

Generative Model을 활용한 몽타주 제작 프레임워크에 관한 연구

황수진, 안미르, 안승연, 양채연, 노장현
명지대학교 융합소프트웨어학부 데이터테크놀로지전공

요약

본 연구는 현재 한국에서 사용하고 있는 몽타주 제작 방법의 한계를 개선하기 위해, Generative Model을 활용하여 목격자의 구체적인 진술뿐만 아니라 추상적인 진술도 효과적으로 반영하는 몽타주 제작 프레임워크를 개발한다. 이를 위해 DALL-E와 Stable Diffusion을 개발하고 각각의 성능을 비교 분석한 결과, Stable Diffusion이 더 효과적인 성능을 보인다. 이 프레임워크는 몽타주 제작자가 보다 완성도 높은 몽타주를 손쉽게 제작하는 데 도움을 줄 것으로 기대된다.

Keywords: composite sketch, generative model, dall-e, vq-gan, stable diffusion

1. 서론

범죄 수사에서 몽타주는 여전히 중요한 증거물 중 하나이다. CCTV의 보급으로 몽타주 제작의 비중이 예전만큼 크진 않지만, CCTV가 없는 곳에서 주로 발생하는 성폭력 사건이나 CCTV 영상이 있더라도 범인의 얼굴을 판독하기 어려운 경우에는 몽타주가 중요한 단서로 부각된다. 더불어, 용의자 몽타주 제작의 비중이 감소하는 대신, 나이 변환 기술의 발전으로 실종 인물을 찾는데 몽타주 제작이 더욱 중요한 역할을 하고 있다.^[1]

몽타주 제작에는 크게 두 가지 원리가 있다. 첫 번째는 개별 특징들을 합성하는 회상(recall)이고, 두 번째는 제시된 얼굴을 보고 비슷한 느낌의 얼굴을 고르는 재인(recognition)이다. 회상에 기반한 몽타주 제작 방식은 재인에 기반한 방식에 비해 인식률이

낮다^[2]. 따라서 몽타주를 제작하거나 얼굴을 재구성하는 인공 신경망 모델에 대한 최근 연구들은 재인에 기반한 기술에 집중하고 있는 추세이다. 예를 들어, 2019년 제안된 Composite Generating GAN (CG-GAN)^[3]은 사용자가 고른 target pictures를 합성하여 composite(몽타주)를 생성한다.

그러나 제시된 얼굴을 보고 고르는 과정은 원래의 기억을 왜곡할 위험이 있다. 따라서 한국의 수사 현장에서는 회상에 기반한 몽타주 제작 방식을 일차적으로 사용한다. 2015년도에는 ‘폴리스캐치’라는 3D 몽타주 제작 프로그램이 도입되어, 몽타주 제작자는 진술자의 구체적 묘사를 기반으로 얼굴형, 눈썹, 눈 등 다양한 이목구비를 데이터베이스에서 선택하여 조합한 몽타주를 제작한다. 이는 얼굴의 구체적인 특징을 조합하는 방식이기 때문에 얼굴 인상에 대한 추상적인 진술은 직접적으로 반영하지 않는다는 한계가 있다. 또한, 이목구비의 모양을 하나하나 찾아야 하기 때문에 시간과 노력이 적지 않게 소요되는 단점이 있다.

따라서 본 연구는 기억의 왜곡이 없는 기존 제작 방식의 원리를 유지하면서 과정을 자동화하여 한계를 보완하고자 한다. 이를 위해 텍스트로부터 이미지를 자동으로 생성하는 Generative Model을 이용한 몽타주 제작 프로그램을 제안한다. Generative Model인 DALL-E와 Stable Diffusion을 각각 개발하고 성능을 비교 분석하여,

- 기존 제작 방식의 정확성을 유지하면서 추상적인 인상을 더 효과적으로 반영하는 모델을 선정한다.

- b. 몽타주 제작에 있어 Generative Model의 한계를 파악한다.

이러한 연구를 통해 기존 방식의 편의성과 효율성 측면에서의 한계를 극복하고 새로운 가능성을 제시하고자 한다.

