Image Guided Inpainting with Parameter Efficient Learning

Sangbeom Lim[1,2]   Seungryong Kim[1,†]

[1]Korea University
[2]NCSOFT

Abstract

Conditional inpainting is the challenging task of generating images that fill in specific regions of an image while preserving the surrounding details, based on an arbitrary binary mask and a specified condition (e.g., text or image). Existing methods for conditional inpainting often struggle to preserve the appearance of the user's subject in the input images and can be computationally expensive to tune for each new condition. In this paper, we propose a novel approach to conditional inpainting that combines an Image Guided Inpainting model with a Denoising Diffusion Probabilistic Model (DDPM). Our approach trains the DDPM model using a small number of user-provided images, enabling users to insert a subject into any scene even if the poses and views were not present in the tuning data. We also propose a parameter-efficient method for training the DDPM that preserves its core performance while reducing the number of retrained parameters. Our experimental results demonstrate that our proposed approach outperforms existing methods in terms of both reconstruction quality and computational efficiency, making it well-suited for use in low-resource environments. Overall, our approach offers a valuable baseline for future research on guided inpainting and personalization.

1. Introduction

Providing personalized user experiences has emerged as a new marketing strategy in many domains, particularly in the field of interior design. However, customers often end up dissatisfied with their purchases due to the difficulty of harmonizing new products with existing room arrangements. Wouldn't it be great if users could preview the actual deployed results of a new product in their room?

Recent advancements in image generation tasks, such as inpainting, style transfer, and image manipulation, have shown great performance. In particular, inpainting tasks, which involve generating an image in the regions specified by a mask, have rapidly advanced. However, recent studies are now focusing on the controllability of the generated regions rather than simply filling masked areas semantically.

While GANs networks currently hold state-of-the-art performance on most image generation tasks, they have clear drawbacks, such as collapsing and the need for carefully selected hyperparameters, which make them hard to train. Instead, Diffusion models have received much attention as an alternative, offering fixed training objectives and the ability to generate diverse images with high quality.

While there have been some efforts to utilize diffusion models for inpainting tasks, prior studies could not reach the performance of generating what users desired due to the large collection of image-caption pairs required and the limitations of bottleneck architecture. Therefore, we present a new approach for providing user experiences by generating user-specific subjects based on any masked region in an image. Our goal is to train an inpainting model for a user's subject with a few given images of the subject, which can then be used to synthesize the subject in any image semantically. Moreover, the model should be not biased towards the supervision image given by the user such as pose, view, and lighting conditions.

Collecting datasets for conditional inpainting tasks is infeasible, and unlike text-guided inpainting, which has large bindings, image-guided inpainting requires semantically synthesized images of source-target-reference. To address this challenge, we automated data annotation using an object detection model, and trained our model in a self-supervised manner that conducts inpainting tasks in object-detected regions, referring to the detected image.

Furthermore, we observed a significant influence of the shape and size of the masking region on the conditional inpainting task. Instances arose where the model faltered in cases where the masked object class and the user's subject class differed. To mitigate this, we augmented the masking area on the training stage, aiming to reduce the model's

[†] Corresponding Author

sensitivity to the mask.

During the model training stage, retraining the whole model parameters could limit its usage due to the huge DDPM size. Therefore, we studied how to train the model efficiently. Motivated by a paper addressing intrinsic rank in large models, we updated only a small amount of parameters, down-sizing the updating parameters up to 0.0002% in our experiment with minimizing model quality.

To the best of our knowledge, our paper is the first to introduce *image synthesis based on inpainting*, where a user's subject is semantically composed to another image using a mask, as shown in Figure 2. Our work shows highly advanced quality over any other comparable works, both qualitatively and quantitatively. Furthermore, while most diffusion models tend to require huge computational resources, our method is also adaptable to limited resource environments. We are looking forward to seeing our method applied in a variety of fields.
Our Contributions are as follows:

- Proposed new approach on subject-driven inpainting, where we synthesize user's subject to image prior[1] utilizing diffusion model.
- We down-sized amount of trainable parameters on training stage, which means our approach is adaptable even on the low computational resource condition.
- Experimental results compared with similar studies, our model outperform any other model on synthesizing user given subject and has made semantically meaningful results.

