



Imagic: Text-Based Real Image Editing with Diffusion Models

□Euron 6기 방선유

목차

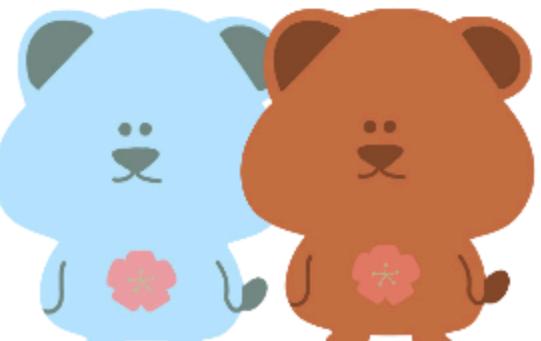
01 Introduction

02 Related Works

03 Imagic: Diffusion-Based Real Image Editing

04 Experiments

05 Conclusions and Future Work



Introduction



01 Introduction

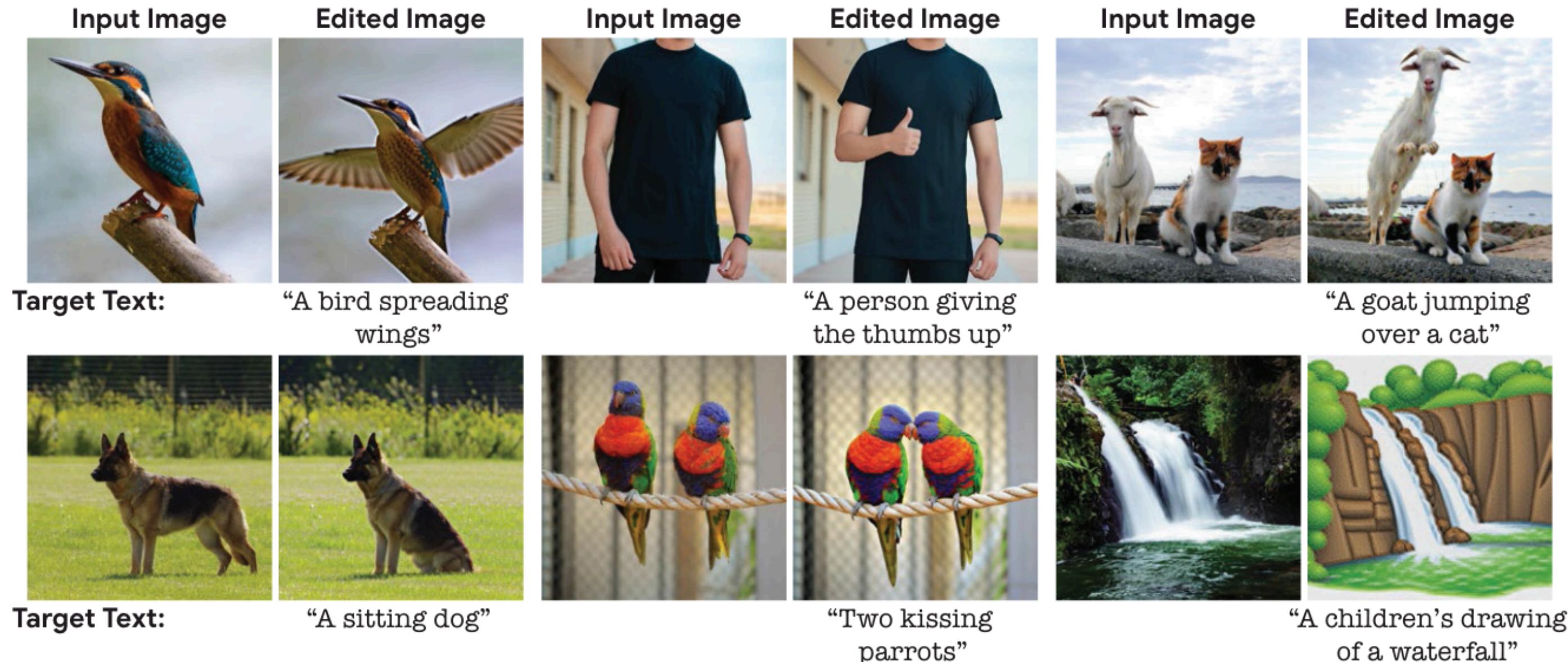
배경

- Text-conditioned image editing은 많은 관심을 끌고 있음
- 기존 방법의 한계점
 - specific editing types (e.g., object overlay, style transfer)
 - synthetically generated images (합성으로 생성된 이미지)
 - 동일한 object에 대한 여러개의 input images를 요구

