Adding Stochastic Features in DDIM Inversion based Image Editing

Wonhark Park Junoh Park Donghwa Kim
Seoul National University
Seoul, Korea
{pwh0515, wnsdh0418, ehd9712}@snu.ac.kr

Abstract

To control images oriented towards user's preference with simple text prompts, various editing methods have been proposed in diffusion-based text-to-image models. Since editing is based on the source image, unedited parts need to be unaffected by the editing prompt. This forces the use of deterministic DDIM inversion process in diffusion models to promise reconstruction performance. Despite their fine performance, stochasticity of diffusion model is totally lost which is one of the largest benefit generative models have.

In this paper, we suggest attachable text generation model for conditioning prompt that adds stochasticity to text-to-image diffusion models. We add variance in 1) null-text embedding or 2) conditional editing text embedding used in classifier-free guidance that gives diversity in generated image. Our methods enable local editing with variation in image features presented through diversity obtained in the text domain.

1. Introduction

Diffusion-based text-to-image models [29, 31, 33, 40] have shown tremendous performance on generating images with semantic meanings described in text features. This property of large-scale language-image(LLI) models made it possible to edit synthesis or real images by adjusting only text prompts [4, 10, 24]. For editing synthesis images through an inverse diffusion process, one can rely on randomness of diffusion latent space expressed in DDPM-based sampling [11] strategies to generate various edited images.

However, such stochasticity not only changes the whole structure of the original image but also changes key features from the original image which are not intended to be edited. Besides, when editing is done on real image, deterministic diffusion inversion and sampling processes must be promised to ensure reconstruction capability such that editing can be done while keeping the unedited features. These deterministic processes removes the stochasticity described above, leaving no variance in generating images.

Although a DDIM inversion [6, 34] has reconstruction ability based on the deterministic diffusion process, it is found lacking when guidance with text prompts is used during the sampling process [12]. In the text-guided diffusion models, guidance is essential to fully express the text prompt in the sampled image. For this reason, a pivotal inversion method [30] was used in diffusion model to fit the latent codes obtained in the classifier-free guidance-based sampling process to the pivots, which correspond to the latent codes obtained from the DDIM inversion process [24], on each time-step by optimizing the null-text embedding.

However, the stochastic feature of the diffusion model is lost by using the deterministic forward and reverse processes. That is, only a single edited image is obtained in a given pair of original and editing prompts. Since the natural language has implications in that a single word can represent an infinite possibility of images, varying images should be generated given an editing prompt. This way, users may choose the favored images while taking away the ones with low-quality or with no preference.

In this paper, we propose two methods of adding stochastic features in text embedding. First, we sample various null-text embeddings from the separate generative model with the variational auto-encoder structure [18], denoted VAE, that ensure reconstruction of the original image. Second, we sample various embeddings of the editing prompt using the VAE. Using the CLIP loss [9, 26], VAE is expected to learn the CLIP space [27] according to a given text prompt and sample various features that the language implicates.

To ensure diverse embedding outputs in both methods, diversity loss [20] is added with KL divergence loss as originally proposed in VAE [18]. Methods that we propose do not affect the diffusion model weights and therefore keep the prior knowledge of the model. These attached-type models make it widely applicable and are novel in that it samples various text embeddings correlated to the image generation model.

2. Related Works

Inversion Process Most current diffusion models used for image editing employ DDPM [11] to achieve diversity sampling. However, this approach presents the challenge of reconstructing the image, that is the output may differ substantially from the input. As a solution, DDIM inversion [6, 34] offers the advantage of deterministic properties, allowing a single image to be sent as a fixed latent code. However, it is known that DDIM inversion exhibits lower reconstruction quality than the VQ auto-encoder method. [7] Furthermore, it has been noted that the reconstruction performance declines significantly when sampling is conducted using classifier-free guidance [12] to improve the fidelity of the sampled image. To resolve this issue, Mokady et al [24] suggest Pivotal Inversion with null-text optimization. In our work, we explore sampling various null-text embeddings rather than a simple empty string from the VAE architecture.

Image Editing Models Large-scale diffusion model, such as Imagen [33], DALL-E 2 [29], Parti [40], Stable Diffusion [31] achieves the state-of-the-art performance in text-to-image synthesis which aims to generate realistic images from a text description. Despite their impressive performance, they are not directly suitable for text-guided image editing tasks [3, 9, 17, 19, 23] that necessitate the controllable editing over desired parts of a given image. Even minor modifications to the textual input can result in a drastically different output image. For example, adding the adjective "classic" to the prompt "car" often results in an image of a car with a completely different composition.