2. 문제 정의 및 관련 연구

몽타주는 사진이나 초상화와 달리 인물에 대한 추상적인 느낌을 전달한다. 따라서 본질적으로 정확하고 구체적이기 어렵다. 현재의 몽타주 제작 방식은 구체적인 진술을 가지고 추상적인 이미지를 구성해야 하기 때문에 정보의 구체성과 추상성 간의 균형을 유지하기 어렵다. 또한 진술에 기초하여 알맞은 이목구비를 고르는 데에 상당한 시간이 소요되어 편의성과 효율성 측면에서 단점이 있다.

본 연구는 이러한 기존 방식의 한계를 극복하고자 다음과 같은 세부 목표를 가지고 새로운 몽타주 제작 프레임워크를 제안한다:

- 진술 텍스트를 입력하면 자동으로 몽타주를 생성한다.
- 얼굴의 구체적인 특징뿐만 아니라 추상적인 인상도 반영한다.
- 30초 이내에 몽타주를 생성한다.
- 텍스트 부분 수정을 통해 이미지를 수정할 수 있다.

이를 위해 다음의 두 Generative Model을 개발한다.

2.1. DALL-E

DALL-E는 transformer를 기반으로 하는 text-to-image 생성 모델이다. 이 모델은 text를 입력으로 받아 해당 내용을 설명하는 image를 생성할 수 있다. DALL-E는 텍스트를 임베딩하는 text encoder와 이미지를 생성하는 image decoder로 구성되어 있다. text encoder는 텍스트를 단어 토큰을 분리하여, 각 토큰을 임베딩 벡터로 변환한다. 그 후, image decoder는 text encoder의

출력을 입력으로 받아 잠재 변수(latent variable)를 생성한 후, 이를 활용하여 이미지를 생성한다.

DALL-E의 text encoder는 영어 모델이기 때문에 한국어 텍스트의 의미를 정확하게 파악하지 못한다. 이에 따라 한국어 데이터에 대해 사전 학습된 transformer 기반의 언어 모델인 klue/roberta-large^[4] 모델을 text encoder로 채택하였다.

또한, DALL-E는 생성된 이미지의 디테일이 부족하거나, 생성된 이미지가 텍스트의 의미를 잘 반영하지 못할 수 있는 한계가 있다.^[5] 이를 개선하기 위해 본 연구는 DALL-E 모델의 image decoder의 구조를 변경하여 VQ-VAE에 비해 생성된 이미지의 디테일과 텍스트의 의미 반영 측면에서 우수한 성능을 보이는 VQ-GAN^[6]을 사용하였다.

2.2. Stable Diffusion

Stable Diffusion은 그림 1과 같은 Latent Diffusion Model(LDM)^[7]의 버전들 중 하나이다. LDM은 Diffusion Model을 메인 아키텍처로 가지고 있고, 그 앞뒤에 Auto Encoder와 Decoder를 추가로 가지고 있다.

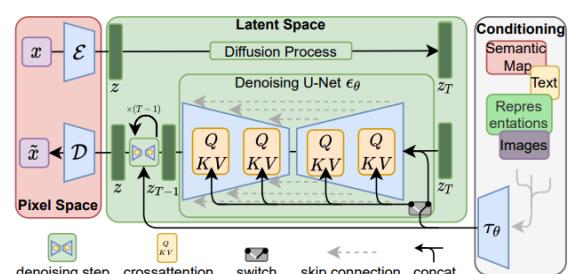


그림 1. LMD 모델 구조^[5]

Stable diffusion은 CLIP이라는 이미 학습된 text autoencoder를 사용한다. CLIP은 텍스트를 latent representation 벡터 z 로 변환한다. 벡터 z 는 점차적으로 노이즈가 추가되는 diffusion process를 거친다. z_T 는 노이즈를 제거하면서 이미지를 복원하는 denoising 과정을 거친다. Denoised latent 벡터 z 는 decoder를 거쳐 최종적으로 pixel space에 표현된 이미지가 된다.

3. 제안하는 방법

3.1. 학습 데이터

몽타주 학습을 위하여 가상 인물 몽타주 학습 데이터셋을 AI Hub에서 다운로드한다. 가상인물 8071명의 이미지, 육안 관찰 인물 스케치, 설명문, 몽타주 스케치로 구성되어 있다.

가상인물 이미지는 컬러로 된 인물 사진이다. 육안 관찰 인물 스케치는 가상인물 이미지를 바탕으로 그린 세밀한 스케치다. 설명문은 가상인물 이미지를 바탕으로 각 얼굴 부분의 구체적인 특징과 인상을 서술한 데이터다. 몽타주 스케치(이하 몽타주)는 설명문을 기반으로 그린 인물의 몽타주다. 설명문과 몽타주는 모두 묘사의 세세한 정도에 따라 상, 중, 하 세 가지 버전이 있다. 표 1은 모델 학습에 사용한 데이터셋을 정리한 것이다.

