

RESEARCH ARTICLE

A Novel Scheme for Generating Context-Aware Images Using Generative Artificial Intelligence

HYUNJO KIM¹, JAE-HO CHOI², AND JIN-YOUNG CHOI¹¹Korea University, Seoul 02841, Republic of Korea²MCCA AI Company Ltd., Republic of Korea

Corresponding author: Jin-Young Choi (choi@formal.korea.ac.kr)

ABSTRACT Humans possess the remarkable capacity to comprehend narratives presented in text and subsequently conjure associated mental images through their imagination. This cognitive ability enhances their grasp of the content and augments their overall enjoyment. Consequently, the development of an automated system aimed at producing visually faithful images based on textual descriptions, often referred to as the text-to-image task, stands as a profoundly meaningful endeavor. For this reason, a variety of text-to-image generating artificial intelligences (AIs) have been devised until now. Nevertheless, the generative AIs introduced thus far encounter an issue wherein they struggle to uphold the coherence of input sentences, particularly when multiple sentences are provided. Within this paper, we present a remedy to this challenge through the application of prompt editing. Furthermore, our experimental results substantiate that our proposed solution more effectively preserves contextual coherence among the generated images in comparison to other preexisting generative artificial intelligence models. The experimental results demonstrate that the proposed scheme improves performance by at least 30 percent in terms of the similarity of the generated image and by 130 percent in terms of $ROUGE_{recall}$.

INDEX TERMS Generative AI, context-aware, text-to-image generation, prompt editing.

I. INTRODUCTION

Humans have the ability to read stories in text and mentally visualize related images in their minds through their imagination, enabling better understanding and enjoyment. Therefore, designing an automatic system that generates visually realistic images from textual descriptions, known as the text-to-image task, is a highly significant endeavor and can be seen as a major milestone towards achieving human-like or general artificial intelligence. With the advancements in deep learning, the task of converting text to images has become one of the most impressive applications in the field of computer vision [1], [2]. Furthermore, there is a growing interest in various applications that leverage Artificial General Intelligence (AGI) [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Renato Ferrero¹.

Artificial Intelligence (AI) that converts text into images can be applied in various fields. For instance, it can automatically generate a movie storyboard by converting a script into images, or it can transform a textual story into a webtoon format. Additionally, product descriptions can be transformed from mundane text into engaging, easy-to-understand images. Due to these diverse applications, various studies have been conducted in the field of converting text into images. GAN (Generative Adversarial Networks)-based methods, autoregressive methods, and diffusion-based methods are typical examples [4], [5]. Although the results of each method are not yet perfect, many users are obtaining satisfactory images. OpenAI's DALL-E [6] and Google's Parti [7] are representative systems.

Although significant progress has been made in converting textual scripts and novels into image content, there are still numerous challenges for general users to utilize effectively.







Input text	Output image 1	Output image 2
1. A man was taking a walk.		
2. The man picked up a bag on the road.		
3. The man sat next to the bag.		

FIGURE 1. The images generated by currently commercialized generative AI platforms.

For instance, consider the following text: “A man was taking a walk. The man picked up a bag on the road. The man sat next to the bag.” In the current generative AI system, the contents of the above three sentences result in three distinct images. Figure 1 illustrates the images generated from currently commercialized generative AI platforms. Unfortunately, images produced in this manner are not easily suitable for use as storyboards or webtoons.

The main challenge in using such generated images is that the context between the images does not align properly. While each picture correctly represents a given sentence unit, the contextual flow present in the original text sentences is lost between the generated images. Due to this ‘context-disappearing problem’, it is currently challenging to utilize this technology for imaging or visualizing various text data, such as novels, scripts, and manuals in the content creation industry. In this paper, we aim to propose a method to address this ‘context-disappearing problem’.

The context-disappearing problem arises when generating images from sentences. Consider creating a children’s book as an example. The user initiates the process with a starting step. If each sentence generates an image through generative AI, and the context is preserved across images generated in each step, the user iterates through the operation until both the desired image and the relational image are processed. However, the challenge arises when it becomes difficult to establish a connection between the generated images, as summarizes the issue as illustrated in Figure 2. As depicted in Figure 2, the majority of current commercial image generation AI outputs yield unsatisfactory images. While the generated images correspond to individual sentences, they

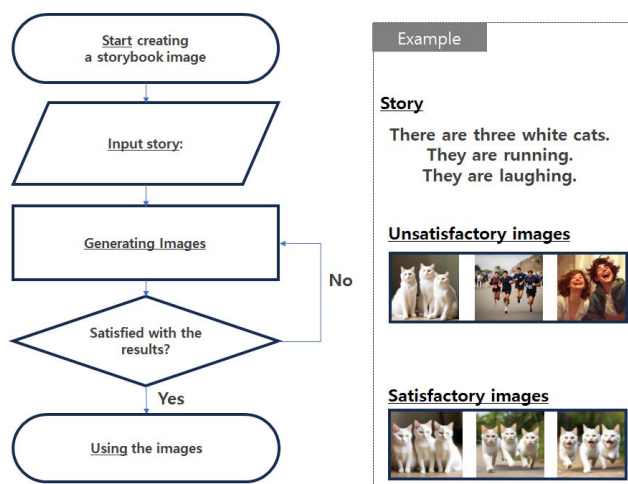


FIGURE 2. Example of image creation process and context-disappearing problem.

fail to fulfill the user’s requirement of preserving context between the generated images. If the generated pictures do not maintain context, the user may need to repetitively engage in the creation process until satisfactory images are achieved. This paper addresses the problem of context loss by presenting a method to preserve input context to the greatest extent possible. Prompt engineering techniques are employed to fully leverage the context of input sentences, enabling better integration with the context of resulting images.

The contributions of this paper are as follows:

- We propose a novel scheme to address the issue of disappearing context when generating images from multiple sentences using generative AI.

- We present four schemes: Initial Context First (ICF), Region Context First (RCF), Single Sentence ICF (SS-ICF), Single Sentence RCF (SS-RCF) to preserve the context between sentences input to generative AI.
- Through the implementation of the proposed schemes and the experiment with the proposed schemes, it has been demonstrated that the proposed schemes outperform existing image generation AI in terms of context preservation, all while maintaining a very low computational overhead.

This paper is structured as follows: In section II, we introduce related works. In section III, we present the proposed processing schemes and the algorithms. Section IV of the paper executes the devised scheme, carries out a series of experiments, and subsequently showcases the obtained results. Finally, in section V, we conclude the paper.

II. RELATED WORKS

A. CONTEXT EXTRACTION AND PRESERVATION

Extensive research has been undertaken in the realms of natural language processing and linguistics to identify and maintain context. Numerous studies have explored the simplification of sentences and vocabulary as a means to comprehend sentence context. Reference [8] conducted a comparative analysis of various studies related to sentence simplification. While the task of simplifying sentences is often employed for summarization purposes, it can also significantly contribute to fields aiming to enhance sentence comprehension and context understanding. Our proposed schemes align with this recognition and were developed accordingly.