One commonly adopted strategy for addressing this limitation involves the use of masks, which can either be provided by the user [1, 2, 25] or generated automatically using an appropriate procedure [4]. However, these mask-based methods may lead to the removal of critical information and the inability to modify complex structures or backgrounds. Hertz et al [10] suggest an intuitive editing technique, named "Prompt-to-Prompt" which utilizes internal cross-attention maps to modify global or local image details via minor adjustments to the text prompt. Besides, ControlNet [41] employs a novel convolution layer called zero convolution as a residual, which facilitates the learning of task-specific input conditions in large-scale diffusion models without disturbing the prior knowledge. As part of our study, we investigate generated images from off-the-shelf diffusion models, conditioned by text embedding produced by our VAE model.

Transformer-based VAE trained by CLIP loss The success of StyleCLIP [26] in text-driven image manipulation has inspired numerous studies to explore the potential of CLIP for addressing various manipulation tasks. [9, 36, 38, 39]. The proposed method involves the mapping of a text prompt to an input-agnostic global direction within StyleGAN's style space [15, 16], In contrast to these text-guided latent manipulation methods, we utilize CLIP loss to acquire a diverse range of meanings present in natural language. The transformer-based VAE [13, 21, 35] is considered as a well-suited architecture for leveraging the information of sequential data in natural language processing. Based on this knowledge, we adopt the same architecture in our approach.

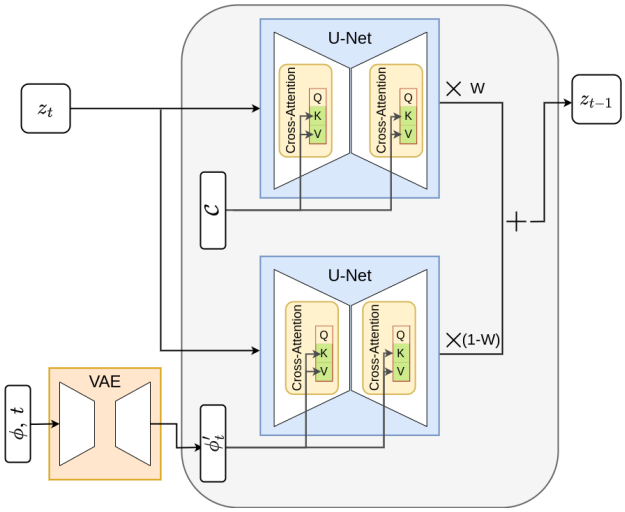

Figure 1. Sampling in arbitrary time step by Sampled Null-Text embedding. Given the null-text embedding ϕ and the time-step t, VAE outputs stochastic version of the null-text embedding, ϕ'_t. Using these null-text embeddings ϕ'_ts, diffusion sampling is advanced via classifier-free guidance.

3. Method

Using the Stable Diffusion model with the CLIP text encoder, we sample various text or null-text embeddings using VAE architecture. When a text or null-text embedding with the time-step used in diffusion process are given as input, VAE encoder produces the mean and standard deviation in lower-space manifold with respect to the given input. When sampling is done via reparameterization trick in the latent space, various text embeddings are produced. We freeze the diffusion model and the CLIP encoder parameters and train only the VAE parameters. Different loss functions are used depending on which embedding the VAE model takes as input; the conditional text embedding or the null-text embedding.

3.1. Sampling Null-text embeddings

The goal is to sample various null-text embeddings that can reconstruct the original image. The fact that the optimized null-text embeddings which fit the latent codes obtained from each DDIM inversion and classifier-free guidance sampling process are not the same with the embed-

ding of the editing prompt [24] implies that there exists various null-text embeddings with reconstruction ability. With different null-text embeddings, we can expect the different output images when sampled via classifier-free guidance. During the classifier-free guidance sampling, the noise prediction with conditional text and the one with null-text are extrapolated to emphasize the power of conditional text embedding. Since the noise prediction with the null-text embedding works as a reference point, changing it will output whole different image. Sampling procedure is depicted in Figure 1.

