IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 October 2024, accepted 24 October 2024, date of publication 28 October 2024, date of current version 6 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3487215

RESEARCH ARTICLE

Prompt-Based Learning for Image Variation Using Single Image Multi-Scale Diffusion Models

JIWON PARK[1], DASOL JEONG[2], HYEBEAN LEE[2], (Graduate Student Member, IEEE), SEUNGHEE HAN[2], AND JOONKI PAIK[1,2], (Senior Member, IEEE)

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Image, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Joonki Paik (paikj@cau.ac.kr)

This work was supported in part by the Field-Oriented Technology Development Project for Customs Administration through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT and Korea Customs Service under Grant 2021M3I1A1097911, and in part by NRF grant funded by Korean Government (MSIT) under Grant NRF-RS2024-00343863.

ABSTRACT In this paper, we propose a novel technique for a multi-scale framework with text-based learning using a single image to perform variations and text-based editing of the input image. Our approach captures the detailed internal information of a single image, enabling numerous variations while preserving the original features. In addition, text-conditioned learning provides a method to combine text and images to effectively perform text-based editing based on a single image. We propose a technique that integrates the diffusion U-Net structure within a multi-scale framework to accurately capture the quality and internal structure of an image from a single image and perform diverse variations while maintaining the features of the original image. Additionally, we utilized a pre-trained Bootstrapped Language-Image Pretraining (BLIP) model to generate various prompts for effective text-based editing, and we fed the prompts that most closely resembled the input image into the training process using Contrastive Language-Image Pretraining (CLIP)'s prior knowledge. To improve accuracy during the image editing stage, we designed a contrastive loss function to enhance the relevance between the prompt and the image. As a result, we improved the performance of learning between text and images, and through various experiments, we demonstrated its effectiveness on text-based image editing tasks. Our experiments show that the proposed method significantly improves the performance of single-image-based generative models and presents new possibilities in the field of text-based image editing.

INDEX TERMS Single image generation, prompt-based learning, text guided image editing.

I. INTRODUCTION

Image generation and editing have made significant strides with the advancement of generative models. Image variations [1], [2], [3], [4], a crucial task in this domain, involves generating new images by applying multiple variations to an original image. The primary goal is to preserve the key features of the original image while broadening its potential applications through these new variations. These variations are essential for fields like digital media, advertising, and virtual reality, where customized imagery can greatly enhance user engagement and satisfaction.

The associate editor coordinating the review of this manuscript and approving it for publication was Olarik Surinta.

However, conventional methods predominantly rely on large datasets and complex training algorithms, requiring significant computational resources. Recent advancements in single-image generative models [2], [5], [6], [7], [8], [9] offer a promising alternative by reducing both data and computational requirements. Among these, Single-image Denoising Diffusion model (SinDDM) [10] introduces a method for training a model using a single image to perform various tasks, significantly lowering the resource needs for diffusion training. SinDDM employs a multi-scale framework to efficiently extract information from images at different scales, supporting a range of image applications and text-based editing tasks. Despite its advantages, this approach encounters difficulties in maintaining high detail

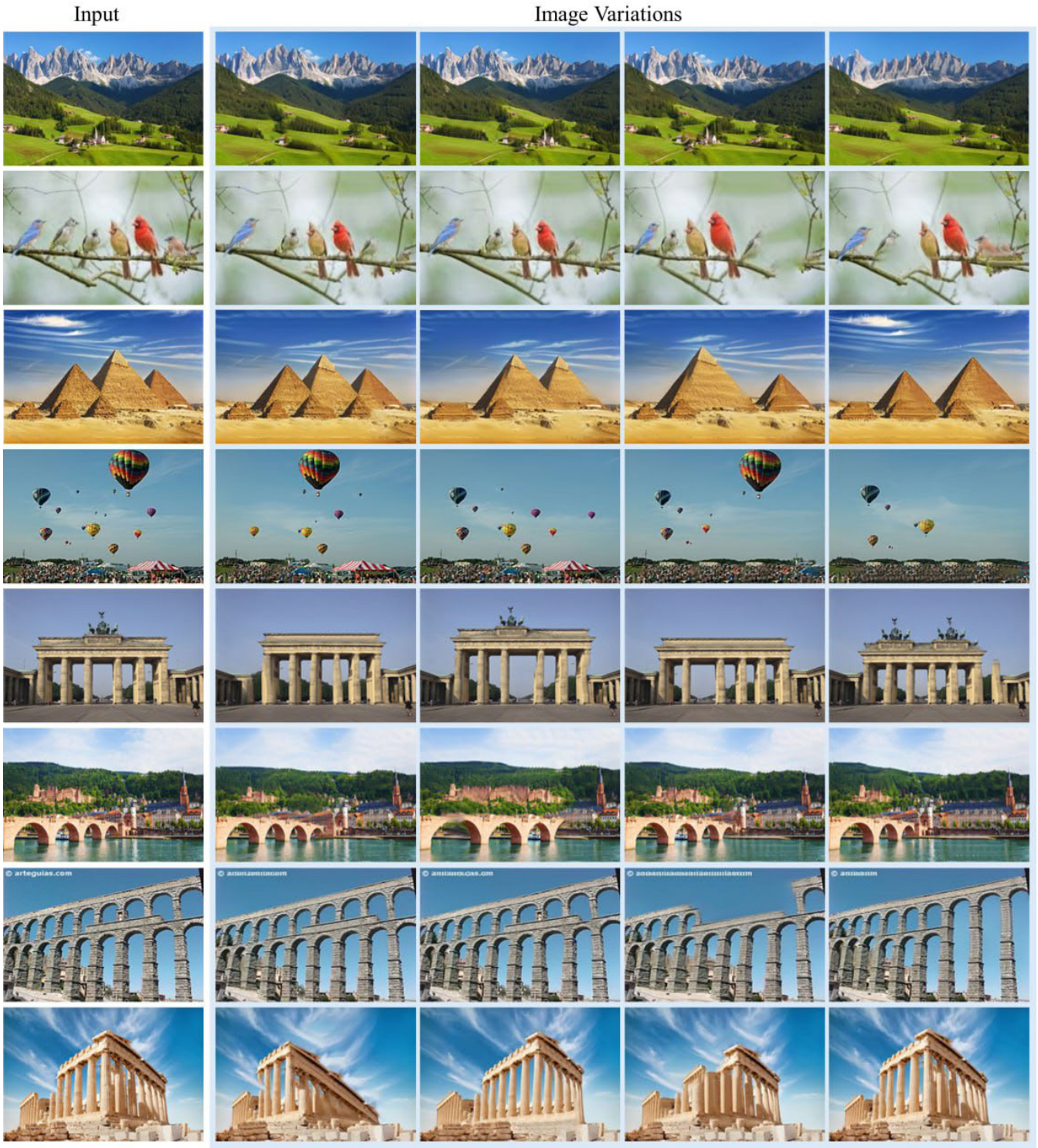


FIGURE 1. Results of image variations generated by our proposed method. Each row shows different variations generated for the same input image, and suggests that the proposed method is effective at generating natural variations while maintaining visual consistency with the structural features of the input image.

quality and preserving the original image's context during variations.

To address these challenges, we propose a novel framework that combines the U-Net architecture [11] with a multi-scale processing approach. The symmetric structure of U-Net enables precise localization and produces high-quality images with minimal data input. This improvement in multi-scale feature extraction significantly enhances both

accuracy and image quality. As shown in Fig. 1, the proposed method effectively generates a variety of variations for different input images such as mountain landscapes, birds, pyramids, and hot air balloons by changing elements such as color, placement, and shape while maintaining the original structural features and identity. This shows a strong generalization ability compared to traditional techniques and implies that it can be effectively applied to a wide range of applications.

This approach is particularly useful in scenarios where it is difficult to gather a wide variety of images, such as military or disaster response situations. In these situations, generating images of the same scene with minor modifications solves the problem of data scarcity and provides a dataset to support model training for these environments. This approach preserves the core structure and essential features of the original image, generating a diverse and consistent dataset that allows the model to train on realistic scenarios with varying levels of detail. As a result, this controlled variation improves the adaptability and reliability of the model, providing practical applications for simulation, risk assessment, and decision support in data-limited environments.

