large text corpus and later fine-tuned for a specific task. The authors note that a similar phenomenon can affect diffusion models, where the model forgets how to generate subjects of the same class as the target subject.

Reduced output diversity is a problem in text-to-image diffusion models. When fine-tuning on a small set of images, it is desirable to generate the subject in novel viewpoints, poses, and articulations. However, there is a risk of reducing the variability in the output poses and views, resulting in a limited range of generated images.

To address these issues, the authors propose a solution called autogenous class-specific prior preservation loss. This method aims to encourage diversity and counter language drift. The idea is to supervise the model with its own generated samples so that it retains the prior (knowledge about the subject class) during few-shot fine-tuning. This allows the model to generate diverse images of the class prior and utilize knowledge about the class prior along with the subject instance.

The proposed loss function includes two terms. The first term measures the reconstruction loss between the generated image and the ground truth image. The second term, the prior-preservation term, supervises the model with its own generated images. The loss function is defined as:

$$L = \mathbb{E}_{x,c,\epsilon,t}\left[w_t\|x - \hat{x}_\theta(\alpha_t x + \sigma_t\epsilon, c) - x\|_2^2 + \lambda_{wt}\|\hat{x}_\theta(\alpha_{t0}x_{pr} + \sigma_{t0}\epsilon_0, c_{pr}) - x_{pr}\|_2^2\right],$$

where $\hat{x}_\theta$ is the generated data using the ancestral sampler on the frozen pre-trained diffusion model, with random initial noise $\epsilon_0 \sim \mathcal{N}(0, I)$ and conditioning vector $c_{pr} := \Gamma\left(\mathbf{f}("a[classnoun]")\right)$. The second term represents the prior-preservation term, which supervises the model with its own generated images. The parameter λ controls the relative weight of this term.

illustrates the model fine-tuning process using the class-generated samples and the prior-preservation loss.

Low-Rank Adaptation (LoRA): LoRA [14] is a finetuning technique which is commonly used to finetune large language models (LLMs). Since finetuning the LLMs to adapt them to particular task

is an expensive process, LoRA freezed the pretrained model weights and injects injects trainable rank decomposition matrices into each layer of the Transformer architecture. Based on recommendations from [4], we apply LoRA to cross-attention layers and use their implementation to train the injected layers while keeping the stable diffusion model frozen. The given technique decompose large cross-attention layer matrices into lower-rank matrices. CloseofSimo [2] observes that training is much faster with much lower computational and memory requirements.

A schematic representation of our approach is illustrated in 2. We benchmark our approach with finetuning the entire stable diffusion model [22] and textual inversion [10].

3.3 Subject guided image editing

Assuming we have an image I, the objective is to edit the given image using the guidance of a target prompt $P*$ to get an edited image $I*$ that retains the content and structure of the original image but corresponds to the target prompt. For instance, if the original image is of and the user wants to replace the horse to the zebra while keeping the appearance and structure of the original image, the user can supply a the text prompt such as "zebra is running". We use cross-attention guidance [11] and text-inversion [18] to perform the text-based edit and avoid the use of mask to guide the model for editing [8] or fine-tuning the model [15]. Using the approach benefits in 3 ways. Firstly, it avoids computational expensive way of fine-tuning the model. Second, this approach preserved the composition and structure of the original image in the edited image. Third, it can also avoids the use of a source prompt and can be conditioned on the input image and target prompt. Since, these properties in the generated image depend on both the random seed and the interaction between the pixels and the text embedding during the diffusion process, it becomes hard to preserve them if we are just conditioning on the edited prompt. The solution is to modify the pixel-to-text interaction that occurs in cross-attention layers to provide image editing capabilities. We would demonstrate that these image editing capabilities are rendered to the diffusion models in a zero-shot way without fine-tuning the diffusion model while preserving the appearance and structure of the conditioned image.

Cross-attention maps In the context of latent diffusion models [9, 22, 7], the models produces text embedding c with the CLIP [21] text encoder. Next, to condition the generation on text, the models computes cross-attention between encoded text and intermediate features of the denoiser θ. The cross-attention maps capture the interaction between the two modalities during the noise prediction, where the embeddings of the visual and textual features are fused using cross-attention layers that produce spatial attention maps for each textual token.

$$Attention(Q, K, V) = M \cdot V, \tag{1}$$

where

$$M = Softmax\left(\frac{QK^T}{\sqrt{d}}\right). \tag{2}$$

Here, $Q = W_Q\phi(x_t)$, $K = W_K c$, and $V = W_V c$ are computed with the learnt projections W_Q, W_K, W_V applied on intermediate spatial features $\phi(x_t)$ of the denoising UNet θ and the text embedding c, and d is the dimension of projected keys and queries. Individual entries of the cross attention map $M_{i,j}$ represent the contribution of the j-th text token towards the i-th spatial location and have a strong understanding of the structure of the image. The cross-attention mask is specific to a timestep, and we get different attention mask M_t for each timestep t. Finally, the cross-attention output is defined to be $\phi^{'}(z_t) = MV$, which is then used to update the spatial features $\phi(z_t)$.