Reference [9] introduces extractive and abstractive text summarization technologies and provides a deep taxonomy of the Automatic Text Summarization (ATS) domain. Similar to the word selection approach in ATS, our scheme also involves selecting words that encompass context. Likewise, [10] conducted a study using more explicit and simpler words not present in the original sentences. The objective of the paper was to decrease content while upholding readability, achieved through the substitution, rearrangement, and insertion of sentences and words. Although the method in the paper effectively enhances sentence context clarity, it was not incorporated into our technique due to computational complexity.

Additionally, within the field of natural language processing, much research has been proposed to preserve context by dividing a sentence into multiple sentences [11], [12]. This approach involves the separation and reprocessing of words and sentences to create clearer expressions that aid in context preservation and understanding. While this technique might prove effective for research papers, academic sentences, and similar contexts, it was not included in our scheme. The sentences generated through this process may be longer than the original sentences or could potentially distract the reader's attention. However, if the application under consideration in the research pertains to academic

papers or technical books, it may be worthwhile to explore its usage.

Recently, with the increased prevalence of artificial intelligence applications, there has been a surge in studies focusing on understanding context through artificial intelligence. Reference [13] conducted a study incorporating optimal sentences containing context within multiple sentences to identify correlations between events. In [13], the objective was not to extract individual contexts from multiple sentences but to extract contexts that represent the entirety of multiple sentences. This study has the potential to complement our research. When dealing with a vast number of input sentences, generating images for each might be inefficient. In such cases, [13] could offer an efficient approach. Similarly, [14] is a related study. The research revolves around identifying sentences that provide the most appropriate answer to a given question. This study is also applicable in scenarios where a sentence must be selected from a large pool of input sentences.

B. IMAGE GENERATIVE AI

Image generative AI has recently garnered significant attention in the realm of content creation. A substantial amount of research is underway to automate the creation of animations and webtoons, endeavors that traditionally depended on manual craftsmanship, through the application of artificial intelligence. AlignDRAW [15] was a pioneering research in the field of generating images based on natural language input. AlignDRAW iteratively draws patches on a canvas while attending to the relevant words in the description. However, it has been found to produce unrealistic results. Recent studies indicate substantial differences in the quality of the generated images [5]. The Text-conditional GAN [16] proposed a discriminative architecture for generating images from text descriptions. Unlike GAN-based methods [17], [18], [19], which are primarily designed for small datasets, the autoregressive method utilizes large-scale data to convert text into images. Notable examples of this approach are OpenAI's DALL-E and Google's Parti. However, the autoregressive method [6] is characterized by high computational costs due to the large number of calculations involved. Additionally, it suffers from the weakness of sequential error accumulation, where errors in the generation process can accumulate and affect the overall quality of the final image.

Recently, a new model called the Diffusion Model (DM) has emerged in the field of text-to-image generation [20], [21], [22], [23]. Diffusion-based text-image synthesis has garnered significant interest in social media and content-related industries. These models have demonstrated superior performance and possess various beneficial properties compared to existing generative AI models. The Diffusion-based model employs a mathematical expression that transforms pixels in an image into noise as the pixels disperse over

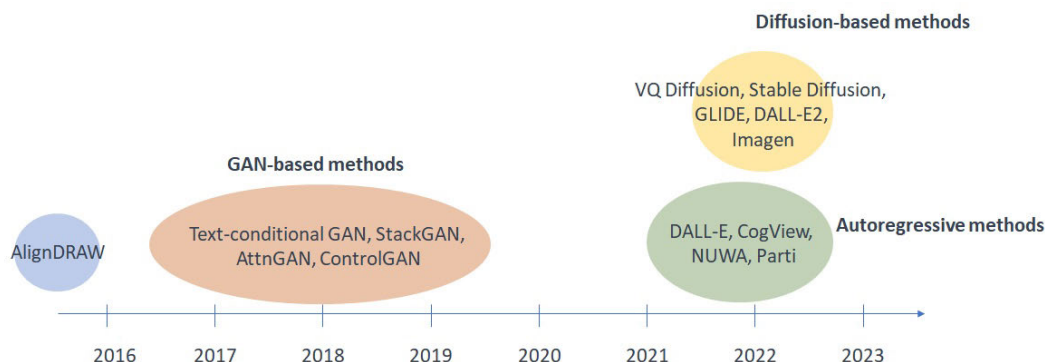


FIGURE 3. Representative works on text-to-image generative AI.

time. In the context of text-to-image conversion applications, diffusion-based models have gained substantial popularity in recent years. In a recently published paper [24], the Diffusion Probability Model (DPM) for efficient remote sensing image Super-Resolution (SR) was introduced. In this paper, unlike previous work that used a heavy UNet for noise prediction, we developed an Efficient Activation Network (EANet) to achieve favorable noise prediction performance through simplified channel attention and simple gate operations. Figure 3 provides a summary of the development flow and key algorithms associated with text-to-image generation AI.

Among the many algorithms, some are released as open source. The Craiyon [25] generator and Glide [20] are open source image generators. The Craiyon generator, also known as DALL-E mini, produces images using the VQGAN [26] encoder, while Glide employs a diffusion model. Apart from these open source image generation AIs, there are well-known commercial counterparts such as DALL-E and Midjourney. While the precise internal structures of these two generative AIs are not fully disclosed, it's common for many other image generative AI models to be constructed using a combination of CLIP [27] and GAN.

The disadvantage of the various text-image generation AI models mentioned above is that they struggle to maintain context between the generated images. While these models are capable of producing suitable images for each input sentence, they face challenges in utilizing them effectively for text-image generation applications. The lack of similarity and consistency among the generated images makes it difficult for users in the application field to use them as storyboards or incorporate them in a cohesive manner. Users typically expect to input text and obtain images that possess a certain degree of context to be useful for their intended purposes.

AI-related algorithms for image generation have been advancing across various fields. Reference [28] introduced an algorithm that utilizes parallel decoding to create images more efficiently and precisely. Reference [29] proposed EEGAN (Edge Enhancement Network), capable of generating images with improved quality. Reference [30]

introduced an algorithm capable of generating images of comparable quality using fewer training data and learning parameters. These studies indicate a shift in research focus from image generation per se to more efficient image generation. Recently, a research result was published in response to a user request for the output of a text-image generation AI model through prompt editing [31], [32]. The study conducted prompt editing to investigate the claims made by the author. The research suggested that stretching the output image, as perceived by the user, has a noticeable impact, although the improvements within the AI engine itself were relatively minor compared to what users reported. Furthermore, our study indicated that context retention can be achieved through prompt editing techniques.

III. PROPOSED SCHEME

In this section, we present our proposed scheme, which involves utilizing prompt editing to adjust the input prompt instead of attempting to control text-image generating AI algorithms which are challenging to manage. The input prompt consists of a context for each sentence, and by linking this context, we connect the context of each input sentence. Through this approach, we introduce a method to preserve the context of the sentences created.