Furthermore, we introduce a prompt-based learning approach for image variations and editing. Prompt-based learning methods [12], [13], [14], [15], initially developed in natural language processing (NLP) to guide models in performing tasks without explicit task-specific training, have recently been extended to computer vision tasks. We employ the Bootstrapped Language-Image Pretraining (BLIP) model [16] to generate contextually relevant prompts from the input image. The proposed approach leverages prior knowledge from the Contrastive Language-Image Pretraining (CLIP) model [17] to improve the alignment between BLIP-generated text prompts and the corresponding images.

We classify the prompts generated by BLIP into positive and negative groups based on their similarity to the input image. For negative messages, select and remove keywords to enable the model to learn the negative information, since negative prompts contain information from the original image. Additionally, we designed a contrastive loss function [18] using the mean of each prompt group to ensure that the model learns without bias toward any specific prompt. This approach enables the generation of high-quality images from limited datasets and supports advanced editing tasks while maintaining a coherent relationship between text and images.

In addition, we applied data augmentation techniques to maximize the effectiveness of variation during text-based editing, further improving performance. This enables the model to perform a wide range of editing tasks with less data. As a result, our proposed method for text-based image editing employs prompt-based learning while retaining the strengths of the U-Net structure to generate diverse images. This approach is anticipated to deliver high-quality image generation and editing solutions for a variety of real-world applications. The primary contributions of this work include:

- We propose a novel multi-scale framework that integrates a diffusion U-Net architecture with prompt-based learning, enabling precise single-image-based variations and editing while preserving the original image features.
- We enhance the alignment between text and image by combining prompt-based learning with the pre-trained Bootstrapped Language-Image Pretraining (BLIP) and Contrastive Language-Image Pretraining (CLIP) models, resulting in more accurate and context-aware text-based image editing.
- We design contrastive loss functions to optimize the generated prompts and enhance learning efficiency, enabling effective image variations and editing with limited data.

II. RELATED WORK

A. SINGLE IMAGE GENERATION MODELS

The typical generative model, Generative Adversarial Network (GAN) [19], learns through the interaction between the generator and discriminator. Traditionally, GANs have been used to train on large datasets to generate diverse images, but recently, learning from a single image has gained popularity. SinGAN [5] is a notable example of this single-image approach, generating new, deformed images from a single input image using a multi-scale pyramid structure. ConSinGAN [6], an extension of SinGAN, improved upon this by training multiple stages simultaneously in a sequential multi-stage manner. It aims to generate more refined images by integrating contextual information into the core GAN structure. However, these GAN-based methods often produce unsatisfactory results due to error accumulation and artifacts that arise during the generation process.

Recently, single-image training with diffusion models [20] has gained popularity. Initially, like GANs, diffusion models required large datasets, but approaches like SinDDM [10], which learn from a single image, have made it possible to produce high-quality images in smaller data environments. SinDDM focuses on generating high-quality images by learning patterns and details from a single image, preserving as much of the original image's detail as possible during the denoising process. SinFusion [8] suggests training a diffusion model using a single image or video. Similarly, SinDiffusion [7] introduces another method to train a diffusion model with a single image, leveraging pre-trained diffusions to generate diverse results from that image.

In addition, recent work introduces a lightweight approach to single image denoising, as seen in Rezvani et al. [21], which explores how to balance model complexity and performance to demonstrate effective denoising with minimal computing resources. This lightweighting strategy aligns with the broader goal of an efficient single image processing framework, providing insights into optimizing image quality with limited resources.

In this paper, we propose a method to generate high-quality images by extracting features directly from a single image through diffusion learning, which eliminates the

need for pre-trained diffusion models such as SinDDM. The multi-scale framework builds on and extends the single-image generation techniques of SinGAN and SinDDM to improve the structural coherence and fine detail of each generated image. In the field of single image generation models, it is possible to generate high-quality images with less reliance on large datasets or pre-trained models.

B. PROMPT BASED LEARNING

Prompt-based learning [22] is a learning approach that has recently gained significant traction in the field of natural language processing (NLP). This approach directs the model to perform specific actions based on a given textual prompt and is commonly used in large language models. By leveraging pre-trained models, this method offers the advantage of adapting to new tasks without the need for extensive fine-tuning. A prime example is GPT-3 [12], which can perform a variety of tasks based on text prompts with minimal fine-tuning, enabling a broad range of text processing applications.

Prompt-based learning has been extended to the field of vision, and these principles are being applied similarly to image generation and editing tasks. In the field of visions, CLIP [17] is important as a pre-trained model for learning associations between text prompts and images. CLIP trains to contrast images and text to learn associations, which can then be used to perform prompt-based tasks.

CoOp [15] is a model that enhances performance by introducing learnable prompts into pre-trained vision-language models like CLIP. Instead of using fixed prompts, CoOp employs learnable prompts, which further optimize CLIP's performance across various vision tasks. Similarly, VLMo [23] and UniCL [24] leverage prompt-based learning built on CLIP's text-to-image association. VLMo is an integrated model that simultaneously processes vision and language data, optimizing the relationship between text and images to excel in tasks such as image classification, generation, and editing. UniCL utilizes a contrastive learning [18] approach that combines images and text, learning text-image pairs to perform a variety of tasks. These studies suggest ways to optimize text prompts to exploit better the interaction between text and images in vision tasks.

In this paper, we utilize prompt-based learning with BLIP and CLIP to strengthen the association between textual prompts and image variations. Specifically, we employ a variety of contextual prompts generated by BLIP and learned CLIP embeddings to ensure that the generated images remain highly relevant to the prompts. Our proposed method provides precise control over image variation and editing with prompts, optimizing text-based editing for a variety of applications, even in data-poor environments.

C. TEXT GUIDED IMAGE EDITING

Text-based image editing [25], [26], [27], [28] is a technique that enables users to modify or transform images directly through natural language. Early research focused on altering simple image properties, but more advanced models are now being developed to comprehend and respond to more complex prompts. For instance, models like TediGAN [29] interpret text prompts to edit images accordingly. These models leverage the GAN architecture to integrate the meaning of the text into the image, producing high-quality editing results. Text2LIVE [28] is a notable example of utilizing diffusion models. Text2LIVE allows users to edit specific parts of an image based on text prompts, enabling them to refine image details naturally. Additionally, models such as DALL-E 2 [25] and Imagen [26] are highly proficient at generating new images or transforming existing ones based on text descriptions.

Previous studies often use text as it is provided, but there are also approaches that modify text for different purposes. For example, Han et al. [30] proposed a method for effectively editing a single remote-sensing image using text guides. They optimized text commands based on user input, demonstrating that precise editing is achievable even for images with unique characteristics. In this study, we demonstrate that text augmentation enables more effective text-based image editing by transforming a single piece of text into multiple perspectives. Our results suggest that effective editing is possible even with small amounts of data.

We demonstrate that text augmentation, which transforms a single text prompt into multiple perspectives, can enhance the effectiveness of text-based image editing. We combine generated prompts with CLIP's learned embeddings to ensure that each generated image is contextually consistent with its prompt. As a result, the details of the image can be fine-tuned to match the text, and the editing results more closely match the user's intent. Our approach allows us to efficiently generate multiple variations even in data-scarce environments, significantly expanding the utility of text-based image editing.

III. PROPOSED METHOD

In this study, we propose a method that integrates text-based learning within a multi-scale framework using the U-Net structure. The proposed approach aims to enhance the quality of image generation while maintaining contextual consistency during text-based image variations. In the first subsection III-A, we describe the overall multi-scale framework structure and specifically address how U-Net can be utilized to effectively generate images. In the second subsection III-B, we outline the proposed text-based learning approach for preserving the context of deformed images. In the third subsection III-C, we present a method for categorizing the prompts generated using the pre-trained BLIP model into positive and negative groups based on CLIP similarity, and we use this categorization to calculate contrastive loss between the prompts. Lastly, subsection III-D, the fourth subsection details the process of prompt augmentation, which enables learning text-based image editing from multiple perspectives with limited data. Through this approach, we demonstrate that effective image variations

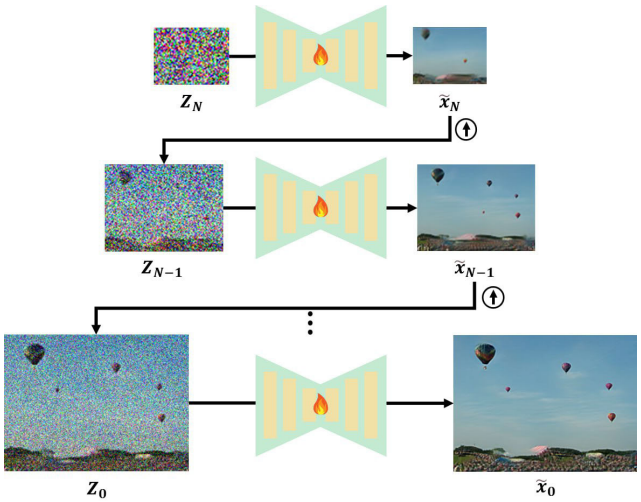


FIGURE 2. **Multi-scale backward diffusion process. At each scale, the noisy input z_n is processed to restore the image $\tilde{x}_n$, which is then upsampled to the next scale. The final result is a high-resolution, clean image $\tilde{x}_0$.**

and text-based image editing can be achieved with strong performance.