2. Related Work

2.1. Image Composition

Synthesizing two images into a single image is a well-established area of research, and many existing studies have focused on harmonizing synthesized images to look more realistic. Several recent models have achieved controllability using a mask image along with the source image, leaving the out-masked area unchanged. For example, [8] and [3] have proposed hierarchical and conditional image harmonization models that can handle the color distribution of each image and achieve high-quality results.

Another related task is image-to-image (img2img) translation, which aims to generate an output image that refers to an input image as a prior. This type of network learns a mapping from the input image to the output image during training and can convert images to different domains while maintaining their structure. However, our task differs in that we focus on generating a subject within a specific scene, with strict control over the generated image.

In addition, semantic image manipulation has been studied extensively, with the advantage of being able to transform user-desired images semantically while maintaining the original structure. Many studies have explored this area due to its potential applications in various fields. [12] showed that unsupervised learning can be used to generate diverse and realistic images, while other studies have used segmentation masks to control the desired label and style. However, these approaches may not be suitable for all domains due to their limitations in terms of model performance.

2.2. Subject-Driven Generation

The user caption-based image generation model[13] is based on the diffusion model of high-resolution images. The detailed description allows us to generate images in close proximity to user intent, and we demonstrate the remarkable performance of generating captions where masked by the ability to inpaint. Thus, Natural language is not limited to domains and has a high degree of freedom, it remains regretful because it has a one-to-many characteristic of words being mapped to multiple objects.

To address this limitation, Recent research has proposed methods for personalized image generation using natural language. Dreambooth[14] and Textual Inversion[5] are two recent studies that generate images based on the degrees of freedom of natural language by mapping learned vocabulary to user-specified subjects. Dreambooth generates images from textual descriptions similar to a GAN-based model, which first maps text to a latent code and then generates an image. Textual Inversion, on the other hand, uses an optimization-based approach to invert the latent representation of a pre-trained image generator based on a given textual description. Both methods allow users to synthesize various images along with natural language descriptions, providing control over generating the subject that the user intended. Thanks to these attempts, users can synthesize images they want in various ways with natural language descriptions. In this paper, We tried here to gain control over image creation and to semantically compose images where the user wants.

3. Preliminaries

3.1. Diffusion model

The diffusion model generates an image with its inverse process from the diffusion process, which sequentially adds Gaussian noise to the original image, and follows a mechanism to generate images from random noise using learned inverse transformation[12]. Forward step, $q(\cdot)$, apply a Gaussian Markov chain sequentially to the original image x, and if the number of times increases enough, x_T follows a complete random Gaussian.