For training, we follow the original training procedure used in DDIM [34] together with the KL divergence loss used in VAE [18]. Mean Squared Error loss is given as a reconstruction loss $\mathcal{L}_{rec}$ and regularization loss $\mathcal{L}_{reg}$ follows the exact KL divergence metric used in VAE [18]. The overall loss function is,

$$\mathcal{L}_{\phi} = \alpha_{\phi}\mathcal{L}_{rec} + \beta_{\phi}\mathcal{L}_{reg} \tag{1}$$

where $\alpha_{\phi}, \beta_{\phi}$ are given as hyperparameters to weight importance of each loss.

3.2. Sampling Conditional text embeddings

Overall architecture is the same with the method in Sec 3.1 and the sampling procedure is shown in Figure 2. However, the training procedure is done on matching a generated output image to the editing text prompt in CLIP embedding space. Therefore, CLIP loss [26] is used rather than direct L2-norm reconstruction loss.

$$\mathcal{L}_{CLIP} = \lambda D_{CLIP}(E_I(\mathcal{I}''), \mathcal{C}') + (1-\lambda)D_{CLIP}(E_I(\mathcal{I}'), \mathcal{C}'') \tag{2}$$

where $\mathcal{I}'$ is the generated image using editing text embedding $\mathcal{C}'$, and $\mathcal{I}''$ is the generated image using stochastic version of editing text embedding $\mathcal{C}''$ with E_I indicating the CLIP image encoder. To guide the image with insecure text embedding, we cross-match the pairs $\{\mathcal{C}', \mathcal{I}''\}$, $\{\mathcal{C}'', \mathcal{I}'\}$ when calculating the CLIP distance. Unlike Method 1 depicted in Sec 3.1, edited images are used in the training process and we take the Prompt-to-Prompt [10] as our baseline editing method. No reconstruction loss is needed since we expect the fixed attention maps will keep the unedited contexts as described in Prompt-to-Prompt. In addition, the diversity loss [20], which forces the variation in the output images, is added. Since we aim to sample images with various features, negative perceptual loss [14] is used in diversity loss. When N diverse images are generated, let the output images $\{\mathcal{I}''_1, \mathcal{I}''_2, ..., \mathcal{I}''_N\}$. Let $\{P_1, P_2, ..., P_N\}$ be the identity ordering and $\{Q_1, Q_2, ..., Q_N\}$ be the random reordering of images $\{\mathcal{I}''_1, \mathcal{I}''_2, ..., \mathcal{I}''_N\}$ satisfying $P_i \neq Q_i$ [20]. The diversity loss is given as follows:

$$\mathcal{L}_{div} = \frac{1}{N}\sum_{i=1}^{N} ||\Phi(P_i) - \Phi(Q_i)||_1 \tag{3}$$

where Φ denotes the VGG-16 perceptual model to collect image features. The overall loss function is,

$$\mathcal{L}_{\mathcal{C}'} = \alpha_{\mathcal{C}'}\mathcal{L}_{CLIP} + \beta_{\mathcal{C}'}\mathcal{L}_{reg} + \gamma_{\mathcal{C}'}\mathcal{L}_{div}, \;\; \gamma < 0 \tag{4}$$

where $\alpha_{\mathcal{C}'}, \beta_{\mathcal{C}'}, \gamma_{\mathcal{C}'}$ are given as hyperparameters.

In the training process, it is impractical to use images sampled through all the DDIM timesteps when training. It is possible to sample latent codes starting from the standard noise using generated text embedding $\mathcal{C}''$ as condition in every training iteration, but it takes much longer time compared to the original training procedure used in DDIM [34]. Therefore, we slightly mimic the original training procedure. First, we keep the latent codes $x_T, x_{T-1}, ..., x_0$ sampled using the original editing prompt $\mathcal{C}'$. Then, as we iterate using gradients, we forward the U-Net architecture only once from the kept latent code x_t in the random timestep t using the generated editing prompt $\mathcal{C}''$ and predict the image x_0 using DDIM sampler. Then, this predicted image can be used when calculating the CLIP loss. Remember that we do not directly use the norm between original edited image $\mathcal{I}'$ and predicted image as reconstruction loss. CLIP loss let the CLIP embedding $\mathcal{C}''$ get close to the corresponding CLIP latent in the image domain.

Since VAE generates from the probability with respect to the given input, it can learn various features of the editing prompt. Therefore, we plan to experiment in two settings: 1) A single (Image, Text) pair editing or 2) Single text(Class) editing.