A. MULTI SCALE FRAMEWORK

We adopted a multi-scale framework to more effectively learn from a single image. Fig. 2 illustrates the overall structure of this multi-scale framework. This approach takes a single image as input and aims to perform various image variations and text-based editing through text-conditioned learning based on multiple scales. The multi-scale framework converts the input image into multiple resolutions, with image restoration performed independently at each resolution. The input image x_0 is downscaled to several resolutions, $0, 1, \ldots, N$, and at each resolution, the image $\tilde{x}_n$ is reconstructed using a diffusion process that leverages the U-Net structure. The diffusion process is divided into forward and reverse stages, with detailed formulas provided in Eqs. (1) and (2).

The forward diffusion process is defined as follows:

$$x_n = \alpha_n x_{n-1} + (1 - \alpha_n)\hat{x}_{n-1} + \sigma_n z, \tag{1}$$

where x_n is the image at the current scale, $\hat{x}_{n-1}$ is the blurred image from the previous scale, α_n and σ_n are parameters controlling the intensity of noise and blur, respectively, and z represents Gaussian noise.

Conversely, the reverse diffusion process is described as:

$$\tilde{x}_n = \uparrow(\tilde{x}_{n+1}) + \psi_n\left(z_n + \uparrow(\tilde{x}_{n+1})\right), \tag{2}$$

where $\tilde{x}_n$ is the restored image at the current scale, and $\tilde{x}_{n+1}$ is the restored image from the next scale. The term z_n represents the noise added at the current scale, while ψ_n is the result obtained through the U-Net structure. This process helps preserve both detail and semantic consistency in the final generated image. Specifically, we emphasize preserving the input image's details at each scale by replacing the convolutional block at each time step in the reverse process with the U-Net structure.

B. PROMPT-BASED LEARNING FOR JOINT TEXT AND IMAGE TRAINING

In this section, we employ prompt-based learning to improve the preservation of contextual information in both text and images. We utilize pre-trained BLIP [16] and CLIP [17] models for this purpose. First, the BLIP model generates a set P of different text prompts corresponding to the input image x_0.

$$P = f_{\text{BLIP}}(x_0), \tag{3}$$

where each prompt contains a different textual description related to the image x_0.

We utilize prior knowledge from CLIP to score the generated prompts as an objective metric. Specifically, we obtain the embedding vectors of the image and text using a pre-trained CLIP model and then calculate the similarity between the input image and the generated prompt using cosine similarity. The embedding vector for image x_0 and the text embedding vector for prompt P_i:

$$\begin{aligned} v_x &= f_{\text{CLIP}}(x_0), \\ v_{P_i} &= f_{\text{CLIP}}(P_i). \end{aligned} \tag{4}$$

Cosine similarity is calculated based on the angle between two vectors v_x and v_{P_i}, and is formulated as follows:

$$cos(v_x, v_{P_i}) = \frac{v_x \cdot v_{P_i}}{\|v_x\| \cdot \|v_{P_i}\|}, \tag{5}$$

where $\cdot$ denotes the inner product of two vectors, and $\|v_x\|$ and $\|v_{P_i}\|$ denote the magnitudes of the vectors v_x and v_{P_i}, respectively. We score each prompt based on this similarity and select the prompt P^* with the highest similarity.

$$P^* = \arg\max_{P_i} cos(v_x, v_{P_i}). \tag{6}$$

The selected prompt P^* is combined with the image during the training process, allowing various editing operations to be performed while preserving the meaning of the image.

In this case, the cross-attention mechanism is critical for aligning semantic features between the selected text prompt and the image. Cross-attention dynamically matches contextual elements of the prompt to corresponding areas within the image to maintain semantic consistency and prevent unintentional semantic shifts during editing. This alignment process associates key semantic aspects of the text with relevant areas of the image to ensure that the generated variation accurately reflects the intended context and meaning specified in the prompt.

As a result, our training progresses while maintaining semantic consistency between images and text.

The cross-attention mechanism consists of an image embedding query Q, a text embedding key K, and a value V:

$$\text{atten}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V, \tag{7}$$

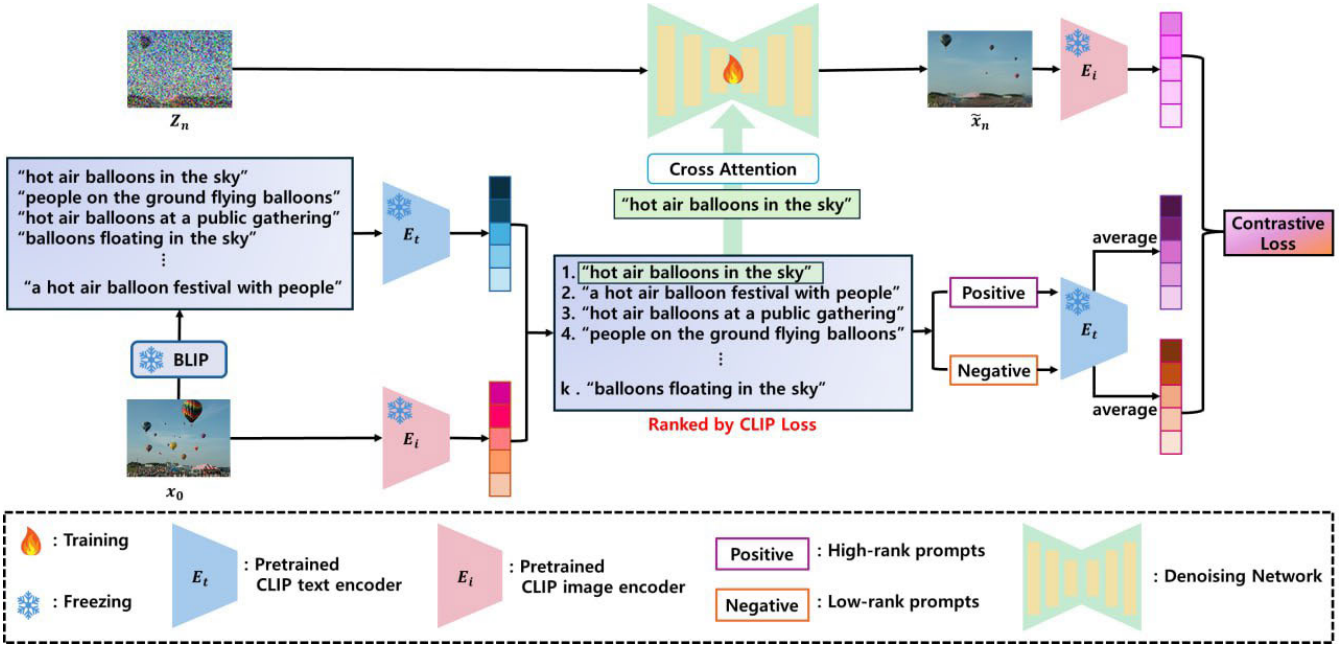


FIGURE 3. Architecture of the scale-specific learning of the proposed method. prompts corresponding to the input image are generated using BLIP, and the generated prompts are ranked using CLIP Loss. Prompts most similar to the image are used in the model's learning by cross attention, while the remaining prompts are separated into Positive and Negative based on their ranking, encoded with the CLIP text encoder, and contribute to the model's learning by contrastive loss.

where $Q = XW_Q$ is the query matrix obtained from the image embedding X, $K = PW_K$ and $V = PW_V$ are the key and value matrices obtained from the text embedding P of the selected prompt. W_Q, W_K, and W_V are the projection matrices being trained, and d_k is the dimension of the key vector, which is used to numerically stabilize the inner product. The selected prompt P^* is combined with the image during the training process, allowing various editing operations to be performed while preserving the meaning of the image.

