IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 November 2023, accepted 23 December 2023, date of publication 1 January 2024,
date of current version 8 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348778

RESEARCH ARTICLE

Optimizing Prompts Using In-Context Few-Shot Learning for Text-to-Image Generative Models

SEUNGHUN LEE[1], JIHOON LEE[1], CHAN HO BAE[1], MYUNG-SEOK CHOI[2], RYONG LEE[2], AND SANGTAE AHN[1,3], (Member, IEEE)

[1]School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]AI Data Research Center, Korea Institute of Science and Technology Information (KISTI), Daejeon 34141, South Korea
[3]School of Electronics Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding authors: Ryong Lee (ryonglee@kisti.re.kr) and Sangtae Ahn (stahn@knu.ac.kr)

This work was supported by the Research and Development Project, Building a Data/AI-Based Problem-Solving System, at the Korea Institute of Science and Technology Information (KISTI), South Korea, under Grant K-23-L04-C05-S01.

ABSTRACT Recently, various text-to-image generative models have been released, demonstrating their ability to generate high-quality synthesized images from text prompts. Despite these advancements, determining the appropriate text prompts to obtain desired images remains challenging. The quality of the synthesized images heavily depends on the user input, making it difficult to achieve consistent and satisfactory results. This limitation has sparked the need for an effective prompt optimization method to generate optimized text prompts automatically for text-to-image generative models. Thus, this study proposes a prompt optimization method that uses in-context few-shot learning in a pretrained language model. The proposed approach aims to generate optimized text prompts to guide the image synthesis process by leveraging the available contextual information in a few text examples. The results revealed that synthesized images using the proposed prompt optimization method achieved a higher performance, at 18% on average, based on an evaluation metric that measures the similarity between the generated images and prompts for generation. The significance of this research lies in its potential to provide a more efficient and automated approach to obtaining high-quality synthesized images. The findings indicate that prompt optimization may offer a promising pathway for text-to-image generative models.

INDEX TERMS In-context few-shot learning, pretrained language model, prompt optimization, text-to-image generation.

I. INTRODUCTION

Recently, there has been a surge in research on multimodal models that simultaneously handle multiple data types, such as sound, images, and biological signals [1]. This surge is attributed to the developing of many high-performance deep-learning models in these domains [2], [3], [4], [5]. One of the most intriguing multimodal models is the text-to-image generative model, which generates images based on text prompts [6]. The quality of synthesized images from text-to-image generative models has significantly improved, and such models can now produce high-quality images comparable to those created by human painters [7], [8]. However, the quality of the generated images heavily relies on the input text provided by the user. If an inappropriate text prompt is used for a text-to-image generative model, the quality of the generated images may be poor and fail to reflect the users' intended results.

Various text prompts are generally tested multiple times to determine the best one based on the generated images. Numerous studies have been conducted to enhance the quality of the generated output by manipulating text prompts, a field commonly called prompt tuning. Prompt tuning has recently gained significant attention because existing generative models are trained on extremely large datasets, resulting in an unexplained high-dimensional latent space [9]. Furthermore, modifying the model structure or conducting additional training with new data can be costly and time-consuming.

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

Consequently, a significant amount of research has been aimed at enhancing the performance of text-to-image models through pretrained language models (PLMs) and prompt tuning [10].

This work aims to build an intelligent system that generates appropriate text prompts for a text-to-image generative model from a chat or conversation. To achieve this, we leveraged in-context few-shot learning techniques using multiple text examples reflecting a text format where users request specific illustrations from illustrators in the real world. We fine-tuned a large-scale PLM to understand the desired text prompt style and convert naive prompts into an optimized text-prompt style tailored to generative models. Instead of a rigid prompt approach that only corresponds to predefined prompt templates, we fine-tuned the language model to understand the format and methodology of optimized prompts and convert the style of the input naive prompts into the proposed prompt format. Additionally, we enhanced the generative model's ability to comprehend the user's intent more clearly by eliminating extraneous information from the naive prompt. Moreover, by transforming a single sentence into a collection of smaller sentence units or sets of words, we facilitated the generative model's understanding, thereby improving the quality of the synthesized images. After fine-tuning the language model, we evaluated its performance using text in the form of chats or conversations. Finally, we collected image results using optimized text prompts as inputs for text-to-image generation models. This approach using text prompts results in high-quality image generation, and performance improvement compared to non-optimized text prompts based on evaluation metrics. The contributions of this work are summarized as follows:

- Our approach optimizes text prompts from naive prompts using in-context few-shot learning within a large PLM, resulting in high-quality images that accurately capture users' intentions without the need for retraining large-scale text-to-image generative models.
- We demonstrate that the proposed prompt optimization method can enhance the performance of image-generative models in a fine-tuned text-to-image generative model for a specific task and other image-generative models for general tasks.
- We apply the proposed optimized prompt technique, which can handle any text prompt input, not limited to predefined templates. To achieve this, we fine-tune a language model by combining the advantages of hard and soft prompt techniques, enabling it to maximize performance regardless of the input prompt. This approach enables the image generation task to achieve a superior performance level.

II. RELATED WORKS

A. TEXT-TO-IMAGE GENERATIVE MODELS

Recently, several pretrained large-scale text-to-image generative models have been released [6], [7], [8], [11]. Representative models are described in the following sections.

1) DALL·E

DALL·E, a variational autoencoder (VAE)-based text-to-image generative model, has significantly contributed to generative models. One of its key strengths is successfully incorporating the transformer architecture widely used in natural language processing into computer vision. This interdisciplinary adaptation has facilitated generating impressive images from textual descriptions. DALL·E has consistently outperformed existing generative adversarial network (GAN)-based generative models regarding performance metrics.

DALL·E 2 successfully demonstrates relatively strong generalization performance through zero-shot learning, overcoming the limitations faced by its predecessor DALL·E in synthesizing images of unseen objects or styles and the inability to generate high-resolution images. Trained on a large dataset of image-text pairs, DALL·E 2 uses a sophisticated training process to establish a shared representation space for images and text, akin to the contrastive language-image pre-training (CLIP) approach. Leveraging the text encoder from CLIP, DALL·E 2 generates text embeddings combined with image embeddings derived from the provided text prompts. The model employs a Gaussian diffusion model decoder to progressively refine the generated images, synthesizing high-level images that closely align with the given text prompts. DALL·E 2 has gained significant attention due to its remarkable ability to generate diverse and visually appealing images tied to the text prompts. While the detailed code for the model has not been publicly disclosed to mitigate potential misuse, its contributions to the domain of generative models remain noteworthy.

2) STABLE DIFFUSION

More recently, stable diffusion [8] was introduced, which can synthesize high-quality images from text prompts using a diffusion training technique that applies noise to an image and restores it to its original state. Previous diffusion-based text-to-image generative models have a trade-off relationship between image quality and computational speed compared to other generative models [12] but stable diffusion addresses this problem by training on a subset of the large-scale image-caption dataset LAION-5B and employing a latent diffusion training technique that offers computational advantages by training in a latent space instead of an embedding space. Additionally, the model leverages relatively smaller text encoders such as U-Net [13] and the frozen CLIP vision transformer (ViT)-L/14 [14], making it more suitable for low-performance computing environments compared to large-scale text-to-image generative models.

Comprising three artificial neural networks (CLIP, U-Net, and VAE-stable diffusion), the model operates as follows. When a user enters text, the text encoder (CLIP) converts it into tokens, a language that U-Net can understand. Then, U-Net denoises the tokens based on generated random noise. As denoising is repeated, a proper image is generated and subsequently transformed into pixels by the VAE. Unlike

previous diffusion probability-based image generation models that require exponentially increasing resources with higher resolutions, stable diffusion incorporates an autoencoder before and after the process. This autoencoder inserts or removes noise in a smaller latent space dimension rather than the entire image, significantly reducing resource usage even when generating relatively high-resolution images. One notable advantage of stable diffusion over early text-to-image generative models is that it enables fine-tuning objects or styles not included in the training process. This flexibility is achieved by employing such techniques as textual inversion [15] or DreamBooth [16]. Furthermore, stable diffusion is publicly available, enabling widespread utilization without retraining.

B. PROMPT TUNING

Text-to-image generative models use user text prompts to generate images. Even when prompts contain identical content, the resulting images can exhibit variations during this process. This variation arises due to the inherent complexity of the generative process and the interpretability of textual descriptions. Prompt optimization or tuning techniques are employed to address this challenge and enhance the generation of high-quality images aligned with the user's intent [10], [17].