4. Experiments

4.1. Training Plan

We begin by conducting experiments to verify the feasibility of our proposed idea using the architectures depicted in Figure 1 and Figure 2. Initially, we use a toy example consisting of a single image and a single text pair to validate our approach. This allows the VAE model to overfit to a single image, simplifying the training process. Subsequently, we expand the scope of the VAE model to include various images, while still limited to a single text prompt. This model can be trained using the ImageNet dataset with one class.

Single image Single text pair editing For the purpose of verifying whether appropriate variance is added to the text embedding, editing is performed for a single image and a single text pair using a toy example. For this, VAE is made using a simple MLP structure. We examine that the

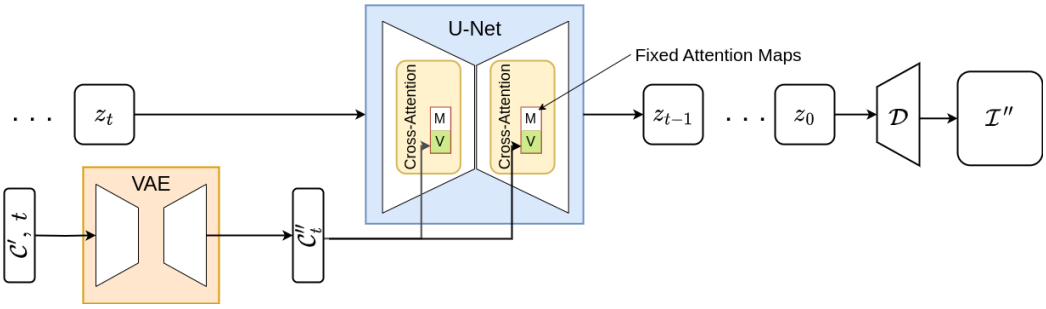

Figure 2. Sampling by Sampled Conditional Text embeddings overview. Given the editing text embedding C' and the time-step t as input, VAE outputs stochastic version of editing text embedding, C''_t. From these C''_ts, edited image I'' is generated.

generated image reflects the modification of a text query while referencing the original image.

Single text Editing ImageNet dataset [5] is used to retrieve the class of the caption [24] where prompt engineering [28] is used to generate text embedding representing the corresponding class. Here, a bank of 80 different sentence templates is used and the CLIP text embeddings are averaged. All images belonging to the corresponding class are used as train dataset. VAE structure is made using MLP or self-attention structure.

Synthesized Image Editing Since training with real image requires diffusion inversion process, training becomes more time-consuming. Therefore, when checking the feasibility of our model and the training process, we use synthesized images as training set instead.

Based on the training plans, we perform training for the null-text embedding and conditional text embedding. However, for null-text embedding case, which aims for reconstruction, we only use real image dataset for training. This means we do not conduct Synthesized Image Editing and focus only on Single image Single text pair editing and Single text Editing.

4.2. Evaluation

Various text-to-image editing methods have been introduced, but are primarily compared qualitatively since confidence numerical evaluation metric is lacking. We get the most out of proposed metrics to quantitatively measure the editing performance of our model with Promp-to-Prompt and Diffedit editing methods as baselines. In addition, we measure the diversity of the generated text embeddings, which is the novel work we make. In the case of sampling various null-text embeddings, reconstruction performance is measured. As editing is done on user's preference, we express qualitative results as well.

Reconstruction Reconstruction loss is used to create null-text embedding with variance. Therefore, it is necessary to verify the quality of reconstruction for various null-text embeddings. For this purpose, we use PSNR and SSIM [37] scores.

Editing To check the editing performance, we will use the CLIP, LPIPS and classification score as evaluation metrics. First, CLIP score, the cosine similarity in the multimodal CLIP space, is calculated between editing text $\mathcal{C}'$ and image $\mathcal{I}''$(CLIP-T) to verify text fidelity or between $\mathcal{I}'$ and $\mathcal{I}''$(CLIP-I) [32]. Second, LPIPS score is calculated between the original image $\mathcal{I}$ and the edited images $\{\mathcal{I}''_1, \mathcal{I}''_2, ..., \mathcal{I}''_N\}$(lower the better) to see whether the edited image $\mathcal{I}''$ contains unedited features of the original image $\mathcal{I}$. Lastly, classification score is verified to ensure that the edited class is well represented even with the modified prompt embeddings.