[1] Where user want subject to fitted

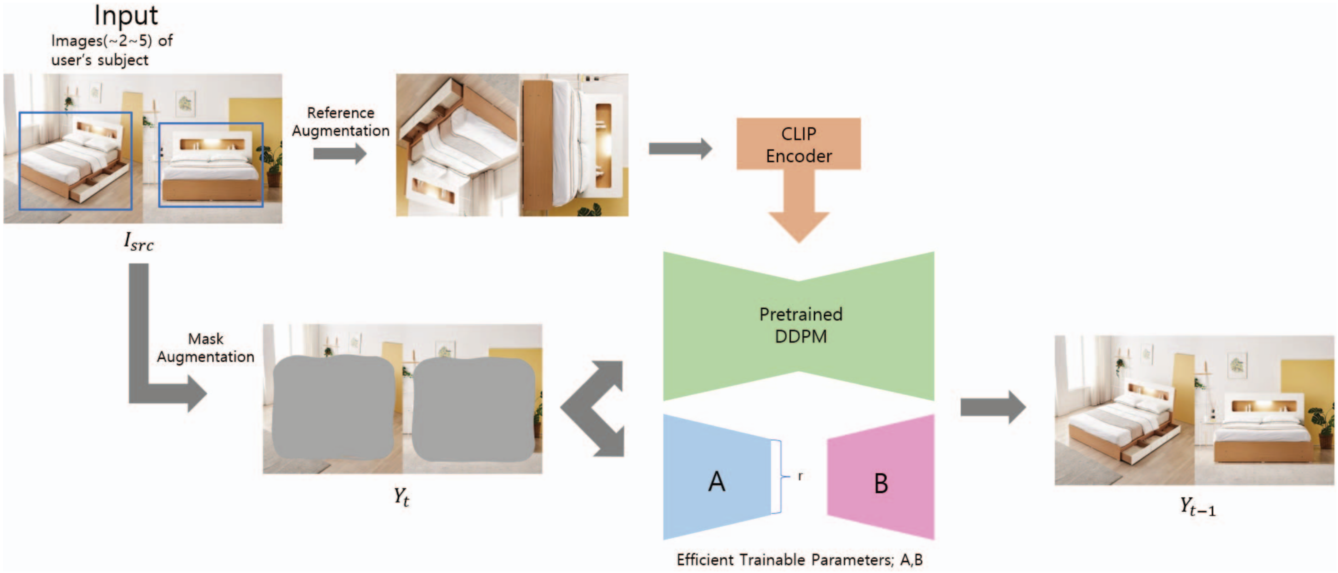

Figure 1: Training Overview

The forward step is defined as

$$q(x_t|x_{t-1}) := N(x_t; \sqrt{1-\beta}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where β_t is pre-defined variance schedule in $\mathbf{T}$ steps. The process of injecting noise can sample x_t at abritrary timestamp *t* is defined as

$$\begin{aligned} x_t &= \sqrt{\alpha_t x_0} + \sqrt{1-\alpha_t \epsilon} \\ \alpha_t &:= \prod_{s=1}^{t}(1-\beta_s), \epsilon \sim N(0, \mathbf{I}). \end{aligned} \tag{2}$$

The Inverse process, $p(\cdot)$, is a process that starts with $x_\mathbf{T}$ following a complete Gaussian and returns it to x_0. Learning this process is the purpose of DDPM, so the inverse process is parameterized by the Gaussian Markov chain parameter $(\mu_\theta, \sum_\theta)$.
Inversion step is defined as

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)) \tag{3}$$

Sampling x_{t-1} can be defined as follows:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t\epsilon. \tag{4}$$

Latent Diffusion model The diffusion model required a lot of computing cost, but this part was relaxed by learning in latent space by Latent Diffusion Model(LDM)[13]. The point of converting representation space, which was the key point in the paper, was that the low-dimension latent space image was much more advantageous in reducing computing cost than the pixel space and that the model's reasoning ability could focus on more semantic points. Pretrained encoder $\boldsymbol{E}$ is used for embedding the image to latent space and reconstruct an image with pretrained decoder $\boldsymbol{D}$, same as auto-encoder. The difference between LDM and DDPM is that LDM uses *z*=***E***(*x*) instead of *x* itself. The sampling process of LDM is as below

$$\begin{aligned} f_\theta(z_t, t) &:= \frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}} \\ z_{t-1} &= \sqrt{\alpha_{t-1}}f_\theta(z_t, t) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t, t). \end{aligned} \tag{5}$$

where $f_\theta(z_t,t)$ is a prediction of z_0 given z_t.

4. Methodology

In this section, we present our methodology for the user subject-driven inpainting task. Firstly, in section 4.1, we describe our training pipeline. To improve application efficiency by reducing the number of diffusion model parameters, we introduce parameter efficient learning in section 4.2. Finally, we discuss various methods we adapted, to enhance model performance such as mask augmentation and reference augmentation in section 4.3.

Premise We begin by defining the source image and reference image as $I_s \in \mathbb{R}^{H\times W\times 3}$ and $I_r \in \mathbb{R}^{H'\times W'\times 3}$, respectively, with *H* and *W* denoting height and width, respectively. Additionally, We use a binary mask $I_m \in 0, 1^{H\times W}$ to guide the inpainting area. The region where *m=0* should remain unchanged compared to I_s, while the same subject in the reference image should be generated semantically in the *m=1* region.

Problem Statement Our objective is to generate a high-resolution synthesized image of the reference image I_r into the merged image $I_s \odot I_m$, when given two $\sim$ four photos of the subject, capturing specific subject's details. Prior works[16, 14, 5] have conducted similar tasks to ours, as mentioned in section 2, but have had weaknesses in reference reconstruction performance and background controllability, respectively. While updating diffusion model parameters[14] showed promising performance in generating reference subject, the inefficiency of training the entire diffusion model[2] has not been resolved.

