일반논문 (Regular Paper)
방송공학회논문지 제29권 제6호, 2024년 11월 (JBE Vol.29, No.6, November 2024)
https://doi.org/10.5909/JBE.2024.29.6.1043
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

일관성 있는 이야기 삽화 생성을 위한 Text-to-image 생성 모델 활용 연구

명 세 민[a], 강 다 빈[a], 송 채 영[a], 홍 정 훈[a], 박 상 효[a)‡]

A study on the Use of a Text-to-image Generative Model to Generate Consistent Story Illustrations

Semin Myeong[a)], Dabin Kang[a)], Chae-yeong Song[a)], Jeong-hun Hong[a)], and Sang-hyo Park[a)‡]

요 약

Text-to-image 생성 모델의 발달로 텍스트에 일치하는 이미지를 만드는 것이 가능하게 되었으나, 이러한 모델을 현실 세계에 바로 적용하기에는 여전히 고려해야 할 요소들이 존재한다. 특히 소설과 같이 긴 문장과 여러 단원으로 구성된 이야기에 대한 삽화를 생성하고자 할 경우 해당 매체가 가지는 특성을 반영하는 것이 중요하다. 따라서 본 논문은 Text-to-image 생성 모델을 활용하여 이야기에 대한 삽화를 생성하는 프레임워크를 제안한다. 프레임워크는 대규모 언어 모델을 사용하여 긴 문장을 요약하고, 자동화된 방식으로 사전에 구축된 장르별 데이터 셋에 기반한 Text-to-image 검색 및 스타일이 일관된 이미지 생성 방법을 통해 이야기에 대한 여러 장의 삽화를 생성한다. 최종적으로, 우리는 제안된 프레임워크와 기존 Text-to-image 생성 모델을 사용하는 경우를 정량적 및 정성적으로 분석하여, 제안된 프레임워크가 이야기에 대한 삽화 제작에 있어 유효함을 입증한다.

Abstract

With the advancement of Text-to-image generation models, it has become possible to generate images that match textual descriptions. However, there are still factors to consider before applying these models directly to real-world scenarios. This is particularly important when generating illustrations for story composed of long sentences and multiple chapters, such as novels, where it is essential to reflect the unique characteristics of the medium. Therefore, this paper proposes a framework for generating illustrations for story using Text-to-image generation model. The framework leverages large language models to summarize long sentences and introduces a method to generate multiple illustrations for stories. This method is based on Text-to-image retrieval from automatically pre-constructed genre-specific datasets and Text-to-image generation that ensure a consistent styles across generated images. Finally, we conduct both quantitative and qualitative analyses of the proposed framework and existing Text-to-image generation models, demonstrating that the proposed framework is effective in generating illustrations for stories.

Keyword : Generative Model, Text-to-image Generation, Deep Learning

Ⅰ. 서 론

Text-to-image 생성 모델은 텍스트를 입력으로 넣어 이에 대응되는 이미지를 생성하는 모델로[1,2,3,4], 최근에는 이러한 생성 모델을 일반적인 이미지 생성을 넘어 특정 분야에 적용하고자 하는 여러 시도들이 나타나고 있다. 구체적으로, Text-to-image 생성 모델을 사용하여 뉴스 삽화를 생성하기 위한 시도[5]나 예술적인 이미지를 생성하기 위한 시도[6] 등이 있으며, 이러한 시도들은 기존 Text-to-image 생성 모델을 조금 더 구체화된 목적을 위해 사용할 수 있음을 보여준다.

이러한 상황에서, Text-to-image 생성 모델을 소설과 같은 긴 이야기에 대한 삽화를 생성하는 데 적용하고자 하는 수요가 존재할 수 있다. 하지만, 기존의 Text-to-image 생성 모델은 이러한 긴 이야기를 다루기에는 적합하지 않다. 비교적 짧은 텍스트에 대한 이미지를 생성하도록 설계된 기존의 Text-to-image 생성 모델을 긴 이야기의 삽화 생성에 사용할 경우, 다음과 같은 문제점들이 존재한다. 첫째, 삽화가 포함된 이야기에서 삽화는 이야기의 일부 내용을 묘사하는 경우가 일반적이다. 또한, 텍스트 기반 생성 모델은 특정 토큰의 수를 기준으로 입력된 텍스트를 자름으로써 입력의 길이를 제한한다[7]. DiffusionDB[8]에서 언급된 Text-to-image 생성 모델인 Stable Diffusion[1]에서 실제 사람에 의해 입력되는 텍스트는 대부분 75개 이하의 토큰으로 구성된다는 점은 이를 뒷받침한다. 따라서, 텍스트 기반 생성 모델을 사용하여 이야기에 대한 이미지를 생성할 때 이야기의 전체 문장 대신 일부 문장을 입력으로 사용해야 한다는 문제가 있다. 둘째, 문장은 맥락에 의해 그 의미가 변하기 때문에, 동일한 문장일지라도 이야기의 전체적인 분위기 혹은 장르에 따라 생성되는 이미지가 달라져야 한다. 텍스트 기반 생성 모델의 입력 텍스트는 전체 맥락이 배제된 독립된 문장일 수 있고, 이 경우 이러한 특성을 제대로 고려하지 못할 수 있다. 따라서, 이야기에 대해 생성된 이미지는 이야기의 전반적인 분위기나 장르를 반영해야 한다는 문제가 있다. 셋째, 여러 단원으로 구성된 이야기에 대한 이미지를 생성할 경우, 여러 장의 이미지를 생성하는 것이 일반적일 수 있다. 이 경우, 생성된 이미지들 사이의 스타일이 일관되지 않으면, 해당 이미지들이 하나의 이야기를 표현하는 것이 아닌 독립적인 이미지로 인식되어 이야기에 대한 몰입을 해칠 수 있다. 따라서, 이야기에 대해 생성된 다수의 이미지들 사이에 화풍과 같은 스타일이 일관되게 유지되어야 한다는 문제가 있다. 위 사항들을 고려할 때, 이야기의 삽화를 생성하기 위해서는 각 단원별로 핵심적인 문장을 추출하여 이미지를 생성하고, 생성된 이미지들 중에서 일관된 이미지들을 선별하는 등 사람의 개입이 불가피할 수 있다.

앞서 언급된 문제를 해결하기 위해, 본 논문은 텍스트 기반 생성 모델을 사용하여 이야기에 대한 삽화를 생성하기 위한 프레임워크를 제안한다. 제안된 프레임워크는 대규모 언어 모델인 ChatGPT[9]를 사용하여 자동으로 문장을 요약하고 장르를 분류하며, Text-to-image 검색(image retrieval)을 사용하여 장르에 맞는 이미지를 가져온 다음, 해당 이미지와 요약된 문장 그리고 Diffusion[10] 기반의 Text-to-image 생성 모델을 활용하여 장르를 반영하면서도 일관된 이미지들을 생성한다. 추가로, 본 논문은 장르에 맞는 이미지 검색을 위한 장르별 데이터 셋을 구축하는 방법을 제안한다. 제안된 방식들은 딥러닝 기반의 자동화된 방법을 통해, 전체 이야기에 대한 입력에 대해 사람의 개입 없이 여러 장의 삽화를 생성하는 것을 가능하게 한다. 우리는 실험을 통해 제안 방법으로 생성된 이미지들이 이야기의 장르적 특성을 잘 반영하고 일관성을 유지한다는 것을 입증한다.