Cross-attention controller: The cross attention map have strong inductive bias of the composition and structure of the image. The ideas is to inject the attention maps M that were obtained from the generation with the original prompt P, into a second generation with the modified prompt P^*. This allows the synthesis of an edited image I^* that is not only manipulated according to the edited prompt, but also preserves the structure of the input image I.

We refer to the algorithm first introduced in prompt-to-prompt [11] for text and image conditioned editing. The goal is to generate an edited image using a source prompt P and a target prompt P^*,

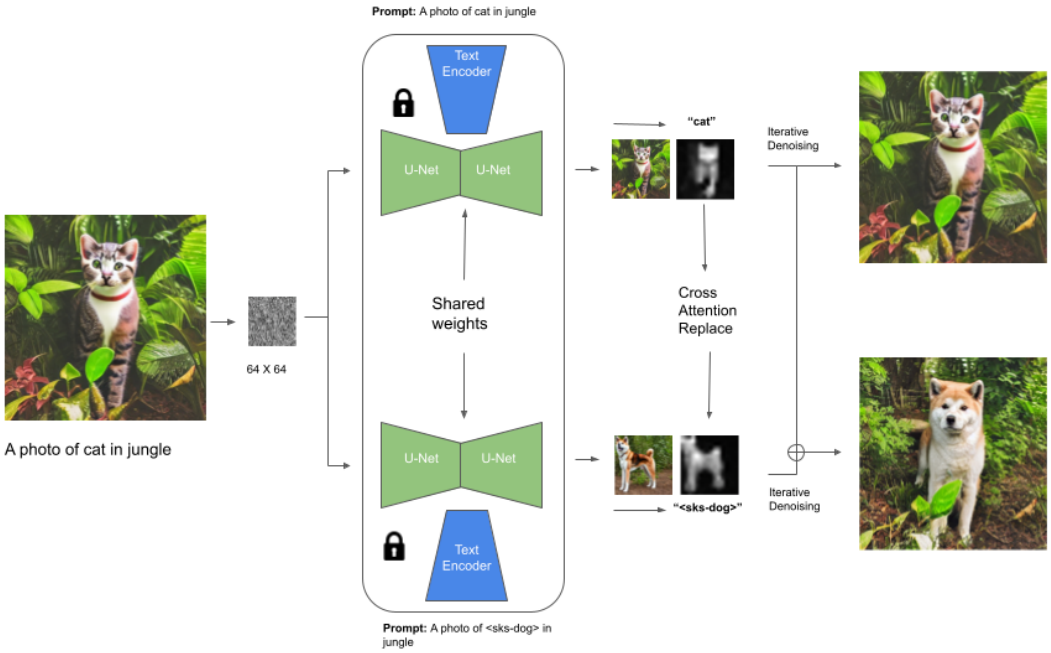

Figure 3: Schematic representation of text-guided image-to-image translation

given a random seed s. The diffusion process is performed simultaneously for both prompts to generate noisy images z_t, and in each iteration, an attention-based manipulation is applied to the images to achieve the desired editing task. In each iteration, it outputs noisy image z_{t-1} and the attention map M_t and also computes the attention map M_t^* for the target prompt P^* and uses the edit function $Edit(M_t, M_t^*t)$ to obtain an attention map $\hat{M}_t$, which represents the desired editing task. The algorithm then performs a diffusion step for the target prompt using z_t and the attention map $\hat{M}_t$, overriding the attention map from the diffusion step for the source prompt. For local editing, the algorithm computes a binary mask α that specifies the editing region using the words w and w^* in the source and target prompts, respectively. The algorithm then applies the mask α to blend the noise from the two diffusion steps and produce the edited image z_t^*. At the end of the diffusion process, the algorithm returns the final images z_0 and z_0^*, which represent the source and edited images, respectively.

A schematic representation of our approach is illustrated in figure 3.

