IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# Generating Diverse Image Variations with Diffusion Models by Combining Intra-Image Self-Attention and Decoupled Cross-Attention

**DASOL JEONG[1], DONGGOO KANG[1], JIWON PARK[2], AND JOONKI PAIK.[1,2,*]**

[1]Department of Image, Chung-Ang University, 84 Heukseok-ro, Seoul, 06974, Korea.
[2]Department of Artificial Intelligence, Chung-Ang University, 84 Heukseok-ro, Seoul, 06974, Korea.

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

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang University)) and This work was supported by Korea Research Institute for defense Technology planning and advancement through Defense Innovation Vanguard Enterprise Project, funded by Defense Acquisition Program Administration(R230106) and This research was supported by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea(NRF) funded by the Ministry of Science & ICT and Korea Customs Service(2021M3I1A1097911).

**ABSTRACT** In this paper, we present a novel integration of Decoupled Cross-Attention and Intra-Image Self-Attention within a diffusion model framework to generate diverse and coherent image variations. Our approach leverages the Decoupled Cross-Attention mechanism from IP-Adapter to align the input image more closely with its textual description, while Intra-Image Self-Attention operates on latent representations extracted through Denoising Diffusion Implicit Models inversion to capture fine-grained dependencies within the image. By utilizing noise interpolation in the diffusion process, we effectively blend the influences of both attention mechanisms, allowing for precise control over global and local features. This integration significantly improves both the semantic fidelity and visual diversity of generated images, making it highly suitable for applications that require detailed and contextually rich image synthesis. Through a three-step process—latent extraction, attention refinement, and noise interpolation—our method demonstrates superior performance compared to traditional models, consistently producing image variations that are visually appealing and aligned with input prompts. Our experiments show that the proposed method significantly outperforms traditional models in generating nuanced image variations, proving its effectiveness and potential for enhancing creative industries and personalized media production.

**INDEX TERMS** Diffusion Models, Intra-Image Self-Attention, Image Variations, Semantic Preservation, Realistic Image Adaptations

## I. INTRODUCTION

DIFFUSION-based approaches have gained substantial traction for their ability to produce highly realistic and diverse variations of input images [1]–[5]. These models are celebrated for their ability to generate high-quality images from textual descriptions, significantly benefiting creative industries, personalized media production, and data augmentation for training machine learning models. Despite the remarkable achievements of these models, producing complex image variations that remain faithful to the original prompts continues to pose a significant challenge. Conventional diffusion models frequently encounter challenges in preserving essential features across diverse variations, particularly within complex scenarios that necessitate the simultaneous maintenance of fine details and global coherence. This limitation restricts their practical applications in fields that require high fidelity and specific content adherence.

To address these issues, various approaches have attempted to leverage advanced attention mechanisms within diffusion models [6]–[12]. The Intra-Image Self-Attention (ISA) method, as seen in the Real-World Image Variation by Aligning the Diffusion Inversion Chain (RIVAL) [8], excels

at preserving the real-world context of the image by capturing dependencies within different regions of the image. Consequently, image variations are produced that preserve the structural and contextual details inherent in the original input. However, ISA exhibits a significant limitation: it does not achieve robust semantic alignment with the text prompt. Consequently, although the generated images preserve the contextual framework of the original input, they may fail to accurately reflect the specific modifications or additions articulated in the textual input. Therefore, in tasks requiring precise text-guided editing, ISA alone may not adequately capture the desired changes.

Conversely, Decoupled Cross-Attention (DCA), as employed within the IP-Adapter framework, provides a more effective alignment between text prompts and images, rendering it particularly advantageous for text-guided editing tasks. DCA ensures that textual instructions are accurately manifested in the visual content, resulting in modifications that closely align with the user's intent. However, a limitation of DCA resides in its potential to compromise the global coherence and realism of the image. By focusing heavily on textual alignment, DCA may alter the image in ways that disrupt its overall consistency and natural appearance, leading to results that, while textually accurate, lack visual plausibility.

These limitations indicate an inherent trade-off in existing methods between preserving the image's context and achieving precise text-to-image alignment. In our approach, we propose a novel method that integrates ISA and DCA to capitalize on the strengths of both mechanisms. While ISA ensures that the real-world context is maintained within the generated image variations, DCA enhances the alignment between textual input and visual content, improving the precision of text-to-image modifications. However, simply combining these mechanisms is not sufficient to achieve both contextual preservation and accurate editing. To address this, we introduce noise interpolation as a critical component that enables the smooth integration of the two attention mechanisms.

Noise interpolation [13] plays a key role in blending the effects of ISA and DCA during the diffusion process. As ISA operates on the latent noise extracted through Denoising Diffusion Implicit Models (DDIM) inversion [14], [15] inversion to preserve the internal structure and context of the image, DCA aligns the text with the visual content, optimizing the editing process. By interpolating between noise levels, our method dynamically adjusts the contribution of both attention mechanisms, allowing for a fine balance between maintaining the global coherence of the image and accurately aligning it with the text. This not only ensures the fidelity of real-world image characteristics but also facilitates the generation of semantically rich and contextually appropriate variations.

Moreover, our approach is training-free, meaning it does not require additional fine-tuning, making it more lightweight and adaptable to different diffusion model architectures. This is particularly important for real-world applications where computational efficiency and scalability are crucial.

**The primary contributions of this work include:**

- We propose a novel method that integrates Intra-Image Self-Attention (ISA) and Decoupled Cross-Attention (DCA) within the diffusion model framework.
- We introduce a dynamic noise interpolation mechanism that modulates the contributions of ISA and DCA during the diffusion process.
- Our method is training-free, meaning it does not require any additional training or fine-tuning of the diffusion models.

Through extensive experiments on multiple datasets, our approach demonstrates significant improvements over existing methods, effectively producing relevant and diverse image variations while preserving the semantic integrity of the original inputs. The overall architecture of our proposed generative model, integrating DCA and ISA mechanisms, is shown in Figure 1, illustrating how these mechanisms interact to create semantically rich and visually coherent image variations.

## II. RELATED WORK

### A. ADVANCEMENTS IN DIFFUSION MODELS AND TEXT-TO-IMAGE GENERATION

Recent advancements in diffusion models have notably improved the capability and flexibility of image generation systems. Furthermore, Diffusion-based image variation models [3], [4], [16], [17] have expanded the boundaries of image transformation by enabling the generation of diverse and contextually rich variations. Accordingly, the application areas of diffusion models have expanded to industry-specific domains such as creative industries and personal media products. For example, diffusion models have been utilized to generate creative interior design videos from texture-free 3D models [18], and to produce aesthetically pleasing and efficient interior designs [19]. Additionally, methods have been developed to generate interior designs directly from textual descriptions, demonstrating the versatility of diffusion models in creative design tasks [20].