Prompt optimization involves strategically adjusting the prompt input to guide the generative model toward synthesizing images that better align with the desired outcome. This optimization process considers various factors, such as the desired content, style, composition, or other specific attributes users want to convey through the generated images [18]. By fine-tuning the prompt input, prompt tuning aims to reduce the discrepancy between user intent and the desired images.

A commonly used technique in prompt tuning is to provide explicit instructions within the prompt to guide the generative model's understanding. For instance, framing the prompt as "A photo of ___" prompts the model to focus on capturing the desired subject matter or scene depicted within the square brackets [19]. This approach helps direct the generative model's attention and improves the likelihood of generating images that align with the user's intended vision [20]. Prompt tuning has shown promise in generating high-quality images, enabling the generative model to grasp the user's intentions more effectively [10]. By optimizing the prompt input, users can exert greater control over the image synthesis process and obtain the desired images. This technique enhances the quality of generated images and enables users to express their creativity and preferences more accurately. For example, VisualChatGPT, which integrates various visual foundation models (VFMs) into one, has recently received considerable attention. When a user enters a text prompt, it matches it to the VFMs that best respond to the user input. Within VisualChatGPT, the prompt manager adds additional information to the user's prompts to increase the probability of matching appropriate VFMs or requests additional information from the user if necessary. While this approach improves performance, the authors manually created prompt templates for every VFM, resulting in heavy prompt tuning and slower computation speed. By relying on human execution, this approach poses reliability concerns. Additionally, it relies on the performance of ChatGPT itself, requiring considerable time to determine the most appropriate VFM. This approach is time-consuming and may not be suitable for building a naturally intelligent system.

Overall, prompt tuning is vital in improving the performance and user experience of text-to-image generative models and empowers users to achieve their desired visual outcomes with greater precision and satisfaction by tailoring the prompt input to align with users' specific requirements and creative objectives.

C. TECHNIQUES FOR PROMPT TUNING

1) DISCRETE AND CONTINUOUS PROMPTS

Effective prompt tuning requires a sophisticated approach to modifying the input text. Two methods commonly considered for this purpose are the discrete and continuous prompts. The discrete prompt method uses predefined fixed sentences or questions to constrain the range of possible responses generated by a natural language processing model. While this method limits the range of answers, its use of fixed-length words or sentences enables fast text processing by the model [21]. However, it may require additional training to manage new tasks. In contrast, the continuous prompt technique generates responses to questions without modifying or structuring the input text. This approach offers greater flexibility because it can manage various tasks without additional training. However, it is more dependent on the capabilities of natural language processing models and may be slower for models with numerous parameters [18].

2) SOFT AND HARD PROMPTS

Hard prompt and soft prompt are techniques related to the template of user input text. In the case of a hard prompt, only correspondence to a fixed template is allowed, and exact text matching is required. To explain better, we provide the following example applied to a generative model: "I want a picture of ___ style with ___ and ___." The user can only input text into the blanks, and other parts cannot be modified. Therefore, the model produces better results through the user's accurate input, but the flexibility is limited because we must define the template [10]. The soft prompt technique offers greater flexibility in responding to input text templates by allowing users to input text prompts freely without predetermined templates. This flexibility is a considerable advantage for a generative model handling various tasks. However, the soft prompt technique tends to have lower performance than the hard prompt technique due to the model's difficulty in finding accurate information as a result of the flexibility in the input [23].

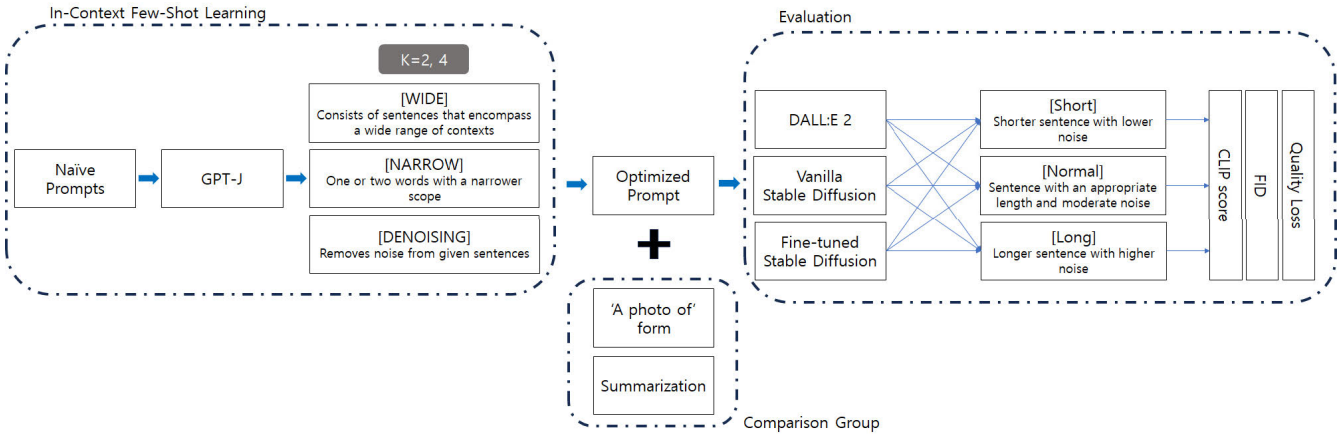

FIGURE 1. Overall pipeline of text-to-image generation using in-context few-shot learning. We used GPT-J as a language model for in-context few-shot learning and evaluated generation using three different generative models (DALL:E 2, vanilla stable diffusion, and fine-tuned stable diffusion). Three evaluation metrics are used (CLIP score, FID, and quality loss).

D. IN-CONTEXT FEW-SHOT LEARNING

In recent years, research has focused on applying few-shot learning to large-scale PLMs to achieve better performance on language-related tasks with limited data. The generative pretrained transformer 3 (GPT-3) [24] has demonstrated competitive performance in natural language understanding using few-shot learning, where properly designed prompts enable effective utilization of limited data without extensive fine-tuning on large datasets. The success of this approach emphasizes the importance of prompt tuning in determining the performance of PLMs in few-shot learning [25]. Manual prompting has been primarily employed to enhance the ability of PLMs to learn desired text styles more effectively [17]. Manual prompt-based few-shot learning enables PLMs to perform text style transfer tasks and generate text in specific desired styles [26]. While several studies have improved the performance of multimodal tasks through prompt-based few-shot learning with PLMs [27], more sophisticated prompt optimization techniques that do not require human effort are needed.

The core concept of in-context few-shot learning involves fine-tuning large-scale PLMs to suit specific tasks. This allows language models to comprehend language in a manner better aligned with the data and objectives of the given task. Our approach effectively operates by providing contextual information when training, using pairs of data, including prompts with noise and data transformed into a format that enhances the performance of the generative model. This applies to both prompts containing noise and data that has been modified to improve the performance of the generative model. Furthermore, in-context few-shot learning finds applications across a spectrum of natural language processing tasks, including machine translation, question-answering, text classification, and text generation. It has been instrumental in driving groundbreaking advancements within the field of natural language processing. In this manner, in-context few-shot learning stands as a pivotal strategy, shaping the cutting-edge technology and research landscape in the domain of natural language processing, by adeptly customizing PLMs for specific tasks, thereby achieving remarkable performance.

E. PRETRAINED LANGUAGE MODELS

Similar to its predecessors, GPT-3 [24] and GPT-2 [28], GPT-J [29] is a publicly available large-scale pretrained model for natural language processing. It operates as an autoregressive model, focusing solely on the decoder component of the standard transformer model. As a causal language model, GPT-J generates predictions based solely on the preceding words and context without relying on context from both directions, as observed in masked language models. The autoregressive nature of GPT-J allows it to use previous predictions as input for subsequent predictions, resulting in more coherent and contextually relevant output. Additionally, GPT-J offers computational efficiency and cost-effectiveness compared to GPT-3, owing to its smaller size. Therefore, we used GPT-J to fine-tune user input prompts in the desired direction.