4 Experiment Details

Dataset: We generate a dataset of Shiba Inu dog by collecting 18 images of the dog breed from Upslash [1] and use a subset of dataset posted on HuggingFace dataset [3]. Since the images are crowdsourced, it ensures a diversity in the conditions (like lightning, pose, environments, scales etc) that the images were captured. We also ensure that we collect high quality samples avoid images with motion blur or low resolution.

Regularization Images: We generate 200 images conditioned on the class using the pretrained stable diffusion model using the prompt "a photo of the [class name]".

Training Process and Hyperparameters: Once the images are collected, they are center cropped to a size of 512 x 512 and we train three different models to learn new concepts of our personal objects

- Textual Inversion: Rather than finetuning the whole model, textual inversion "learns to generate specific concepts, like personal objects or artistic styles, by describing them using new "words" in the embedding space of pre-trained text-to-image models.
- Regular Finetuning: We perform a full finetuning of the pretrained Stable Diffusion v1.4.

- Finetuning with LoRA: We finetune only the additional trainable layers injected into U-Net of Stable Diffusion v1.4 while the pretrained model remains frozen.

Hyperparameters For full fine-tuning, we observed that learning rate and training steps has a huge impact on the end results. We learned that if we had to use a lower learning rate, we had to train the model little longer. In our experiments, model trained with 800-1000 and learning rate of $5e^{-6}$ gave results with similar fidelity compared to a model trained with 1300-1500 steps and learning rate of $1e^{-6}$. We didn't train the text encoder owing to huge memory requirements. For finetuning with LoRA, we observe that a much larger learning rate of $1e^{-4}$ is feasible compared to $1e^{-6}$ for regular finetuning. For textual inversion, trained the model for 2000 steps with a leanring rate of $5e^{-4}$. In all our experiments, we have used a batch size of 2.

Evaluation Metrics: We use three different metrics for quantitative evaluation of the diffusion model.

- CLIP Score: CLIP score [12] measures the compatibility of image-caption pairs. Higher CLIP scores imply higher compatibility. The CLIP score is a quantitative measurement of the qualitative concept "compatibility". Image-caption pair compatibility can also be thought of as the semantic similarity between the image and the caption. CLIP score was found to have high correlation with human judgement. For our evaluation, we generate images from a list of 20 prompts carefully curated of our subject in diverse environments, performing different activities etc. Further, we pass a generator with a common seed to each of the pipelines for all the prompts.
  - *a <sks-dog> in the jungle*
  - *a <sks-dog> in the snow*
  - *a <sks-dog> on the beach*
  - *a <sks-dog> on top of a wooden floor*
  - *a <sks-dog> with a mountain in the background*
  - *a <sks-dog> with a city in the background*
  - *a <sks-dog> with a blue house in the background*
  - *a <sks-dog> wearing a red hat*
  - *a <sks-dog> wearing a yellow shirt*
  - *a <sks-dog> sitting near the window*
  - *a <sks-dog> on top of green grass with sunflowers around it*
  - *a <sks-dog> wearing a tuxedo*
  - *a <sks-dog> walking on a red carpet on the beach*
  - *a <sks-dog> on top of the sidewalk in a crowded city*
  - *a <sks-dog> with the Taj Mahal in the background*
  - *a <sks-dog> getting a bath*
  - *a <sks-dog> surfing on a wave*
  - *a <sks-dog> floating in space*
  - *a <sks-dog> jumping in the air*
  - *a <sks-dog> sleeping on a bed*
- Fréchet Inception Distance: FID is a metric that is used to assess the similarity between the distribution of generated images and the distribution of real images. A lower FID score indicates a better match between the distribution of generated and real images.
- Inception Score: IS is a metric that is used to assess the quality and diversity of generated images from a generative model. Usually, a higher IS typically indicates better quality and diversity in generated samples.

However, please note that using metrics such as IS and FID may not be a right thing to do since our latent diffusion model was pre-trained on a large image-captioning dataset (the LAION-5B dataset). This is because underlying these metrics is an InceptionNet (pre-trained on the ImageNet-1k dataset) used for extracting intermediate image features. The pre-training dataset of Stable Diffusion may have limited overlap with the pre-training dataset of InceptionNet, so it may not be a good candidate here for feature extraction.