Recent advances in diffusion models have introduced approaches that utilize inversion processes to manage and control image changes [6], [8], [21]. Zhang et al. [8] introduces a method to enhance real-world image variations by aligning the inversion process within diffusion models. Focused on maximizing control over the generative process, Cao et al. [6] presents a framework that allows users to dictate specific attributes of the generated images through controlled manipulation of the model's latent space.

Despite these successes, current diffusion models face significant challenges. A primary issue is the difficulty in maintaining semantic alignment between generated images and their corresponding text prompts while simultaneously preserving both fine-grained details and global coherence in complex scenarios. Traditional diffusion models often

struggle to control finer details when tasked with preserving specific attributes of the input images, such as identity in portraits or objects. Our research builds upon these findings, aiming to enhance the control and fidelity of image variations using novel attention mechanisms.

### B. ATTENTION MECHANISMS AND IMAGE GENERATION

Recent advancements in attention mechanisms have further enhanced deep learning models across various tasks [13], [22]–[24]. For instance, the Cascaded Attention Transformer Network [22] employs a cascaded attention mechanism to improve feature representation and has demonstrated remarkable results in image recognition tasks. Similarly, the Cascaded Visual Attention Network [24] integrates visual attention mechanisms in a cascaded fashion to refine feature maps progressively, enhancing object detection accuracy. These advancements underscore the importance of sophisticated attention mechanisms and innovative network architectures in improving image generation and editing tasks.

The integration of attention mechanisms in diffusion models [7], [14], [25]–[28] has significantly improved the control over, and the detail and relevance of, generated images. Zhang et al. [8] leverages self-attention within the image domain to capture dependencies between different regions of the image, thereby preserving structural and contextual details. This approach excels at maintaining the real-world context of the image, ensuring that generated variations remain consistent with the intrinsic features of the input. However, a significant limitation of self-attention is its lack of strong semantic alignment with textual prompts, making it less effective for tasks requiring precise text-guided modifications. Conversely, Ye et al. [7] introduces the IP Adapter mechanism, which applies decoupled cross-attention to align textual descriptions more accurately with their corresponding images. This alignment significantly refines the image synthesis process, allowing for more precise and contextually accurate image variations based on text prompts. Nonetheless, IP Adapter may compromise the global coherence and realism of the image by focusing heavily on textual alignment, potentially disrupting the original image's structure and natural appearance.

To address these limitations, our work proposes noise interpolation, a novel method that integrates attention mechanisms in diffusion models to balance the trade-off between preserving the original image context and achieving precise text alignment. By combining the strengths of self-attention and cross-attention through noise modulation, our approach ensures that the generated images remain faithful to both their structural integrity and textual guidance.

### C. NOISE INTERPOLATION APPROACHES

Denoising techniques have been widely studied in image processing, especially for enhancing the quality and fidelity of generated images. For instance, Rezvani et al. [29] introduces a novel approach to image denoising that effectively reduces noise without sacrificing image quality. While their work focuses primarily on single-image denoising, the concept of efficiently managing noise is central to our use of noise interpolation in diffusion models. Our approach leverages noise interpolation not only to refine the generated images but also to balance the integration of attention mechanisms, ensuring both global coherence and local detail. This approach allows for finer control over the noise reduction steps in the diffusion process, enabling more precise interventions at different stages of the image generation.

StyleGAN [30] developed by Karras et al., utilizes noise interpolation to transition between different learned styles smoothly, enabling the generation of highly realistic and diverse facial images. SDEdit [31] proposed by Meng et al., performs image editing by adding controlled noise to an input image and then denoising it conditioned on a new textual or visual prompt. Similarly, Blended Diffusion [32] introduced by Avrahami et al. leverages noise interpolation to seamlessly blend new content into existing images based on text descriptions. The noise interpolation allows for smooth integration of new elements, minimizing artifacts and preserving the overall quality of the image. Zheng et al. [33] introduces advanced noise correction techniques for image interpolation, utilizing noise interpolation in diffusion models to achieve more precise and visually coherent transitions between different images.

Despite their effectiveness, these approaches often struggle to fully balance image content preservation with precise semantic alignment to text prompts. They may have difficulty maintaining fine-grained details or global coherence when significant edits are required, and might not accurately capture complex dependencies between the text and various image regions.

These limitations underscore the need for an integrated approach that combines noise interpolation with advanced attention mechanisms. Our proposed method addresses this gap by integrating noise interpolation with both ISA and DCA within the diffusion framework. This integration facilitates the balancing of preserving the contextual integrity of images with achieving precise text alignment, thereby mitigating the trade-offs inherent in existing methodologies and advancing text-guided image generation and editing.

## III. PROPOSED METHOD

In this section, we present a novel training-free method that integrates Intra-Image Self-Attention (ISA) and Decoupled Cross-Attention (DCA) within the diffusion model framework, enhanced by a noise interpolation mechanism. Our approach does not require additional training or fine-tuning, making it highly efficient while still achieving a balance between image context preservation and precise text-to-image alignment. This integration effectively balances image context preservation and precise text-to-image alignment, overcoming the inherent trade-offs in existing methods.

The integration process consists of three main steps:

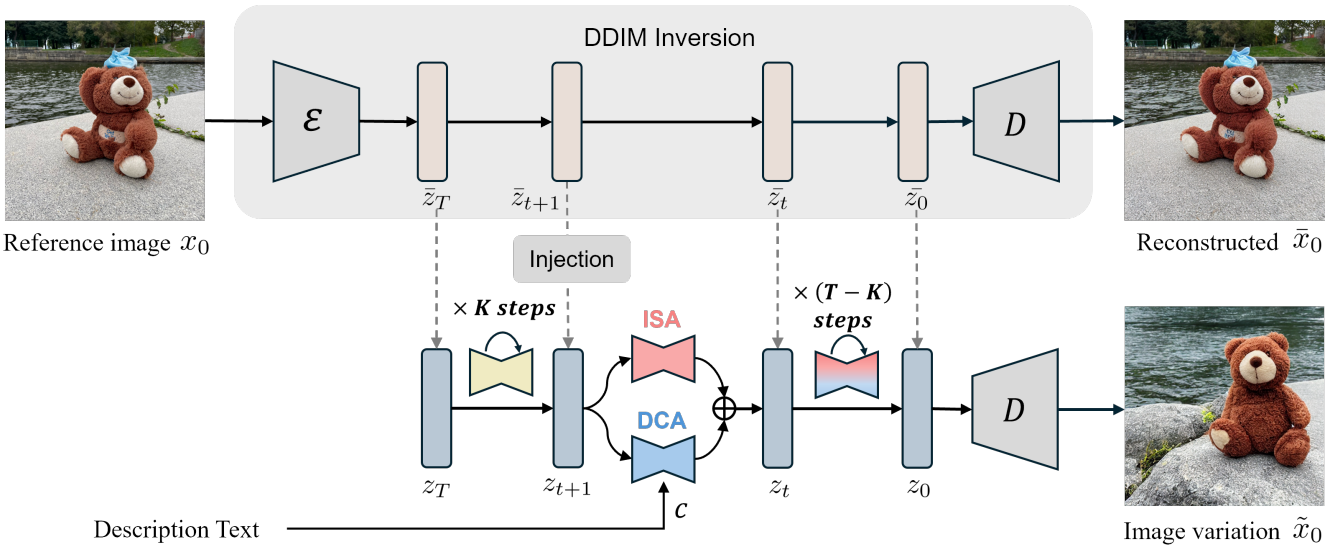


**FIGURE 1. Overview of the Proposed Architecture.** The process begins with the DDIM inversion of a reference image $x_0$ to extract latent representations $\tilde{z}_t$. The reconstructed image $\bar{x}_0$ is generated by reversing the diffusion process without any modifications, serving as a baseline. During the diffusion process, we inject ISA and DCA to modify the latent representations in a balanced manner. ISA preserves the structural and contextual coherence of the image, while DCA aligns the generated image with the given description text $c$. Noise interpolation dynamically balances the contributions of both attention mechanisms, ensuring that the image variations maintain both fidelity to the original and alignment with the textual prompt. The final output $\tilde{x}_0$ demonstrates how the model generates a new image variation based on the description text.

1) Initial latent extraction via DDIM Inversion: Utilizes controlled diffusion steps to invert the reference image into sequential latent representations
2) Parallel Processing via ISA and DCA: Once the latent representations are obtained, they are concurrently processed through two distinct attention mechanisms: Internal Spatial Attention (ISA) and Decoupled Cross-Attention (DCA)
3) Combining Results with Noise Interpolation: This critical step uses noise level manipulation within the diffusion process to integrate the influences of both attention mechanisms effectively.

Fig. 1 illustrates the overall architecture of the proposed method.

### A. PREMININARIES ON DIFFUSION MODELS

Conditional latent diffusion models [2], [34], [35] are a class of generative models designed to synthesize images conditioned on given inputs, such as textual descriptions or specific attributes. These models iteratively convert a random noise distribution into a structured output that aligns with the conditioned inputs. The objective function for the latent diffusion process [1], [36], [37] in these models is expressed as:

$$L(\theta) = \mathbb{E}_{\mathcal{E}(x_0),\epsilon,t}\left[|\epsilon - \epsilon_\theta(z_t, t, c)|_2^2\right], \tag{1}$$

where $x_0$ is the original image from the data distribution. $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise sampled from a standard normal distribution. $t \in 1, \ldots, T$ represents the timestep in the diffusion process. $c$ denotes the conditioning information, such as textual prompts. $z_t$ is the noisy latent representation of the image at timestep $t$, obtained by adding noise to $x_0$. $\epsilon_\theta(z_t, t, c)$ is the noise prediction function parameterized by $\theta$, typically implemented as a neural network.

The forward diffusion process gradually adds noise to the original image to produce $z_t$. This is defined by:

$$z_t = \sqrt{\alpha_t} z_{t-1} + \sqrt{1 - \alpha_t}\epsilon_t, \tag{2}$$

where $z_{t-1}$ is the data at the previous timestep, $\alpha_t$ is a predetermined variance schedule, and $\epsilon_t$ represents Gaussian noise. The reverse diffusion process aims to reconstruct the original image by iteratively denoising $z_t$ using the learned noise prediction function. In the case of Denoising Diffusion Implicit Models (DDIM) [38], the reverse process is defined to provide more deterministic and efficient sampling:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}}\epsilon_\theta(z_t, t, c)\right) + \sigma_t \cdot z, \tag{3}$$

where $\alpha_t$ denotes a predetermined variance schedule, $t, \epsilon_\theta(z_t, t, c)$ is the learned noise prediction function, $\sigma_t$ is the standard deviation of noise added at each step, and $z$ is noise sampled from a standard normal distribution.

DDIM introduces a non-random, deterministic mapping in the reverse process by eliminating the stochastic noise component present in other diffusion models. This modification allows for more controlled and predictable image generation, which is crucial for maintaining high-level features and semantic integrity when generating images from textual descriptions. To utilize DDIM in the context of

image editing and variation generation, we employ DDIM Inversion [8], [25]. DDIM inversion maps a given reference image $x_0$ back into its corresponding latent representations $\bar{z}_t$ through the forward diffusion process, but using the learned noise prediction function to estimate the added noise at each step. The inversion is performed using:

$$\bar{z}_{t-1} = \sqrt{\alpha_{t-1}} x_0 + \sqrt{1-\alpha_t}\epsilon_\theta(\bar{z}_t, t, c), \tag{4}$$

where $\alpha_t$ dictates the variance schedule of the noise model, which decreases progressively throughout the diffusion process. The noise term $\epsilon_\theta(\bar{z}_t, t, c)$, estimated by the neural network, is pivotal for effectively reversing the diffusion path. This precise estimation allows for the methodical reduction of noise and enhancement of image details at each step, facilitating a high-fidelity reconstruction of the original image for specific conditions, denoted by $c$. Moreover, DDIM inversion enhances the stability and predictability of the denoising path, which is particularly effective for maintaining consistency across generated image variations and enabling meaningful semantic interpolations. By leveraging the capabilities of DDIM and its inversion process, our method establishes a robust foundation for integrating advanced attention mechanisms, such as ISA and DCA, to achieve superior text-guided image generation and editing.

### B. INTRA-IMAGE SELF-ATTENTION (ISA)

The Intra-Image Self-Attention (ISA) mechanism enhances the model's ability to preserve structural and contextual details of the original image by capturing dependencies between different spatial regions. This mechanism is crucial for maintaining fine-grained details and global coherence in the generated images, ensuring that variations remain consistent with the intrinsic features of the input.