The challenging points of our task are: First, the model should be able to reconstruct the subject of the user's image with the same performance for perspectives not seen during the learning stage. Second, finetuning the model should be parameter-efficient or of an acceptable size for application. Third, the model needs to interpret the objects being generated appropriately. The textual style and arrangement of the subject should semantically correspond with I_s.

4.1. User's Subject Training

To train our model, we need to annotate pairs of images as $(I_s, I_r, I_m), y$ from user-provided images. We also require the bounding box of the desired subject to generate a binary mask I_m that guides the model to understand the user's subject during the training stage. However, collecting I_s, y, and I_r individually to obtain the user's subject is impossible. Thus, we use a training method that crops the reference image from the same image as the desired subject, i.e., $I_r = I_m \odot I_s$.

We implemented our method using the LDM architecture[13], which is computationally efficient and generates images with various conditions such as text embedding. We applied a conditioning method to the image embedding similar to a recent task[16]. Our model loss function is defined as follows:

$$\mathcal{L} = \mathbb{E}_{z\sim\varepsilon(x),y,\epsilon\sim N(0,1),t}\left[\|\epsilon - \epsilon_\theta(I_s, I_r, t, c)\|_2^2\right] \quad (6)$$

where ε is an encoder that maps images into a latent space, c and t represent the condition and time step, respectively. The objective of Equation 6 is to remove the noise added to the latent image. We leverage condition c and ϵ_θ when producing the masked region ($m = 1$). We used the pre-trained CLIP image encoder[11] as a conditioning model.

Our model should be capable of reconstructing the user's subject on any given scene using a mask image. Additionally, Since our training method is based on self-supervised learning, a limited number of scenes are provided, training sequence should consider the model from overfitting to one photograph and limit the opportunity to learn various scenes. To mitigate this limitation, we adopted a pre-train stage on a large domain dataset to provide the model with strong pretrain knowledge. Strong pretrain knowledge would help the model interpret the user's subject and develop the ability to adapt the subject to various scenes semantically. Explicitly, we employed Stable Diffusion[13] for our prior knowledge, which is trained with the LAION dataset[15]. Stable Diffusion would help our model generate high-resolution images and optimize in the initialization stage.

4.2. Efficient training

State-of-the-art models for natural language and computer vision tasks are typically pre-trained on large datasets. BERT[4] and Stable Diffusion are two popular pre-trained models for NLP and CV, respectively. However, retraining a large model for a specific task can be challenging, as the more tasks a model is trained on, the more parameters it requires for each task, proportionally.

While the Diffusion model is effective, it has a large inference time and model size compared to GANs. This can be problematic for our task, and retraining models with independent subjects can be difficult to store as the model and subject need to be matched one-to-one. For this reason, we need to ensure that the model size can be stored and used in a database with only a small capacity. The model should also have comparable storage space with a singular image or else. Large models can be updated with the intrinsic rank when retrained[7]. If it is possible to update the entire parameter without updating it with only a small amount of parameters, it must be applicable even on the industrial service.

In updating large models, dense layers that perform matrix multiplication typically have a full-rank. However, previous research has found that these layers have a low "intrinsic dimension" and can still learn efficiently even with random projections into smaller subspaces. Low-Rank-Parameterized Update Metrices (LORA)[7] have shown similar or even better performance on fine-tuning tasks by only updating weights that are low-rank decomposed. Inspired by recent findings on NLP fine-tuning tasks, we adapted an efficient fine-tuning method as follows:

$$h = W_0x + \Delta Wx = W_0x + BAx \quad (7)$$

where the pre-trained weight matrix is $W_0\epsilon\mathbb{R}^{d\times k}$, $B\epsilon\mathbb{R}^{d\times r}$, $A\epsilon\mathbb{R}^{r\times k}$, and the rank $r \ll min(d,k)$. For the experiment details, we used the same initialization methods as the LORA paper[7].

[2] Storing the whole parameters of the diffusion model takes up about 13GB in our condition.

4.3. Additional Applied Methods

One of the challenging aspects of our task is that the model needs to be able to generate subjects under untrained conditions (such as different viewpoints, lighting, and pose). To achieve this, the model needs to have a full understanding of the subject, including its position, shape, and mask image. Augmenting the mask image and the subject image can help the model reconstruct the subject even under unsupervised conditions.