본 연구는 가상인물 이미지는 제외하고 육안 관찰 인물 스케치, 설명문 그리고 몽타주를 학습에 이용한다.

표 1. 학습 데이터셋 정리(AI Hub 참고)

수집(수량)		가공(수량)		
가상인물 이미지	8071	(육안 관찰) 인물 스케치		8071
		설명문 (상)	8071	몽타주 스케치 (상)
		설명문 (중)	8071	몽타주 스케치 (중)
		설명문 (하)	8071	몽타주 스케치 (하)

스케치와 모든 몽타주 데이터를 8:2 비율로 나누어 train, validation data로 활용하며, hyperparameter는 해당 논문에서 제시한 기본 설정을 따라 학습한다.

모델 학습은 NVIDIA GeForce RTX 3090 24GB 두 개의 GPU로 약 5일 동안 135 epoch 수행한다. overfitting을 방지하기 위해 Autoencoder, Generator, Discriminator 등의 loss를 고려하여 각 학습 단계마다 가장 우수한 성능을 나타내는 모델 3개를 선택한다. 135 epoch까지 학습을 진행하였음에도, 105 epoch 모델이 가장 우수한 성능을 보여주었기 때문에 DALL-E 모델의 image decoder로 105 epoch 모델을 사용한다.

DALL-E

본 연구는 기존의 DALL-E와 달리 klue/roberta-large를 text encoder로, VQ-GAN을 image decoder로 사용한다. 모델 학습은 NVIDIA GeForce RTX 3090 24GB 한 대의 GPU에서 진행한다.

표 2. DALL-E 학습에 사용한 hyperparameter

Hyperparameter	
batch size	24
learning rate	3.0e-5
text sequence length	256
depth	16
attention type	full

표 2의 hyperparameter를 포함한 그 외의 조건들은 모두 고정하고 epoch 단위로 모델 성능을 확인하며 세 단계로 나누어 학습을 진행한다. 이 과정에서는 특별한 성능 평가 지표를 사용한 것이 아니라 직접 눈으로 결과 샘플들을 확인하면서 성능을 판단한다.

표 3. DALL-E 학습 단계별로 조정한 input

epoch	input image	input text
0~9	인물 스케치, 몽타주 상, 중, 하	인상
10~12	인물 스케치, 몽타주 상	인상

3.2. 모델 학습

3.2.1. DALL-E

VQ-GAN

DALL-E를 학습하기 전, DALL-E의 image decoder로 사용될 VQ-GAN을 먼저 학습한다. 인물

13~19	인물 스케치, 몽타주 상	특징
-------	------------------	----

표 3은 학습 단계별 입력들을 정리한 것이다. 처음에는 인물 스케치와 모든 몽타주 데이터를 활용하여 학습을 시도하였으나, 낮은 품질의 몽타주 중, 하 데이터로 인해 오히려 모델의 성능까지 하락하는 문제가 발생하였다. 이에 따라 이후 학습 단계에서는 인물 스케치와 몽타주 상의 데이터만을 입력 이미지로 사용한다. 이로써 낮은 품질의 데이터로 인한 모델의 성능 하락을 방지하고자 한다. 또한, 단순히 추상적인 얼굴 인상만을 학습하는 것이 아니라, 얼굴의 특징까지 학습시켜 구체적인 진술도 반영하는 모델을 구축하고자 한다.

3.2.2. Stable Diffusion

본 연구는 huggingface¹에 공개되어 있는 사전 학습된 한국어 stable diffusion 모델인 my-korean-stable-diffusion-v1-5를 fine tuning 하여 사용한다. 한정된 컴퓨팅 자원에서 수행이 가능하도록 하기 위해 LoRA 방식을 이용한다.

LoRA는 Low-Rank Adaptation of Large Language Models으로, 사전 학습된 모델의 일부 파라미터를 조정하여 fine tuning을 수행하는 방법이다. 이 방법을 사용하면 전체 모델을 fine tuning 하는 것에 비해 컴퓨팅 자원을 절약할 수 있다.^[8]

본 연구는 LoRA 방식을 적용하기 위해 다음과 같은 하이퍼파라미터를 설정한다. 기본적으로 선행 연구^[8]에서 쓰인 값과 동일하지만 random flip을 TRUE로 하지 않았다는 점이 다르다.