F. TEXT STYLE TRANSFER

Text style transfer (TST) is a technique that leverages language models to transform the style of text. This technology enables the transformation of given text into different styles, including modifications to writing style, tone, and language style. It utilizes language models such as GPT to comprehend the context and style of the text, aiding in the reconfiguration of input text into the desired style. One notable feature of this technology is its compatibility with few-shot learning techniques. In our approach, we trained the GPT-J model with a limited amount of training data to generate text in the desired style. Since we used a small amount of training data, high-quality data was crucial. To address this, we created a dataset by referencing text requesting illustrations from illustrators, ensuring the dataset closely resembled the task at hand.

III. METHODS

A. IN-CONTEXT FEW-SHOT LEARNING

1) DENOISING TEXT PROMPT

We considered a conversational-style text, resembling informal conversation and chat, to have potential noise when used as text prompts for text-to-image generative models. For instance, when a user requests "Hello! Can you create a ___ picture for me ___?" the conversational elements, such as "Hello!" and "Can you create a ___" can introduce noise that hampers the model's understanding of the user's intentions. To address this challenge, we employed the TST task to guide the language model in transforming the conversational text into a more appropriate prompt format. The proposed approach used a language model with few-shot learning capabilities and established an experimental setup comprising a limited number of data pairs in the desired style to facilitate the desired TST.

2) PROMPT STYLE TRANSFER

The task aims to convert conversation-style text to a prompt for a text-to-image generative model to generate images that accurately reflect users' intentions. This task poses unique challenges due to its nonuniversal nature and the scarcity of available datasets. Training such models requires careful consideration of various constraints. Previous studies indicated the effectiveness of prompt-based learning, highlighting its flexibility [30]. To address this task, we adopted a prompt-based learning approach. We manually generated a few text examples and incorporated them into the prompts, facilitating the PLM to learn TST using a simple yet effective, manually crafted prompting method. By leveraging prompt-based learning, we empowered the PLM to better understand and capture the desired style, enabling the generation of high-quality images that align with user preferences.

3) PRETRAINED LANGUAGE MODELS

Generation-based TST is commonly implemented using sequence-to-sequence architectures in PLMs. Employing prompt-based few-shot learning during training makes it feasible to perform style transfer using only few-shot data through the decoder of PLMs. For this study, we considered several PLMs for candidates, including GPT-J [29], T5 [31], and GPT-2 [28]. After careful evaluation, we selected GPT-J as the target PLM due to its ample model parameters that facilitate effective few-shot learning. Therefore, we employed GPT-J to fine-tune user input prompts in the desired direction.

B. TEXT-TO-IMAGE GENERATIVE MODELS

As part of the evaluation process, we used DALL·E 2 to assess the generalization performance of the proposed approach. By leveraging the strengths of this state-of-the-art model, we aimed to ascertain the effectiveness and robustness of the prompt optimization technique.

To compare the performance of the proposed method with conventional text prompt input, we used vanilla and fine-tuned stable diffusion models. To assess the effectiveness of the prompt optimization approach for a specific task in a text-to-image generative model, we performed fine-tuning on stable diffusion using a fairy-tale image dataset. We gathered approximately 2,100 fairy-tale book copies from publicly available sources on the web and applied a set of refinement rules. Within the dataset, we excluded images containing text, real-life photographs in fairy-tale books, illustrations deemed unsuitable for children, and illustrations with unclear styles. After applying these refinement rules, we obtained approximately 700 images, further categorized into three distinct fairy-tale styles: 260 traditional-style images, 155 cartoon-style images, and 250 illustration-style images. To ensure style consistency across the categories, we conducted a thorough review, determining that illustrations used for fairy-tale book covers exhibited the most consistent style and best aligned with the objective. Therefore, we proceeded with fine-tuning the model using the illustration style. We employed the textual inversion technique during the fine-tuning process using a limited number of images to learn novel concepts not addressed by stable diffusion. This technique involves the embedding of text prompts into an encoder, where it is tokenized, and then proceeds to discover new embeddings that represent specific visual concepts provided by the user. These embeddings are subsequently associated with pseudo-words and can be integrated into new sentences like any other words. This process can be described as finding an appropriate latent space within the frozen model's text-embedding space. Through this technique, we fine-tuned the stable diffusion model, enabling it to consistently generate images in the style of a children's book.

C. PROMPT TUNING

1) CONTINUOUS PROMPTS

Considering the limitations of the discrete prompt in producing the desired results and the need for adaptability in scenarios with unknown user input, we determined that the continuous prompt approach, with its flexibility and creativity, while minimizing human intervention, is a more suitable method. Additionally, we selected GPT-J as a language model that balances performance and cost trade-offs and further fine-tuned it using in-context few-shot learning.

2) SOFT PROMPTS

We used the soft prompt technique to enable flexible responses to user inputs, increasing model flexibility. It allowed the model to manage a wide range of user inputs. Further, we performed additional fine-tuning to enhance model performance, resulting in significant improvements. We were able to generate better results through these efforts.

D. IMAGE STYLE TRANSFER

1) OBJECT IMAGE STYLE TRANSFER

One of the major tasks in text-to-image generative models is image style transfer, which translates source images to target

images based on text prompts. To translate from an object in an image to another object (object image style transfer) using text prompts, we constructed the following pipeline: object detection, binary masking, and inpainting. We adopted Grounding DINO [32] for object detection based on bounding boxes and the detected bounding boxes were binarized to 0 (black) or 255 (white) for inpainting. After that, we used text prompts to translate the detected object to another using stable diffusion.

2) BACKGROUND IMAGE STYLE TRANSFER

Image style transfer for an object is relatively straightforward since detecting an object (object detection) is a widely used task. Another task is translating the background to another and this task may be more difficult since sophisticated object segmentation is needed. To translate the background in an image, we constructed the following pipeline: object segmentation, binary masking, inversion, and inpainting. We adopted the Grounded-SAM [32], [33] for object segmentation and performed binary masking and inversion to capture the background. Then, we performed inpainting to translate the background to another with text prompts using stable diffusion.

3) WHOLE IMAGE STYLE TRANSFER

To translate the whole image into another, we adopted ControlNet [34]. Based on the canny edge detection, we captured the whole image structure and performed inpainting using stable diffusion.

E. CLIP SCORE

The CLIP [19] pretrained multimodal model achieves joint representation between text and images using a contrastive methodology. It was trained on 400 million image-text pairs on the web and evaluated similarity in the joint embedding space between the image and text for use in various multimodal tasks in a zero-shot manner. The CLIP score [35] uses CLIP to compute the cosine similarity score between images and text in the joint embedding space to evaluate the ability of text-to-image models to generate images close to the input prompt. The computation process of the CLIP score involves the following steps:

- Text embedding: The textual description provided is encoded using the CLIP model, resulting in the generation of a high-dimensional vector known as a 'text embedding'. This vector captures the underlying semantic meaning of the text description.
- Image embedding: CLIP also possesses the capability to encode images into embeddings. These image embeddings represent the images as high-dimensional vectors.
- Mutual computation: To calculate the CLIP score, a measurement of similarity is performed between the text embedding and the image embedding. This similarity assessment often involves mathematical operations such as dot products or other similarity metrics. The outcome of this computation quantifies the degree of compatibility between the given text and image.
- Comparison and ranking: Subsequently, CLIP employs the computed CLIP Score to conduct a comparison and ranking of multiple images relative to the provided text. Images with higher CLIP scores are considered to be more closely aligned with the text description.

We used the CLIP score as the evaluation metric for how much the proposed model improves compared to the original naive (noisy) text prompt:

$$\text{CLIP score}(\mathbf{p}, \mathbf{i}) = \omega^* \max(\cos(\mathbf{p}, \mathbf{i}), 0) \tag{1}$$

where $\mathbf{p}$ and $\mathbf{i}$ are the embeddings for text prompts and generated images, respectively. We set ω as 2.5.