Ⅱ. 관련 연구

1. 생성 모델

딥러닝 기술을 기반으로 하여 이미지 혹은 텍스트를 생

a) 경북대학교 컴퓨터학부(School of Computer Science and Engineering, Kyungpook National University)
‡ Corresponding Author : 박상효(Sang-hyo Park)
E-mail: s.park@knu.ac.kr
Tel: +82-53-950-6373
ORCID: https://orcid.org/0000-0002-7282-7686
※ 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2022-00167169, 이동형 로봇 기반 실사 메타버스 실감형 비디오의 획득 및 처리 기술 개발).
· Manuscript October 18, 2024; Revised October 25, 2024; Accepted October 25, 2024.

성하기 위해 많은 연구들이 진행되었다. 이미지 생성의 경우, Variational Autoencoder (VAE)[11]는 오토 인코더 구조에 기반하여 정규화 기법이 적용된 변분 추론을 통해 다층 퍼셉트론 신경망 기반의 이미지 생성이 가능함을 보였다. 또한, Generative Adversarial Network (GAN)[12]는 생성자와 판별자 사이의 경쟁적인 학습을 통하여 이미지를 생성하는 것이 가능함을 보였다. 이후, VAE는 생성되는 이미지의 품질이 낮고 GAN은 학습이 불안정하다는 문제점을 보여 Diffusion[10] 기반의 생성 모델에 대한 연구가 큰 관심을 받게 되었다. 텍스트 생성의 경우, 최근에는 ChatGPT[9]와 같은 대규모 언어 모델이 자연어 처리 영역에서 매우 뛰어난 성능을 보였다. ChatGPT는 텍스트 쿼리를 입력으로 받아 해당 쿼리에서 요구되는 내용을 생성하는 것이 가능하기 때문에, 본 논문에서는 이러한 점을 사용하여 긴 문장을 요약하고 장르를 분류한다.

2. Text-to-image 생성 모델

Text-to-image 생성 모델은, 앞서 언급한 생성 모델에 텍스트를 입력으로 주어 텍스트에 대응되는 이미지를 생성하는 모델이다. 대표적으로, Stable diffusion[1]은 Diffusion 기반의 구조에 텍스트 정보를 Attention 메커니즘을 통해 조건으로 넣어 텍스트에 대응되는 이미지를 생성하였다. 대규모 Text-to-image 생성 모델들[1,2,3,4]은 우수한 성능을 보여주고 있으나, 이러한 모델들을 직접 학습시키는 것은 많은 비용이 든다. 따라서, 최근에는 일반화 성능이 뛰어난 대규모 생성 모델을 미세 조정하는 여러 방법들이 주목을 받고 있다. DreamBooth[13]는 Stable Diffusion이나 Imagen[2]과 같은 대규모 생성 모델을 미세 조정하여, 특정 객체를 나타내는 이미지를 생성하는 방법을 제시하였다. 이와 유사하게, StyleAligned[14]는 별도의 학습 없이 추론 과정에서 생성 모델의 Attention을 공유함으로써 특정 이미지와 비슷한 스타일을 나타내는 이미지를 생성하는 방법을 제시하였다. 본 논문에서 제안하는 프레임워크는 StyleAligned를 이야기에 대한 삽화를 생성하기 위한 일부분으로 사용한다.

Ⅲ. 제안 방법

이 논문에서 제안하는 프레임워크는 크게 사전 작업 단계와 추론 단계로 구분된다. 사전 작업 단계에서는, 미리 지정된 장르에 대해 각각의 장르에 적합한 이미지들을 포함하는 장르별 이미지 데이터 셋을 구축한다. 추론 단계는 세 가지 하위 단계로 구분되며, 각 하위 단계의 역할은 다음과 같다: (1) 입력된 문장에 대한 요약 및 이야기의 장르를 분류한다. (2) 요약된 문장과 분류된 장르에 기반하여 사전 구성된 장르별 이미지 데이터 셋에서 해당 문장 및 장르에 적합한 이미지를 가져온다. (3) 요약된 문장과 (2)에서 가져온 이미지를 모두 사용하여 장르에 적합하면서도 일관된 스타일의 이미지를 생성한다. 이 장의 각 절에서는 사전 작업 단계와 추론의 각 하위 단계에 대해 서술하며, 마지막 절에서는 삽화 생성을 위한 추가적인 프롬프트에 대해서도 간략하게 언급한다.

1. 사전 작업 단계: 장르별 이미지 데이터 셋 구축

장르별 이미지 데이터 셋을 구축하기 위한 방식으로, 이 논문에서는 대규모 오픈소스 데이터 셋인 DiffusionDB[8]를 정제하는 방법을 사용한다. DiffusionDB는 텍스트 기반 생성 모델인 Stable Diffusion으로 생성한 14,000,000장의 이미지, 생성에 사용한 프롬프트, 생성 시 사용자가 입력한 하이퍼 파라미터를 포함하는 데이터 셋이다[8]. 본 논문에서는 DiffusionDB에서 제공하는 하위 집합에 해당하는 데이터 셋을 사용하는데, 이 하위 집합의 경우 2,000,000개의 데이터를 포함하고 있다. 해당 데이터 셋에 포함된 이미지는 장르별로 분류되어 있지 않으므로, 먼저 우리는 이 이미지를 장르별로 분류한다. 이 과정에서 많은 수의 이미지를 사람이 일일이 확인하여 분류하는 방식은 비용이 많이 들기 때문에, 사전 학습된 CLIP[15]을 통해 자동으로 이미지를 분류한다. 구체적으로, 데이터 셋 내 각 이미지를 i, 사전에 정의한 장르의 집합을 $G(|G| = n)$, 장르 집합 내 장르를 $g_k(g_k \in G, k = 1, 2, \cdots, n)$라고 할 때, 각 장르에 대해 프롬프트 함수 F를 적용하여 일관된 프롬프트 형식 $F(g_k)$를 얻는다. $F(g_k)$는 “a photo that correspond to a g_k illustration”에 해당하는 형태가 된다. 예를 들어, 사전 정의한 장르 집합 {Adventure, Dystopian, Fantasy, Horror,

Mystery, Romance, Science Fiction, Thriller} 내의 장르 Fantasy에 대해 프롬프트 함수 F를 적용하면 “a photo that correspond to a fantasy illustration”에 해당하는 프롬프트가 생성된다. 이렇게 생성된 프롬프트 $F(g_k)$와 이미지 i를 각각 CLIP의 텍스트 인코더와 이미지 인코더에 입력으로 넣어 이에 대한 특징인 $E_v(i)$와 $E_t(F(g_k))$를 추출한다. 이후 추출된 특징을 사용하여 이미지와 장르 프롬프트 사이의 유사도 $S(E_v(i), E_t(F(g_k)))$를 측정한다. 유사도를 계산할 때는 CLIP을 따라 두 특징 사이 코사인 유사도를 사용한다. 그 다음, 코사인 유사도가 가장 높게 나오는 장르로 각 이미지를 분류함으로써 일차적으로 장르별 이미지 데이터 셋 $I_{g_1}', I_{g_2}', \cdots, I_{g_n}'$을 구축한다. 이후, 일차적으로 분류된 각 장르별 이미지 데이터 셋 내에서 코사인 유사도가 높은 5,000개의 이미지를 추가로 필터링하여 최종적인 장르별 이미지 데이터 셋 $I_{g_1}, I_{g_2}, \cdots, I_{g_n}$을 구축한다. 그림 1은 이 과정을 나타낸다.