01 Introduction

⇒ 논문에서 제안하는 방법: 복잡한 text (구조화되지 않은) 기반의 의미 편집을 단일한 real image에 적용

ex) 원래 특징은 유지하면서 한가지 또는 여러가지 object에 대해 포즈와 구성을 바꿀 수 있음



01 Introduction

1. 최초의 텍스트 기반 의미론적 이미지 편집 기술, **Imagic**

: 전체 구조와 구성을 유지하면서 단일한 실제 입력 이미지에 대해 복잡하고 비구조적인 편집을 허용

2. 두 텍스트 임베딩 시퀀스 간의 의미론적 Linear interpolation 시연,

text-to-image diffusion models의 강력한 구성 능력을 밝힘.

3. TEdBench 소개

: 다양한 텍스트 기반 이미지 편집 방법을 비교할 수 있는 새롭고 도전적인 복잡한 이미지 편집 벤치마크

Related Works



02 Related Works

- 최근 이미지 합성 품질의 발전에 따라, 많은 작업에서는 사전 훈련된 GAN의 잠재 공간을 활용하여 다양한 이미지 조작을 수행했음
- 실제 이미지에 이러한 조작을 적용하기 위한 다양한 기법이 제안됨
 - : optimization 기반, 인코더 기반, 입력별 모델 조정 방법, GAN 기반 방법, 딥러닝 기반 시스템 등

02 Related Works

- 최근에는 유사한 이미지 편집 작업에 diffusion 모델 활용
 - SDEdit: 이미지에 중간 노이즈를 추가한 다음 원하는 편집에 따라 diffusion 프로세스를 사용, 노이즈를 제거
 - DDIB: 소스 클래스(또는 텍스트)를 입력받고 DDIM 반전을 사용하여 입력 이미지를 인코딩하고 대상 클래스에 맞게 다시 디코딩하여 편집된 버전을 얻음
 - DiffusionCLIP: Language-Vision Model gradient, DDIM 반전, fine-tuning을 활용하여 도메인별 diffusion 모델을 사용하여 이미지를 편집
 - Textual Inversion, DreamBooth: 주제에 대한 3~5개의 이미지와 대상 텍스트로 주어진 object에 대한 새로운 view를 합성

02 Related Works

⇒ **Imagic**

단일한 실제 이미지에서 작동,

높은 fidelity(충실도)를 유지,

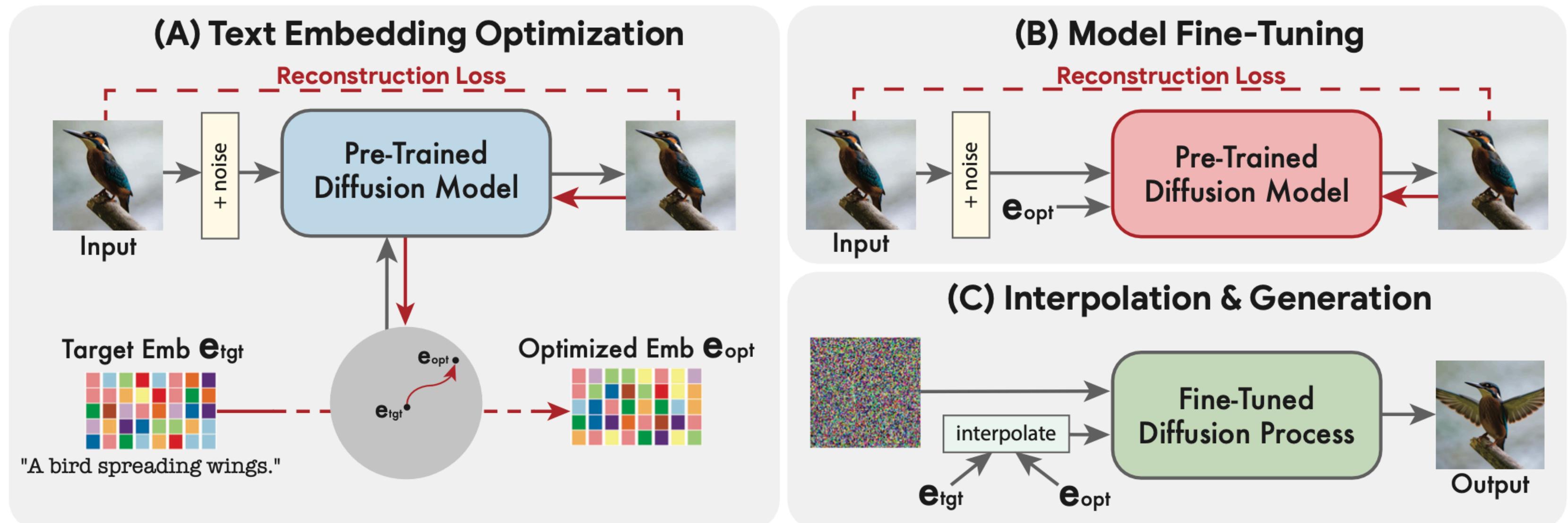
단일한 자유 형식의 자연어 텍스트 프롬프트가 주어지면 고정되지 않은 편집을 적용하는