F. FRECHET INCEPTION DISTANCE

The Frechet inception distance (FID) [36] serves as a pivotal metric for assessing image quality in the context of image generation models. It accomplishes this task by quantifying the disparity between the distribution of generated images and that of real images, capitalizing on the hidden feature layer embedded within the Inception-v3 [37]. FID holds a prominent role in the evaluation of image generation models, with particular emphasis on its utility in appraising GANs and other deep learning-based generative models. One of FID's key strengths lies in its ability to furnish an objective and precise performance measure, extending its value as a guiding compass for refining and optimizing model development processes. Furthermore, FID's multi-faceted approach involves the examination of a hidden feature layer, assessing image quality, computing the Frechet distance between feature vector distributions of generated and real images, and interpreting the resulting scores. Lower FID scores signify enhanced visual and semantic congruence between generated and real images, while elevated scores denote a greater divide, suggesting diminished quality in the generated image set. We utilized the FID as an evaluation metric for assessing the quality of images generated by the DALL·E 2, vanilla stable diffusion, and finetuned stable diffusion models under both naive prompts and prompts optimized through our proposed technique. The calculation of FID is as follows:

$$\mathbf{FID} = \|\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{Y}}\|^2 - \mathbf{Tr}(\sum_X + \sum_Y - 2(\sum_X \sum_Y)^{1/2}) \tag{2}$$

The symbols $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ denote the mean feature vectors for the sets of images generated from naïve and optimized prompts, respectively. The covariance matrices for these sets of images are represented by $\sum_X$ and $\sum_Y$, again corresponding to the naïve and optimized prompt sets. Lastly, the symbol $\boldsymbol{Tr}$ indicates the trace of a matrix, which is the sum of the diagonal elements.

G. QUALITY LOSS

We propose a novel evaluation metric (quality loss), which serves as a form of human evaluation. Text-to-image

generation models are heavily influenced by the prompts provided by users. Particularly, in the case of sentences containing significant noise, it has been observed that generated images may include characters or text elements. To address this, we conducted a manual count of instances where characters or text were present in the generated images, dividing this count by the total number of generated images to calculate the probability of characters or text being present in generated images (PC in equation (3)). Subsequently, we multiplied this value by the CLIP score, resulting in the final evaluation conducted through the quality loss. This metric assesses how similar the output image is to the prompt, and whether characters or text are present in the generated image.

$$\text{Quality Loss} = \text{CLIP score}(\mathbf{p}, \mathbf{i}) * \text{PC} \tag{3}$$

where $\mathbf{p}$ and $\mathbf{i}$ are the embeddings for text prompts and generated images, respectively. We set ω as 2.5.

IV. EXPERIMENTS

We conducted a series of experiments to evaluate the effectiveness of prompt optimization based on in-context few-shot learning in improving the quality and performance of text-to-image generative models.

A. IN-CONTEXT FEW-SHOT LEARNING

We used GPT-J as the baseline model for the task. To tailor few-shot learning to our specific objectives, we defined the number of sentences used in in-context learning as K and trained the GPT model to align with our objectives using K sentences. Thus, we performed few-shot learning with $K = 2$ and $K = 4$. The objective was to reduce the noise in the input text prompts; thus, we added noise in the text for training.

1) PROMPT TEMPLATE

To enhance the performance of prompt optimization, we applied training data with a consistent format: [Noise] + [Request to draw a picture] + [Noise] + [Description of background] + [Description of specific objects or situations] + [Noise].

2) TASKS

We designed three distinct tasks to determine optimal prompts. The first task, *[WIDE]*, consists of sentences that encompass a wide range of contexts. The second task, *[NARROW]*, involves selecting one or two words with a narrower scope. The last task, *[DENOISING]*, removes noise from given sentences. We conducted experiments using these three tasks. The following task examples were used.

Input:

"Hello. This is my first request, so it may be difficult. It's soon New Year's Day and the background is a full moon. And…and there's a black rabbit wearing a traditional Korean hanbok standing there. Could you please draw this for me as soon as possible?"

Output:

[WIDE] New Year's Day, background is a full moon, black rabbit wearing a traditional Korean hanbok standing there.

[NARROW] New Year's Day, full moon, background, black rabbit, traditional Korean hanbok, standing.

[DENOISING]

The background is a full moon and draw a black rabbit wearing a traditional Korean hanbok, standing on New Year's Day.

3) IN-CONTEXT LEARNING TEXT TEMPLATE

We replicated the text template used in a prior successful research study on in-context few-shot learning. The text template is "Here is some text: {naive text prompt}. Here is a rewrite of the text, which is simpler: {optimized text prompt}." This approach was employed to leverage the GPT-J model for TST.

4) PARAMETERS

For the experiment, we trained the GPT-J model for TST using in-context few-shot learning, and the parameters used during that time are as follows. First, the parameter 'Temperature', which regulates the probability distribution of the language model, is typically set to a positive value of less than 1. When it approaches to 0, the probability distribution becomes sharper, encouraging the selection of tokens with higher probabilities and resulting in sentences with consistent features. When it approaches to 1, the probability distribution flattens, increasing randomness and yielding more diverse and creative sentences. To receive soft prompts and achieve more human-like results, we set this parameter to 0.8 during training. Additionally, we set the 'do_sample' value to 'True', indicating that the model should sample tokens probabilistically, in contrast to the 'False' setting where it always selects the token with the highest probability. This choice increases diversity in generated outputs.

B. EVALUATION

To assess the robustness of the proposed approach, we conducted experiments using different prompt templates. We tested four complete sentences divided into three cases: 1) *[NORMAL]*, representing a sentence with an appropriate length and moderate noise; 2) *[LONG]*, representing a longer sentence with higher noise; and 3) *[SHORT]*, representing a shorter sentence with lower noise. We calculated the average CLIP score for each case and determined the optimal K value. The following sentences provide examples of the evaluation:

[NORMAL]

"I'm not sure if it's okay to ask this, but I had a dream last night. Two rabbits were fighting with laser swords in a space background on a spaceship. And the black rabbit won. Can you draw a picture of this scene for me?"

[LONG]

"Hello, nice to meet you. I want to make an illustration for a webcomic cover, can you draw it for me? Well, about the webcomic, it's about secrets involving treasure in a dungeon,

TABLE 1. Clip score in DALL·E 2, fine-tuned and vanilla stable diffusion.

Model	DALL·E 2	*Fine-tuned stable diffusion*	*Vanilla stable diffusion*	*Average score*
Naïve prompt (noisy)	31.10	29.35	28.82	29.76