2. 문장 요약 및 장르 분류

입력된 문장에 대한 요약 및 장르 분류를 위해 우리는 대규모 언어 모델인 ChatGPT[9]를 사용한다. ChatGPT는 텍스트 쿼리를 입력으로 받아 이에 대응되는 텍스트를 결과물로 생성하는 것이 가능하기 때문에, 이 방식을 이용하여 문장을 요약하고 장르를 분류한다. 구체적으로, 먼저, 전체 이야기에 해당하는 문장 T가 입력으로 들어올 때 이를 생성하고자 하는 이미지의 수 m에 맞추어 분할한다. 본 논문에서는 챕터를 기준으로 문장을 분할한다. 예를 들어, 전체 이야기가 10개의 챕터로 구성되어 있을 경우, $m = 10$이 된다. 이후, 분할된 각 부분과 해당 부분을 한 문장으로 요약해달라는 쿼리 q_s를 ChatGPT에 입력으로 사용하여 요약된 문장 $t_1, t_2, \cdots, t_m$을 얻는 방식을 사용한다. 챕터를 기준으로 이야기를 분류하는 경우, t_k는 이야기의 k번째 챕터를 한 문장으로 요약한 내용이 된다. 장르 분류의 경우,

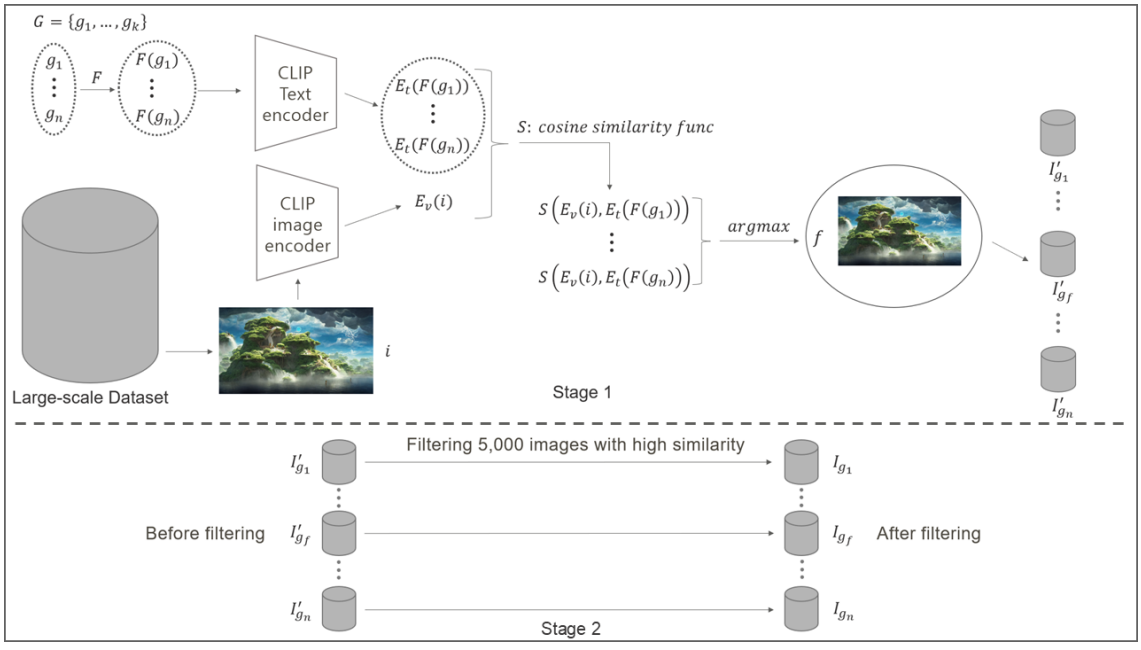

그림 1. 본 논문에서 제안하는 사전 단계인 장르별 이미지 데이터 셋 구축 방법. 사전 단계에서는 우선 각 장르에 프롬프트 함수를 적용하고, 데이터 셋 내 모든 이미지에 대해 유사도를 계산함으로써 일차적으로 데이터 셋을 구축한다. 이후 일차적으로 구축된 데이터 셋 내에서 유사도가 높은 5,000 개의 이미지를 필터링 하여 최종적인 장르별 데이터 셋을 구축한다.

Fig. 1. The preprocessing step proposed in this paper for building a genre-specific image dataset: In the preprocessing step, prompt functions are first applied to each genre, and a dataset is initially constructed by calculating similarity for all images within the dataset. Then, from the initially constructed dataset, 5,000 images with the high similarity are filtered to build the final genre-specific dataset.