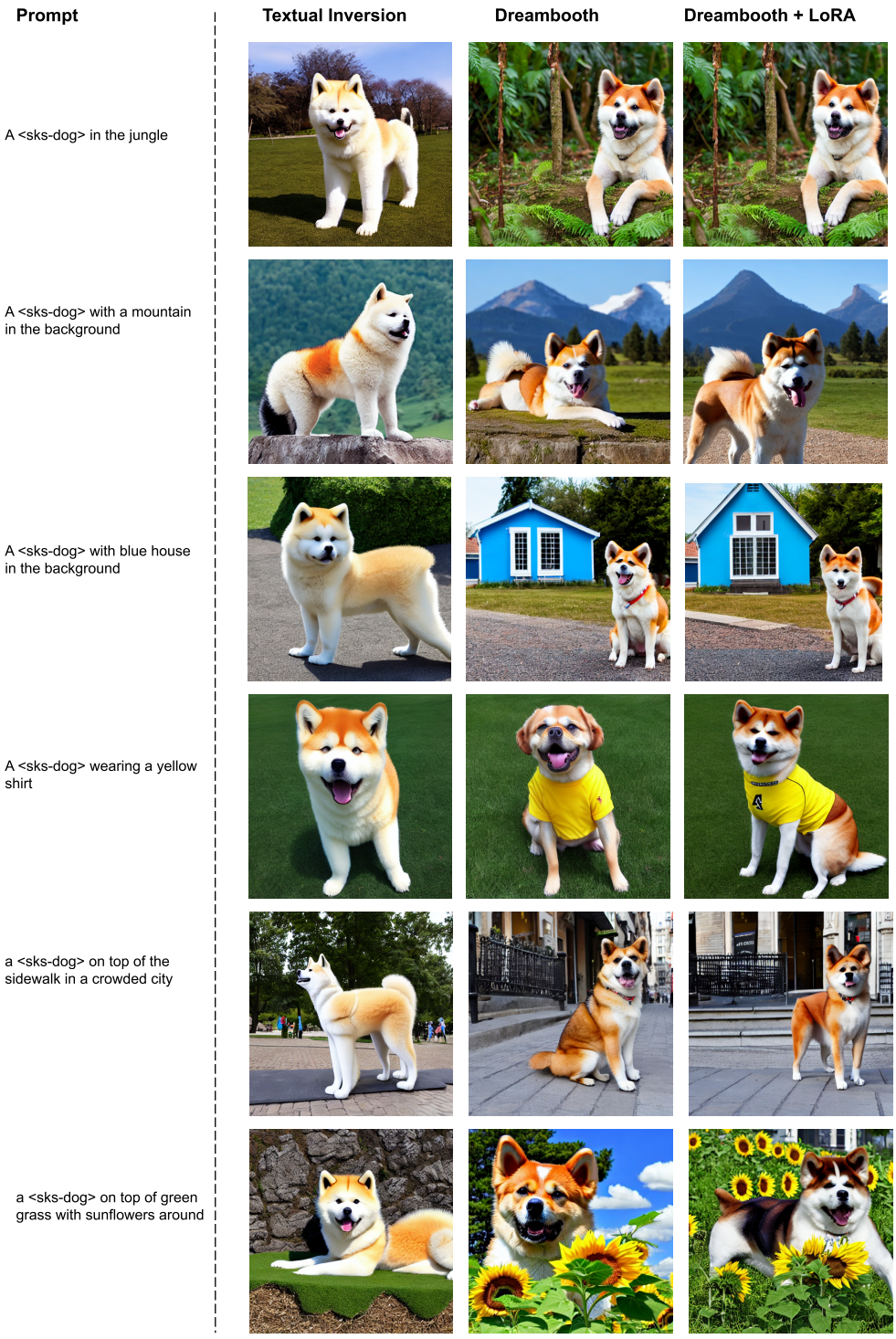

Figure 4: Comparison of generated samples from the trained models given a prompt

5 Results

5.1 Qualitative Assessment

Figure 4 showcases the results of comparing DreamBooth models with Textual Inversion. The images are generated using a selection of prompts and demonstrate the models' ability to learn new concepts and generate images in new scenarios. Upon qualitative inspection, it is observed that both the regularly finetuned and LoRA-finetuned models are capable of generating images that align with the given text. However, they fall short in generating realistic images, which is a known limitation of diffusion models.

Moreover, the usage of LoRA shows promising results, as it allows for training with significantly fewer weights compared to the original model size while still achieving excellent results. On the other hand, textual inversion struggles to effectively learn the given concept and faces challenges in generating coherent and visually appealing images.

5.2 Quantitative Assessment

The evaluation of the diffusion model was carried out using three different metrics: CLIP score, Fréchet Inception Distance (FID), and Inception Score. The CLIP score measures the compatibility or semantic similarity between image-caption pairs, with higher scores indicating better compatibility. For the evaluation, a list of 20 prompts was carefully curated, representing diverse environments and activities for a dog. Images were generated using a common seed for all the prompts across different pipelines.