Variance In contrast to previous models, we incorporate diversity in diffusion-based editing. Since image features are well represented through convolution networks, we once again use the LPIPS score to evaluate variance of features in edited images. Generally, LPIPS score is used to evaluate fidelity between the original image and the output image as depicted in Subsection **Editing**; lower the better. For diversity, however, higher is the better [22, 32].

$$LPIPS(\mathcal{I}, \mathcal{I}') \geq LPIPS(\mathcal{I}', \mathcal{I}'') \tag{5}$$

To evaluate variance in editing, we calculate the LPIPS score between $\mathcal{I}'$ and $\mathcal{I}''$ both created from the same conditional text embedding $\mathcal{C}'$. As $\mathcal{I}$ and $\mathcal{I}'$ are generated from different conditional text embeddings $\mathcal{C}$ and $\mathcal{C}'$, respectively, $LPIPS(\mathcal{I}, \mathcal{I}')$ must be bigger than $LPIPS(\mathcal{I}', \mathcal{I}'')$, as described in Equation 5. The difference of the features would be bigger when compared in different text prompts, implying that diversity would be largest when the equality holds. That is, $LPIPS(\mathcal{I}, \mathcal{I}')$ acts as the upper bound and if $LPIPS(\mathcal{I}', \mathcal{I}'')$ gets closer to the upper bound, variance is proved.

When training VAE for sampling various null-text embeddings ϕ'_t, we conduct evaluation based on the aforementioned metrics.

When stochastically sampling the conditional editing text embeddings C'', however, only editing is done and there is no need to evalute the reconstruction ability. Therefore, editing performance and variance of generated images are measured.

5. Results

First, the results of the single-image-single-text toy example are depicted below. We expect the variation the VAE should have will be learnt through the diversity diffusion model has when using clip loss as the training measure. For sampling null-text embeddings, the diversity of the sampled images will be automatically satisfied through classifier-free guidance when the null-text embeddings are in the manifold that reconstruct the latents well in each timestep. That is, when the VAE learns the region the null-text embeddings show probable capability of reconstructing original image, classifier-free guidance will extrapolate the latent noise to diverse space, showing various image features at last.

To check the validity of adding variation in generated images through text conditions, we first add random noise to the replaced token for editing using Prompt-to-Prompt [10] method. Since our method changes the whole conditional CLIP text embeddings, the variation found in this process would be the lower bound of what we expect. As described in Figure 3. , we can find various dog images which are not constrained to a single identity. It is therefore reasonable to disturb conditional text embeddings for generating various images using diffusion model.

5.1. Analysis of Null-text embeddings

Using the training process described in section 3.1, the semantic meanings or the overall structure of the original prompt are effectively showcased, although complete reconstruction is lacking in representation. (See Figure 4)

However, when sampling is done via editing text embedding using Prompt-to-Prompt [10] method, some semantics remain but most features previously shown in the original image are lost. We suspect two main reasons in this issue. First, modified null-text embedding losses text features that CLIP text encoder provides. The output of the VAE model may reconstruct the image latent with the generated null-text condition ϕ' well, but the null-text condition itself does not have the text modality that CLIP formulates. Therefore, we assume it would be better to take the text embedding through VAE before processing through CLIP text encoder.

Figure 5 is the result from the order of CLIP text encoder to VAE, but sequential semantics that CLIP provides will be kept when we reverse the forwarding order: The VAE first, then the CLIP text encoder.

Second, due to shortage of time, experiments with various model architectures have not yet been made. Also, the training process lacks an adequate number of iterations so far. As training is done for only one time-step in DDIM sampler process in each iteration, much more iterations are required for training the VAE model along with the time-step positional embedding.

5.2. Analysis of Conditional text embeddings

CLIP loss ablation study For the CLIP loss, we weighted sum the $D_{CLIP}(E_I(\mathcal{I}''), \mathcal{C}')$ and $D_{CLIP}(E_I(\mathcal{I}'), \mathcal{C}'')$. However, while $D_{CLIP}(E_I(\mathcal{I}''), \mathcal{C}')$ uses loss gradient backproped through diffusion model, $D_{CLIP}(E_I(\mathcal{I}'), \mathcal{C}'')$ only uses gradient from the text VAE model. Therefore, adversary example is likely to take in place for the use of $D_{CLIP}(E_I(\mathcal{I}'), \mathcal{C}'')$. This causes generated image to be unrealistic while the CLIP cosine similarity is very high.

Pre-training VAE We conducted VAE Pretraining based on the assumption that if a VAE is capable of reconstructing various texts like "A photo of a {class}", it would be able to learn text features. Therefore, we adopted the zero-shot ImageNet prompt engineering approach [28] for VAE pretraining. During this process, we utilized cyclic scheduling [8] to prevent the KL divergence loss from converging to 0.