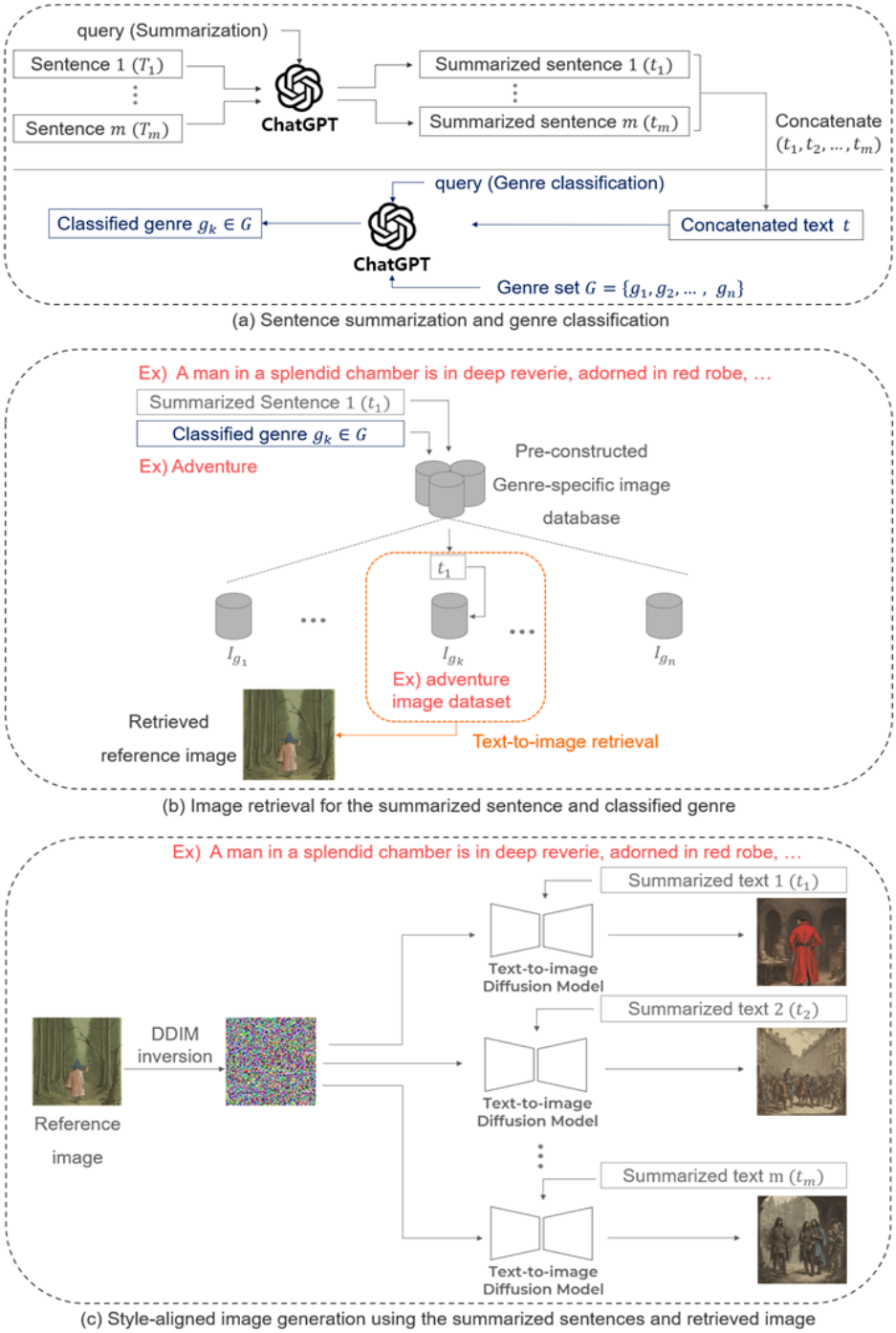

그림 2. 본 논문에서 제안하는 프레임워크. 해당 프레임워크는 크게 세가지 요소로 구분된다: (a) 문장 요약 및 장르 분류, (b) 요약된 문장 및 분류된 장르에 적합한 이미지 검색, (c) 요약된 문장과 검색된 이미지를 사용한 일관된 이미지 생성

Fig. 2. The framework proposed in this paper. The framework is divided into three main components: (a) Sentence summarization and genre classification, (b) image retrieval for the summarized sentence and classified genre, (c) style-aligned image generation using the summarized sentence and retrieved image

요약된 $t_1, t_2, \cdots, t_m$들을 모두 연결한 t, 사전 정의된 장르 집합 G, 요약된 전체 이야기 t에 대한 장르를 G에 포함된 장르 중 하나로 분류해 달라는 쿼리 q_c를 ChatGPT에 입력으로 사용하여 분류된 장르 g_c를 얻는 방식을 사용한다. 그림 2의 (a)는 이 과정을 나타낸다.

3. 요약된 문장 및 분류된 장르에 적합한 이미지 검색

요약된 문장 $t_1, t_2, \cdots, t_m$과 분류된 장르 g_c를 사용하여 이에 적합한 이미지 $i_{g_{c\hat{l}}}$를 얻기 위해, CLIP을 사용한 이미지 검색을 수행한다. 먼저, g_c에 상응하는 이미지 데이터 셋 I_{g_c} 내의 모든 이미지 $i_{g_{cl}}(i_{g_{cl}} \in I_{g_c}, l = 1, 2, \cdots, 5000)$들과 요약된 첫 문장인 t_1에 CLIP 이미지 인코더와 텍스트 인코더를 사용하여 특징을 추출한다. 그 다음, 추출된 특징들 사이의 유사도 $S(E_v(i_{g_{cl}}), E_t(t_1))$를 측정하며, 이 역시 CLIP을 따라 코사인 유사도를 사용한다. 유사도 측정 이후, I_{g_c}의 모든 이미지들 중 요약된 첫 문장인 t_1과 유사도가 가장 높은 이미지 $i_{g_{c\hat{l}}}$와 $i_{g_{c\hat{l}}}$을 생성할 때 사용한 프롬프트를 가져온다. 이 과정을 예를 들어 설명하자면, 특정 이야기가 Adventure 장르로 분류되었을 때 해당 이야기의 첫 번째 챕터를 요약하고, 요약된 첫 챕터와 사전에 구축된 Adventure 장르 데이터 셋의 모든 이미지 사이 유사도를 계산하여, 가장 유사한 이미지를 찾는 방식이 된다. 그림 2의 (b)는 이 과정을 나타내며, $i_{g_{c\hat{l}}}$을 검색하는 과정을 수식으로 표현하면 다음과 같다:

$$i_{g_{c\hat{l}}}(\hat{l} = \underset{l}{argmax}\, S(E_v(i_{g_{cl}}), E_t(t_1)) \quad (l = 1, 2, \cdots, 5000)) \tag{1}$$

4. 요약된 문장과 검색된 이미지를 사용한 일관된 이미지 생성

마지막으로 요약된 문장들 $t_1, t_2, \cdots, t_m$과 이미지 검색을 통해 얻은 $i_{g_{c\hat{l}}}$를 사용하여 $i_{g_{c\hat{l}}}$의 스타일을 유지하면서 샘플들 사이의 일관성이 보장된 이미지 $i'_1, i'_2, \cdots, i'_m$을 생성한다. 이미지 생성 모델은 StyleAligned[14]의 방식을 따라 사용한다. 해당 모델은 Attention을 사용하는 텍스트 기반 이미지 생성 Diffusion 모델에서 레퍼런스 샘플과 타겟 샘플 사이의 Attention을 공유한다. 여기에서 의미하는 Attention은 이미지에 대응되는 특징을 쿼리 Q, 키 K, 밸류 V로 투영 시의 Q와 K의 차원의 수가 d_k일 경우 Scaled dot-product attention이다:

$$\text{Attention}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}V) \tag{2}$$

StyleAligned는 Adaptive normalization operation (AdaIN)[16]을 통해 타겟의 쿼리 Q_t, 키 K_t 각각을 레퍼런스의 쿼리 Q_r, 키 K_r로 Normalize하고, 이 상태로 Scaled dot-product attention을 적용한다:

$$Q_t \Leftarrow \text{AdaIN}(Q_t, Q_r) \quad K_t \Leftarrow \text{AdaIN}(K_t, K_r) \tag{3}$$[14]

$$\text{AdaIN}(x, y) = \sigma(y)(\frac{x - \mu(x)}{\sigma(x)}) + \mu(y) \tag{4}$$[16]