Among the results, the Textual Inversion approach obtained a CLIP score of 24.123, indicating a relatively lower compatibility between the generated images and the provided captions. This suggests that the generated images may not align well with the intended concepts or activities described in the captions. The FID score for Textual Inversion was 0.103, indicating a reasonably close similarity between the distribution of the generated and real images. However, the Inception score was 1.50, suggesting that the generated images had limited quality and diversity.

On the other hand, Regular finetuning yielded better results. It achieved a higher CLIP score of 30.1439, indicating better compatibility or semantic similarity between the generated images and the captions compared to Textual Inversion. The FID score for Regular finetuning was 0.030, indicating a closer similarity between the distribution of the generated and real images. Additionally, the Inception score was 1.777, suggesting higher quality and diversity in the generated samples compared to Textual Inversion.

LoRA finetuning showed comparable results to Regular finetuning. It obtained a CLIP score of 30.1661, similar to Regular finetuning, indicating a similar level of compatibility between the generated images and the captions. The FID score was 0.051, suggesting a reasonably close similarity between the distribution of the generated and real images. However, the Inception score was slightly lower at 1.592, indicating slightly lower quality and diversity compared to Regular finetuning.

Overall, the evaluation metrics provided insights into the performance of the diffusion model and its variations. Regular finetuning achieved the best results with higher CLIP scores, lower FID scores, and a higher Inception score. This indicated better compatibility, closer similarity to real images, and higher quality and diversity in the generated images. On the other hand, Textual Inversion exhibited the lowest performance in terms of CLIP score, while LoRA finetuning performed slightly lower in terms of Inception score compared to Regular finetuning.

Method	Textual Inversion	Regular finetuning	LoRA finetuning
CLIP score	24.123	30.1439	30.1661
FID score	0.103	0.030	0.051
Inception score	1.50	1.777	1.592

Table 1: Quantitative assessment and comparison of trained models

6 Conclusion

In this work, we propose text-guided image generation by fine-tuning on personal objects using combination of DreamBooth and LoRA techniques and text-guided editing using cross-attention guidance. We demonstrate how diffusion models can be extended to generate personalized and realistic images of specific subjects or concepts with a few examples, even if these have not been seen during the initial pre-training. Our approach utilizes techniques such as DreamBooth and cross-attention guidance to introduce new concepts into the model's understanding and generate coherent images that correspond to natural language prompts.

We find that finetuning the entire model using DreamBooth yields the best performance in generating diverse images that align with the text prompts. The generated images demonstrate the model's ability to learn new concepts and generate them in new scenarios. However, finetuning a large model can be computationally expensive. As an alternative, we apply Low-Rank Adaptation to only finetune certain layers of the model while keeping the majority of parameters frozen. We observe that this approach can achieve comparable results with significantly lower compute requirements.

In the future, one might extend the work to video generation and editing. Applying techniques such as DreamBooth and cross-attention guidance to videos can enable the generation and manipulation of personalized short clips and scenes. One might also use techniques such as temporal cross attention layers and guidance for video generation and editing, that may yield temporally coherent sequence of frames in the video. Overall, our work demonstrates the potential for diffusion models to generate personalized image and videos using text prompts.

References

[1] Beautiful free images & pictures | unsplash. `https://unsplash.com/`.

[2] cloneofsimo/lora: Using low-rank adaptation to quickly fine-tune diffusion models. `https://github.com/cloneofsimo/lora`. (Accessed on 06/11/2023).

[3] Datasets. `https://huggingface.co/docs/datasets/index`. (Accessed on 06/11/2023).

[4] Using lora for efficient stable diffusion fine-tuning. `https://huggingface.co/blog/lora`. (Accessed on 06/11/2023).

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

[6] Paramanand Chandramouli and Kanchana Vaishnavi Gandikota. Ldedit: Towards generalized text guided image manipulation via latent diffusion models, 2022.

[7] Saurabh Saxena† Lala Li† Jay Whang† Emily Denton Seyed Kamyar Seyed Ghasemipour Burcu Karagol Ayan S. Sara Mahdavi Rapha Gontijo Lopes Tim Salimans Jonathan Ho† David Fleet† Mohammad Norouzi* Chitwan Saharia*, William Chan*. Imagen: unprecedented photorealism × deep level of language understanding, 2022.

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022.

[9] Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents, 2022.

[10] 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, 2022.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[15] 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, 2023.

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022.

[17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2022.

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

[19] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.

[20] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[24] 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. 2022.

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

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[27] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022.

[28] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022.