However, far from our expectation, the simple VAE model was not sufficient enough to learn the text features. This can be seen from the fact that output image from the pre-trained VAE model is not any different from the one using initialized VAE model from scratch in that it had no semantics; consisting of pure noise texture implying no semantic meanings indicated in the conditional text embedding. In addition, for simple MLP-based VAE model, validation performance while pre-training showed lacking robustness. Therefore, the transformer architecture is needed to learn text modality and one example to validate the efficacy of this transformer architecture is training some portion of the pre-trained CLIP encoder transformer. Another example is to use knowledge distillation for transferring CLIP text encoder knowledge to smaller model. These are left to future research. It is important to note that VAE does not need to have full generation capability of text embedding; it only needs to make diversity given the input data.

VAE model architecture Multi layer perceptron architectures are compared in model complexities; *token-level* MLP and *sequence-level* MLP. The token-level MLP applies the same linear operation on every tokens, whereas the sequence-level MLP operates on the whole flattened sequence. The token-level MLP leads to a problem where the image texture becomes uniform, as shown in Figure 6 As a result, the VAE model struggled to perform accurate reconstruction. Therefore, a more complex structure, was necessary, which involved distinguishing and operating on individual tokens. To solve this problem, instead of apply-

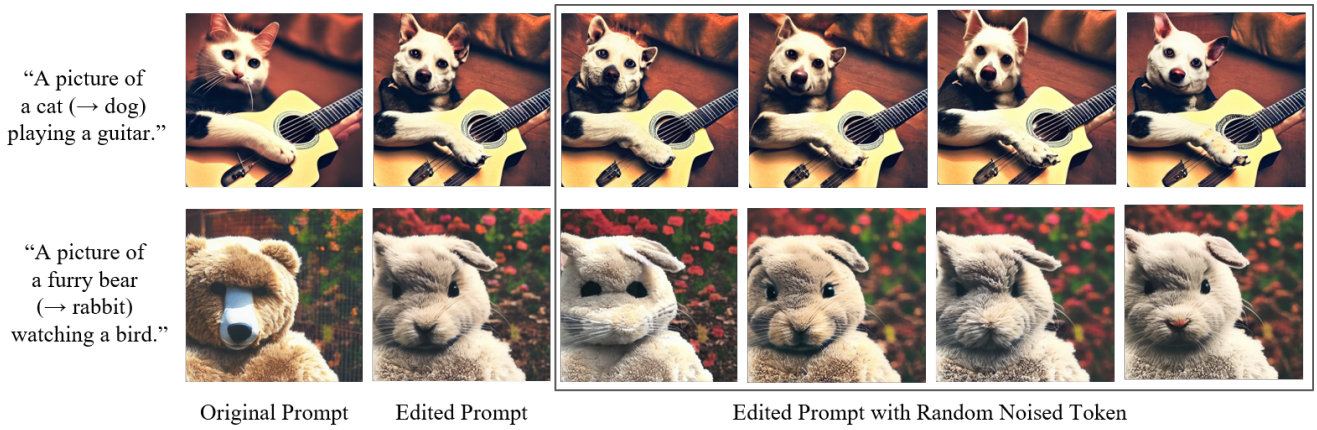

Figure 3. Edited images from a prompt with a random noise token. The images in the first and second column are generated from the original prompt and edited one, respectively, by applying Prompt-to-Prompt [10] method. The last four columns show images from text embedding, which is added random noise to the replace token ("dog").

Figure 4. Generated Images at different iterations. It shows that the quality of the reconstruction exhibits enhancement with the progressing of the iterations.

ing the same MLP-based VAE to each token, We flattened embeddings and conducted MLP operations. Additionally, to overcome memory constraints, we utilized 1x1 convolutions to reduce the number of channels.

Training limitation Besides, conducting training from scratch with a combination of CLIP loss and reconstruction loss reveals the ability to preserve the semantic meanings derived from edited text prompts. However, to optimize the CLIP loss, it is necessary to perform backpropagation in a pixel-level rather than in the U-Net latent space. Consequently, significant GPU memory is required for computing the decoder part of the stable-diffusion auto-encoder. This limitation also prevents us from conducting training experiments in a diverse manner of VAE architecture. The training time substantially increases when clearing cache midway to conserve memory due to the impact of Prompt-to-Prompt method. We presumed the resulting text embedding from the VAE is not noticeably different from the original one. Therefore, during training, only reconstruction loss is employed and the reconstruction capability is shown in Figure 6. For further reconstruction and variation capability, CLIP loss will be incorporated at a later stage. The result is yet obtained and more training iterations are needed to see the result. Extensive ablation study with different VAE architectures along with diverse hyperparameter for training is required to fully examine the idea of adding stochasticity in text condition when editing via diffusion model.