StyleAligned는 이 방식을 통해 생성된 두 샘플의 스타일을 일치시키고, 생성된 이미지들 사이의 일관성을 보장한다. StyleAligned에서는 레퍼런스 샘플로 이미지를 사용할 경우 해당 이미지에 DDIM inversion[17]을 적용하여 이미지에 대응되는 노이즈를 얻고, 이 노이즈를 생성에 사용한다. 이 방식을 따라 우리는 검색된 이미지를 레퍼런스 이미지로 사용하여, m개의 요약된 문장들에 대한 이미지를 생성한다. 그림 2의 (c)는 이 과정을 나타내며, DDIM inversion에는 레퍼런스 샘플이 되는 이미지를 생성하는 데 사용된 프롬프트도 함께 사용한다. 해당 프롬프트는 DiffusionDB 내에 이미지와 쌍을 이루어 저장되어 있으므로 이를 이용한다.

5. 삽화 생성을 위한 추가적인 프롬프트

이전 연구들에 따르면[18,19], 이미지 생성 모델을 사용할 때 입력 프롬프트에 변화를 주는 것으로 생성되는 이미지

의 스타일이 달라질 수 있다. 이러한 발견에 영감을 받아, 우리는 이미지 생성 시 삽화를 위한 프롬프트를 추가한다. 해당 프롬프트는 "~, illustration style"이 되며 이미지 생성에 사용되는 요약된 텍스트의 뒷부분에 추가된다. 예를 들어 생성에 사용하는 텍스트가 "the adventurer discovered a hidden cave, its entrance glimmering with ancient crystals"일 때, 이미지 생성에 사용되는 최종적인 텍스트는 "the adventurer discovered a hidden cave, its entrance glimmering with ancient crystals, illustration style"이 된다.

Ⅳ. 실 험

1. 평가 데이터

실험에 앞서, 이미지 생성을 위한 이야기의 경우 정제된 벤치마크 데이터 셋을 구하기 어렵다는 문제점이 존재한다. 따라서, 우리는 프로젝트 구텐베르크 (PG)[20]에 의해 업로드 및 카테고리별로 분류된 전자책들 중 사전 정의한 장르 집합에 속하는 카테고리의 이야기를 이미지 생성에 사용한다. 실험에서 장르 집합은 {Adventure, Dystopian, Fantasy, Horror, Mystery, Romance, Science Fiction, Thriller}로 정의하였고, 해당 장르 집합에 대해 데이터 셋을 구축한 이후 PG에서 'Dystopian'과 'Thriller' 장르에 해당하는 카테고리를 찾기 어려워 이 두 장르를 제외한 나머지 장르에 대해 실험을 진행하였다. Romance 장르의 경우에는 Romantic Fiction, Mystery 장르의 경우에는 Mystery Fiction 카테고리에 해당하는 이야기를 사용하였다. 각 이야기들은 여러 챕터로 구성되어 있어, 우리는 각 챕터를 요약하여 챕터별로 1장의 이미지를 생성하였으며, 장르마다 최근 30일 동안 다운로드 수가 많은 상위 25개의 이야기들 중 2개 이상의 챕터로 구성되어 있는 이야기를 5개씩 중복되지 않도록 선정하였다. 표 1은 실험에 사용한 각 장르별 5가지 이야기의 제목을 나타낸다.

2. 평가 지표

앞서 언급한 평가용 데이터를 사용하여, 실험에서는 제안 방법의 유효성을 3가지 측면에 대하여 검증한다: (1) 생성된 이미지들이 장르적 특성을 잘 반영하는가? (2) 생성된 이미지들 사이 일관성이 보장되는가? (3) 생성된 이미지들이 문장의 의미를 충분히 반영하는가? 먼저, 장르적 특성에 대한 반영 정도를 평가하기 위해 각 장르 g_k에 해당하는 프롬프트와 생성된 이미지들 사이 CLIP Score[21]를 사용한다. 하나의 이야기에 대해 여러 장의 이미지가 생성되므로, 각 이야기에 대해 생성된 이미지들과 각 장르 g_k사이 CLIP Score의 평균을 구하여 이를 이야기별 CLIP Score로 사용한다. 장르 내 모든 이야기별 CLIP Score의 평균을 장르별 CLIP Score로 정하고 이를 CLIP-G라고 명명한다. 다음으로, 생성된 이미지들 사이 일관성을 평가하기 위해 DreamBooth[13]를 따라 ViT-S/16 DINO[22] 임베딩 값들 사

표 1. 실험에 사용된 각 장르별 이야기의 제목들
Table 1. Titles of stories for each genre selected in the experiments

Genre	Title				
Adventure	Around the World in Eighty Days	The Count of Monte Cristo	The Three Musketeers	The Call of the Wild	Twenty Years After
Fantasy	Le Morte d'Arthur (Volume 1)	The Mabinogion	The Merry Adventures of Robin Hood	The Story of the Volsungs (Volsunga Saga)	The Wonderful Wizard of Oz
Horror	Carmilla	Dracula	Metamorphosis	The King in Yellow	The Strange Case of Dr.Jekyll and Mr.Hyde
Mystery	The Moonstone	The Mystery of Edwin Drood	The Secret Agent: A Simple Tale	The Thirty-Nine Steps	The Woman in White
Romance	Kate Vernon: A Tale (Volume 1)	Kate Vernon: A Tale (Volume 2)	Only a Girl's Love	Star of India	Wastralls: A Novel
Science Fiction	A Journey to the Centre of the Earth	The island of Doctor Moreau	The Time Machine	The War of the Worlds (Volume 1)	Twenty Thousand Leagues under the Sea

이 코사인 유사도를 사용하는 DINO 지표를 사용한다. 첫 단원에 대해 생성된 이미지와 이후 단원들에 대해 생성된 이미지들 사이 DINO의 평균 값을 이야기별 DINO 값으로 사용하며, 이야기별 DINO 값의 평균을 장르별 DINO 값으로 사용한다. 마지막으로, 문장에 대한 반영 정도를 평가하기 위해 생성된 이미지와 생성에 사용된 텍스트 사이 CLIP Score를 계산한다. 하나의 이야기 내의 각 이미지와 텍스트 사이 CLIP Score를 모두 계산하고, CLIP-G와 유사한 방식으로 이에 대한 평균을 사용하여 장르별로 계산하여 CLIP-T라고 명명한다. CLIP-G는 장르에 대한 반영 정도를 평가하고 CLIP-T는 문장에 대한 반영 정도를 평가한다는 점에서 다르다.