Our scheme is largely divided into three steps. Firstly, the input script is classified by sentence, and important context is extracted from each sentence. Secondly, a context sequence is constructed using the extracted context. Lastly, using the configured context sequence, sentence-by-sentence matching is performed, and the input script is edited accordingly.

A. SYSTEM OVERVIEW

Figure 4 presents an overview of the operational methodology proposed in this study. The user provides a sentence or multiple sentences of text, which the system subsequently divides into individual sentence units. Following this division, the context elements are automatically extracted from each sentence segment and incorporated into the input sentence. This enriched text is then utilized as the input for

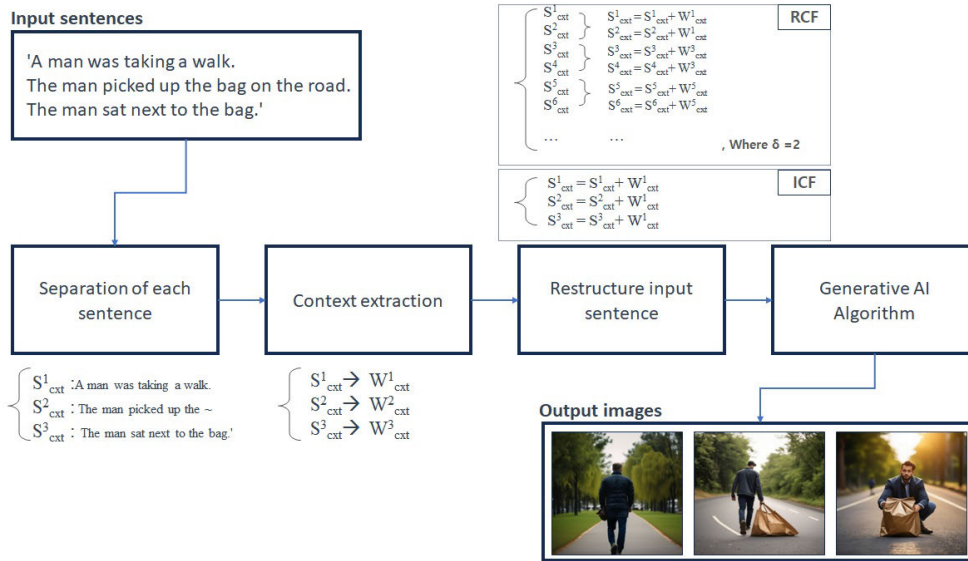


FIGURE 4. Overview of the proposed scheme's operation.

the image generative AI. All these steps occur seamlessly within the system, hidden from the user's awareness. Figure 4 illustrates examples of the operational methods for the two schemes proposed in this paper. The most significant distinction between the two proposed schemes lies in the scope of influence by context. Instead of altering a complex image-generating AI algorithm prone to numerous modifications, we opted to preserve context through prompt editing of the input text. Irrespective of whether the user inputs multiple sentences or one sentence at a time, the approach advocated in this study ensures the continuity of context that remains applicable for various applications.

B. CONTEXT EXTRACTION

In this paper, we assume a scenario where users of text-image generation AI input multiple sentences. For instance, consider the following input: 'A man was taking a walk. The man picked up the bag on the road. The man sat next to the bag.' Such a situation finds practical applications in various fields, including converting scripts into storyboards, novels into webtoons, and poems into images. To proceed with our study, we define the input sentence and the sentence, S_{input} and S as follows:

- $S_{input} = \{S \mid S \text{ is a complete sentence with a subject and a verb}\}$
- $S = \{W \mid W \text{ is a word}\}$

An input sentence in this paper comprises multiple sentences, each containing a subject and a verb. These sentences are composed of individual words. For context extraction, we opted for a straightforward approach by using words that serve as subjects and verbs, rather than employing complex processing techniques. While various methods exist for extracting the core content of a sentence, this paper's primary emphasis lies in demonstrating the

practical utilization of the extracted context. Therefore, the approach is simplified, focusing on the use of subject and verb words for context extraction from the input sentence.

To establish the process of extracting context from the input sentence using this method, the sentence was redefined as follows:

- $S = \{W_{ctx} \cup W_{non} \mid W_{ctx} \text{ is a word that contains context, and } W_{non} \text{ is a word that does not contain context}\}$

When the user inputs sentences, the sentences are first separated. From each separated sentence, context information is extracted through morpheme analysis, and this extracted context information is then appended to each sentence. Each sentence, now including its associated context information, is defined as follows:

- $S_{ctx} = \{(S, W_{ctx}) \mid W_{ctx} \text{ are words extracted from } S\}$

S_{ctx} is a tuple that contains the sentence and context entered by the user. It is automatically generated from the user input sentences and subsequently used for prompt editing, ensuring the preservation of context.

C. PROMPT EDITING

We divided the user's input sentences into individual sentences using the following format and generated data by incorporating context information into each sentence. Equation 1 mathematically shows how to decompose input sentences.

$$S_{input} \rightarrow \begin{cases} S_{ctx}^i \\ S_{ctx}^{i+1} \\ \dots \\ S_{ctx}^n \end{cases} \quad (1)$$

Each separated sentence, denoted as S_{ctx} , includes context information. Our approach aimed to uphold the context of the

Algorithm 1 Algorithm of ICF Processing

- 1: if users input n sentences;
- 2: for($i = 1; i < n + 1; i + +$)
- 3: Separate sentences into unit sentences, S_{cxt}^i ;
- 4: Extract W_{cxt}^i from S_{cxt}^i ;
- 5: Add context, W_{cxt}^1 , to S_{cxt}^i ;
- 6: Reconstruct S_{input} ;
- 7: Input S_{input}^{cxt} into generative AI;
- 8: endif

generative AI output by modifying this context information within each sentence.

1) INITIAL CONTEXT FIRST (ICF)

Our first prompt editing scheme is Initial Context First(ICF).The first scheme centers around the observation that thoughts or contexts commonly transition within a paragraph. According to recent papers in linguistics [33], a paragraph is typically composed of a topic sentence, body sentences, and a concluding sentence. The primary content of the paragraph is often conveyed in the initial sentence, and the scheme that implements this for context extraction is the ICF scheme. It recognizes that sentences constituting a paragraph typically amount to around five sentences. If the sentence isn't lengthy, the primary context often resides within the initial sentence, exerting influence over the entire paragraph. Consequently, within the framework of the first scheme, the context from the first sentence was duplicated and appended to all sentences within the paragraph. This can be succinctly summarized as Equation 2:

$$S_{input} \rightarrow \begin{cases} S_{cxt}^i \\ S_{cxt}^{i+1} + W_{cxt}^i \\ \dots \\ S_{cxt}^n + W_{cxt}^{n-1} \end{cases}$$

where, $S_{cxt}^i, S_{cxt}^{i+1}, \dots, S_{cxt}^n$ are within a paragraph. (2)

In ICF, the context of the first sentence is considered the most crucial. Context information, W_{cxt}^i , is derived from the initial sentence entered by the user and represented as words. Utilizing this information, the system proceeds to process subsequent user input sentences by incorporating this context information into them. Algorithm 1 describes the basic scheme, ICF. In Algorithm 1, S_{input}^{cxt} refers to the modified user input sentences.