Reference Augmentation To help the model learn the variety of views of the subject, we augmented ground truth image, I_s. We adopt several image augmentation methods, such as flip, rotation, and blur methods. These augmentation steps would drive the model to generate the subject whether the given generation condition is not seen in the training step.

Mask Augmentation Our mask region is based on the bounding box of the subject, and during the training step, the model is forced to reconstruct the subject to fit as closely as possible within the mask region. However, this approach may not be practical in the real world, as the user may give a mask size that is semantically larger than the actual subject.

5. Experiments

5.1. Experimental Settings

Datasets We additionally conducted pretraining on the LSUN-Bedroom dataset over Stable Diffusion. The LSUN-Bedroom dataset contains a large number of images of various bedroom states. To ensure efficiency, we sampled 300,000 images from this dataset.

For finetuning, We collected interior objects, such as beds, and couches, on commerce[3] website especially containing more than two images of the object. Therefore, we could collect 12 objects of interior objects and annotate them with facebook/detr-resnet-50[1], an object detection model to automate building mask image.

We evaluated our model with 20 background scenes which are randomly collected on a commerce website, in other words, synthesize 20 backgrounds with 12 objects of interior objects.

Implementation Details We trained our model using the AdamW optimizer with a learning rate of 3e-5 for training the user's subject and a learning rate of 1e-6 for training lsun-bedroom data for pretrain model. We used a batch size of 16 and trained the models for 40 epochs. To prevent overfitting, we applied dropout with a rate of 0.2 during training. We used NVIDIA A100.

Method	Training Time ↓	Storage Amount ↓
Single Photo	NONE	12MB∼48MB[4]
Textual Inversion	8:00	3KB
Dreambooth	25:00	7.2GB[5]
Ours (+LORA)	17:30	**2MB∼40MB**

Table 1: efficiency comparison table

Method	DINO Score ↑	CLIP Score ↑	FID ↓
Baseline	66.8	71.4	170.3
Ours(Whole Params Update)	**72.7**	**74.7**	**155.4**
Ours (+LORA)	71.9	71.9	160.6

Table 2: Performance Comparison

Evaluation Metrics To evaluate the performance of our model, we used the following evaluation metrics:

- DINO[14]: Average pairwise cosine similarity between the ViT-S/16 DINO embeddings[2] of generated region and reference images.

- CLIP[11]: We used a pretrained CLIP model to compare the similarity between the reference image and the generated region. The higher the score, the better the model's performance in terms of generating images that are semantically similar to the user's intended subject.

- FID[6]: Measure to compute similarity between the distribution of real images and generated images. FID score is combination of Inception Score and Fréchet Distance.

- We carefully conducted Qualitative Evaluations of our experiment results.

5.2. Quantitative Evaluation

Table 2 shows a point that our methods outperform the prior approach on image editing, inpainting reference image to source image. Updating whole model parameters with few reference images made the best performance on DINO, CLIP, FID scores.

Based on table 1, Applying the LORA method made our model highly efficient, thus taking a few performance loss. It is feasible that a single photo takes 12MB∼48MB Storage, however, our method takes less than 40MB storage for each subject with the ability to inpaint the user's subject on any photo.

[3] https://shopping.naver.com/home

[4] https://support.apple.com/en-us/HT211965

[5] https://huggingface.co/CompVis/stable-diffusion-v-1-4-original/resolve/main/sd-v1-4-full-ema.ckpt