표 4. LoRA를 적용한 Stable Diffusion fine tuning hyperparameter

Hyperparameter	
random flip	FALSE
center crop	TRUE
train batch size	1
gradient accumulation steps	4
learning rate	1.00E-04
learning rate scheduler	"cosine"

표 4에서, Random flip은 이미지를 랜덤으로 상하 또는 좌우 반전한다. Center crop은 학습 이미지의 중심 부분을 훈련에 이용한다. Train batch size는 한번에 학습할 데이터의 양을 지정한다. Gradient accumulation steps는 gradient를 몇 step마다 업데이트할 것인지 지정한다. Learning rate scheduler는 learning rate를 조정할 스케줄러를 지정한다.

앞서 DALL-E 모델 학습 때와 마찬가지로 위의 하이퍼파라미터를 포함한 그 외의 조건들은 모두 고정하고 몇 epoch 단위로 모델 성능을 확인하며 훈련을 세 단계로 나누어 진행한다. 이 과정에서는 특별한 성능 평가 지표를 사용한 것이 아니라 샘플 여러 가지를 뽑아 판단한다.

표 5. Stable Diffusion 훈련 단계별로 조정한 파라미터

epoch	image	input text	resolution
0 - 8	몽타주 상, 중, 하	인상	256
9 - 23	몽타주 상	특징	256
23 - 33	몽타주 상	인상 + 특징	512

초기에는 몽타주 상, 중, 하 데이터 모두를 가지고 8 epochs를 훈련한다. 원본 데이터의 해상도는 512*720이나 훈련 시에는 한정된 자원 환경에서 빠르게 학습하기 위해 256*256으로 낮춘다.

하지만 중, 하 몽타주 데이터의 엄청난 묘사로 인해 일정한 품질의 이미지가 생성되지 않아 이후 추가로 15 epochs는 상 데이터만으로 훈련한다. 이후 전단계의 문제는 해결되었으나 얼굴이 부자연스럽고 다소 정형적인 결과를 보인다.

¹ 데이터 사이언스와 머신 러닝을 위한 오픈 소스 플랫폼

따라서 마지막 10 epochs는 해상도를 원래에
가깝게 512*512로 올려 진행한다. 이후에는 더 이상
모델의 성능이 개선되지 않아 훈련을 중단한다.

Stable diffusion의 fine tuning은 Google
Colab에서 진행한다. GPU는 NVIDIA T4, 메모리
16GB, 저장 공간은 500GB SSD이다.

3.3. 시스템 아키텍처

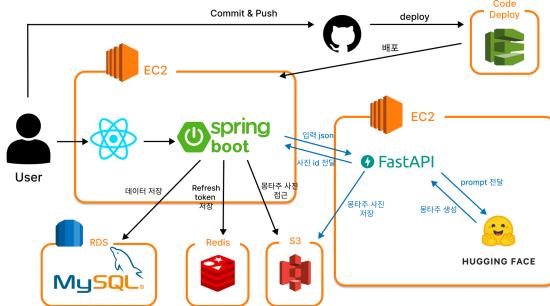


그림 2. 시스템 아키텍처

그림 2는 본 연구에서 설계 및 개발한 시스템 아키텍처 그림이다. 프론트엔드에서는 React.js와 Typescript를 사용한다. 백엔드에서는 웹 서버를 위해 Spring boot로 서비스를 개발하며, 데이터베이스로는 MySQL을 사용하고 Refresh Token을 관리하기 위한 데이터베이스로는 Redis를 사용한다. 또한, Open Feign을 이용해 데이터 모델 서버의 api를 호출할 수 있도록 구현한다. 추가로, Github Actions를 활용한 CI/CD 파이프라인을 구축하여 새로운 기능 개발 시 빠르게 테스트 및 배포되도록 한다.