After executing the DDIM inversion process using a reference image, we extract sequential latent representations $\bar{z}_t$, we incorporate these representations into the diffusion process to reinforce features of the original images. At each timestep $t$, we concatenate the inverted latent representation $\bar{z}_t$ with the current noisy latent state $z_t$. The ISA mechanism computes self-attention over $\bar{z}_t \oplus z_t$ to update the latent representation, allowing the model to consider relationships between different regions within the image. Specifically, we project $\tilde{z}_t$ to obtain the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$:

$$\mathbf{Q} = W_Q(z_t), \mathbf{K} = W_K(\bar{z}_t \oplus z_t), \mathbf{V} = W_V(\bar{z}_t \oplus z_t), \tag{5}$$

where $\oplus$ denotes the concatenation operation along the feature dimension. $W_Q$, $W_K$, and $W_V$ are learnable weight matrices for the query, key, and value projections, respectively.

The self-attention output is then computed as:

$$z_t{}' = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{6}$$

where $d_k$ is the dimensionality of the key vectors. The scaling factor $\sqrt{d_k}$ stabilizes the dot-product values, and the softmax function ensures that the attention weights sum to one, effectively computing a weighted sum of the value vectors.

This self-attention operation allows each position in the latent representation to attend to all other positions, capturing long-range dependencies within the image. By integrating information from $\bar{z}_t$, the model reinforces the structural and contextual features of the original image, aiding in preserving important details during the generation process. The updated latent representation $z_t'$ thus encapsulates both the original image context and the necessary adjustments to facilitate high-quality image synthesis.

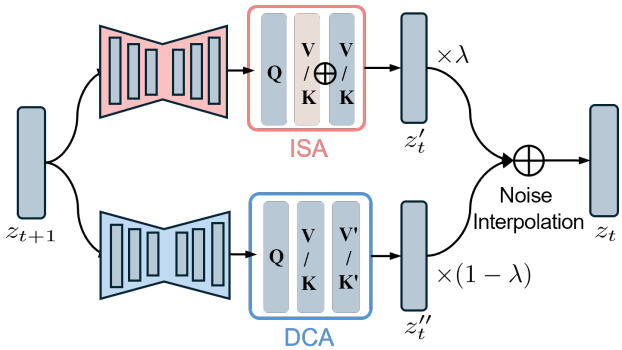


**FIGURE 2. Integration of ISA and DCA through Noise Interpolation.** This diagram illustrates the Noise Interpolation process Between Intra-Image Self-Attention (ISA) and Decoupled Cross-Attention (DCA). At each step $z_{t+1}$, latent representations are processed through ISA to preserve the image's structural and contextual information and through DCA to enhance text-image alignment. The outputs from both attention mechanisms, $z_t'$ and $z_t''$, are weighted using a noise interpolation factor $\lambda$ to control the contributions of each mechanism. Noise interpolation dynamically blends the two results to achieve a balance between maintaining image coherence (via ISA) and ensuring semantic alignment with the text prompt (via DCA). The final interpolated result $z_t$ represents the refined latent representation used for image generation.

### C. DECOUPLED CROSS-ATTENTION

The Decoupled Cross-Attention (DCA) mechanism is employed to enhance the semantic alignment between the latent image representations and the associated textual descriptions. DCA integrates textual information into the image generation process, allowing the model to accurately reflect the semantic content of the text prompt in the synthesized images.

At each timestep $t$ in the diffusion process, we compute the queries $\mathbf{Q}$ from the current latent representation $z_t$ and derive keys and values from the textual embeddings. Specifically, the DCA mechanism operates by processing the latent representation and textual embeddings through separate attention pathways, enabling the model to consider both the primary text prompt and additional contextual information simultaneously. The mathematical formulation of DCA is described as follows:

$$\begin{aligned} Q = W_Q(z_t),\ K = W_K(c_p),\ V = W_V(c_p), \\ K' = {W_K}'(c_i),\ V' = {W_V}'(c_i), \end{aligned} \tag{7}$$

where $Q$ is composed of queries derived from latent representation using the weight matrix $W_Q$. $K$ and $V$ are generated

from prompt-embedded features $c_p$, processed through the transformation matrices $W_K$ and $W_V$, respectively. Additionally, to enrich the model's capability to integrate image-specific contexts, $K'$ and $V'$ are generated from alternative textual or contextual embeddings $c_i$ using distinct transformation matrices ${W_K}'$ and ${W_V}'$. This dual-path setup allows the model to refine its focus on both text-based and image-derived features concurrently.

The DCA mechanism computes attention scores between the latent representation and both sets of textual embeddings, effectively allowing the model to attend to multiple sources of textual information. The cross-attention outputs are computed as:

$$z_t{}'' = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V + \text{softmax}\left(\frac{QK'^\top}{\sqrt{d_k}}\right)V'. \quad (8)$$

The updated latent representation $z_t''$ integrates semantic information from multiple textual sources by summing the attention outputs from both pathways. This integration enhances the model's ability to generate images that are semantically rich and accurately aligned with the text prompts. The DCA mechanism effectively decouples the attention processes for different types of textual information, allowing the model to handle complex and nuanced instructions. By processing $c_p$ and $c_i$ separately and then combining their influences, the model can capture both the primary directives and additional contextual cues provided in the text. Incorporating $z_t''$ into the diffusion process allows the model to produce images that not only align closely with the semantic content of the text prompts but also maintain the structural coherence of the original image. This approach enhances the diversity and relevance of the generated images, enabling precise text-guided modifications without compromising visual quality.

### D. INTEGRATION USING NOISE INTERPOLATION

We incorporate an interpolated noise denoising technique to further enhance the effectiveness of our integrated approach combining Intra-Image Self-Attention (ISA) and Decoupled Cross-Attention (DCA). To leverage the complementary strengths of ISA and DCA, we interpolate the latent representations obtained from both mechanisms. This interpolation is controlled by a factor $\lambda$, which determines the balance between the contributions of ISA and DCA.

Initially, we extract the latent representations $z_t$ from the input image using the DDIM inversion process. These representations contain inherent noise that needs to be addressed to improve the image quality. For each timestep $t$, we interpolate the noise between two consecutive latent representations. ISA output ($z_t'$) emphasizes intra-image consistency and preserves structural details by capturing dependencies within the image. DCA output ($z_t''$) enhances semantic alignment with the text prompt by integrating textual information into the latent representation. To integrate these representations, we perform interpolation controlled by an interpolation factor $\lambda_t \in [0, 1]$, which can be dynamically adjusted based on the timestep or other criteria. The interpolated latent representation $\tilde{z}_t$ is computed as:

$$\tilde{z_t} = \lambda {z_t}' + (1 - \lambda){z_t}'', \quad (9)$$

where $\tilde{z_t}$ is the final interpolated latent representation at timestep $t$, and $\lambda$ determines the relative contributions of ISA and DCA at timestep $t$. After obtaining the interpolated latent representation $\tilde{z}_t$, we proceed with the denoising step using the diffusion model's denoising function, as defined by Eq. 3 The updated latent representation is computed as:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\tilde{z}t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(\tilde{z}_t, t, c)\right) + \sigma_t \cdot z, \quad (10)$$

where $\epsilon_\theta(\tilde{z}_t, t, c)$ is the predicted noise component, and $\sigma_t \cdot z$ accounts for the stochasticity in the diffusion process, with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. This interpolation process is iteratively applied at each timestep $t$ in the reverse diffusion process. The iterative refinement ensures that the latent representations progressively integrate the benefits of both attention mechanisms, resulting in synthesized images that maintain structural coherence and exhibit precise semantic alignment with the text prompt.