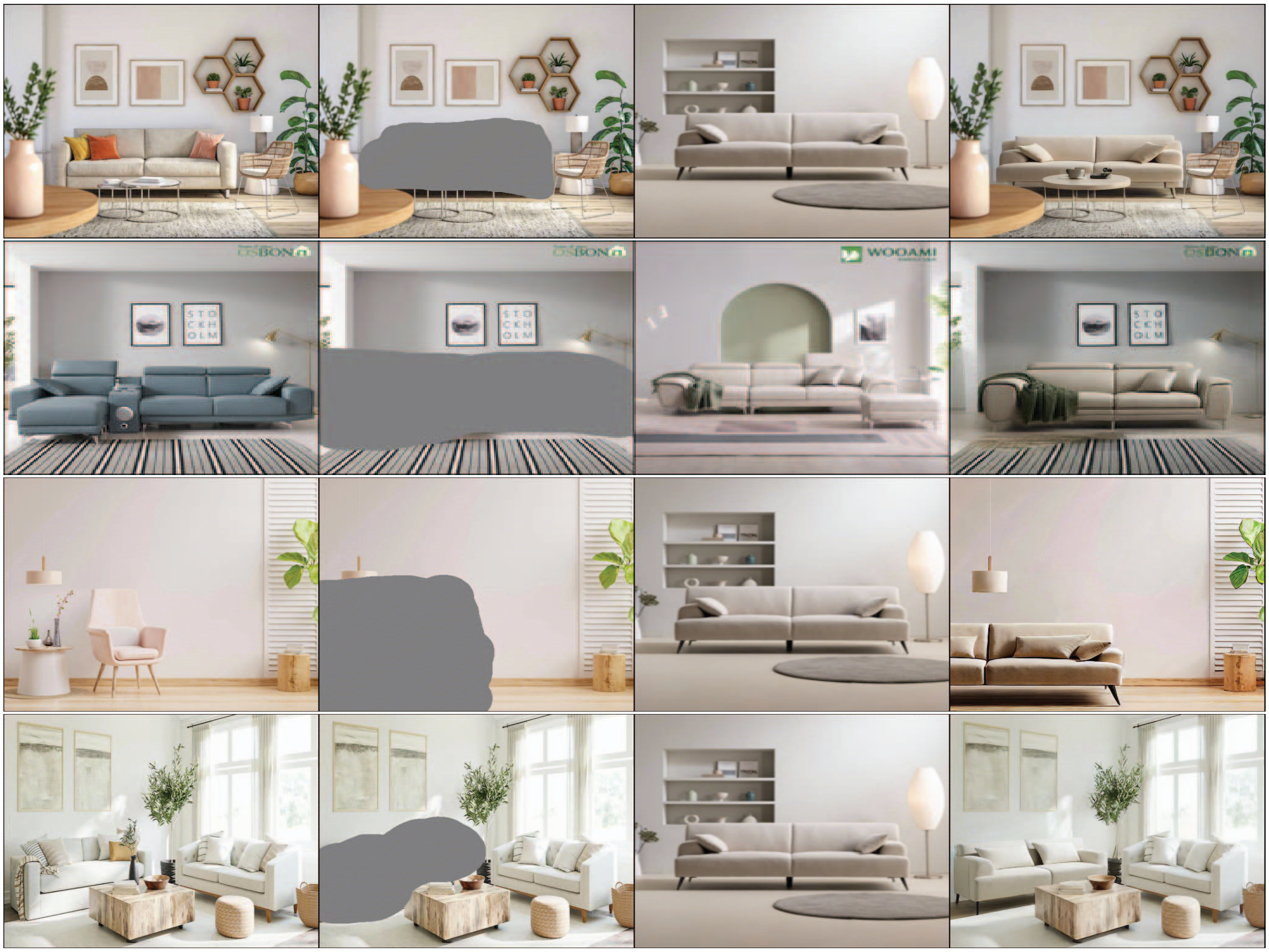

Figure 2: **Example of fine-tuning method pretrained with LSUN-Bedroom Dataset.** Each column represent original image, original image with mask region, reference image, generated inpaint image using our methods, respectively.

The Quantitative results support our claim that our method generates high quality images as well as reconstructing the subject by incorporating with the background image.

5.3. Qualitative Evaluation

Figure 2 demonstrates the qualitative results of our proposed method compared to Paint-by-Example and Stable diffusion inpainting models for different subjects. For Stable Diffusion setting, we used the img-to-prompt model[9] to convert the reference image to a prompt. As shown in the figure, our proposed method successfully reconstructs the user's intended subject in various conditions, with the ability to generate views, poses, and lighting conditions that were not provided in the training stage.

In addition, Figure 3 shows generated image trained with Dreambooth, which has limitations on the remainder of the background. Our proposed method has the advantage of generating highly personalized images of the user's intended subject on any source image, leaving the background untouched. This provides greater controllability and specificity over the generated results compared to other text-guided generation models.