C. PROMPT CONTRASTIVE LOSS

In this section, we describe the contrastive loss function [18] designed based on the prompts generated in Section III-B to strengthen the association between the text and images used in training. We adopted a method to separate the prompts generated by the BLIP model into positive prompts P^+ and negative prompts P^- to maintain semantic consistency between text and images and to enable the model to learn different text variations. To evaluate the similarity between prompts, we employed cosine similarity, and the classification process is shown in Equation (8).

The BLIP model generates various text prompts for the input image, offering diverse textual representations associated with the image. However, not all prompts are closely related to the image content. Therefore, we evaluate these prompts based on their similarity scores and subsequently categorize them into positive P^+ and negative P^- prompts.

$$\begin{aligned} P^+ &= P_i \mid \cos(v_x, v_{P_i}) \geq M, \\ P^- &= P_i \mid \cos(v_x, v_{P_i}) < M, \end{aligned} \tag{8}$$

where M denotes the median of $\cos(v_x, v_{P_i})$, representing the median similarity score among all prompts, and is used as a measure of the center of the overall similarity distribution. By using M as a threshold, prompts with similarity scores above the median are categorized as positive, while prompts below the median are categorized as negative. This categorization allows the model to learn which text has a strong semantic association with an image. In this way, the model develops the ability to maximize similarity to positive prompts and minimize similarity to negative prompts.

Specifically, positive prompts P^+ consist of texts with high relevance to the image, enabling the model to learn text that aligns with the main features of the image. In contrast, negative prompts P^- are composed of texts with low relevance to the image. To ensure the model distinctly learns negative associations, key words related to the image are removed from these prompts, defined as follows:

$$P^-_{\text{key}} = P^- \setminus \text{key word}, \tag{9}$$

where key word refers to the textual representation directly associated with the image, selected as the most frequently occurring word across all prompts. The "key words" are identified through statistical frequency analysis and are considered critical text elements in maintaining the image's relevance. By removing key words, the model more clearly learns examples of mismatched text and images, allowing it to better recognize discrepancies between text and images.

We divide the prompts into positive P^+ and negative P^-, and each prompt is converted into an embedding vector by the

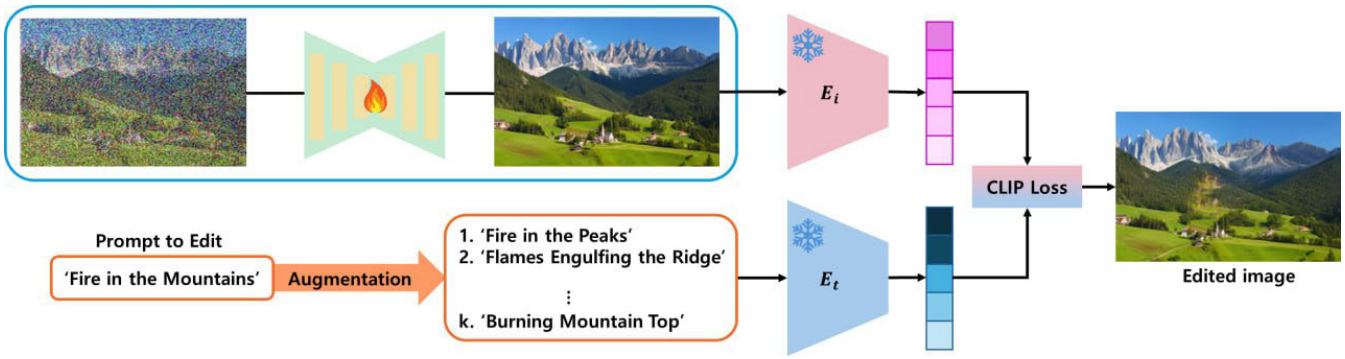


FIGURE 4. Text guided editing task. During inference, we conduct the text-guided editing task by supplying the specified text and converting it into multiple prompts. The image is then updated using CLIP loss, calculated between the image embedding and the average embedding of these transformed prompts.

CLIP text encoder. The converted vectors are averaged by group and used as input to the final contrastive loss function. The contrastive loss function is defined as follows:

$$\mathcal{L}_{\text{contra}} = \lambda_1 \cdot cos(v_p^+, v_{\tilde{x}}) - \lambda_2 \cdot cos(v_p^-, v_{\tilde{x}}), \tag{10}$$

where $v_{\tilde{x}}$ is the embedding vector of the generated image, $v_{p_i}^+$ is the average embedding vector of the positive prompts, and $v_{p_i}^-$ is the average embedding vector of the negative prompts. The λ_1 and λ_2 are the weighting factors of the calculated cosine loss function, both set to 1 in this paper. The contrastive loss function trains the model to maximize similarity with positive prompts and minimize similarity with negative prompts, allowing the model to create more consistent associations between text and images.

Additionally, L1 loss is used to measure the discrepancy between the noise prediction and the reconstructed image, defined as follows:

$$\mathcal{L}_{\text{L1}} = \mathbb{E}[\|\epsilon - \hat{\epsilon}\|], \tag{11}$$

where ϵ is the noise vector, and $\hat{\epsilon}$ is the noise vector predicted by the model.

Finally, the combined contrastive and L1 loss, termed the final loss function, is expressed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{contra}} \cdot \mathcal{L}_{\text{contra}}, \tag{12}$$

where λ_{contra} is the parameter adjusting the weight of the contrastive loss, set to 0.1 in this study.

We designed the contrastive loss to be applied at specific intervals, rather than every iteration, to maintain the generalization power of the model and use computational resources efficiently. This is critical for speeding up training on large models. As a result of this contrastive loss, the model is trained on a wider variety of text-image associations.

D. TEXT GUIDED IMAGE EDITING WITH PROMPT AUGMENTATION

In this section, we describe the data augmentation techniques employed to maximize the effectiveness of variations during text-based editing. The proposed approach is shown in Fig. 4, which aims to improve image editing performance by varying the original text prompts in multiple ways. We utilize language databases such as WordNet [31] to generate various versions of the prompt T' by replacing key words in the original text prompt with synonyms or altering the sentence structure.

$$T' = \{t'_1, t'_2, \ldots, t'_k\}. \tag{13}$$

This variations are illustrated in Fig. 4, where, for example, the original text "fire in the mountains" can be transformed into expressions such as "fire in the peaks," "flames devouring the ridge," or "burning mountain tops." This approach helps to reflect different contexts while maintaining the meaning of the original text.

Each generated prompt is converted into an embedding vector by the CLIP text encoder, and these vectors are averaged and combined into the final prompt embedding.

$$v_{\text{avg}} = \frac{1}{k} \sum_{i=1}^{k} v_{t'}^i, \tag{14}$$

where $v_{t'}^i$ is the embedding vector of each transformed prompt $t'i$. The combined prompt embedding vector v_{avg} is trained with the embedding vector of the original image via CLIP Loss:

$$\mathcal{L}_{\text{CLIP}} = \text{sim}(v_{\text{avg}}, v_x), \tag{15}$$

where $\text{sim}(v_{\text{avg}}, v_{\text{img}})$ represents the similarity between the averaged prompt embedding vector and the image embedding vector. As a result, the model trains on stronger contextual associations between text and images, which can further improve the performance of text-based image editing.