2) REGION CONTEXT FIRST (RCF)

The second approach to prompt editing is the Region Context First (RCF) scheme. The most significant limitation of our initial scheme is that when numerous input sentences, denoted as S_{cxt} , are present, the context toward the end may feel disconnected. With an increasing number of sentences, adjustments to the context in the latter portion become necessary. However, the challenge lies in maintaining context

consistently, as the same context is appended throughout. To address this concern, we introduce a solution involving the establishment of a sentence threshold. This threshold dictates the span over which context continuity should be preserved. When the threshold is exceeded, context reconfiguration is implemented. The rationale behind our introduction of the second scheme based on region is connected to the number of sentences constituting the paragraph discussed above. Initially, various writing guides advise constructing paragraphs with a length between five and ten sentences [34], yet numerous articles deviate from this recommendation. As previously explained, issues emerge when the number of sentences increases in such articles. Through experiments, we found that images generated using the ICF scheme in the latter sentences of the entire paragraph often fail to align with the context. Consequently, we designed the RCF to address and complement this discrepancy.

Our second scheme can be applied in diverse manners. The threshold itself can be tailored within a range of 5 to 8 sentences, reflecting the typical composition of a paragraph. Alternatively, it can be employed to reset the threshold whenever a paragraph transition occurs. The flexibility of the threshold usage allows it to be adjusted according to the specific characteristics of the application it's being applied to. The second scheme can be succinctly expressed as follows:

$$S_{input} \rightarrow \begin{cases} S_{cxt}^i \\ S_{cxt}^{i+1} \begin{cases} (If \ i+1 < Threshold, \delta) \\ S_{cxt}^{i+1} + W_{cxt}^i \\ (else) \\ S_{cxt}^{i+1} + W_{cxt}^{i+1} \end{cases} \\ \dots \\ S_{cxt}^n + W_{cxt}^{i+\alpha} \end{cases} \quad (3)$$

Input sentences are divided into multiple sentences, each equal to the threshold value. If the total number of sentences is less than the threshold, RCF produces the same results as ICF. However, if the total number of sentences exceeds the threshold, the sentences are divided by the threshold value(i.e., number of sentences). For the divided sentences, context is extracted and utilized for each unit. Algorithm 2 explains how the proposed RCF scheme operates. The most significant difference between ICF and RCF is the scope of influence of the selected context. In RCF, the nearby context is considered more important.

3) SINGLE SENTENCE ICF AND RCF (SS-ICF AND SS-RCF)

The third scheme involves leveraging the previous two schemes but can be adapted to different preconditions. The third scheme amalgamates the preceding two methods, presenting an algorithm designed to address scenarios in which the user inputs one sentence at a time. This scheme is predicated on the notion that users frequently seek to uphold context while entering one sentence and then proceeding to input another within a specific timeframe.

Algorithm 2 Algorithm of RCF Processing

```

1: if users input  $n$  sentences;
2:   for( $i = 1; i < n + 1; i++$ )
3:     Separate sentences into unit sentences,  $S_{cxt}^i$ ;
4:     Extract  $W_{cxt}^i$  from  $S_{cxt}^i$ ;
5:     if ( $n > \delta$ )
6:       Add context,  $W_{cxt}^{i \bmod \delta}$ , to  $S_{cxt}^i$ ;
7:       Reconstruct  $S_{input}$ ;
8:     else
9:       Add context,  $W_{cxt}^1$  to  $S_{cxt}^i$ ;
10:      Reconstruct  $S_{input}$ ;
11: Input  $S_{input}^{cxt}$  into generative AI;
12: endif

```

Algorithm 3 Algorithm of SS-ICF and SS-RCF Processing

```

1: if ( $\Delta T < T_{input}$ )
2:   Add  $W_{cxt}$  according to ICF or RCF;
3:   Use the generated sentence as input to the generative
   AI;
4: else
5:   Use the input sentence as input to the generative AI;
6: endif

```

Many commercially used image-generating AIs currently only support the input of one sentence at a time. Additionally, in certain situations, users may prefer to input one sentence at a time rather than the entire sentence at once. The schemes devised to address these scenarios are SS-ICF and SS-RCF. These two schemes generate and process context utilizing the ICF and RCF schemes. However, the time interval between input instances serves as a criterion for determining whether a sentence is part of a continuous sequence. In this third scheme, a designated time ΔT functions as a threshold. If an additional sentence is entered within a time shorter than this threshold, the context from the preceding sentence is automatically appended to the input value to facilitate prompt editing. Equation 3 summarizes the operation of the proposed third scheme. The accumulation of the context value utilizes the ICF and RCF techniques respectively. Consequently, the third scheme consists of two methods, denoted as SS-ICF (Single Sentence ICF) and SS-RCF (Single Sentence RCF) correspondingly. The operational algorithm's detailed process aligns with Algorithm 3.

In Algorithm 3, T_{input} represents the time difference between the user's creation of the current query and the immediately preceding query. To facilitate the functioning of Algorithm 3, a storage capacity is necessary to retain each context W_{cxt} . For SS-ICF, a single context storage space is needed, whereas SS-RCF requires storage space to accommodate contexts based on the established threshold. In both scenarios, the context is represented as a word, resulting in a very modest storage space requirement due to the compact nature of word-based representations.

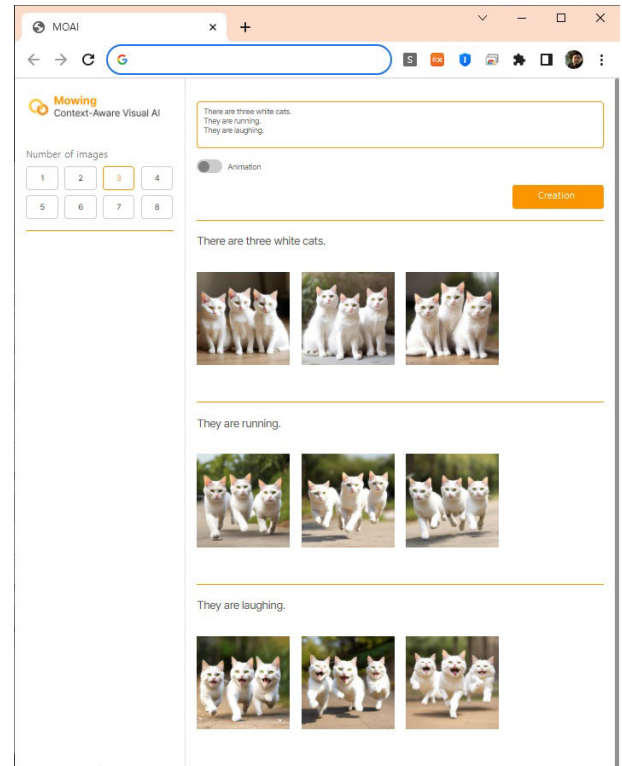


FIGURE 5. The system implementing the schemes proposed.