4. 실험 및 결과 분석

DALL-E, Stable Diffusion 모델의 성능을 비교하기 위해 test data에서 임의로 선택한 9개의 텍스트를 사용한다. 구체적인 진술과 추상적인 진술에 대한 종합적인 평가를 위해 3개는 구체적인 표현만으로, 다른 3개는 추상적인 표현만으로, 그리고 나머지 3개는 구체적인 표현과 추상적인 표현의 혼합으로 구성한다. 원래는 30개의 텍스트를 고려하였으나 사용자 평가 시 설문자의 피로를 줄이기 위해 9개로 축소한다. 평가 지표로는 TIFA와 사용자 평가를 사용한다.

다음은 성능 평가에 이용된 샘플 이미지로, 두 모델에 제시된 텍스트를 각각 입력하여 생성한다. 그림 3-1은 구체적 텍스트 입력 결과 이미지이고, 그림 3-2은 구체적, 추상적 표현 혼합 텍스트 입력 결과 이미지이다.



그림 3-1. 구체적 텍스트 입력 결과 이미지
(좌 DALL-E, 우 Stable Diffusion)

“남성. 50대. 얼굴은 둥글고 턱은 둥근형이다. 광대가 나왔다. 짧은 머리이고 눈썹이 흐리고 미간은 넓다. 작은 눈에 코가 크고 입술은 얇다. 눈가주름이 있다. 팔자주름이 있다.”



그림 3-2. 구체적, 추상적 표현 혼합 텍스트 입력
결과 이미지 (좌 DALL-E, 우 Stable Diffusion)

“남성. 30대. 얼굴은 둥글고 턱은 둥근형이다. 볼이
통통하다. 짧은 머리이고 왼쪽가르마를 탔다. 눈썹이
흐리고 미간은 좁다. 인중은 길다. 팔자주름이 있다.
깔끔하게 내려온 짧은 머리와 굴곡없는 동그란
얼굴은 깔끔하고 자기 관리를 잘하는 사람으로
보이며 표정이 차분해 보여 진지하고 꼼꼼한
이미지로도 느껴진다. 매사에 침착하고 성실한
사람으로도 보인다.”

4.1. TIFA

TIFA는 text-to-image 모델의 성능을 평가하기 위한 지표로 생성된 이미지에 텍스트가 얼마나 잘 반영되었는지를 0부터 1의 스케일로 나타낸다.

TIFA는 입력 텍스트로부터 채점 기준이 될 질문들을 생성한다. 그리고 생성된 이미지를 인식하여 각 질문의 요구사항을 충족하는지를 1(예) 또는 0

(아니오)으로 채점하고 점수들의 평균을 최종 점수로 제출한다.^[9]

4.2. 사용자 평가

사용자 평가는 설문지를 통해 텍스트와 생성된 이미지가 얼마나 유사한지를 사용자가 1부터 5의 스케일로 답변하도록 하였으며, 총 26명이 참여하였다.

4.3. 실험 결과

표 6. 두 모델의 성능 평가 결과

	TIFA	사용자 평가
StableDiffusion	0.50 / 1	3.17 / 5
DALL-E	0.53 / 1	2.96 / 5

표 6은 DALL-E, Stable Diffusion의 성능 평가 결과다. TIFA 결과에서는 DALL-E가 약간 앞서는 듯 했으나, 사용자 평가에서는 Stable Diffusion이 더 앞서는 결과를 보였다. 봉타주 분야에서는 사용자의 평가가 더 신뢰할 수 있다고 판단하여, 최종적으로는 사용자 평가 점수가 더 높은 Stable Diffusion을 선택하였다.

그림 4-1부터 4-7은 본 연구에서 fine tuning을 완료한 Stable Diffusion의 이미지 생성 결과들이다. 그림 4-1의 프롬프트 텍스트를 기준으로 텍스트 일부를 수정하여 그림 4-2부터 4-7을 생성하였다. 이를 통해 텍스트 수정이 이미지에 어떻게 반영되는지 확인할 수 있다.



그림 4-1.

"남성. 40대. 얼굴은 사각형이고 턱은 각진형(사각)이다. 광대가 나왔다. 짧은 머리이고 원쪽가르마를 탔다. 눈썹이 진하고 작은 눈에 코가 크고 인중은 길다. 입이 작다. 팔자주름이 있다. 관리하지 않은 헤어나 피부에서 상당한 피로감이 보인다. 짜려보는 듯 째진눈에서 날카로운 통찰을 할 것 같은 느낌이 있고, 형사 또는 건축의 총괄을 할 것 같은 차분하면서 치밀한 성격의 이미지이다."