IV. EXPERIMENTAL RESULTS AND EVALUATION

In this section, we describe in detail the implementation and evaluation of the proposed method, including the experimental setup, comparative analysis, and a discussion of the results. Our experiment was designed to demonstrate the effectiveness of a multi-scale structure that incorporates

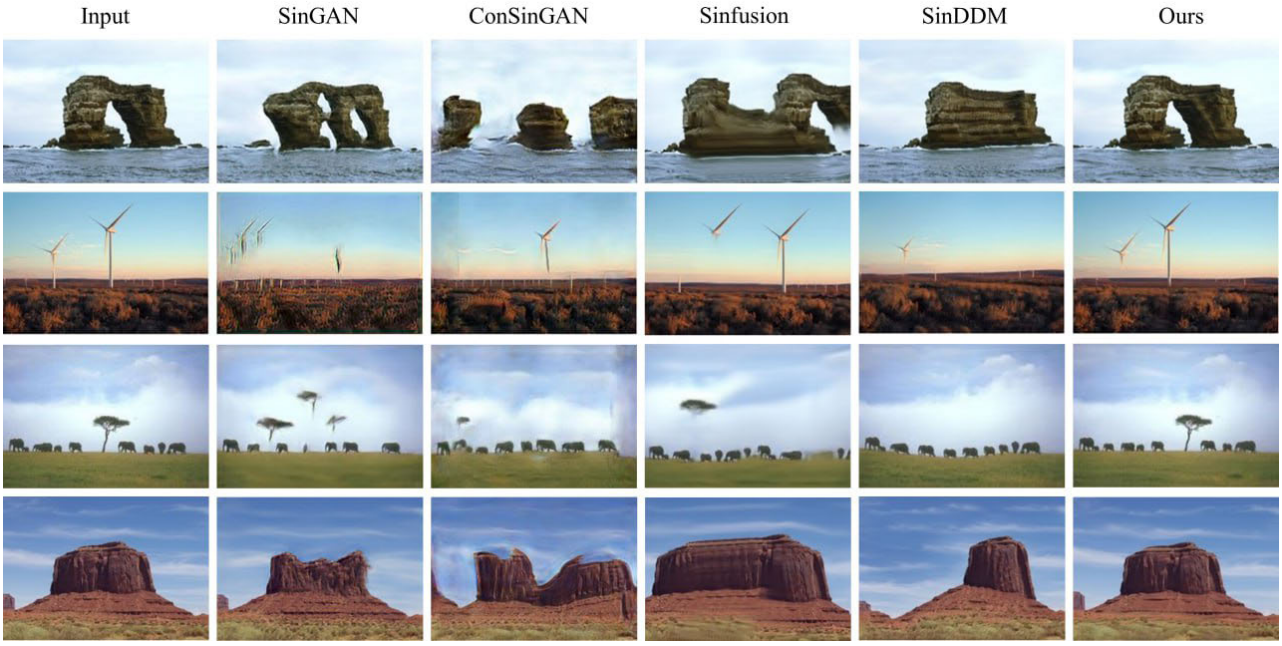


FIGURE 5. Qualitative comparison of single image generation results. The first column is the input image, the second column is the results of SinGAN [5], the third column is the results of ConSinGAN [6], the fourth column is the results of Sinfusion [8], the fifth column is the results of the SinDDM [10] method, and the last line is the results of our proposed method.

text-conditioned U-Nets to perform various image variations and text-based image editing effectively.

TABLE 1. Quantitative comparison. Comparison of image variations methods SinGAN, ConSinGAN, SinFusion, SinDDM, and the proposed method using CLIP similarity, SSIM, NIMA, and LPIPS metrics. To highlight important results, the best results are in bold.

Model	CLIP sim.↑	SSIM↑	NIMA↑	LPIPS↓
SinGAN	0.91±0.05	0.48±0.10	5.47±0.97	0.33±0.02
ConSinGAN	0.86±0.09	0.42±0.11	5.02±1.32	0.40±0.06
SinFusion	0.87±0.04	0.37±0.13	5.30±1.08	0.43±0.03
SinDDM	0.90±0.03	0.44±0.10	**7.45±0.75**	0.34±0.03
Ours	**0.98±0.01**	**0.70±0.10**	5.77±0.66	**0.16±0.03**

A. DATASETS AND METRICS

In this study, we utilized the datasets from SinGAN [5] and SinDDM [10] to ensure smooth comparative experiments with the Places50 dataset [32]. The Places50 dataset consists of images covering a wide variety of locations, making it well-suited for evaluating the generalization ability of our model.

We used the following metrics to quantitatively evaluate our experimental results. First, CLIP Similarity [37] measures the similarity between the generated image and the corresponding text, which is critical for assessing the consistency between images and text. This allows us to evaluate how well the proposed model generates images that reflect the meaning of the text. Secondly, SSIM (Structural Similarity Index Measure) [38] evaluates the structural similarity between the generated and target images, with higher values indicating greater structural fidelity. Third, Neural Image Assessment (NIMA) [39] is used to evaluate the perceptual quality of an image, with higher scores indicating more aesthetically satisfying images. Lastly, Learned Perceptual Image Patch Similarity (LPIPS) [40] is a metric that evaluates the difference between two images in a way similar to human visual perception and is useful for perceptually measuring the similarity of images. These metrics are used to comprehensively evaluate the quality, structural similarity, aesthetic value, and perceptual similarity of the generated images.

B. QUANTITATIVE COMPARISON

The comparative results presented in Table 1 provide a comprehensive overview of the performance of various image variations methods, including our proposed approach, SinGAN [5], ConSinGAN [6], SinFusion [8], and SinDDM [10]. We assess the performance using widely adopted metrics such as CLIP Similarity [37], Structural Similarity Index Measure (SSIM) [38], Neural Image Assessment (NIMA) [39], and Learned Perceptual Image Patch Similarity (LPIPS) [40]. Additionally, we used CLIP Similarity to gauge the conceptual alignment of the generated images with textual descriptions, where higher scores indicate a better match.

The distinct advantage of our proposed method lies in its exceptional performance in CLIP Similarity and SSIM, highlighting its ability to generate images that are not only visually appealing but also structurally aligned

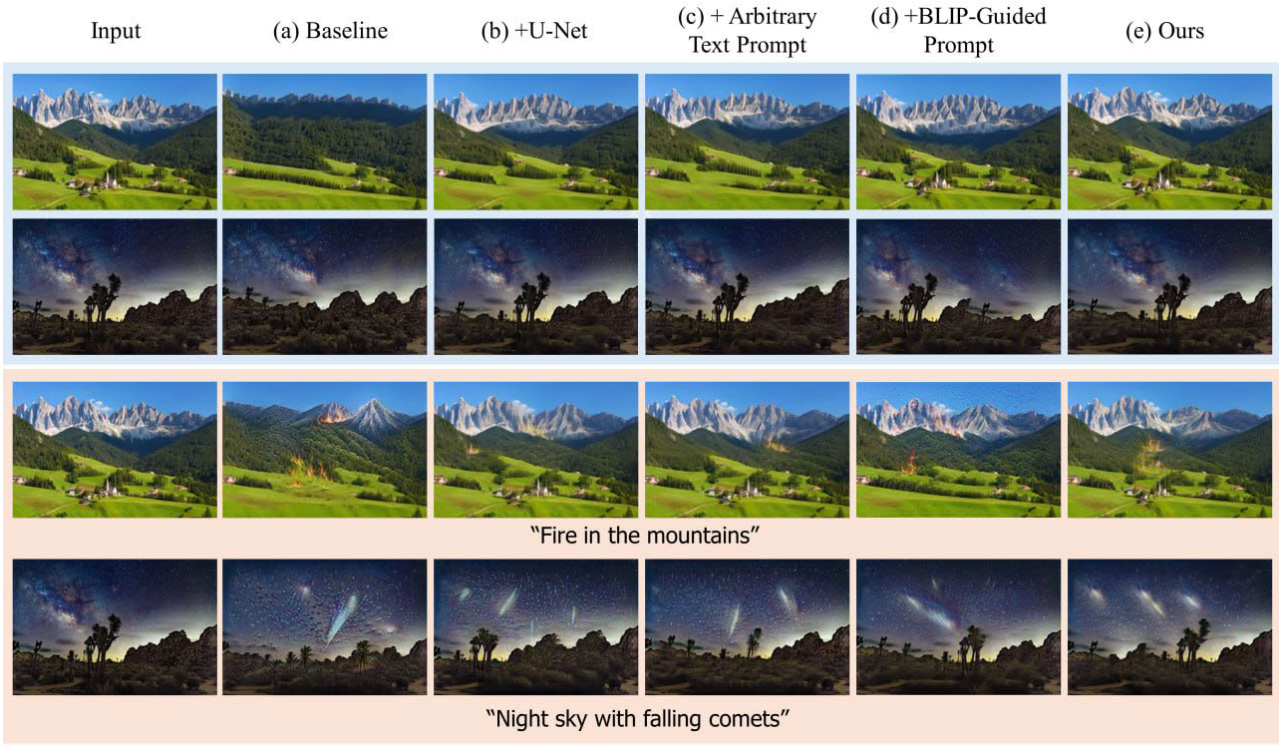


FIGURE 6. Results of ablation study. This figure shows the results of an experiment comparing the performance of our proposed method in steps. The contribution of each component is evaluated and the overall performance improvement is visualized. Our results allow us to validate the effectiveness of the proposed method step by step.

with their textual descriptions. Specifically, our method achieved the highest CLIP Similarity score, demonstrating excellent consistency with textual content, which is particularly beneficial in applications such as digital marketing, where images must closely reflect textual content to effectively convey a message. Moreover, the proposed method demonstrates competitive performance in the NIMA metric, indicating that the generated images possess high perceptual quality. Lastly, the proposed method achieved the lowest LPIPS score, indicating that the generated images exhibit high perceptual similarity to the reference images. This strongly supports that the proposed approach offers a robust and effective solution for diverse image variations need, affirming its overall excellence in the experimental outcomes.