Through the meticulous integration of latent representations processed by both ISA and DCA, our method leverages the detailed and context-rich information available in the latent space. The resultant images are not only a reflection of the input's semantic integrity but also demonstrate the model's ability to innovate visually within the constraints of the given context. This innovative method highlights the potential of combining and interpolating latent representations obtained from different attention mechanisms, paving the way for more sophisticated and effective generative image models.

### E. ALGORITHMIC FRAMEWORK

To address the challenges identified in the synthesis of image variations, we developed a novel algorithmic approach, encapsulated in Algorithm 1. This approach leverages both Decoupled Cross-Attention (DCA) and Intra-Image Self-Attention (ISA) to enhance the fidelity and coherence of generated images.

The core steps of the algorithm, detailed in Algorithm 1, systematically integrate these attention mechanisms to process and refine image features effectively. This integration is pivotal in optimizing the alignment between the visual features and the associated textual descriptions, ensuring that the generated images are both contextually relevant and visually coherent.

## IV. EXPERIMENTAL RESULTS

This section details the implementation and evaluation of our proposed method, outlining the experimental setup, comparative analyses, and discussions surrounding the outcomes. The experiments are designed to demonstrate the effectiveness

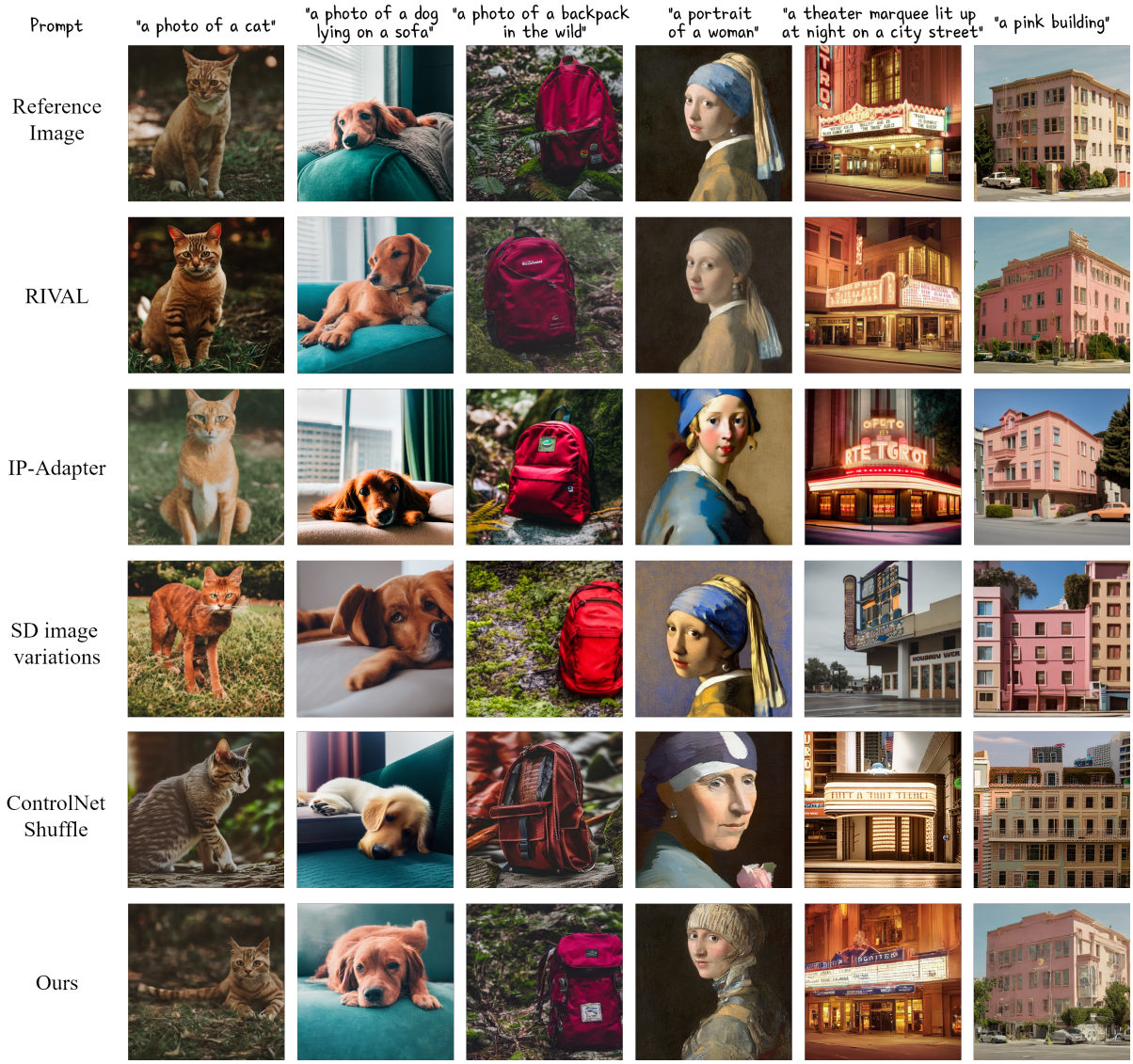


**FIGURE 3. Qualitative comparison of image variation results.** The reference image in the first row is transformed according to the input prompts listed above each column. Our model demonstrates improved fidelity and alignment to the input prompts compared to RIVAL, IP-Adapter, SD image variations, and ControlNet Shuffle.

of integrating Decoupled Cross-Attention (DCA) and Intra-Image Self-Attention (ISA) within diffusion models for generating image variations.