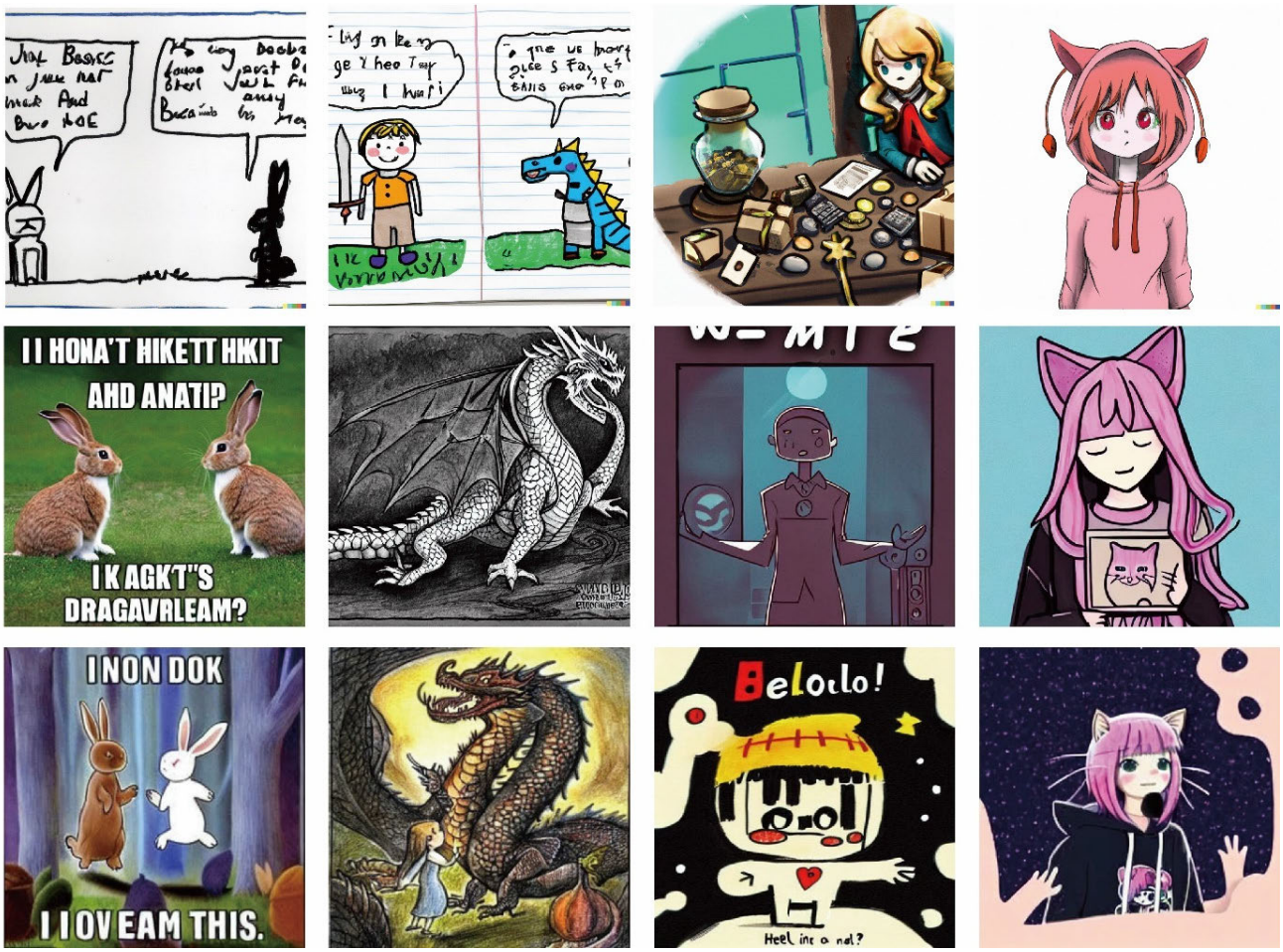

FIGURE 2. Generated images in DALL·E 2 (top row), fine-tuned stable diffusion (middle row), and vanilla stable diffusion (bottom row) from naive text prompts with noise.

and…you know why already, um…the illustration I need is of a 15-year-old blonde girl holding a sword while passing through a dungeon.''

[SHORT]

''Hello, please draw an illustration. Um…and…a pink-haired girl wearing a hoodie with cat ears, around the age of 16, in the style of Japanese animation. Please.''

V. RESULTS

We evaluated the performance of the prompt optimization in the three text-to-image generative models: DALL·E 2, vanilla stable diffusion, and fine-tuned stable diffusion.

A. EVALUATION FROM NAÏVE TEXT PROMPTS

We conducted experiments using the DALL·E 2 model to assess its performance in a general scenario. The average CLIP score for naive text prompts with noise was 31.10 (Table 1). The scores for fine-tuned and vanilla stable diffusion were 29.35 and 28.82, respectively. DALL·E 2 achieved the highest score among the naive text prompts. The fine-tuned stable diffusion displayed a slight improvement compared to the vanilla version. Examples of generated images from naive text prompts in DALL·E 2 are presented in Fig. 2 (top row).

B. EFFECT OF K VALUES

In in-context few-shot learning, selecting the value of $\boldsymbol{K}$ is critical. In the experiments, we focused on two values of $\boldsymbol{K}$ (2 and 4) to demonstrate that prompt optimization is possible even with minimal training data. Generally, the performance of 4-shot learning was superior to 2-shot learning across models (Table 2). However, in the *[SHORT]* and *[NORMAL]* cases, 2-shot learning performed better than 4-shot learning. Thus, a smaller value of $\boldsymbol{K}$ is sufficient when dealing with

TABLE 2. Clip score in DALL·E 2, fine-tuned and vanilla stable diffusion.

Model	DALL·E 2	*Fine-tuned stable diffusion*	*Vanilla stable diffusion*	*Average score*
K=2 [WIDE]	34.90	34.00	33.57	34.16
K=2 [NARROW]	**35.88**	**34.52**	**34.84**	**35.08**
K=2 [DENOISING]	35.63	33.10	33.17	33.97
K=4 [WIDE]	35.34	34.53	34.50	34.79
K=4 [NARROW]	35.30	33.98	33.68	34.32
K=4 [DENOISING]	36.80	34.37	33.35	34.84

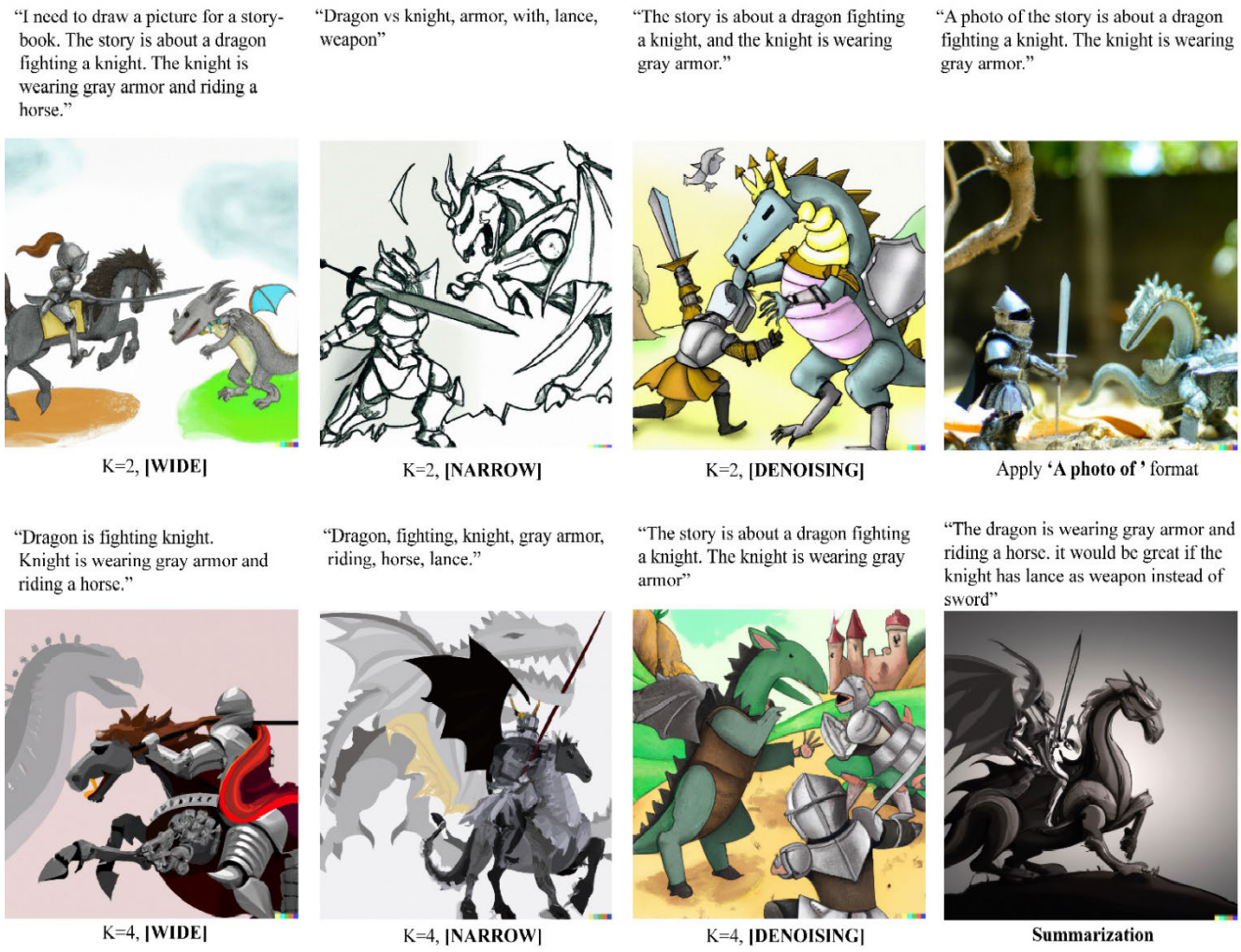

FIGURE 3. Generated images using prompt optimization based on the in-context few-shot learning technique of the following naive text: "How are you? To be honest, I need to draw a picture for a storybook. The story is about a dragon fighting a knight. The knight is wearing gray armor and riding a horse. Oh, and it would be great if the knight has a lance as a weapon instead of a sword. Please let me know as soon as possible."

low noise levels. In the *[LONG]* case, 4-shot learning outperformed 2-shot learning because more noise and the need to extract multiple keywords necessitated more training examples for effective few-shot learning. In other words, when sentences with higher noise levels require optimization, the amount of noise to be removed increases, and a higher $\boldsymbol{K}$ value becomes necessary. Examples of generated images from prompt optimization with $\boldsymbol{K} = 4$ in the *[NARROW]* task using DALL·E 2 are illustrated in Fig. 3 (lower row).