These results demonstrate that the proposed approach excels in generating and editing images without compromising visual or structural integrity, proving its superiority across various image generation tasks.

C. QUALITATIVE COMPARISON

1) SINGLE IMAGE RECONSTRUCTION

To further demonstrate the effectiveness of the proposed method, we qualitatively compare the images generated by our approach with SinGAN [5] and ConSinGAN [6], two representative models for single image generation, and SinFusion [8] and SinDDM [10], which utilize diffusion models for single image generation. Fig. 5 presents a visual comparison of the images generated by the different models for each input image, clearly highlighting the differences in performance.

The visual comparison reveals that the proposed approach more accurately reproduces the structural features and details of the input images, suggesting it produces higher-quality outputs compared to the other models. In particular, our method effectively preserves complex patterns and fine details, significantly enhancing the similarity to the input image. These results demonstrate that the proposed approach performs strongly in single-image generation.

2) IMAGE VARIATIONS

In Fig. 1, we experimentally demonstrate that the proposed method produces visually diverse variations while maintaining the structure and realism of the original image in a variety of image variation situations. In particular, for complex structures such as architectural images, we observe that the method is effective in generating variations while preserving the original layout.

We also experimented with the variability of the generated image variations by adjusting the Load Milestone (LM) parameter. The LM parameter is responsible for loading the weights at a specific point in time during the model training process. In our study, a checkpoint is saved every 10,000 training runs out of a total of 120,000 training runs,

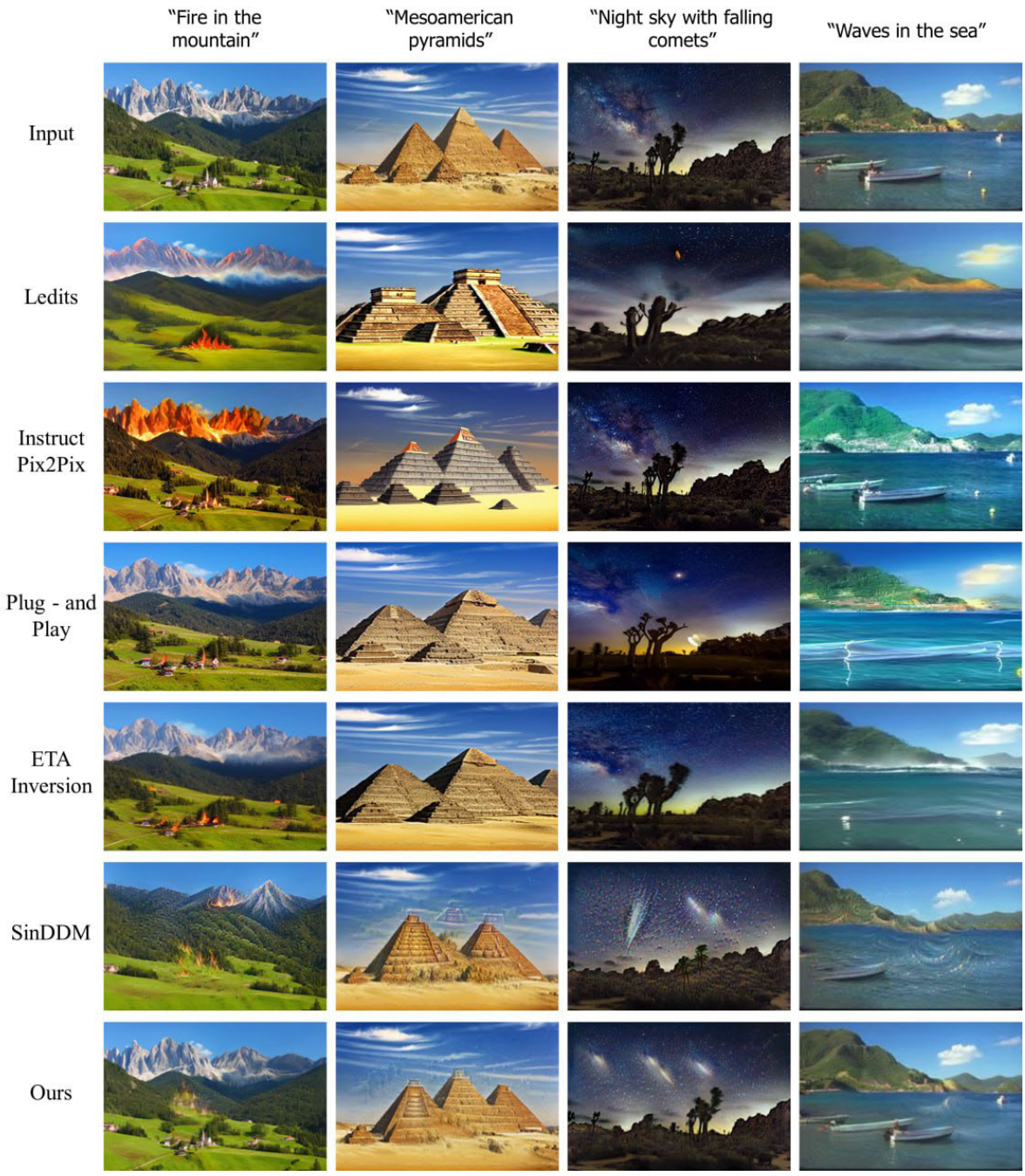


FIGURE 7. **Qualitative comparison of Text-guided image editing results. The first row shows the input images, the second row shows the results of Ledits [33], the third row shows the results of InstructPix2Pix [34], the fourth row shows the results of Plug and Play [35], the fifth row shows the results of ETA Inversion [36]. The sixth row shows the results of SinDDM [10] and the bottom row shows the results of our proposed method.**

and depending on the LM value, we load the weights from one of them. For example, an LM value of 2 would load the weights from the 20,000th training checkpoint, while an LM value of 12 would load the weights from the 120,000th training checkpoint.

Fig. 8 shows the results using LM values of 2, 8, and 12. A lower LM value, such as 2, allows more changes to be made to the image by applying the weights from the early learning steps, while a higher LM value, such as 12, allows the weights from later learning to be used to add only subtle variations while preserving as much of the structure of the original image as possible.

The LM parameter thus allows the user to control the degree of visual variation. This flexibility is particularly useful in scenarios that require different datasets within a consistent context, such as military and disaster response

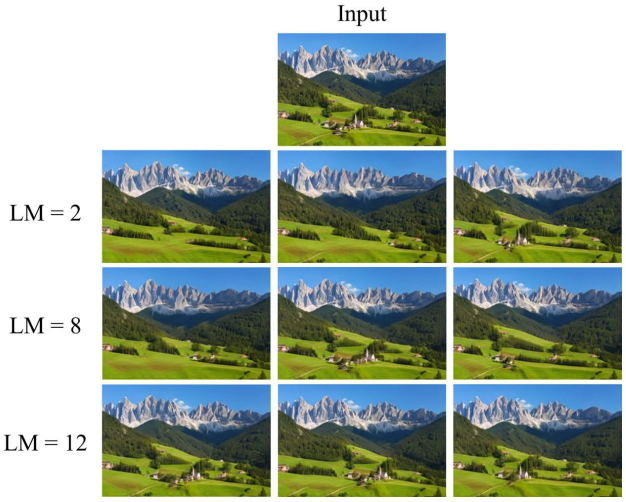


FIGURE 8. Results showing the effect of different load milestone (LM) values on image variations. The input image is shown in the first column, while the subsequent rows display generated images with LM set to 2, 8, and 12. Lower LM values produce more noticeable variations in color and texture, while higher values emphasize structural similarity to the original image, demonstrating the model's flexibility in controlling image variation.

training, and allows for controlled changes that support effective model training.