### A. IMPLEMENTATION DETAILS

In our experiments, we utilized Stable Diffusion V1.5 [39] as the foundational model, consistent across all comparisons with RIVAL and IP-Adapter frameworks to ensure uniformity in our evaluations. The experiments were conducted using a high-quality dataset of images, selected to represent a diverse range of scenes and objects from the RIVAL [8] and DreamBooth [40]. We conducted image inversion and generation using DDIM sample steps ($T = 50$) per image. After extensive empirical evaluation, we set the classifier-free guidance scale to $m = 7$. This value was chosen to optimize the trade-off between maintaining the structural integrity of the original image and ensuring precise semantic alignment with the text prompts, following prior studies [41]. Lower guidance scales resulted in insufficient incorporation of text-based modifications, while higher scales led to excessive alterations that compromised image fidelity. All computations were performed on an NVIDIA RTX 4090 GPU.

### B. QUANTITATIVE COMPARISON

The comparative results presented in Table IV-A provide a comprehensive overview of the performance of various image variation methods, including our proposed approach, RIVAL [8], IP-Adapter [7], SD image-variations [42] and ControlNet Shuffle [43]. We assess the performance using widely adopted metrics including: Learned Perceptual Image

**TABLE 1. Performance Comparison of Image Variation Methods.** This table compares the performance of different image variation methods (RIVAL, IP-Adapter, SD image-variations, ControlNet Shuffle and Ours) based on four key metrics: LPIPS (lower is better), SSIM (higher is better), SIFID (lower is better), and CLIP similarity (higher is better). To highlight important results, the best results are in **bold**

| Metric | LPIPS ↓ | SSIM ↑ | SIFID ↓ | CLIP sim. ↑ |
|---|---|---|---|---|
| RIVAL [8] | **0.4488** | **0.5851** | **0.0133** | **0.8772** |
| IP-Adapter [7] | 0.5329 | 0.2390 | 0.0228 | 0.7526 |
| SD image-variations [42] | 0.6401 | 0.5248 | 0.0212 | 0.8545 |
| ControlNet Shuffle [43] | 0.5410 | 0.5458 | 0.0307 | 0.8029 |
| Ours | **0.5413** | **0.5740** | **0.0150** | **0.8822** |

**Algorithm 1** Integrating DCA and ISA

**Input:** Reference image $x_0$, prompt embedding $\mathcal{P}$
**Output:** Synthesized image variations $\tilde{x}_0$
Compute the intermediate results $\bar{z}_T, \ldots, \bar{z}_0$ using DDIM inversion over $x_0$
Initialize $z_T \leftarrow \bar{z}_T$
**for** $t = T$ to 1 step $-1$ **do**
  **if** $t > K$ **then**
    *Only perform attention if* $t > K$
    *Intra-Image Self-Attention (ISA)*
    $Q_{isa} \leftarrow W_Q(z_t)$
    $K_{isa}, V_{isa} \leftarrow W_K(\bar{z}_t \oplus z_t), W_V(\bar{z}_t \oplus z_t)$
    $z'_t \leftarrow \text{softmax}\left(\frac{Q_{isa}K_{isa}^\top}{\sqrt{d_k}}\right) V_{isa}$
    *Decoupled Cross-Attention (DCA)*
    $Q_{dca} \leftarrow W_Q(z_t)$
    $K_{dca}, V_{dca} \leftarrow W_K(\mathcal{P}), W_V(\mathcal{P})$
    $z''_t \leftarrow \text{softmax}\left(\frac{Q_{dca}K_{dca}^\top}{\sqrt{d_k}}\right) V_{dca}$
    *Combine results using noise interpolation*
    $\lambda \leftarrow \text{adjust\_interpolation\_factor}(t, T)$
    $\tilde{z}_t \leftarrow \lambda z'_t + (1 - \lambda) z''_t$
  **else**
    *Skip attention mechanisms*
  **end if**
**end for**
$\tilde{x}_0 \leftarrow \mathcal{D}(z_T)$
**Return** final image $\tilde{x}_0$

Patch Similarity (LPIPS) [44], Structural Similarity Index Measure (SSIM) [45], and Single Image Fréchet Inception Distance (SIFID) [46]. In addition, we incorporated the CLIP similarity score [47] to gauge the conceptual alignment of the generated images with textual descriptions, where higher scores reflect a better match.

The distinct advantage of our proposed method lies in its exceptional performance in CLIP similarity, which underscores its ability to generate images that are not only visually appealing but also contextually aligned with their descriptions. This feature is particularly beneficial in applications such as digital marketing, where images must closely reflect the textual content to convey a message to the target audience effectively. While our method consistently performs at a high level across LPIPS, SSIM, and SIFID metrics, achieving competitive results in perceptual quality and structural integrity, CLIP similarity stands out as the key differentiator. This demonstrates that our approach excels in producing contextually relevant and meaningful image variations, which is the primary objective in image variation tasks. Moreover, the relatively higher LPIPS score can be seen as an indication of greater diversity and creativity in the generated images, offering a wider range of variations while maintaining strong alignment with the text.

### C. QUALITATIVE COMPARISON

To further demonstrate the effectiveness of our proposed method, we present a qualitative comparison of the image variation results generated by our approach and two state-of-the-art methods, RIVAL [8], IP Adapter [7], SD image-variation [42], and ControlNet Shuffle [43]. As illustrated in Fig. 3, we use a series of diverse text prompts applied to a reference image to evaluate the alignment between the generated images and the input prompts. The prompts range from simple object descriptions (e.g., "a photo of a cat") to more complex scenes (e.g., "a theater marquee lit up at night on a city street"). Our model consistently outperforms competing methods in terms of preserving the core characteristics of the reference image while adhering to the given text prompt. RIVAL tends to generate images that retain too much of the reference image's identity, failing to fully adapt to the new prompt context. IP-Adapter shows a moderate adaptation but struggles with fine details, often producing images that only loosely align with the prompt. SD image variations display inconsistency in maintaining visual fidelity to the original image, particularly in texture and object coherence. ControlNet Shuffle introduces noticeable artifacts in more complex prompts and demonstrates limited capacity for precise control over the transformation process. In contrast, our model successfully captures both the high-level semantic meaning of the prompt and the essential visual features of the reference image. Our approach preserves the contextual relevance of the original image while presenting a transformed result of the prompt's target.


**FIGURE 4.** Ablation study results showing the effect of different noise interpolation coefficients $\lambda$ on the generated image variations. The first row displays the input images. The second and third rows show the generated images with interpolation coefficients $\lambda = 0.5$ and $\lambda = 0.7$, respectively. The fourth row presents the results of using a different prior model with an interpolation coefficient of $\lambda = 0.7$. This study demonstrates how varying the interpolation coefficient and changing the prior model affect the diversity and coherence of the generated images.