IV. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENTS

To assess the context similarity of the generated images, we devised the experiment in the following manner. Given the absence of established metrics for evaluating context similarity of generated images in previous research, this study adopted the assumption that context preservation is relatively effective when the similarity between the image generated in the preceding sentence and the one generated in the subsequent sentence is high. Building upon these premises, the experiment was structured into two phases as outlined below. Initially, outcomes from both the target generative AI and the generative AI employing our scheme were produced. The extent of similarity between the generated images was then quantified. In order to validate the correlation between the computed similarity and the actual output results, we presented both the actual generated image results and the similarity result graphs side by side.

The experiment followed this structure: Firstly, diverse input texts were used to generate images through generative AI systems. The produced images were then indexed using HNSW (Hierarchical Navigable Small World graphs) in Faiss, a vector similarity search algorithm developed by Facebook. Subsequently, the similarity was measured and assessed using the indexing outcome. The comparison targets for generative AI encompassed DALL-E, Stable Diffusion, and MidJourney. Input values consisted of text composed of 3 to 5 sentences each. The similarity measure is represented

TABLE 1. The text generated for scenario 1.

DALL-E	StableDiffusion
blue building, man, cell phone, dark pants, road	a man, park, tree, grey sky,
black t-shirts, black pants, black plastic bag, road, a man	yellow plastic bag, red cap, blue jumper, car, a man
blue check shirts, pink bag, sitting man, hat, road	dog, bench, sitting men, red bag, blue jacket
Midjourney	ICF
big bag, a man, car, city, building.	green tree, road, a man, waking, sky
a man, hat, big bag, people, building.	green tree, road, a man, waking, big bag
road, a sitting man beggar, bag, sky, cloud	green tree, road, a sitting man, waking, big bag

as a value ranging between 0 and 1, where 0 signifies no similarity whatsoever, and 1 indicates a perfect match. The system implementing the schemes proposed in this paper was established and put into practice, as depicted in the figure 5.

Initially, we initiated a comparison involving the commercially available product groups under examination. The fundamental algorithm used for comparison was ICF. Subsequent to the comparison with commercial products, an assessment was carried out between our algorithms. Notably, the primary output results for the third scheme mirror those of ICF and RCF, so it was omitted from the experimental results.

To complement the described performance evaluation method, we conducted a quantitative performance evaluation using the performance metric proposed in ROUGE [35]. To utilize ROUGE for performance evaluation, we slightly modified the process as follows. Initially, text was generated to summarize the generated image. The text creation process involved ten users who examined each image and generated words with five modifiers. They then combined these words and selected the five most frequently occurring ones among the commonly created words. The text generated according to Scenario 1 is presented in Table 1. The words listed in Table 1 are derived from each generated image, as illustrated in Figure 7. For instance, the words generated for the image corresponding to the first sentence, produced by the ICF scheme, include “green tree, road, a man, walking, sky”. Based on the generated text, a ROUGE-N performance evaluation was conducted using the formula below (Equation (4)).

$$ROUGE - N_{recall} = \frac{\text{Number of overlapped words}}{\text{Total words in two images}} \quad (4)$$

In this paper, similar words among the generated words were considered the same. For example, “men” and “man” were treated as the same word, and “bag” and “big bag” were also treated as the same word. This is because, unlike text, individual differences in describing similar images are considered in the evaluation process of images.

B. EXPERIMENTAL RESULTS

1) CONTEXT SIMILARITY

Figure 6 and Figure 7 present the experimental results. Figure 6 illustrates a graph showcasing performance in terms

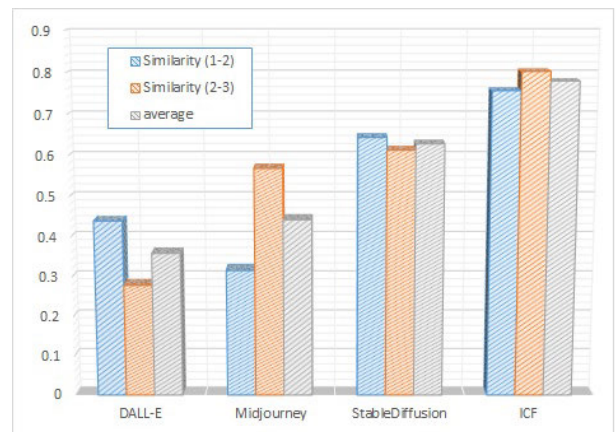


FIGURE 6. Comparison of the similarity among the results from the first scenario.

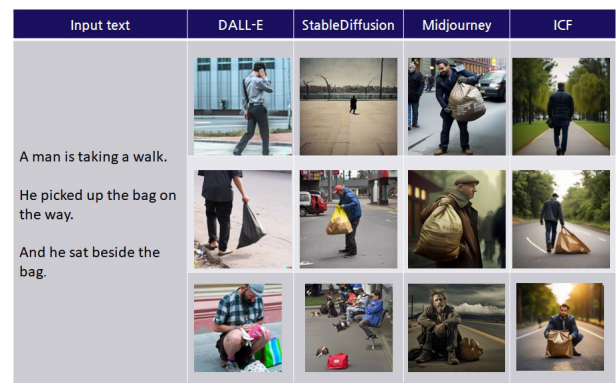


FIGURE 7. Comparison of the generated images (scenario 1).

TABLE 2. Comparison of ROUGE - N_{recall} (scenario 1).

Scenario 1	DALL-E	StableDiffusion	Midjourney	ICF
ROUGE-N recall (image 1&2)	0.4	0.2	0.6	0.8
ROUGE-N recall (image 2&3)	0.4	0.4	0.4	1

of similarity. The sentence utilized in the experiment is composed of three sentences, as indicated in the input text within Figure 7. In Figure 6, “Similarity (1-2)” signifies the likeness between the image generated from the first sentence and the one from the second sentence. Likewise, “Similarity (2-3)” quantifies the resemblance between the

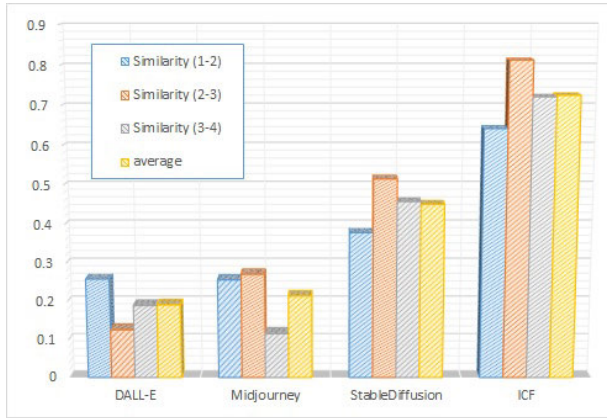


FIGURE 8. Comparison of the similarity among the results from the second scenario.