It is important to note that our proposed method's strengths lie in its ability to generate personalized images of the user's intended subject while maintaining the fidelity of the original image. In addition, our model demonstrates flexibility and robustness in handling a wide range of input images and conditions.

5.4. Ablation Study

We performed an ablation study to validate the effectiveness of our proposed method. Our ablation study consisted of four steps: Stable diffusion prior, augmentation,

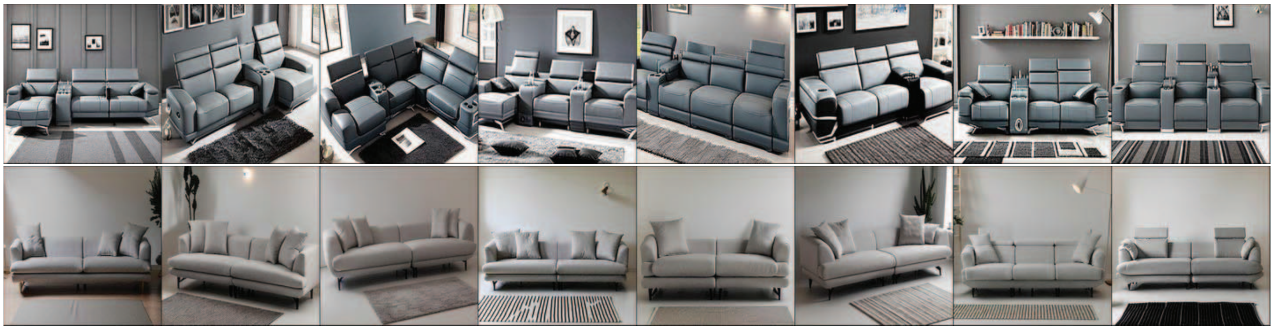

Figure 3: Dreambooth Finetune Examples with prompt “A photo of [V] sofa”