C. EVALUATION OF THREE TASKS

We designed three tasks to identify the most effective prompt optimization method. Upon analyzing the experimental results, the *[NARROW]* and *[DENOISING]* tasks

TABLE 3. Clip score in DALL·E 2, fine-tuned and vanilla stable diffusion. - apply "a photo of" form.

Model	DALL·E 2	*Fine-tuned stable diffusion*	*Vanilla stable diffusion*	*Average score*
"A photo of" form	37.01	34.37	33.35	34.84

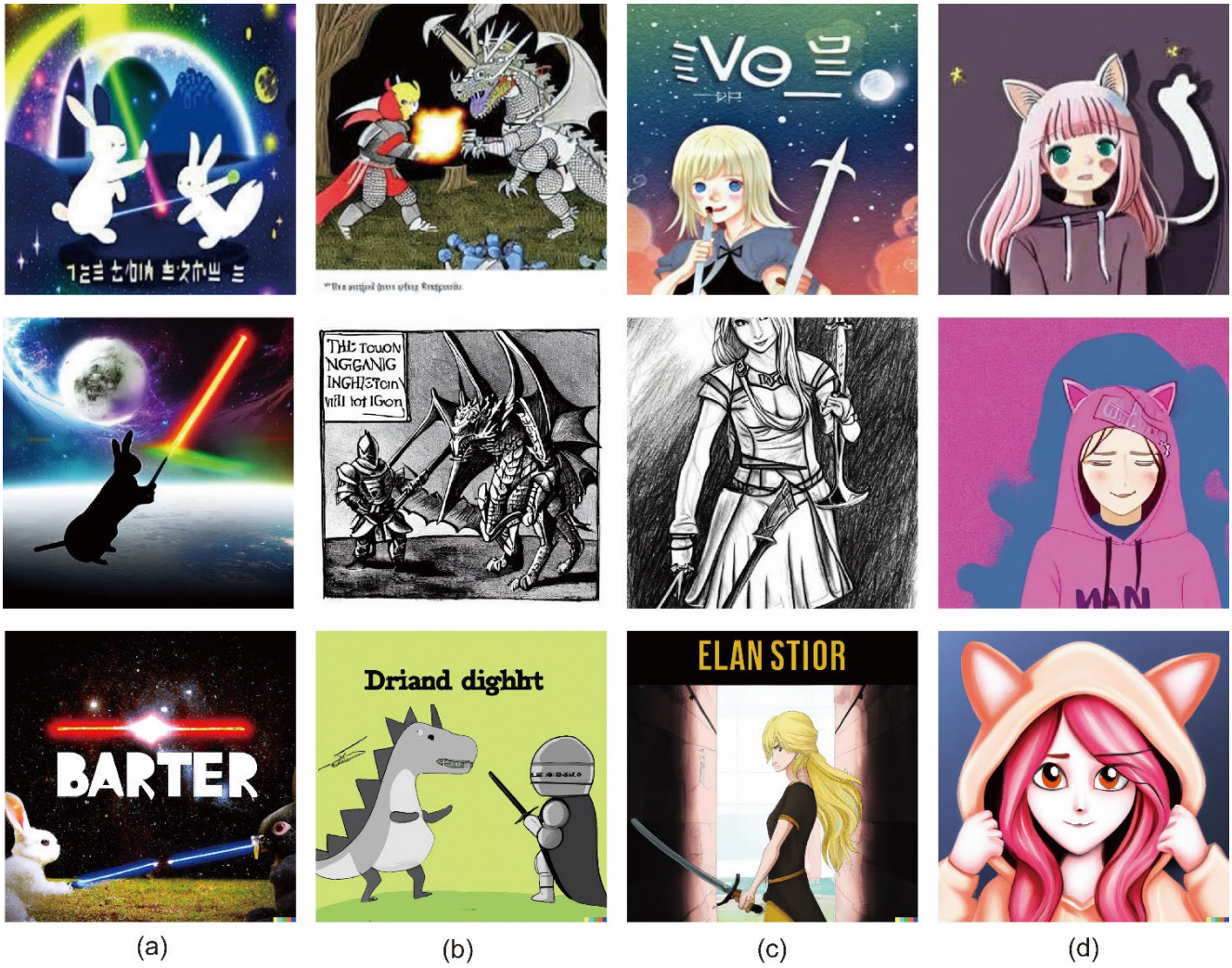

FIGURE 4. Generated images from (a) "a photo of [ ]" form, (b) [NORMAL], (c) [LONG], (d) [SHORT] when K = 4 in [DENOISING], DALL·E 2 (top row), fine-tuned stable diffusion (middle row), and vanilla stable diffusion (bottom row).

demonstrated clear strengths across all models. However, when $K = 2$, the performance of *[DENOISING]* was not as strong as *[NARROW]*. Overall, *[DENOISING]* with $K = 4$ performed well. However, there were occasional problems where letters appeared in the generated images, which we speculate may be due to the input sentence leading to dialogue as part of the *[DENOISING]* process. Considering the experimental results discussed in the previous section, we concluded that the prompt optimization method must perform well even in the *[LONG]* task, which involves long sentences with significant noise. Thus, we determined that *[NARROW]* achieved the best overall performance. Fig. 3 provides examples of generated images from prompt optimization across different K values and tasks using DALL·E 2.

D. EFFECT OF "A PHOTO OF" FORM

Several studies have demonstrated that using the format "a photo of ___" can improve performance. Hence, we applied this format in these experiments (Table 3). The optimal configuration was achieved with $K = 4$ and *[DENOISING]* for the "a photo of" prompt. However, for the *[NORMAL]* and *[SHORT]* cases, the best performance was observed with $K = 4$ and *[DENOISING]* without using the "a photo of"

TABLE 4. Clip score in DALL·E 2, fine-tuned and vanilla stable diffusion compared with summarization.

Model	DALL·E 2	*Fine-tuned stable diffusion*	*Vanilla stable diffusion*	*Average score*
T5	33.4	33.1	32.90	33.13