### D. COMPUTATIONAL EFFICIENCY ANALYSIS

We compared the computational time required by our method and the traditional approach for processing a single image. The results show that our method takes an average of 8.6096 seconds per image, while the traditional method takes 8.3894 seconds per image, resulting in a slight increase of 0.2202 seconds per image, or 2.63%. Despite this marginal increase, the substantial improvements in image quality and semantic alignment achieved by our method make this trade-off in computational time worthwhile.

Additionally, as our approach is training-free, it avoids the need for the extra computational resources and time required for model training or fine-tuning in other methods. This makes our method highly efficient and practical for real-world applications.

### E. ABLATION STUDY

To evaluate the effectiveness of our proposed method, we conducted an ablation study focusing on the impact of different noise interpolation coefficients $\lambda$ and the use of a different prior model for the prompt encoder. Fig. 4 illustrates the results of this study.

The first row in Fig. 4 shows the input images used for generating variations. These images serve as the baseline for evaluating the quality and diversity of the outputs produced by different configurations.

The second row displays the generated images with a noise interpolation coefficient of $\lambda = 0.5$. At this interpolation level, the generated images exhibit a moderate level of adherence to the original input characteristics while introducing some degree of variation. However, the structural integrity and specific features of the original images are partially preserved.

The third row presents the results with a higher interpolation coefficient of $\lambda = 0.7$. As the coefficient increases, the generated images maintain the original input's characteristics more effectively. This indicates that higher values of $\lambda$ lead to better preservation of the semantic and visual features of the original images, resulting in variations that are more coherent and closely aligned with the input.

The fourth row showcases the results using a different prior model [48], specifically utilizing the Kandinsky prompt encoder, with an interpolation coefficient of $\lambda = 0.7$. This configuration demonstrates a significant improvement in the semantic quality of the generated images. The use of the Kandinsky prompt encoder effectively captures and reflects the semantic meaning of the input prompts, producing variations that are not only visually coherent but also semantically rich.

Overall, the ablation study highlights two key findings:

Increasing the noise interpolation coefficient $\lambda$ facilitates enhanced preservation of the original image's characteristics, thereby resulting in more coherent variations. Employing the Kandinsky prompt encoder as the prior model significantly improves the semantic consistency of the generated images, ensuring that the variations accurately reflect the intended meaning of the input prompts. These results validate our approach, demonstrating the importance of careful selection of interpolation coefficients and prompt encoders in achieving high-quality, diverse, and semantically meaningful image

variations.

## V. LIMITATION AND FUTURE WORK

Although our method demonstrates significant improvements, it has some limitations that open opportunities for future research. First, the approach relies on pre-trained diffusion models and attention mechanisms, inheriting potential limitations and biases from these models, which could impact the quality, fairness, or diversity of the generated images. Our method also faces challenges when applied to highly complex scenes or detailed textual prompts, potentially affecting the fidelity and coherence of the generated images in such cases. Furthermore, the current approach is primarily designed for image synthesis, and significant adaptations would be required to extend the framework to other data modalities, such as video or 3D content generation.

Future research could focus on optimizing computational efficiency through the development of more efficient attention mechanisms or model compression techniques. Moreover, exploring the extension of our framework to other modalities, such as video generation or 3D modeling, could broaden its application scope. Investigating alternative strategies for integrating ISA and DCA, such as adaptive weighting schemes or novel architectures, may further enhance performance and flexibility. Addressing inherited biases from pre-trained models and implementing fairness-aware techniques will also be crucial for improving the ethical aspects of the generated content. Lastly, conducting comprehensive user studies to evaluate the perceptual quality of the generated images, and enhancing robustness across a broader range of inputs, could provide valuable insights for future refinements of the method.

## VI. CONCLUSION

In this paper, we introduced a novel training-free method for image variation, which integrates Decoupled Cross-Attention (DCA) and Intra-Image Self-Attention (ISA) within the diffusion model framework, enhanced by a noise interpolation mechanism. Our approach leverages the critical challenge of balancing image context preservation with precise text-to-image alignment in diffusion-based image generation and editing tasks.

By leveraging ISA, our method preserves the structural integrity and fine-grained details of the original image, capturing dependencies within different spatial regions. Concurrently, DCA enhances semantic alignment between the latent image representations and associated textual descriptions, enabling accurate reflection of the text prompts in the synthesized images. The introduction of noise interpolation dynamically balances the contributions of ISA and DCA during the denoising process, allowing for a fine-tuned integration that overcomes the inherent trade-offs present in existing methods. One of the significant advantages of our approach is that it operates without the need for additional training or fine-tuning, making it highly efficient and adaptable to different diffusion model architectures. This training-free characteristic is particularly valuable for real-world applications where computational resources and scalability are crucial considerations.

Experiments demonstrate that our method outperforms baseline approaches in terms of both image fidelity and semantic alignment. The generated images not only maintain the global coherence and fine details of the original content but also accurately reflect the semantic nuances of the textual prompts. This balance between visual quality and semantic accuracy is essential for applications in creative industries, personalized media production, and specialized fields. In this paper, we introduced a novel training-free method for image variation, which integrates DCA and ISA within the latent space of pre-trained diffusion models. Our methodology leverages the unique strengths of these attention mechanisms to enhance the fidelity and coherence of generated images, demonstrating significant improvements over existing methods.

## REFERENCES

[1] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," arXiv preprint arXiv:2011.13456, 2020.

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in neural information processing systems, vol. 35, pp. 36 479–36 494, 2022.

[3] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint arXiv:2112.10741, 2021.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, vol. 1, no. 2, p. 3, 2022.

[5] S. Witteveen and M. Andrews, "Investigating prompt engineering in diffusion models," arXiv preprint arXiv:2211.15462, 2022.

[6] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22 560–22 570.

[7] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," arXiv preprint arXiv:2308.06721, 2023.

[8] Y. Zhang, J. Xing, E. Lo, and J. Jia, "Real-world image variation by aligning diffusion inversion chain," Advances in Neural Information Processing Systems, vol. 36, 2024.

[9] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-p2p: Video editing with cross-attention control," arXiv preprint arXiv:2303.04761, 2023.

[10] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15 932–15 942.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[12] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18 392–18 402.

[13] S. Rezvani, M. Fateh, and H. Khosravi, "Abanet: Attention boundary-aware network for image segmentation," Expert Systems, vol. 41, no. 9, p. e13625, 2024.