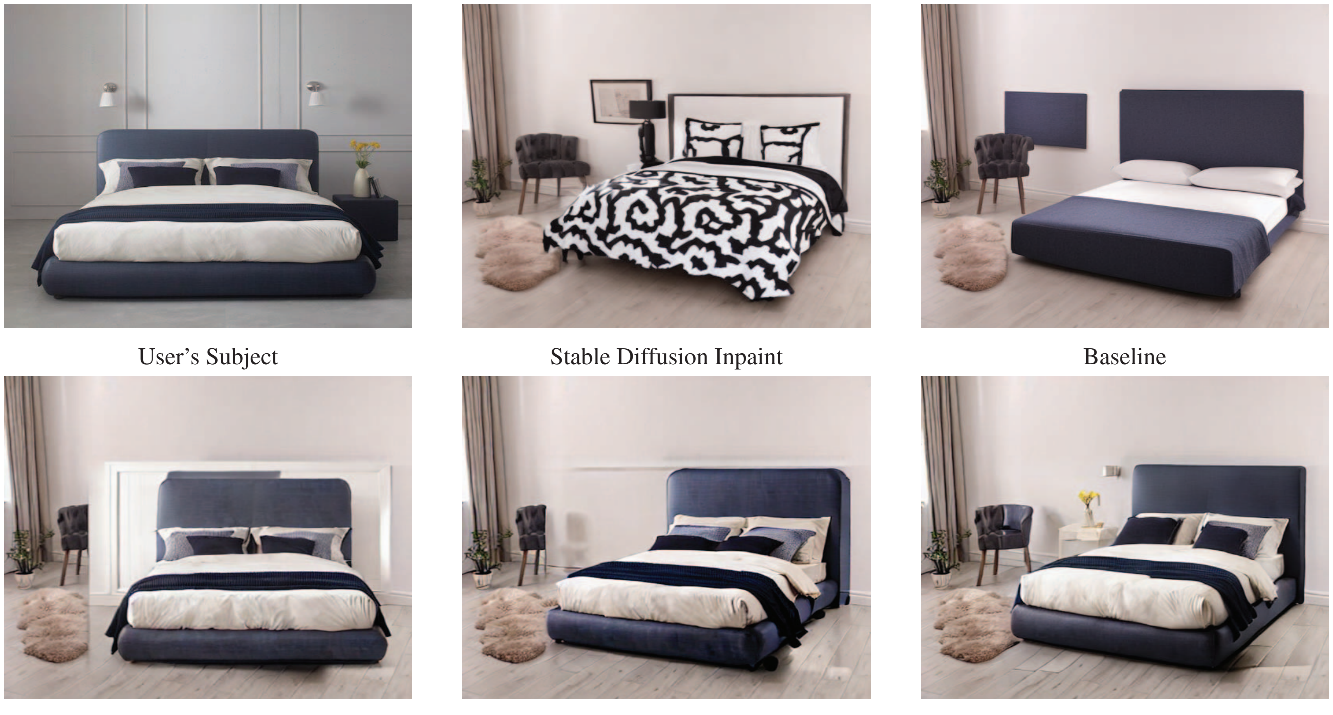

Figure 4: Examples on Ablation Study

Method	DINO Score ↑	CLIP Score ↑	FID ↓
Baseline	71.4	66.8	170.3
+ LORA	70.9	71.3	164.8
+ Reference Augmentation	**71.6**	**71.5**	165.8
+ Mask Augmentation	71.1	71.3	**161.3**

Table 3: Ablation Study

and LORA.

First, we compared the Stable diffusion inpainting model as our baseline architecture with a modified conditioning method for text-to-image generation. Specifically, we converted User’s Subject to prompt using image captioning model[10]. Second, we leveraged the Stable Diffusion Pretrained Prior as an initialization for our model. Third, we implemented the LORA method to ensure that our method can be applied to any condition, even in a low computational environment. Lastly, we adapted our augmentation methods on previous methods, sequentially.

Figure 3 shows the gradual performance improvement achieved by applying our proposed methods in the order of Stable diffusion prior augmentation, and LORA. This figure demonstrates that our proposed method is effective and that each proposed method contributes to the overall improvement in performance.

Additionally, We Tried tuning model without Stable diffusion Prior, the model showed poor quality of generation results and nonsemantic results. Utilizing large dataset knowledge of stable diffusion brought model generate high quality of images, however user's subject seem less relevant reason of unnaturality. Our method seem to generate well on similar condition of supervised setting, some results showed bias results when non-supervised settings. Implementing Augmentation method helped model from bias problems and also from trivial problem caused from filling entire region. Finally, Our efficient training method showed remarkable performance while on reduced trainable parameter setting.

6. Limitation

While our proposed model has demonstrated impressive performance in generating user-specific images, there are still some limitations to our approach. One of the primary limitations is that our model currently requires the training of a one-to-one matching of a single model and a subject. To overcome the capacity limitation of the model, it would be ideal to have a model that can generate subject images in a zero-shot setting, without requiring additional explicit supervision. Although this presents a significant challenge, it could greatly enhance the applicability of our approach, especially in scenarios where it may be impractical to train a separate model for each user. We leave the development of such a model as a potential avenue for future research.

7. Conclusion

In this paper, we presented a novel approach for synthesizing any given user's subject into an image, given a corresponding mask condition, by training the model with only a few images of the subject. Our key idea is to use self-supervised learning to train a conditional inpainting model that can generate high-quality images. To overcome the limitations of the baseline model, which produced biased and non-semantic generated results, we incorporated Stable Diffusion Prior and augmentation techniques into the training process. We conducted a careful evaluation of our model on a variety of measures and demonstrated its remarkable performance both quantitatively and qualitatively. Furthermore, we recognized the heavy size of the DDPM model, which is inapplicable in low-resource environments. Therefore, we applicated an efficient training method that allows for the model to be trained in any condition with parameters mapped to a subject, which can be reused at any time. Our proposed task opens up new opportunities for image synthesis by enabling users to generate high-quality images without requiring an extensive dataset. We believe that our model can serve as a baseline for future research in this field and look forward to further advances in this area.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[3] Haoxing Chen, Zhangxuan Gu, Yaohui Li, Jun Lan, Changhua Meng, Weiqiang Wang, and Huaxiong Li. Hierarchical dynamic image harmonization. *arXiv preprint arXiv:2211.08639*, 2022.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[7] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[8] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 690–706. Springer, 2022.

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[10] NLP Connect. vit-gpt2-image-captioning (revision 0e334c7), 2022.

[11] 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.

[12] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

[14] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[15] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[16] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022.