3) TEXT GUIDED IMAGE EDITING

To further demonstrate the effectiveness of the proposed method, we qualitatively compare our approach with text-based image editing results generated by Ledits [33], Instruct Pix2Pix [34], Plug and Play [35], ETA Inversion [36], and SinDDM [10]. Fig. 7 presents a visual comparison of each input image transformed to match the editing prompts, allowing us to observe the performance differences between the models.

The visual comparison reveals that the proposed method preserves the structural features and details of the input images while naturally performing the user-specified edits. In particular, compared to the other models, our method captures the context of the original image without introducing unnecessary artifacts and naturally represents changes in color and lighting.

Furthermore, the proposed method consistently delivers high-quality results across various editing prompts, maintaining alignment with the intended edits. Our findings demonstrate that the proposed method performs effectively in text-based image editing.

D. ABLATION STUDY

1) PROPOSED METHOD ABLATION STUDY

To evaluate the contribution of each component in our proposed method, we performed an ablation study, and the corresponding results are illustrated in Fig. 6. The first column presents the input images alongside the text prompts utilized for the text-based image editing task.

TABLE 2. Quantitative comparison of the ablation study. Results showing the impact of each component on model performance across CLIP similarity, SSIM, and LPIPS metrics.

Model	CLIP sim.↑	SSIM↑	LPIPS↓
Baseline	0.90±0.03	0.44±0.10	0.34±0.03
+U-Net	0.97±0.02	0.44±0.09	0.29±0.06
+Arbitrary Text Prompt	0.97±0.02	0.43±0.07	0.30±0.08
+BLIP-Guided Prompt	0.97±0.03	0.44±0.09	0.29±0.09
Ours	**0.98±0.01**	**0.46±0.13**	**0.27±0.03**

Fig. 6 (a) shows the image generated by the baseline model, SinDDM. While the baseline effectively extracts and generates image features through a multi-scale structure, it suffers from degradation in quality when preserving fine details and during text-based editing. Fig. 6 (b) shows the outcome of integrating a U-Net-based multi-scale framework. This incorporation leads to enhanced image quality, particularly in terms of fine detail preservation and denoising. The U-Net architecture better captures intricate image features, improving overall visual fidelity. Fig. 6 (c) shows the results of training with arbitrary text prompts for text-based learning. By incorporating textual information, the model produces more detailed images and retains some original detail in text-based editing. However, in some cases, biased information from a single text prompt skewed the generation results. Fig. 6 (d) shows the outcome of employing the BLIP model to generate prompts corresponding to the image and selecting the most similar one. Here, we use a loss function to reduce the similarity between the embedding of the original image and the average embedding of the generated prompts, without distinguishing between positive and negative prompts. While the generated images preserved more details, issues with background artifacts during editing tasks persisted. Finally, in Fig. 6 (e), we present the results of applying our contrastive loss function. Contrastive loss trains the model to enhance the semantic consistency between text and images. This approach significantly reduced background artifacts in text-based editing, resulting in more natural and consistent image editing.

Table 2 provides a quantitative summary of the ablation study, emphasizing the contribution of each component to model performance. The base model shows basic performance on the CLIP similarity, SSIM, and LPIPS metrics, with incremental improvements in semantic consistency, image quality, and detail preservation as each component of the U-Net structure, random text prompts, BLIP guided prompts, and contrast loss is added. The final structure with all components achieves the strongest performance on these metrics, indicating the effectiveness of the proposed approach.

The results of the ablation study demonstrate that each proposed component made a substantial contribution to model performance. The U-Net structure significantly improved

the quality of image generation, and the inclusion of text conditions increased the diversity of generated images. The BLIP model required accurate text descriptions to ensure semantic consistency between text and image, and the addition of Contrastive Loss further strengthened this consistency, ultimately enhancing text-based image generation and editing.

TABLE 3. Loss ablation study. Ablation study on loss functions. The table shows the performance impact of adding BLIP Prompt Loss and Contrastive Loss on the baseline L1 loss model. Metrics include CLIP similarity, SSIM, LPIPS, and PSNR. The results indicate that each additional loss component improves specific aspects of image quality and structural consistency.

Model	CLIP sim.↑	SSIM↑	LPIPS↓	PSNR↑
L1 loss	0.90	0.31	0.39	15.27
+ BLIP Prompt Loss	0.98	0.50	0.22	19.62
+ Contrastive Loss	**0.99**	**0.55**	**0.19**	**20.68**

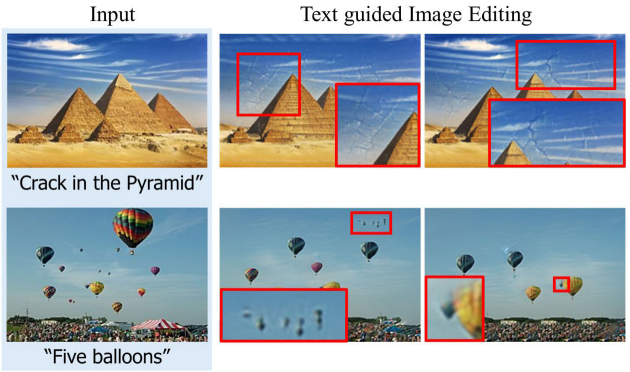


FIGURE 9. Limitations. The first column shows input images with specific prompts ('Crack in the Pyramid' and 'Five balloons'). The second and third columns display generated results where inconsistencies appear in accurately capturing the prompt's semantic details, such as the exact number or precise location of objects. This highlights challenges in retaining fine-grained semantic information, suggesting areas for future improvement in prompt interpretation and object localization.

2) LOSS ABLATION STUDY

In this section, we analyze the contribution of each loss component to model performance through a loss ablation study. As shown in Table 3, we experimented with three configurations: a baseline model using only L1 loss, a model with L1 loss and BLIP prompt loss, and a model combining L1 loss, BLIP prompt loss, and contrastive loss.

The L1 loss represents the default diffusion loss without any additional prompt-related loss. The BLIP prompt loss introduces a loss component that measures the similarity between the prompt and the generated image using BLIP. Finally, the contrastive loss—proposed in this paper—enhances the model's ability to distinguish between prompts categorized as positive and negative, promoting alignment with relevant semantic features. The evaluation metrics include CLIP similarity, SSIM, LPIPS, and PSNR, which reflect various aspects of image quality and structural consistency.

The results demonstrate that incorporating BLIP prompt loss significantly improves the model's ability to maintain semantic alignment with text prompts, as evidenced by increases in CLIP similarity and SSIM. Adding contrastive loss further strengthens this alignment, yielding improvements in SSIM and PSNR, while also reducing perceptual error as indicated by LPIPS. These progressive enhancements indicate that each loss component contributes uniquely to improving image quality and structural fidelity.

V. LIMITATION AND FUTURE WORK

Although our study proposes a method that utilizes text-based learning and augmentation to enhance the effectiveness of prompt-based image editing, there is a limitation remain, presenting opportunities for future research. Our approach is valuable for generating a variety of prompts to extend the range of image variations and edits, but it has difficulty accurately reflecting detailed semantic information, such as the number or location of specific objects. For example, in situations that require the quantity or location of a specific object, as in Fig. 9, the augmented prompt may not fully reflect the original intent, potentially obscuring the meaning. Such a limitation is caused by the dilution of semantic information during prompt augmentation, which prevents the model from accurately reflecting detailed requirements.

To address this limitation, future work needs to focus on developing techniques that can more accurately analyze and retain important semantic information within prompts. Specifically, we can explore ways to combine semantic parsing and object recognition techniques into the model to ensure that the model accurately identifies and maintains the key objects and relationships required by the textual prompts. This approach will strengthen the association between the prompt and the generated image, allowing for more fine-grained text-based editing.

In addition, research is also needed to explore adaptive learning methods or new architectures that can weight the information that should be emphasized for a particular prompt, so that models adapt to more complex semantic requirements. The advances will ultimately make it possible to maintain high levels of precision and control in text-based image editing even in data-limited environments, providing reliable editing capabilities for a wide range of real-world applications.