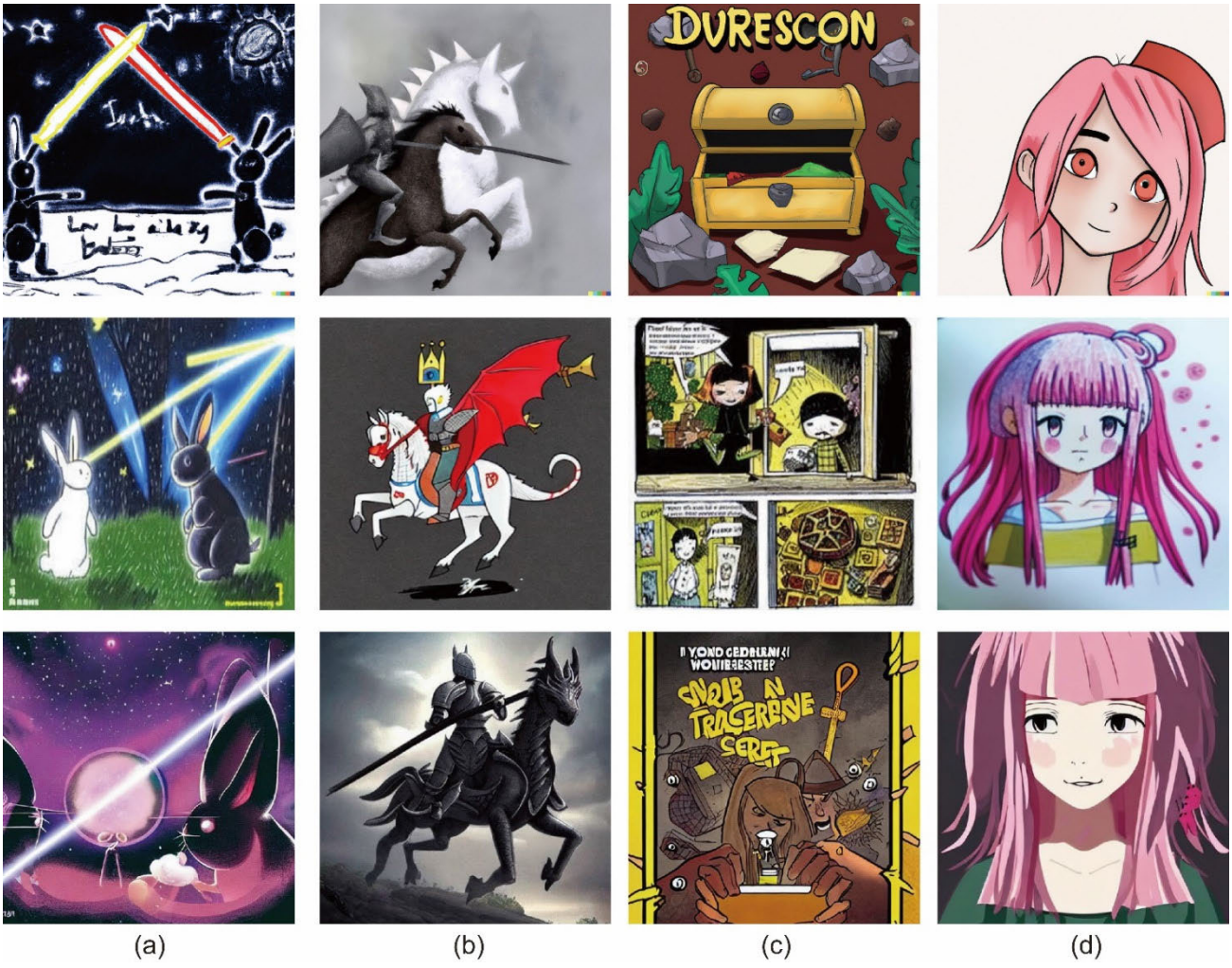

FIGURE 5. Generated images from sentence summarization using the T5 base model. Letters are displayed in (a) and (c). In (b), the intended prompt was "gray armored knight fighting against a dragon" but "fighting against a dragon" was omitted. In (d), the prompt was "a girl with pink hair wearing a hoodie and having cat ears" but "hoodie" and "cat ears" were not displayed. DALL·E 2 (top row), fine-tuned stable diffusion (middle row) and vanilla stable diffusion (bottom row).

form. In the case of fine-tuned stable diffusion, the performance was better with $K = 2$ and *[NARROW]* than with the "a photo of" prompt. Comparable results were observed in vanilla stable diffusion. Furthermore, when using the prompt "a photo of + *[DENOISING]*," we encountered a problem where letters appeared in the generated images (Fig. 4). These findings suggest that applying this form in prompt optimization can negatively affect image quality.

E. SENTENCE SUMMARIZATION

We compared prompt optimization and simple sentence summarization using the T5 base model. The results revealed that prompt optimization outperformed the naive text prompt (29.76) but had a lower score (33.13) than the lowest performance achieved with prompt optimization (34.16) in Table 4. This outcome indicates that simple sentence summarization fails to effectively remove noise from sentences and may exclude important keywords, resulting in suboptimal image generation. Examples of generated images from sentence summarization are presented in Fig. 5.

F. PROBABILITY OF LETTERS PRINTED ON IMAGES

We calculated the probability of letters appearing in the generated images and observed that the lowest probability occurred when $K = 4$ in *[NARROW]* (Table 5). Additionally, we calculated the quality loss by multiplying the

TABLE 5. Quality Loss in DALL·E 2, fine-tuned and vanilla stable diffusion.

Case	Naïve	K=2, [WIDE]	**K=2, [NARROW]**	K=2, [DENOISING]	K=4, [WIDE]	K=4, [NARROW]	K=4, [DENOISING]	"a photo of" Form	Summarization
Probability of letter appearance	13.28%	7.71%	**10.4%**	5.00%	0.78%	0.63%	1.41%	0.94%	7.34%
Quality Loss	3.95	2.63	**0.37**	1.70	0.27	0.21	0.49	0.33	2.43

TABLE 6. FID in DALL·E 2, fine-tuned and vanilla stable diffusion.

Model	DALL·E 2	*Fine-tuned stable diffusion*	*Vanilla stable diffusion*	*Average score*
FID	39.74	14.21	22.78	25.57

TABLE 7. FID in image style transfer for three tasks (object, background, whole).

	Object	*Background*	*Whole*	*Average*
FID	6.88	27.35	22.47	18.90

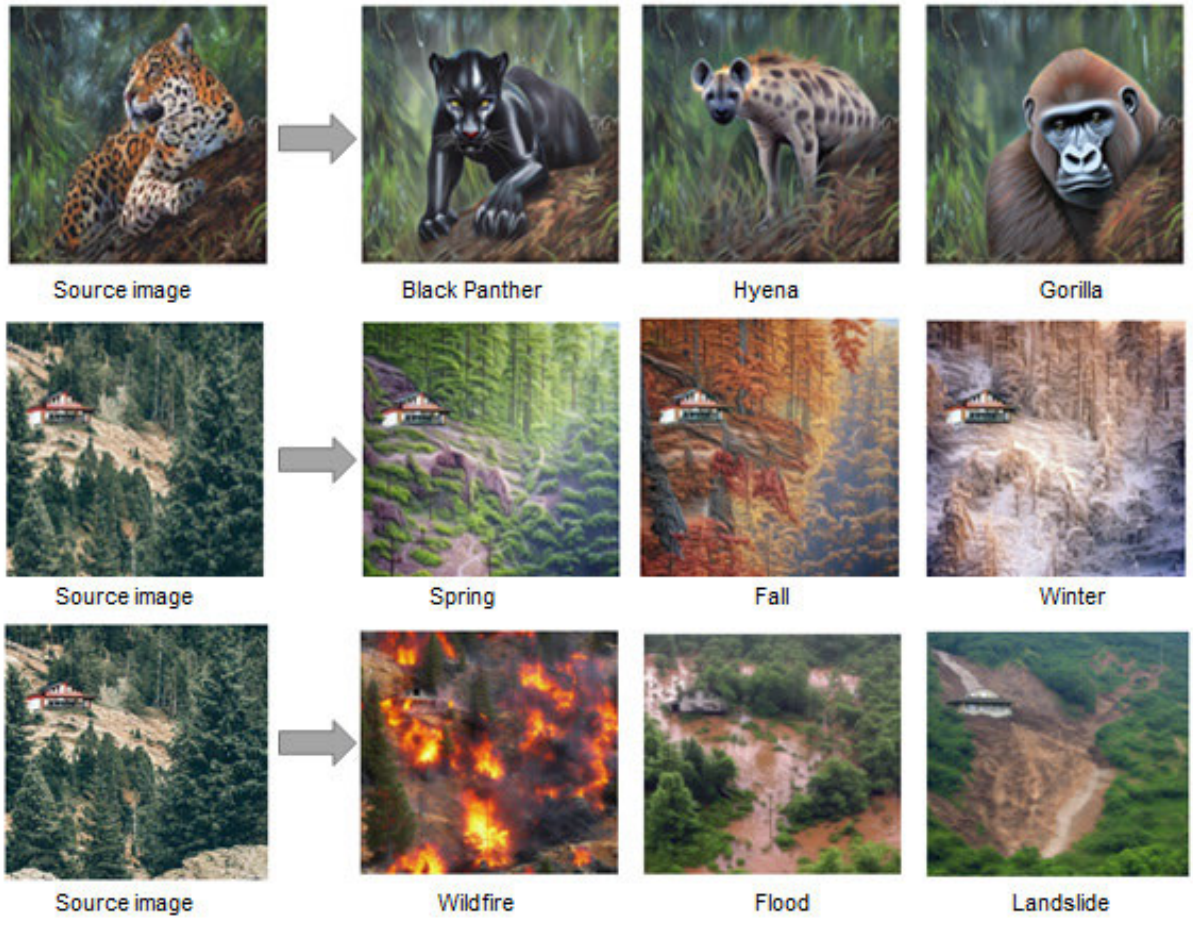

FIGURE 6. Examples of image style transfer for three tasks (top: object, middle: background, bottom: whole image style transfer). Source images are translated to target images by text prompts.

probability of letter appearance by the average CLIP score. The results indicated that the case with $K = 4$ in *[NARROW]* exhibited the lowest quality loss. This finding suggests that four-shot learning with short words (i.e., *[NARROW]*) produces high-quality images using prompt optimization.