그림 4-2는 "40대"를 "20대"로 수정했을 때, 그림 4-3은 "남성"을 "여성"으로 수정했을 때, 그림 4-4는 얼굴과 턱을 각각 "사각형"에서 "계란형"으로 "각진형"에서 "둥근형"으로 수정했을 때, 그림 4-5는 "왼쪽 가르마"를 "오른쪽 가르마"로 수정했을 때, 그림 4-6은 "쌍꺼풀이 있다"를 추가했을 때, 그림 4-7은 눈썹이 "진하고"를 "흐리고"로 수정했을 때 생성된 이미지로 변경된 프롬프트 텍스트를 모델이 잘 반영하는 것을 확인할 수 있다.

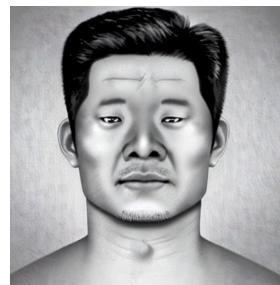


그림 4-2.
"40대" → "20대"

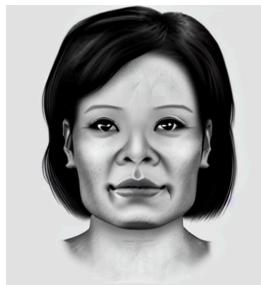


그림 4-3.
"남성" → "여성"



그림 4-4.
얼굴 "사각형" → "계란형" "왼쪽" → "오른쪽" 가르마
턱 "각진형" → "둥근형"

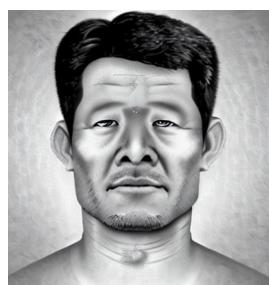


그림 4-5.



그림 4-6.
"쌍꺼풀이 있다" 추가 눈썹이 "진하고" → "흐리고"



그림 4-7.

5. 결론

본 연구를 통해 개발된 몽타주 제작 프로그램은 한국 수사 현장의 몽타주 제작 과정과 같이 회상 원리에 기반하되, 구체적인 진술뿐만 아니라 추상적인 진술도 반영할 수 있는 가능성을 제시하였다. 또한, 30초 이내에 몽타주 생성을 가능하게 함으로써 효율성을 향상시킬 수 있는 가능성을 제시하였다. 따라서 본 연구는 진술을 다각적으로 반영하여 몽타주 제작에 도움을 줄 수 있는 방법을 제시한 점에서 의의가 있다. 이는 기존의 몽타주 제작 방식의 장점을 유지하면서도 범죄 수사에 보다 효과적으로 활용될 수 있는 몽타주 제작 프로그램을 개발하는 데 기여할 것으로 기대된다.

그러나 Generative Model 특성상 때로 무너진 이미지가 생성될 수 있는 점, 긴 텍스트의 경우 모든 내용을 완벽하게 반영하지 못할 수 있는 점, 텍스트 수정에 따른 이미지 수정 기능이 완벽하지 않다는 점에서 한계가 있었다.

참고 문헌

- [1] 이연주. “그놈 얼굴을 찾아라’ 결정적 단서 그리는 몽타주 담당 수사관의 세계.” *jobsN*. 2017.01.31.
- [2] Zahradníkova, Barbora, Zuzana Sutova, and Peter Schreiber. "Interactive evolution of facial composites." *IFAC-PapersOnLine* 50, no. 1 (2017): 11776-11781.
- [3] Zaltron, Nicola, Luisa Zurlo, and Sebastian Risi. "Cg-gan: An interactive evolutionary gan-based approach for facial composite generation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, pp. 2544-2551. 2020.
- [4] Park, Sungjoon, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyo Han, Jangwon Park, Chisung Song et al. "Klue: Korean language understanding evaluation." *arXiv preprint arXiv:2105.09680* (2021).
- [5] Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark

- Chen, and Ilya Sutskever. "Zero-shot text-to-image generation." In *International Conference on Machine Learning*, pp. 8821-8831. PMLR, 2021.
- [6] Esser, Patrick, Robin Rombach, and Björn Ommer. "Taming transformers for high-resolution image synthesis." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873-12883. 2021.
- [7] Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684-10695. 2022.
- [8] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).
- [9] Hu, Yushi, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. "Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering." *arXiv preprint arXiv:2303.11897* (2023).