VI. CONCLUSION

In this paper, we present a novel method for image variations and text-based editing, integrating a multi-scale framework with prompt-based learning for single-image generation. Our approach leverages the U-Net structure to produce diverse image variations while accurately preserving the fine details of the original image. Additionally, we utilize prompts generated by the BLIP model and pre-trained

CLIP to enhance text-image consistency. To further optimize model performance, we introduce a contrastive loss function, enabling the model to learn semantic differences between positive and negative prompts effectively.

The proposed method demonstrates its capacity to maintain visual consistency across various editing tasks while preserving the structural integrity and details of each image. Experimental results show that our approach outperforms existing models, including SinGAN, ConSinGAN, SinFusion, and SinDDM, in terms of image quality, alignment with textual prompts, and overall visual coherence. Specifically, our method excels in text-based image editing, delivering natural variations that reflect the original image's context without introducing unnecessary distortions.

This work significantly advances the performance of single-image generation models, offering a robust solution for producing high-quality images even in data-limited environments. Furthermore, it introduces a new approach to harmonizing text and images, achieving creative and consistent results across a range of applications. Our method fosters a deeper interaction between text and images, paving the way for innovative image variation techniques in various domains.

REFERENCES

[1] X. Xu, Z. Wang, E. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7754–7765.

[2] Z. Zhang, Y. Liu, C. Han, H. Shi, T. Guo, and B. Zhou, "PetsGAN: Rethinking priors for single image generation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 3408–3416.

[3] J. Chen, Q. Xu, Q. Kang, and M. Zhou, "MOGAN: Morphologic-structure-aware generative learning from a single image," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 4, pp. 2021–2033, Apr. 2024.

[4] Y. Jiang, L. Yan, X. Zhang, Y. Liu, and D. Sun, "TcGAN: Semantic-aware and structure-preserved GANs with individual vision transformer for fast arbitrary one-shot image generation," 2023, *arXiv:2302.08047*.

[5] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4570–4580.

[6] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "Improved techniques for training single-image GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1300–1309.

[7] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "SinDiffusion: Learning a diffusion model from a single natural image," 2022, *arXiv:2211.12445*.

[8] Y. Nikankin, N. Haim, and M. Irani, "SinFusion: Training diffusion models on a single image or video," 2022, *arXiv:2211.11743*.

[9] Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren, "SINE: SINgle image editing with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6027–6037.

[10] V. Kulikov, S. Yadin, M. Kleiner, and T. Michaeli, "SinDDM: A single image denoising diffusion model," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 17920–17930.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Munich, Germany: Springer, 2015, pp. 234–241.

[12] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[13] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," 2021, *arXiv:2102.07350*.

[14] P. Zhu, X. Wang, L. Zhu, Z. Sun, W.-S. Zheng, Y. Wang, and C. Chen, "Prompt-based learning for unpaired image captioning," *IEEE Trans. Multimedia*, vol. 26, pp. 379–393, 2023.

[15] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Jul. 2022, doi: 10.1007/s11263-022-01653-1.

[16] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[18] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[21] S. Rezvani, F. Soleymani Siahkar, Y. Rezvani, A. Alavi Gharahbagh, and V. Abolghasemi, "Single image denoising via a new lightweight learning-based model," *IEEE Access*, vol. 12, pp. 121077–121092, 2024.

[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

[23] H. Bao, W. Wang, L. Dong, Q. Liu, O. Khan Mohammed, K. Aggarwal, S. Som, and F. Wei, "VLMo: Unified vision-language pre-training with mixture-of-modality-experts," 2021, *arXiv:2111.02358*.

[24] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," 2022, *arXiv:2204.03610*.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[26] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.

[27] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2085–2094.

[28] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2LIVE: Text-driven layered image and video editing," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 707–723.

[29] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-guided diverse face image generation and manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2256–2265.

[30] F. Han, L. Si, H. Dong, L. Zhang, H. Chen, and B. Du, "Exploring text-guided single image editing for remote sensing images," 2024, *arXiv:2405.05769*.

[31] C. Fellbaum, "Wordnet," in *Theory and Applications of Ontology: Computer Applications*. Springer, 2010, pp. 231–243.

[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[33] L. Tsaban and A. Passos, "LEDITS: Real image editing with DDPM inversion and semantic guidance," 2023, *arXiv:2307.00522*.

[34] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18392–18402.

[35] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-Play diffusion features for text-driven image-to-image translation," 2022, *arXiv:2211.12572*.

[36] W. Kang, K. Galim, and H. Il Koo, "Eta inversion: Designing an optimal eta function for diffusion-based real image editing," 2024, *arXiv:2403.09468*.

[37] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," 2021, *arXiv:2104.08718*.

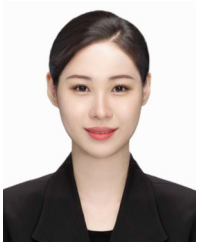
[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[39] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018, doi: 10.1109/TIP.2018.2831899. http://dx.doi.org/10.1109/TIP.2018.2831899

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

JIWON PARK was born in Busan, South Korea, in 2000. She received the B.S. degree in mathematics and data science from Gyeongsang National University, South Korea, in 2023. Currently, she is pursuing the M.S. degree in artificial intelligence with Chung-Ang University. Her research interest includes image generation and editing.

DASOL JEONG received the B.S. degree in electrical and electronic engineering from Ulsan University, South Korea, and the M.S. degree in AI imaging from Chung-Ang University, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in AI imaging. Her research interests include generative models and human re-identification.



HYEBEAN LEE (Graduate Student Member, IEEE) was born in Pohang-si, South Korea, in 1998. She received the B.S. degree in statistics from Cheongju University, South Korea, in 2021. Currently, she is pursuing the M.S. degree in artificial intelligence with Chung-Ang University. Her research interest includes image generation and editing.

SEUNGHEE HAN was born in Daegu, South Korea, in 2000. She received the B.S. degree in AI and big data engineering from Daegu Catholic University, South Korea, in 2023. Currently, she is pursuing the M.S. degree in AI imaging with Chung-Ang University. Her research interests include person re-identification and image generation.

JOONKI PAIK (Senior Member, IEEE) was born in Seoul, South Korea, in 1960. He received the B.S. degree in control and instrumentation engineering from Seoul National University, in 1984, and the M.S. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, USA, in 1987 and 1990, respectively.

He began his career with Samsung Electronics, from 1990 to 1993, where he played a key role in designing image stabilization chipsets for consumer camcorders. In 1993, he joined the Faculty of Chung-Ang University, Seoul, South Korea. From 1999 to 2002, he was a Visiting Professor with the Department of Electrical and Computer Engineering, The University of Tennessee, Knoxville. Since 2005, he has been the Director of the National Research Laboratory, South Korea, specializing in image processing and intelligent systems. He held the position of the Dean of the Graduate School of Advanced Imaging Science, Multimedia, and Film, from 2005 to 2007. Concurrently, he was the Director of the Seoul Future Contents Convergence Cluster. In 2008, he took on the role of a full-time Technical Consultant with the Systems LSI Division, Samsung Electronics. Here, he developed various computational photographic techniques, including an extended depth of field systems. He is currently a Professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film, Chung-Ang University. He has had a notable influence in scientific and governmental circles in South Korea.

Dr. Paik is a member of the Presidential Advisory Board for Scientific/Technical Policy with Korean Government and serves as a Technical Consultant for computational forensics with the Korean Supreme Prosecutor's Office. His accolades include being a two-time recipient of the Chester-Sall Award from the IEEE Consumer Electronics Society. He has also received the Academic Award from the Institute of Electronic Engineers of Korea and the Best Research Professor Award from Chung-Ang University. He has actively participated in various professional societies. He served the Consumer Electronics Society for IEEE in several capacities, including a member for the Editorial Board, the Vice President for the International Affairs, and the Director for the Sister and Related Societies Committee. In 2018, he was appointed as the President of the Institute of Electronics and Information Engineers. Since 2020, he has held the position of the Vice President of the Academic Affairs, Chung-Ang University. In an exceptional move in 2021, he simultaneously assumed the roles of the Vice President of Research and the Dean of the Artificial Intelligence Graduate School, Chung-Ang University, for a one-year term. Expanding his scope of responsibilities in 2022, he accepted a five-year appointment as the Project Manager of the Military AI Education Program under Korea's Department of Defense. With a career spanning over three decades, he has made significant contributions to the fields of image processing, intelligent systems, and higher education.