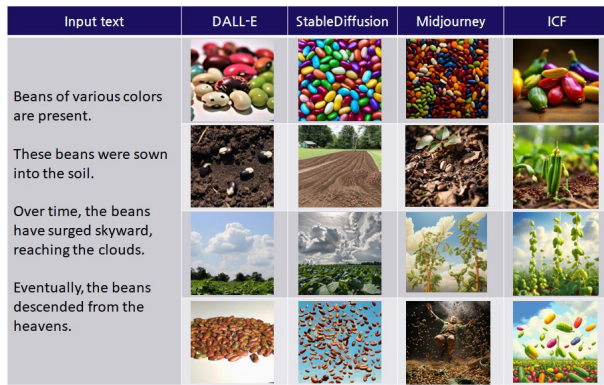


FIGURE 9. Comparison of the generated images (scenario 2).

second sentence and the third sentence. The ‘‘Average’’ value represents the mean of all the measured similarities. As evidenced by Figure 6, the image produced using our ICF scheme demonstrates the highest similarity. This observation suggests that minimal changes occur between images, implying effective context preservation.

The resulting images are depicted in Figure 7. The rightmost image corresponds to the result generated by the method proposed in this paper. Comparisons were performed for each of the three input sentences against existing market-available image-generating AIs. As demonstrated in Figure 7, our ICF scheme produces results that retain the most context. Analyzing the images presented in Figure 7, it’s evident that all generative AIs are producing images that align with the input sentences. However, the ICF scheme’s output notably excels in maintaining context coherence among the results.

As shown in Table 2, it is evident that our scheme outperforms other state-of-the-art commercial programs in terms of $ROUGE - N_{recall}$. The performance evaluation measured the $ROUGE - N_{recall}$ between images 1 and 2, and the $ROUGE - N_{recall}$ between images 2 and 3. The closer the number is to 0, the less the context matches, and the closer it is to 1, the more the context matches.

Figure 8 and Figure 9 illustrate the outcomes of images generated from the second input sentence. As depicted

TABLE 3. Comparison of $ROUGE - N_{recall}$ (scenario 2).

Scenario 2	DALL-E	StableDiffusion	Midjourney	ICF
ROUGE-N recall (image 1&2)	0.4	0.0	0.2	0.4
ROUGE-N recall (image 2&3)	0.0	0.4	0.2	0.6
ROUGE-N recall (image 3&4)	0.0	0.2	0.2	0.6

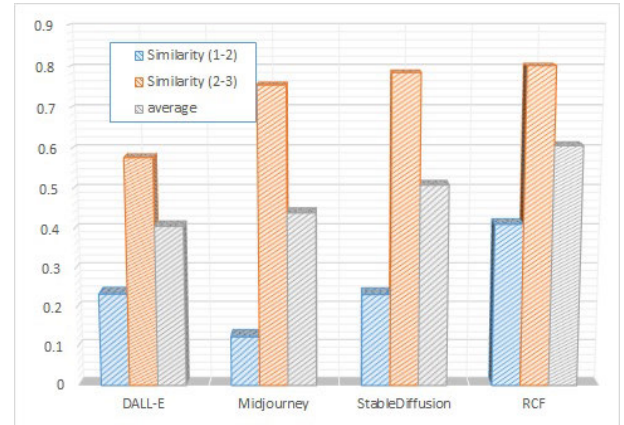


FIGURE 10. Comparison of the similarity among the results from the third scenario.

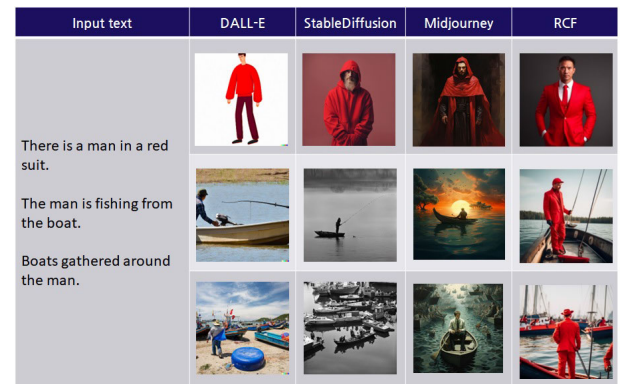


FIGURE 11. Comparison of the generated images (scenario 3).

in Figure 8, the ICF scheme’s results exhibit the highest level of similarity. It’s important to note that unconditional preservation of extensive context cannot be asserted solely based on high similarity. Nevertheless, in a general sense, if the first and second images differ substantially, users might perceive a lack of maintained context. Hence, a higher level of similarity can suggest better context preservation compared to relatively lower similarity. As observed in Figure 8, our proposed scheme demonstrates superior performance in terms of similarity when compared to the outcomes of other systems.

This understanding can be gleaned from the actual image results. Examining the outputs for the fourth sentence, it’s apparent that the beans in the other images are consistently of a single color. However, when inspecting the outcomes