3. 정량적 평가

우리는 일반화 성능이 뛰어난 것으로 알려진 기존 여러 디퓨전 기반의 생성 모델과 제안 방법을 비교한다. 비교 대상이 되는 기존 모델들은 GLIDE[3], Stable Diffusion (SD)[1], Stable Diffusion XL (SDXL)[4]이다. 제안 방식은 SDXL을 기반 모델로 사용하기에, SDXL과 제안 모델의 차이는 이미지 검색 및 검색된 이미지를 사용하여 Style-Aligned[14]를 적용한다는 부분에 있다. 표 2는 CLIP-G, DINO, CLIP-T에 대한 각 모델의 결과를 보여준다. CLIP-G와 DINO의 경우 제안 방법이 기존 모델들에 비해 정량적으로 우수한 성능을 보인다. CLIP-T의 경우 Adventure, Fantasy, Science fiction에 대해서는 SDXL이 가장 우수한 성능을 보이고 Horror, Mystery, Romance에 대해서는 제안 방법이 가장 우수한 성능을 보인다. 기존 모델과 비교하였을 때, 제안 방법이 SDXL과 비슷한 수준으로 문장에 대한 정보를 반영하면서도, 장르적 특성을 반영하는 능력과 이미지 사이의 일관성을 유지하는 능력이 가장 뛰어남을 알 수 있다. 실험에 사용된 기존 모델들 또한 요약된 문장을 사용하여 이미지를 생성하였다.

4. 정성적 평가

정량적 평가에 이어, 우리는 제안 방법인 데이터 셋 구축과 이미지 생성의 유효성을 시각적으로 검증한다. 먼저, 사전에 구축된 장르별 이미지 데이터 셋에 대해, 그림 3을 통해 자동화된 방법으로 분류된 각 이미지가 해당 장르의 특성을 충분히 반영하고 있음을 확인할 수 있다. 다음으로, 이미지 생성에 대해, 그림 4는 제안 방법으로 생성된 이미지들이 SDXL로 생성된 이미지들에 비해 더 일관된 스타일을 유지하고 있음을 보여준다. 그림 4의 모든 이미지는 “The Story of the Volsungs (Volsunga Saga)” 이야기의 동일한 챕터에 대해 생성되었다.

5. 삽화 생성을 위한 추가적인 프롬프트 적용 결과

마지막으로, 이 절에서는 삽화 생성을 위한 추가적인 프롬프트가 이미지 생성에 미치는 영향을 분석한다. 표 3은 CLIP-G, DINO, CLIP-T 세 평가 지표에 대해, 제안 방법에서 “~, illustration style”에 해당하는 추가적인 프롬프트를 사용할 경우 모든 지표 상에서 성능이 향상됨을 보여준다.

표 2. 기존 모델들과 제안 방법의 CLIP-G, DINO, CLIP-T 결과
Table 2. CLIP-G, DINO, and CLIP-T results of existing models and the proposed method

Metric (↑) / Genre	CLIP-G				DINO				CLIP-T			
	GLIDE	SD	SDXL	Ours	GLIDE	SD	SDXL	Ours	GLIDE	SD	SDXL	Ours
Adventure	0.195	0.188	0.197	0.215	0.155	0.386	0.379	0.475	0.217	0.282	0.305	0.298
Fantasy	0.198	0.205	0.217	0.231	0.124	0.481	0.428	0.612	0.212	0.284	0.311	0.308
Horror	0.210	0.204	0.211	0.224	0.182	0.453	0.356	0.541	0.218	0.270	0.285	0.286
Mystery	0.206	0.195	0.208	0.225	0.121	0.356	0.277	0.522	0.212	0.273	0.276	0.277
Romance	0.196	0.197	0.213	0.227	0.149	0.407	0.300	0.532	0.210	0.274	0.287	0.296
Science fiction	0.210	0.210	0.227	0.235	0.145	0.287	0.251	0.444	0.240	0.281	0.300	0.298

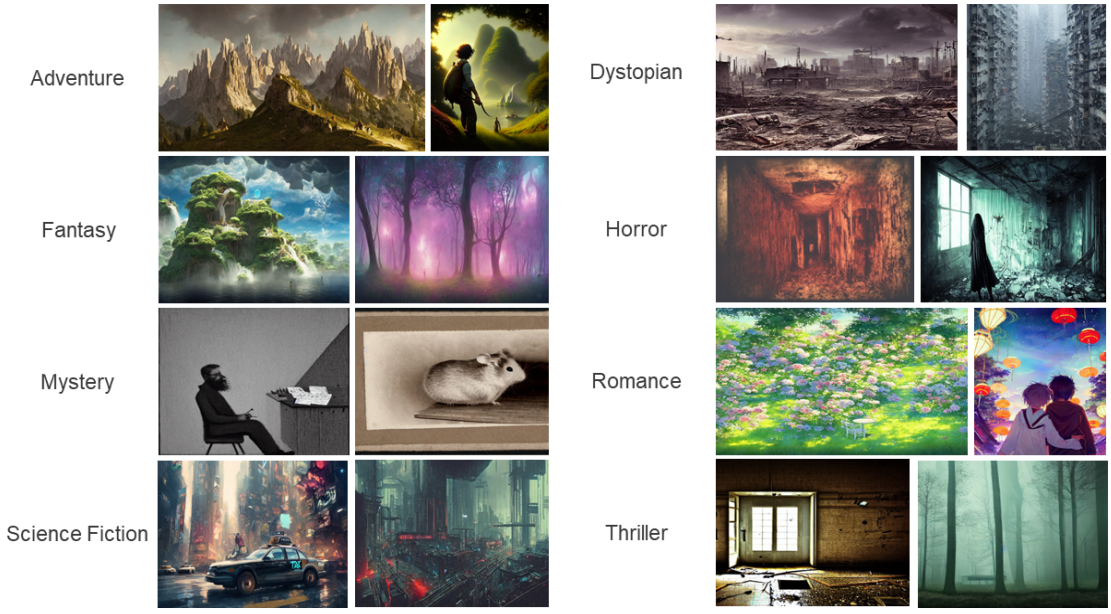

그림 3. 제안 방법을 통해 구축된 장르별 이미지 데이터 셋 내의 샘플들
Fig 3. Samples within the genre-specific image dataset constructed following the proposed method