[14] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6038–6047.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," Advances in neural information processing systems, vol. 34, pp. 8780–8794, 2021.
[16] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in European Conference on Computer Vision. Springer, 2022, pp. 89–106.
[17] D. Li, J. Li, and S. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," Advances in Neural Information Processing Systems, vol. 36, 2024.
[18] Z. Shao, J. Chen, H. Zeng, W. Hu, Q. Xu, and Y. Zhang, "A new approach to interior design: Generating creative interior design videos of various design styles from indoor texture-free 3d models," Buildings, vol. 14, no. 6, p. 1528, 2024.
[19] J. Chen, Z. Shao, X. Zheng, K. Zhang, and J. Yin, "Integrating aesthetics and efficiency: Ai-driven diffusion models for visually pleasing interior design generation," Scientific Reports, vol. 14, no. 1, p. 3496, 2024.
[20] J. Chen, Z. Shao, and B. Hu, "Generating interior design from text: A new diffusion model-based method for efficient creative design," Buildings, vol. 13, no. 7, p. 1861, 2023.
[21] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 10 146–10 156.
[22] W. Zhang, G. Chen, P. Zhuang, W. Zhao, and L. Zhou, "Catnet: Cascaded attention transformer network for marine species image classification," Expert Systems with Applications, vol. 256, p. 124932, 2024.
[23] W. Zhang, Z. Li, G. Li, P. Zhuang, G. Hou, Q. Zhang, and C. Li, "Gacnet: Generate adversarial-driven cross-aware network for hyperspectral wheat variety identification," IEEE Transactions on Geoscience and Remote Sensing, 2023.
[24] W. Zhang, W. Zhao, J. Li, P. Zhuang, H. Sun, Y. Xu, and C. Li, "Cvanet: Cascaded visual attention network for single image super-resolution," Neural Networks, vol. 170, pp. 622–634, 2024.
[25] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," arXiv preprint arXiv:2208.01626, 2022.
[26] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," ACM Transactions on Graphics (TOG), vol. 42, no. 4, pp. 1–10, 2023.
[27] D. Epstein, A. Jabri, B. Poole, A. Efros, and A. Holynski, "Diffusion self-guidance for controllable image generation," Advances in Neural Information Processing Systems, vol. 36, pp. 16 222–16 239, 2023.
[28] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in ACM SIGGRAPH 2023 Conference Proceedings, 2023, pp. 1–11.
[29] S. Rezvani, F. S. Siahkar, Y. Rezvani, A. A. Gharahbagh, and V. Abolghasemi, "Single image denoising via a new lightweight learning-based model," IEEE Access, 2024.
[30] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
[31] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," arXiv preprint arXiv:2108.01073, 2021.
[32] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18 208–18 218.
[33] P. Zheng, Y. Zhang, Z. Fang, T. Liu, D. Lian, and B. Han, "Noisediffusion: Correcting noise for image interpolation with diffusion models beyond spherical linear interpolation," in The Twelfth International Conference on Learning Representations, 2024.
[34] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," arXiv preprint arXiv:2307.01952, 2023.
[35] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein et al., "Muse: Text-to-image generation via masked generative transformers," arXiv preprint arXiv:2301.00704, 2023.
[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.
[37] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 4, pp. 4713–4726, 2022.
[38] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," arXiv preprint arXiv:2010.02502, 2020.
[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695.
[40] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22 500–22 510.
[41] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.
[42] J. Pinkney, "Experiments with stable diffusion," https://huggingface.co/lambdalabs/sd-image-variations-diffusers.
[43] "Controlnet 1.1 shuffle." [Online]. Available: https://github.com/lllyasviel/ControlNet-v1-1-nightly
[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.
[46] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Advances in neural information processing systems, vol. 30, 2017.
[47] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08718, 2021.
[48] A. N. V. A. I. P. A. K. D. D. Arseniy Shakhmatov, Anton Razzhigaev, "kandinsky 2.2," 2023. [Online]. Available: https://huggingface.co/kandinsky-community/kandinsky-2-2-prior

DASOL JEONG received a B.S degree in Electrical and Eletrocnic Engineering from Ulsan University, South Korea. Also, he received the M.S. degree in AI Imaging at Chung-Ang University, South Korea, in 2020. Currently, he is pursuing a Ph.D. degree in AI Imaging at Chung-Ang University. Her research includes generative models and human re-identification.

DONGGOO KANG was born in Seoul, Korea, in 1992. He received the B.S. degree in Financial Economics from Seokyeong University, South Korea, in 2018. Also, he received the M.S. degree in AI Imaging at Chung-Ang University, South Korea, in 2020. Currently, he is pursuing a Ph.D. degree in AI Imaging at Chung-Ang University. His research interests include computational photography and human-object interaction discovery.

JIWON PARK was born in Busan, Korea, in 2000. She received a B.S degree in Mathematics and Data Science from Gyeongsang National University, South Korea in 2023. Currently, she is pursuing an M.S. degree in Artificial Intelligence at Chung-Ang University. Her research interests include image generation and editing.

JOONKI PAIK was born in Seoul, Korea, in 1960. He completed his BS degree in Control and Instrumentation Engineering from Seoul National University in 1984. He continued his education in the United States, earning MS and Ph.D. degrees in Electrical Engineering and Computer Science from Northwestern University in 1987 and 1990, respectively.

Dr. Paik began his career at 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 at Chung-Ang University in Seoul, Korea. He is currently a professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film at the university.

From 1999 to 2002, he served as a visiting professor in the Department of Electrical and Computer Engineering at the University of Tennessee, Knoxville. Since 2005, Dr. Paik has been the director of a national research laboratory in Korea specializing in image processing and intelligent systems. He held the position of Dean for the Graduate School of Advanced Imaging Science, Multimedia, and Film from 2005 to 2007 and concurrently served as the director of the Seoul Future Contents Convergence Cluster.

In 2008, Dr. Paik took on the role of a full-time technical consultant for the Systems LSI Division of Samsung Electronics. Here, he developed various computational photographic techniques, including an extended depth of field system.

Dr. Paik has had a notable influence in scientific and governmental circles in Korea. He is a member of the Presidential Advisory Board for Scientific/Technical Policy with the 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 of the IEEE in several capacities, including as a member of the Editorial Board, Vice President of International Affairs, and Director of Sister and Related Societies Committee. In 2018, he was appointed as the president of the Institute of Electronics and Information Engineers.

Since 2020, Dr. Paik has held the position of Vice President of Academic Affairs at Chung-Ang University. In an exceptional move in 2021, he simultaneously assumed the roles of Vice President of Research and Dean of the Artificial Intelligence Graduate School at Chung-Ang University for a one-year term. Expanding his scope of responsibilities in 2022, Dr. Paik accepted a five-year appointment as Project Manager for the Military AI Education Program under Korea's Department of Defense.

With a career spanning over three decades, Dr. Joonki Paik has made significant contributions to the fields of image processing, intelligent systems, and higher education.