G. FID SCORE IN GENERATIVE MODELS

We compared the FID values of images generated using the initial naive prompts provided by users with those generated using our proposed optimized prompts. This served as a metric to assess whether the user's intent was well-preserved

before and after prompt optimization. This was necessary because the [WIDE], [NARROW], and [DENOISING] techniques may alter the original sentence structure and could potentially return sentence structures different from the user's intent. The experimental results (Table 6) showed that all three models consistently exhibited low average FID values (25.57). This demonstrated that even when breaking down a single long sentence into smaller parts, prompt optimization effectively captures the user's intent, affirming our ability to optimize prompts.

H. FID SCORE IN IMAGE STYLE TRANSFER

To investigate how the proposed method applies to image style transfer, we calculated FID scores for three different tasks (object, background, and whole image style transfer). The first task was to detect objects with image style transfer using the proposed method (details in Methods), object detection, and segmentation. For image style transfer, we had three different tasks: object image style transfer, background image style transfer, and whole image style transfer. From the original image datasets including ImageNet, MS-COCO, and LAION, we performed data augmentation using image style transfer and evaluated FID scores between the original image datasets and style-transferred images. We found FID scores of 6.88 for the object image style transfer, 27.35 for the background image style transfer, and 22.47 for the whole image style transfer. On average, we obtained an FID score of 18.90 (Table 7). Examples of image style transfer for the three tasks are shown in Fig. 6. These findings indicate that our proposed method can be used as a data augmentation method for image generation.

VI. DISCUSSION AND CONCLUSION

This study proposes a prompt optimization method to enhance the performance of text-to-image generative models and effectively capture users' intentions. Through the experiments, encompassing various tasks and sentence lengths, four-shot in-context learning, particularly when the text prompts consist of a few words, yields superior results. Compared to conventional methods, such as simple sentence summarization, the proposed prompt optimization technique outperforms others in terms of noise removal and inclusion of critical keywords, resulting in more accurate and visually appealing image generation.

The significance of the findings lies in the potential application of prompt optimization with large PLMs in various domains. By allowing users to fine-tune the image generation process and achieve the desired visual output, the proposed approach offers new possibilities for intelligent image generation systems. Whether in creative arts, advertising, or other fields that rely on visually compelling content, the proposed prompt optimization method offers a valuable tool for users to express their creativity and meet their specific requirements. In conclusion, this study demonstrates the effectiveness of prompt optimization in improving the performance of text-to-image generative models. We believe that this research contributes to the advancement of intelligent image generation systems and inspires further exploration in the field of natural language processing and computer vision.

REFERENCES

[1] J. Summaira, X. Li, A. M. Shoib, S. Li, and J. Abdul, "Recent advances and trends in multimodal deep learning: A review," 2021, *arXiv:2105.11087*.
[2] W. Zhang, X. Sun, L. Zhou, X. Xie, W. Zhao, Z. Liang, and P. Zhuang, "Dual-branch collaborative learning network for crop disease identification," *Frontiers Plant Sci.*, vol. 14, pp. 1–14, Feb. 2023, doi: 10.3389/fpls.2023.1117478.
[3] W. Zhang, L. Zhou, P. Zhuang, G. Li, X. Pan, W. Zhao, and C. Li, "Underwater image enhancement via weighted wavelet visual perception fusion," *IEEE Trans. Circuits Syst. Video Technol.*, p. 1, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10196309, doi: 10.1109/TCSVT.2023.3299314.
[4] Z. Pang, L. Zhao, Q. Liu, and C. Wang, "Camera invariant feature learning for unsupervised person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 1–12, Sep. 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9891821
[5] Z. Pang, C. Wang, L. Zhao, Y. Liu, and G. Sharma, "Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, p. 1, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10234457
[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. (PMLR)*, Jul. 2021, pp. 8821–8831. [Online]. Available: https://proceedings.mlr.press/v139/ramesh21a.html?ref=journey-matters
[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
[8] 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, doi: 10.1109/cvpr52688.2022.01042.
[9] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, pp. 1–13, *arXiv:1702.08608*.
[10] Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein, "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery," 2023, pp. 1–15, *arXiv:2302.03668*.
[11] 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.
[12] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Adv. Neural Inf. Process Syst.*, vol. 11, 2021, pp. 8780–8794.
[13] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016, *arXiv:1602.02410*.
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[15] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022, *arXiv:2208.01618*.
[16] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–21.
[17] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–46, 2021.
[18] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, vol 1, no. 1, Apr. 2022, pp. 1–27, doi: 10.1145/3491102.3501825.
[19] 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," 2021, *arXiv:2103.00020*.
[20] N. Pavlichenko, F. Zhdanov, and D. Ustalov, "Best prompts for text-to-image models and how to find them," 2022, *arXiv:2209.11711*.

[21] M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, and Z. Hu, "RLPrompt: Optimizing discrete text prompts with reinforcement learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 3369–3391.

[22] Q. Zhu, B. Li, F. Mi, X. Zhu, and M. Huang, "Continual prompt tuning for dialog state tracking," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2022, pp. 1124–1137, doi: 10.18653/v1/2022.acl-long.80.

[23] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3045–3059, doi: 10.18653/V1/2021.EMNLP-MAIN.243.

[24] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, May 2020, pp. 1877–1901.

[25] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," 2021, *arXiv:2103.10385*.

[26] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, and J. Wei, "A recipe for arbitrary text style transfer with large language models," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2022, pp. 837–848, doi: 10.18653/v1/2022.acl-short.94.

[27] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," 2022, *arXiv:2212.09611*.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[29] B. Wang and A. Komatsuzaki, "GPT-J-6B: A 6 billion parameter autoregressive language model," Tech. Rep., 2021. [Online]. Available: https://academictorrents.com/details/feb5891fd364f357b03a9ebbf3b7d83a0aabe1ec

[30] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–37, 2023.

[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, Oct. 2019.

[32] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," 2023, *arXiv:2303.05499*.

[33] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.

[34] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[35] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 7514–7528, doi: 10.18653/v1/2021.emnlp-main.595.

[36] 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," in *Proc. Adv. Neural Inf. Process Syst. (NIPS)*, Dec. 2017, pp. 6627–6638, doi: 10.18034/ajase.v8i1.9.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

SEUNGHUN LEE received the B.S. degree in electronic engineering from Kyungpook National University, Daegu, South Korea, in 2022, where he is currently pursuing the M.S. degree with the School of Electronic and Electrical Engineering. His current research interests include expansive language models and image generation models, in particular, he is also conducting research on prompt engineering, involving fine-tuning large language models for their integration as inputs into image generation models.

JIHOON LEE received the B.S. degree in electronic engineering from Kyungpook National University, Daegu, South Korea, in 2023, where he is currently pursuing the M.S. degree with the School of Electronic and Electrical Engineering. His research interests include artificial intelligence and particularly in computer vision. Within this domain, he focuses on research related to generative models, with a specific emphasis on text-to-image synthesis.

CHAN HO BAE received the B.S. degree in metallurgical engineering from Pukyong National University, Busan, South Korea, in 2022. He is currently pursuing the M.S. degree with the School of Electronic and Electrical Eengineering, Kyungpook National University, Daegu, South Korea. His current research interests include computer vision and multimodal technologies utilizing artificial intelligence.

MYUNG-SEOK CHOI received the B.S., M.S., and Ph.D. degrees from the Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1996, 1998, and 2005, respectively. Since 2005, he has been with the Korea Institute of Science and Technology Information (KISTI), where he is currently the Director of the Department of Machine Learning Data Research. His research interests include open science, research data management, and artificial intelligence.

RYONG LEE received the B.S. degree from the School of Electronics, Telecommunication and Computer Engineering, Korea Aerospace University, South Korea, in 1998, and the M.S. and Ph.D. degrees from the Department of Social Informatics, Kyoto University, Japan, in 2001 and 2003, respectively. From 2003 to 2008, he was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), South Korea. Since 2013, he has been with the Korea Institute of Science and Technology Information (KISTI), South Korea, where he is currently a Senior Researcher with the Research Data Sharing Center, and a Professor with the University of Science and Technology, South Korea. His research interests include AI, big data analysis, spatial data analysis, and the Internet of Things.

SANGTAE AHN (Member, IEEE) received the B.S. degree in electrical engineering and computer science from Chungnam National University, Daejeon, South Korea, in 2010, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2012 and 2016, respectively. From 2016 to 2020, he was a Postdoctoral Research Associate with The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He is currently an Assistant Professor with the School of Electronics Engineering, Kyungpook National University, Daegu, South Korea. His research interest includes the development of brain-inspired and generative models using machine learning approaches.