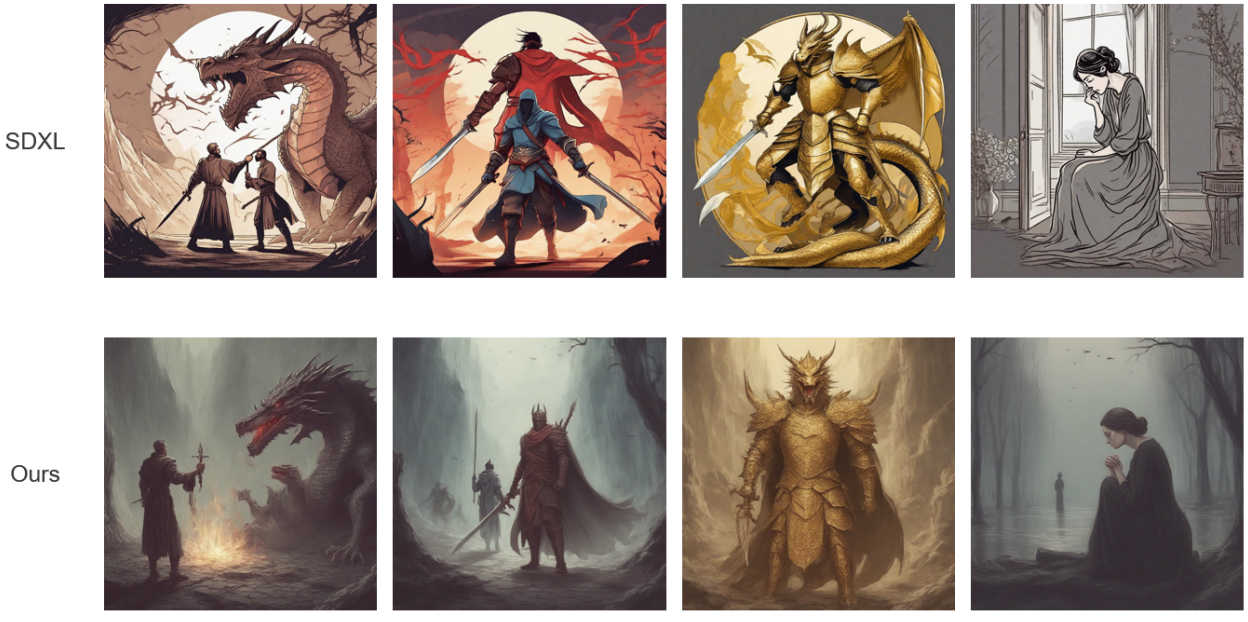

그림 4. “The Story of the Volsungs (Volsunga Saga)” 이야기에 대해 SDXL과 제안 방법으로 생성한 이미지들
Fig. 4. Images generated by SDXL and the proposed method for the story “The Story of the Volsungs (Volsunga Saga)”

표 3. 추가적인 프롬프트 유무에 따른 제안 방법의 성능 비교
Table 3. Performance comparison of the proposed method with and without additional prompts

	Genre	Adventure	Fantasy	Horror	Mystery	Romance	Science fiction
CLIP-G	Ours w/o ", illustration style"	0.211	0.227	0.220	0.216	0.213	0.231
	Ours w/ ", illustration style"	0.215	0.231	0.224	0.225	0.227	0.235
DINO	Ours w/o ", illustration style"	0.424	0.562	0.445	0.435	0.484	0.367
	Ours w/ ", illustration style"	0.475	0.612	0.541	0.522	0.532	0.444
CLIP-T	Ours w/o ", illustration style"	0.294	0.298	0.277	0.276	0.289	0.295
	Ours w/ ", illustration style"	0.298	0.308	0.286	0.277	0.296	0.298

특히 이미지의 일관성을 평가하는 지표인 DINO에서의 향상 폭이 크게 나타나며, 이를 통해 추가적인 프롬프트 사용이 일관된 이미지 생성에 도움을 준다는 것을 알 수 있다. 그림 5는 추가적인 프롬프트를 사용한 경우 그렇지 않은 경우에 비해 더 일관된 이미지를 생성 가능함을 보여준다. 구체적으로 추가적인 프롬프트가 없을 경우 3, 4번째 이미지인 관과 집 이미지에 대해 2번째 이미지와 달리 현실풍의 이미지가 나오지만, 추가적인 프롬프트를 사용할 경우 2번째 이미지와 더 비슷한 이미지가 생성된다.

Ⅴ. 한계점

제안 방법은 이미지 검색을 통해 레퍼런스 이미지를 얻어 해당 이미지와 유사한 스타일을 유지하도록 하여 생성된 이미지가 장르를 반영하고 일관된 스타일을 유지할 수 있게 한다. 따라서, 이미지 검색 결과에 따라 생성되는 이미지의 품질 등이 달라질 수 있다. 그림 6은 이에 대한 예시를 보여준다. 이러한 문제를 해결하기 위해서는 레퍼런스 이미지의 분위기나 스타일을 유지하면서 사용자가 의도한 품

그림 5. 추가적인 프롬프트 유무에 따른 제안 방법의 "Dracula" 이야기에 대한 이미지 생성 결과
Fig. 5. Image generation results for the story "Dracula" using the proposed method with and without additional prompts

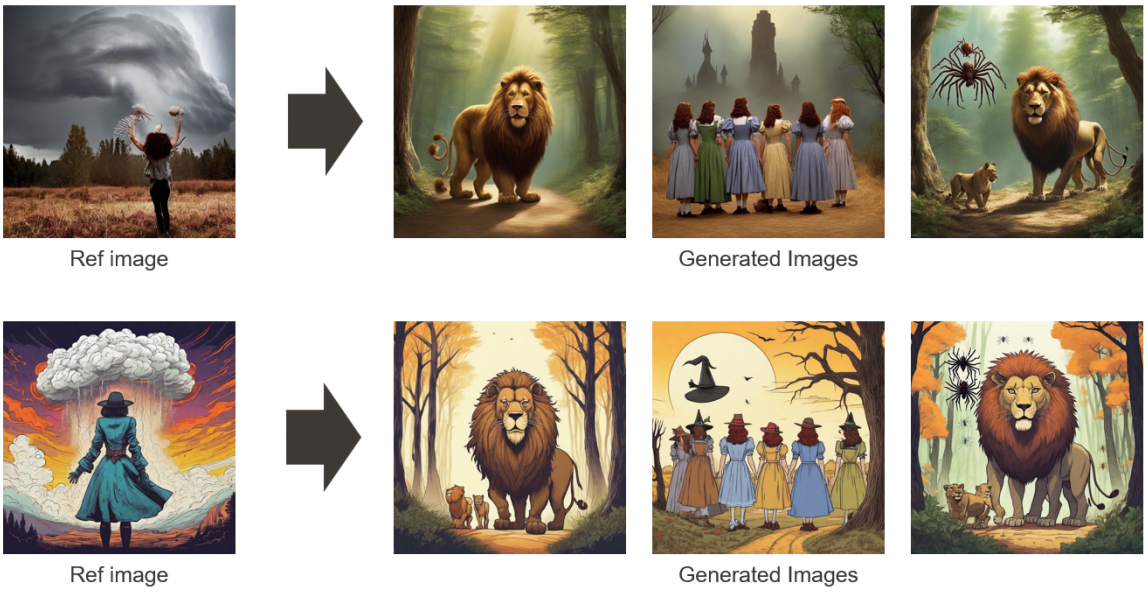

그림 6. 레퍼런스 이미지에 따른 “The Wonderful Wizard of Oz” 이야기에 대한 이미지 생성 결과
Fig. 6. Image generation results for the story “The Wonderful Wizard of Oz” using the proposed method based on reference image

질대로 이미지를 생성할 수 있는 방안이 연구되어야 한다.