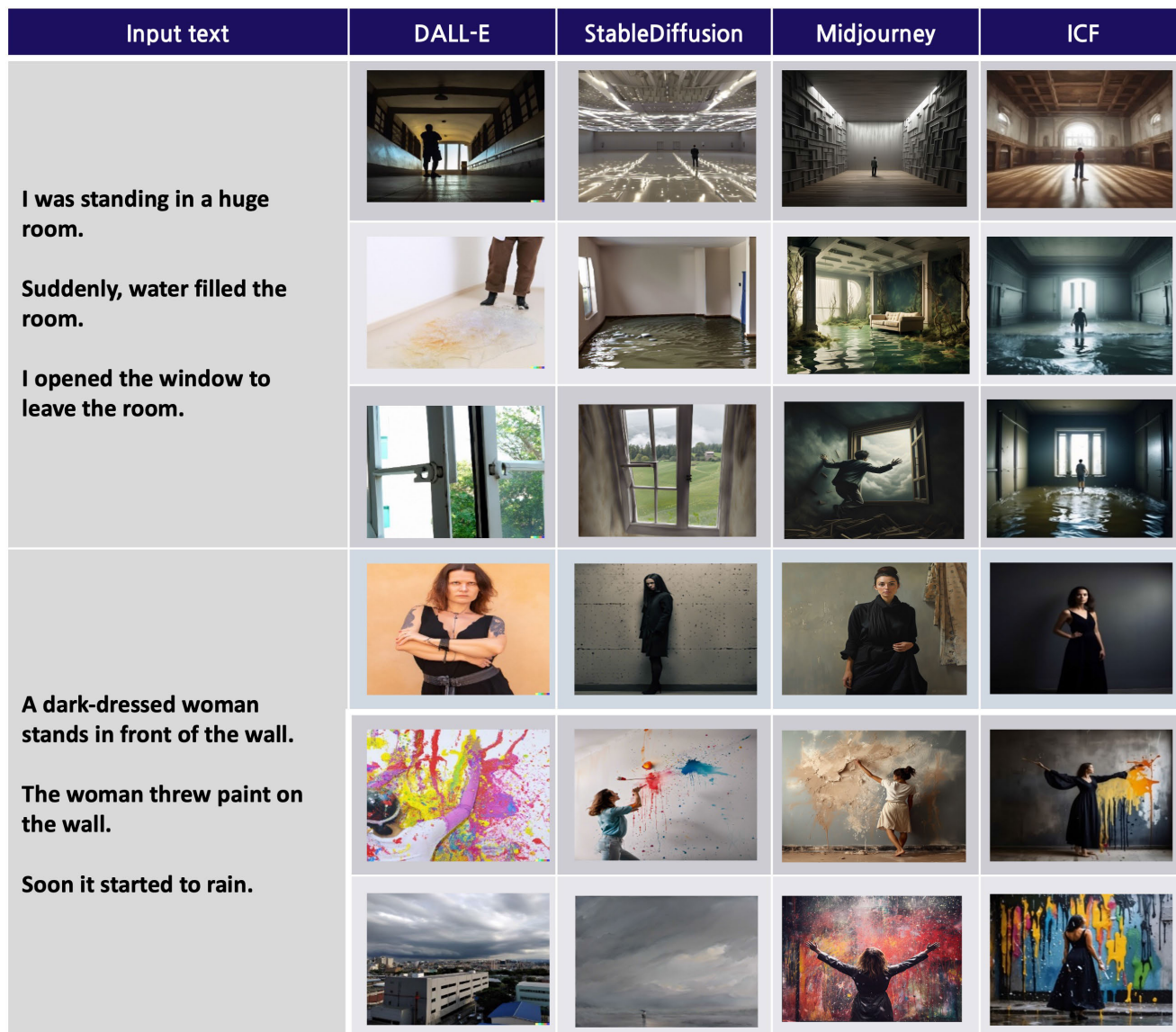


FIGURE 12. Comparison of images generated under various scenarios (scenario 4, 5).

TABLE 4. Comparison of ROUGE – N_{recall} (scenario 3).

Scenario 3	DALL-E	StableDiffusion	Midjourney	ICF
ROUGE-N recall (image 1&2)	0.2	0.2	0.2	0.4
ROUGE-N recall (image 2&3)	0.4	0.4	0.6	0.8

produced using the ICF scheme, there is observable variation in the bean colors. This divergence is likely due to the fact that the context from the first sentence has an impact on the final sentence. Apart from our results, the performance of the MidJourney model turned out to be the most promising among the alternatives. Table 3 shows performance in terms of ROUGE – N_{recall} . As observed in Table 3, our technique exhibits the best performance.

Figure 10 and Figure 11 present the performance comparison outcomes for images generated from the third input sentences. In this particular scenario, the RCF scheme was employed. As observed in earlier experiments, the proposed scheme yields the most favorable results in terms of similarity. Additionally, it’s evident that the proposed scheme excels in maintaining context coherence, as depicted by the output results. As highlighted in Figure 11, only the output from the proposed scheme consistently depicts the man in red attire, indicating the sustained preservation of context. Table 4 shows performance in terms of ROUGE – N_{recall} . As observed in Table 4, our technique exhibits the best performance.

Figure 12 and Figure 13 illustrate the outcomes of comparing images generated by employing different scenarios. Our

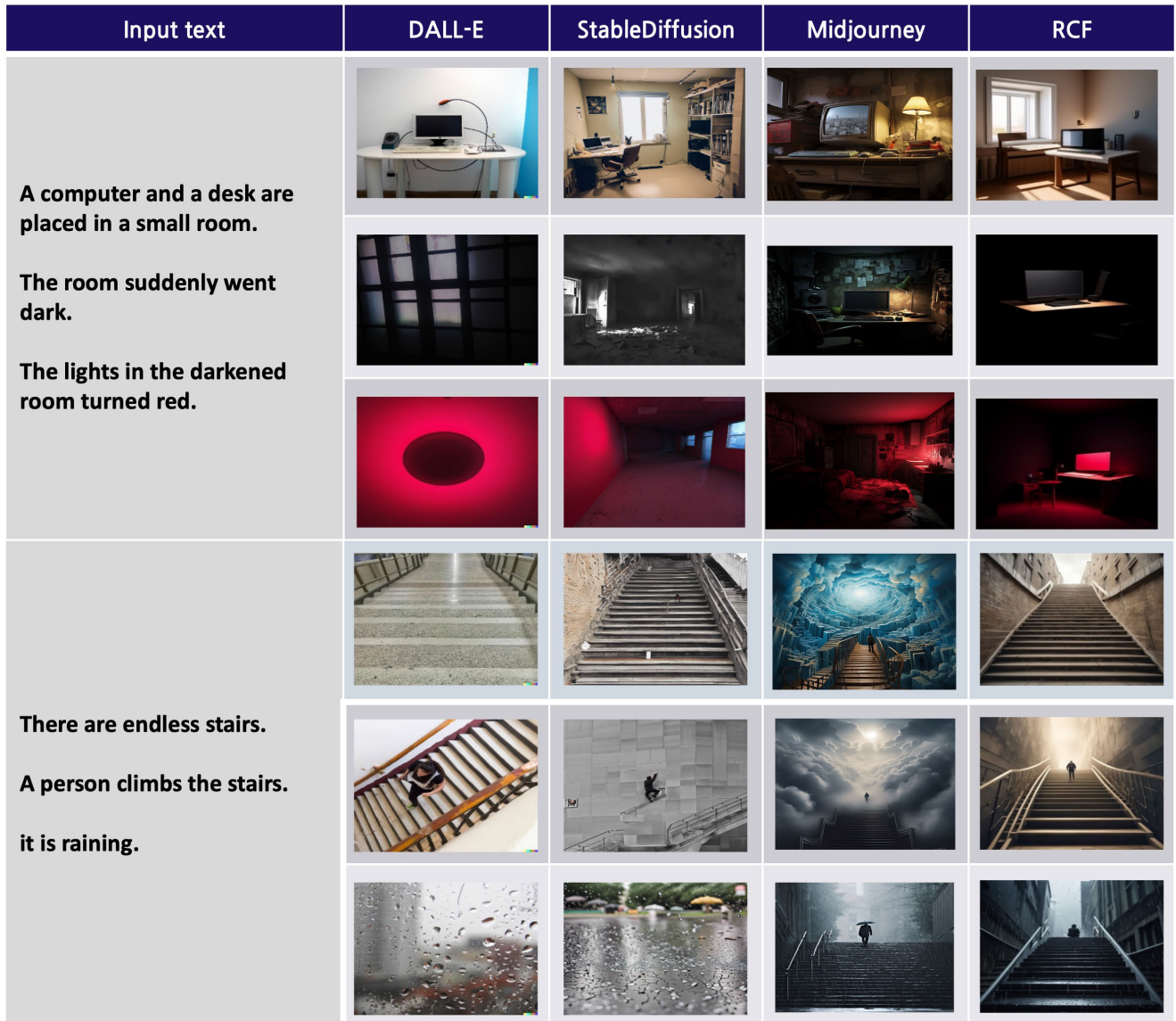


FIGURE 13. Comparison of images generated under various scenarios (scenario 6, 7).