6. Conclusion

To the best of our knowledge, there has not been an attempt to add stochasticity in text-to-image editing process. We bring up the problem of the ODE solver restricting the sampling ability, which is one of the most important feature of generative models, when faced in the situation of the fixed latent and seed environment. Such process is indispensable when editing is done via changing latent in diffusion process. In this paper, we make an attempt to solve the deterministic text-to-image editing methods in multimodality aspect by adding stochasticity in conditional text embeddings.

We examined the possibility of our model by injecting random noise and conducted experiments using a simple MLP VAE model. Additionally, we reported the results based on different model size, indicating the ability

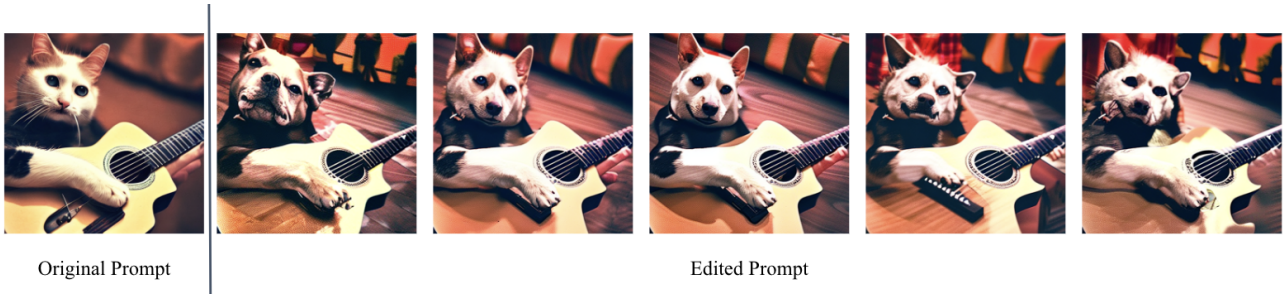

Figure 5. Edited Images from various Null-text embeddings sampled from VAE. The image in the first column is generated from the original prompt with the original null-text embedding and the images in the remaining columns are generated from editing text embedding with the generated null-text embeddings.

Figure 6. A Comparison of edited images at different diffusion steps using two simple MLP architectures. The reconstruction quality improves as the diffusion steps increase, with the lower steps depicted on the left and the higher steps on the right.

to achieve better performance than initially proposed.

Successful outcomes of our proposed experiments will lead us to develop a powerful editing technique that can generate diverse images by conditioning them on a single editing prompt. The embeddings produced by our attachable VAE model can be applied to any existing diffusion model and provide a means of guiding the generator to synthesize images with stochastic properties. We anticipate that our approach will have a positive impact on the field of text-guided image synthesis.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

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

[4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 1, 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[8] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019. 5

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 2

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 3, 5, 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[13] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 516–520. IEEE, 2020. 2

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution.

In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 3

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[17] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3

[19] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 2

[20] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3920–3928, 2017. 1, 3

[21] Danyang Liu and Gongshen Liu. A transformer-based variational autoencoder for sentence generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019. 2

[22] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16377–16386, 2021. 4

[23] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[24] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1, 2, 3, 4

[25] 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. 2

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1, 2, 3

[27] 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. 1

[28] 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. 4, 5

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

[30] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 1

[31] 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. 1, 2

[32] 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. 4

[33] 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. 1, 2

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3

[35] Shiqi Sun, Shancheng Fang, Qian He, and Wei Liu. Design booster: A text-guided diffusion model for image translation with spatial layout preservation. *arXiv preprint arXiv:2302.02284*, 2023. 2

[36] Jianan Wang, Guansong Lu, Hang Xu, Zhenguo Li, Chunjing Xu, and Yanwei Fu. Manitrans: Entity-level text-guided image manipulation via token-wise semantic alignment and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10707–10717, 2022. 2

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[38] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022. 2

[39] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 2256–2265, 2021. 2

[40] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 2

[41] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2