Ⅵ. 결 론

우리는 텍스트 기반 이미지 생성 모델을 소설과 같은 긴 문장과 여러 챕터로 구성된 이야기에 대한 삽화 생성에 활용하는 프레임워크를 제안한다. 제안 방법은 크게 사전 단계와 추론 단계로 구분된다. 사전 단계에서는 장르별 이미지 데이터 셋을 구축한다. 추론 단계에서는 텍스트 요약 및 장르 분류 후 사전 구축된 데이터 셋을 통해 이미지를 검색하고, 검색된 이미지를 활용하여 최종적으로 이미지를 생성한다. 또한, 우리는 삽화 생성을 위한 추가적인 프롬프트를 제시한다. 실험을 통해, 우리는 제안 방법을 사용할 경우 기존 모델들에 비해 장르를 잘 반영하며 일관된 이미지들을 생성할 수 있음을 보인다. 마지막으로, 우리는 레퍼런스 이미지에 따라 이미지의 품질이 변화된다는 제안 방법의 한계를 언급하며, 이를 극복하기 위해 레퍼런스 이미지의 분위기나 스타일을 유지하면서 사용자가 의도한 품질대로 이미지를 생성할 수 있는 연구가 필요함을 주장한다.

참 고 문 헌 (References)

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp.10684-10695, 2022. doi: https://doi.org/10.1109/CVPR52688.2022.01042

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. Karagol Ayan, S. S. Mahdavi, R. Gontijo-Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” *Advances in Neural Information Processing Systems*, pp.36479-36494, 2022. doi: https://doi.org/10.48550/arXiv.2205.11487

[3] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” *arXiv preprint arXiv*:2112.10741, 2021. doi: https://doi.org/10.48550/arXiv.2112.10741

[4] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” *International Conference on Learning Representations*, Vienna, Austria, 2024. doi: https://doi.org/10.48550/arXiv.2307.01952

[5] V. Liu, H. Qiao, and L. Chilton, “Opal: Multimodal Image Generation for News Illustration,” Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (Bend, OR, USA), *Association for Computing Machinery*, New York, NY, USA, Article 73, pp.1-17, 2022.

doi: https://doi.org/10.1145/3526113.3545621

[6] R. Rombach, A. Blattmann, and B. Ommer, “Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models,” *arXiv preprint arXiv*:2207.13038, 2022.
doi: https://doi.org/10.48550/arXiv.2207.13038

[7] P. Von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf, “Diffusers: Stateof-the-art diffusion models”, https://github.com/huggingface/diffusers, 2022.

[8] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, “DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, pp.893 – 911, 2023.
doi: https://doi.org/10.18653/v1/2023.acl-long.51

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, pp.1877 – 1901, 2020.
doi: https://doi.org/10.48550/arXiv.2005.14165

[10] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” *Advances in Neural Information Processing Systems*, pp.6840 – 6851, 2020.
doi: https://doi.org/10.48550/arXiv.2006.11239

[11] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *arXiv preprint arXiv*:1312.6114, 2013.
doi: https://doi.org/10.48550/arXiv.1312.6114

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,” *Communications of the ACM*, Vol 63, Issue 11, pp.139-144, October 2020.
doi: https://doi.org/10.1145/3422622

[13] 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,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp.22500-22510, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.02155

[14] A. Hertz, A. Voynov, S. Fruchter, and D. Cohen-Or, “Style Aligned Image Generation via Shared Attention,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.4775-4785, 2024.
doi: https://doi.org/10.48550/arXiv.2312.02133

[15] A. Radford, J. 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,” *International Conference on Machine Learning*, PMLR, Virtual Only, pp.8748 – 8763, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[16] X. Huang, S. Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization,” *Proceedings of the IEEE/CVF international conference on computer vision*, Venice, Italy, 2017.
doi: https://doi.org/10.1109/ICCV.2017.167

[17] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” *International Conference on Learning Representations*, Vienna, Austria, 2021.
doi: https://doi.org/10.48550/arXiv.2010.02502

[18] S. Witteveen, M. Andrews, “Investigating Prompt Engineering in Diffusion Models,” *arXiv preprint arXiv*:2211.15462, 2022.
doi: https://doi.org/10.48550/arXiv.2211.15462

[19] Y. Hao, Z. Chi, L. Dong, and F. Wei, “Optimizing Prompts for Text-to-Image Generation,” *Advances in Neural Information Processing Systems*, pp.66923-66939, 2023.
doi: https://doi.org/10.48550/arXiv.2212.09611

[20] “Project Gutenberg,” Project Gutenberg, www.gutenberg.org.

[21] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, “CLIPScore: A Reference-free Evaluation Metric for Image Captioning,” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, pp.7514 – 7528, 2021.
doi: https://doi.org/10.18653/v1/2021.emnlp-main.595

[22] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging Properties in Self-Supervised Vision Transformers”, *Proceedings of the IEEE/CVF international conference on computer vision*, Montreal, QC, Canada, pp.9650-9660, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00951

저 자 소 개

명 세 민

- 2020년 ~ 현재 : 경북대학교 컴퓨터학부 학사과정
- ORCID : https://orcid.org/0009-0001-5893-8240
- 주관심분야 : 기계학습, 컴퓨터 비전, 생성모델, 멀티모달 러닝

강 다 빈

- 2024년 2월 : 경북대학교 컴퓨터학부 학사
- 2024년 3월 ~ 현재 : 경북대학교 컴퓨터학부 석사과정
- ORCID : https://orcid.org/0009-0000-8242-797X
- 주관심분야 : Multi-modal learning, Text-to-video retrieval, Video question & answering, Knowledge distillation, 3D scene understanding

송 채 영

- 2024년 2월 : 경북대학교 경영학부 학사
- 2024년 3월 ~ 현재 : 경북대학교 컴퓨터학부 석사과정
- ORCID : https://orcid.org/0009-0006-0192-336X
- 주관심분야 : Model compression, 3D scene graph, 3D gaussian splatting, Multi-modal learning

홍 정 훈

- 2023년 8월 : 경북대학교 컴퓨터학부 학사
- 2023년 9월 ~ 현재 : 경북대학교 컴퓨터학부 석사과정
- ORCID : https://orcid.org/0009-0004-9868-3432
- 주관심분야 : Text-to-video retrieval, Multi-modal learning

박 상 효

- 2011년 2월 : 한양대학교 컴퓨터전공 학사
- 2017년 8월 : 한양대학교 컴퓨터·소프트웨어학과 박사
- 2017년 5월 ~ 2018년 2월 : 전자부품연구원 지능형영상처리센터 Post-doc
- 2018년 3월 ~ 2018년 12월 : 연세대학교 바른ICT연구소 연구원
- 2019년 2월 ~ 2020년 1월 : 이화여자대학교 전자전기공학과 박사후연구원
- 2020년 3월 ~ 현재 : 경북대학교 컴퓨터학부 부교수
- ORCID : https://orcid.org/0000-0002-7282-7686
- 주관심분야 : HEVC, VVC, Encoding/Decoding Complexity, Omnidirectional Video, Deep Learning