TABLE 5. Context creation time and image creation time for each scenario.

Scenario	Context creation time	Image generation time
2	2ms	32,100ms
3	1ms	22,500ms
4	1ms	20,600ms
5	1ms	21,400ms
6	1ms	23,000ms
7	1ms	20,800ms

scheme incorporates both ICF and RCF. These images are the results of scenarios 4, 5, 6, and 7. As shown in Figure 12 and Figure 13, the proposed scheme demonstrates exceptional performance in terms of context preservation.

2) PROCESSING OVERHEAD

Our schemes go through context extraction and sentence reconstruction processes before image generation

AI processing. To achieve this, the user input sentences are read once before image creation, and the context for each sentence is extracted. The extracted context is then used to recreate each input sentence, and this processing occurs in two stages. The time complexity of extracting context and reconstructing sentences is $O(N)$, where N is the number of sentences, which is relatively short compared to the processing time required for image creation.

Table 5 represents the time measurements for creating context and generating images using RCF in our scheme, for scenarios 2 to 7. Note that the time complexity for generating and processing context with both RCF and ICF is similar and the time is measured in milliseconds. As shown in Table 5, the time required for context creation is notably small compared to the image creation time. While the image creation time may vary based on hardware

performance, even considering this variability, the time taken to create context, process it, and modify each sentence is remarkably short. Therefore, we can confirm that our technique has a minimal temporal impact on image generation processing.

V. CONCLUSION AND FUTURE WORK

This paper introduces an image generation scheme that preserves context using text-to-image generation AI. Our approach centers on prompt editing, which is designed to maintain the context of input sentences when users input multiple sentences. Through the assessment of image similarity for each sentence, we've demonstrated that our proposed scheme results in more similar images compared to those generated using DALL-E, StableDiffusion, and Midjourney. Furthermore, the tangible results obtained from the actual generated images emphasize its superior context preservation performance compared to the most recent results.

It's worth noting that the proposed scheme isn't an algorithm that directly modifies the generative AI algorithm itself. Instead, it's a methodology that maintains context by automatically revising the input prompt. Although the proposed algorithm introduced an RCF scheme to handle situations where many sentences are input, it remains a challenging problem to address cases where a very large number of sentences are input in a structured manner. As such, certain challenges remain in retaining the text within the image and ensuring consistency with the same character. In the coming time, we intend to further our research to address these challenges by enhancing the algorithm itself.

REFERENCES

- [1] J. Clune, "AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence," 2019, *arXiv:1905.10985*.
- [2] R. Fjelland, "Why general artificial intelligence will not be realized," *Humanities Social Sci. Commun.*, vol. 7, no. 1, pp. 1–9, Jun. 2020.
- [3] V. C. Müller and N. Bostrom, "Future progress in artificial intelligence: A survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*, 2016, pp. 555–572.
- [4] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," *Neural Netw.*, vol. 144, pp. 187–209, Dec. 2021.
- [5] C. Zhang, C. Zhang, M. Zhang, and I. So Kweon, "Text-to-image diffusion models in generative AI: A survey," 2023, *arXiv:2303.07909*.
- [6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8821–8831.
- [7] J. Yu, Y. Xu, J. Yu Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. Karagol Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," 2022, *arXiv:2206.10789*.
- [8] M. Shardlow, "A survey of automated text simplification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 1, pp. 58–70, 2014.
- [9] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021.
- [10] T. A. Cohn and M. Lapata, "Sentence compression as tree transduction," *J. Artif. Intell. Res.*, vol. 34, pp. 637–674, Apr. 2009.
- [11] S. Narayan and C. Gardent, "Hybrid simplification using deep semantics and machine translation," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 435–445.
- [12] S. Narayan and C. Gardent, "Unsupervised sentence simplification using deep semantics," 2015, *arXiv:1507.08452*.
- [13] H. M. D. Trong, N. N. Trung, L. van Ngo, and T. H. Nguyen, "Selecting optimal context sentences for event-event relation extraction," in *Proc. AAAI Conf. Artif. Intell. Intell.*, 2022.
- [14] C. Tan, F. Wei, Q. Zhou, N. Yang, B. Du, W. Lv, and M. Zhou, "Context-aware answer sentence selection with hierarchical gated recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 540–549, Mar. 2018.
- [15] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," 2015, *arXiv:1511.02793*.
- [16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.
- [17] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [18] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
- [19] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5908–5916.
- [20] 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," 2021, *arXiv:2112.10741*.
- [21] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. K. 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.
- [22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
- [23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
- [24] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.
- [25] A. Panton, "Intimate crip self-portraits brought to life in partnership with crayon (DALL-E Mini)," *Can. J. Theol., Mental Health Disab.*, vol. 2, no. 2, pp. 143–156, 2022.
- [26] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12868–12878.
- [27] 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. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [28] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan, "Muse: Text-to-image generation via masked generative transformers," 2023, *arXiv:2301.00704*.
- [29] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
- [30] M. Tao, B.-K. Bao, H. Tang, and C. Xu, "GALIP: Generative adversarial CLIPs for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14214–14223.
- [31] Y. Wang, S. Shen, and B. Y. Lim, "RePrompt: Automatic prompt editing to refine AI-generative art towards precise expressions," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2023, pp. 1–29.

- [32] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2022, pp. 1–23.
- [33] O. Wali and A. Q. Madani, "The importance of paragraph writing: An introduction," *Organization*, vol. 3, no. 7, pp. 44–50, 2020.
- [34] P. Creme and M. Lea, *Writing at University: A Guide for Students*. New York, NY, USA: McGraw-Hill, 2008.
- [35] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Jul. 2004, pp. 74–81.



HYUNJO KIM received the B.S. degree in industrial engineering and the M.S. degree in information security from Korea University, South Korea, in 2014 and 2015, respectively, where he is currently pursuing the Ph.D. degree in information security. His research interests include formal methods, secure software engineering, software security, and software assurance.



JAE-HO CHOI received the M.S. and Ph.D. degrees in computer science from Korea University, South Korea, in 2005 and 2011, respectively. From 2011 to 2013, he was a Research Professor with Yonsei University. Currently, he is with MCCAII Company Ltd., an AI research and development company. His research interests include generative AI, image recognition AI, and data analysis.



JIN-YOUNG CHOI received the B.S. degree in computer engineering from Seoul National University, South Korea, in 1982, the M.S. degree in computer science from Drexel University, Philadelphia, USA, in 1986, and the Ph.D. degree in computer science from the University of Pennsylvania, Philadelphia, in 1993. From 1994 to 1995, he was a Research Assistant with the University of Pennsylvania. From 1996 to 1999, he was an Assistant Professor with the Computer Science Department, Korea University, South Korea, where he has been a Professor, since 2004. He is currently a Professor with the Information Security Department, Korea University.

...