최초의 텍스트 기반 의미론적 이미지 편집 도구

Imagic

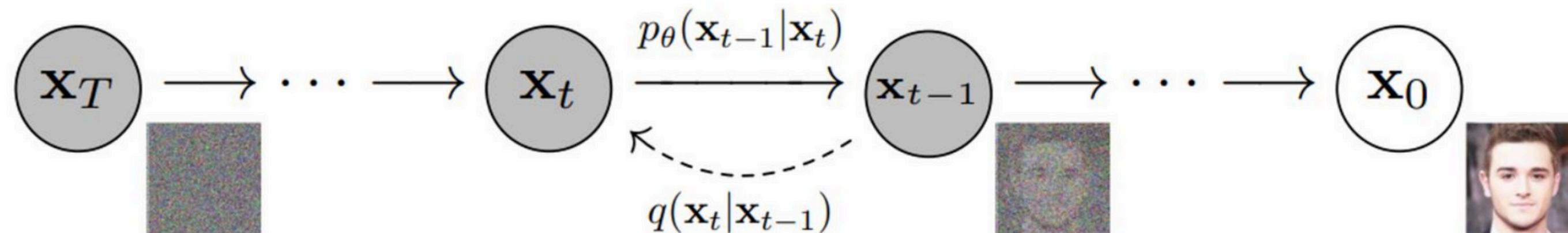


03 Imagic: Diffusion-Based Real Image Editing



03 Imagic: Diffusion-Based Real Image Editing

1. Preliminaries



[Figure 1] Diffusion process



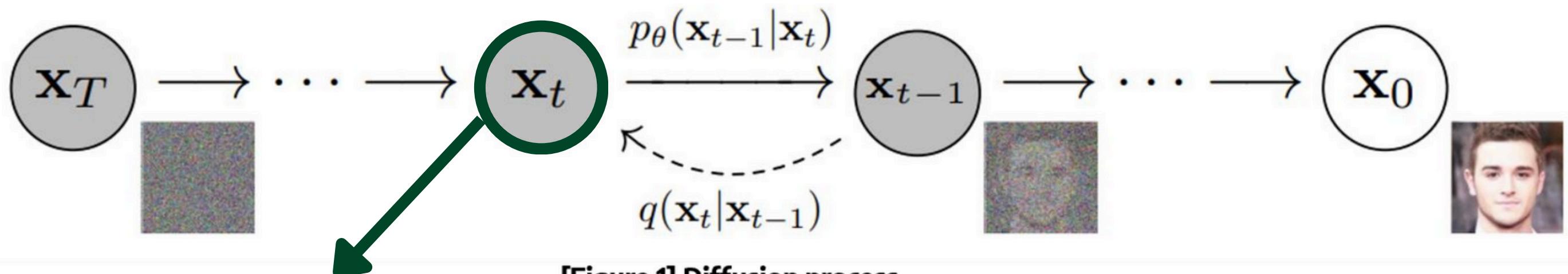
[Figure 2-1] Backward process



[Figure 2-2] Forward process

03 Imagic: Diffusion-Based Real Image Editing

1. Preliminaries



$$\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_t,$$

$$\mathbf{x}_t \xrightarrow[f_\theta(x_t, t)]{} \mathbf{x}_{t-1}$$

03 Imagic: Diffusion-Based Real Image Editing

1. Preliminaries

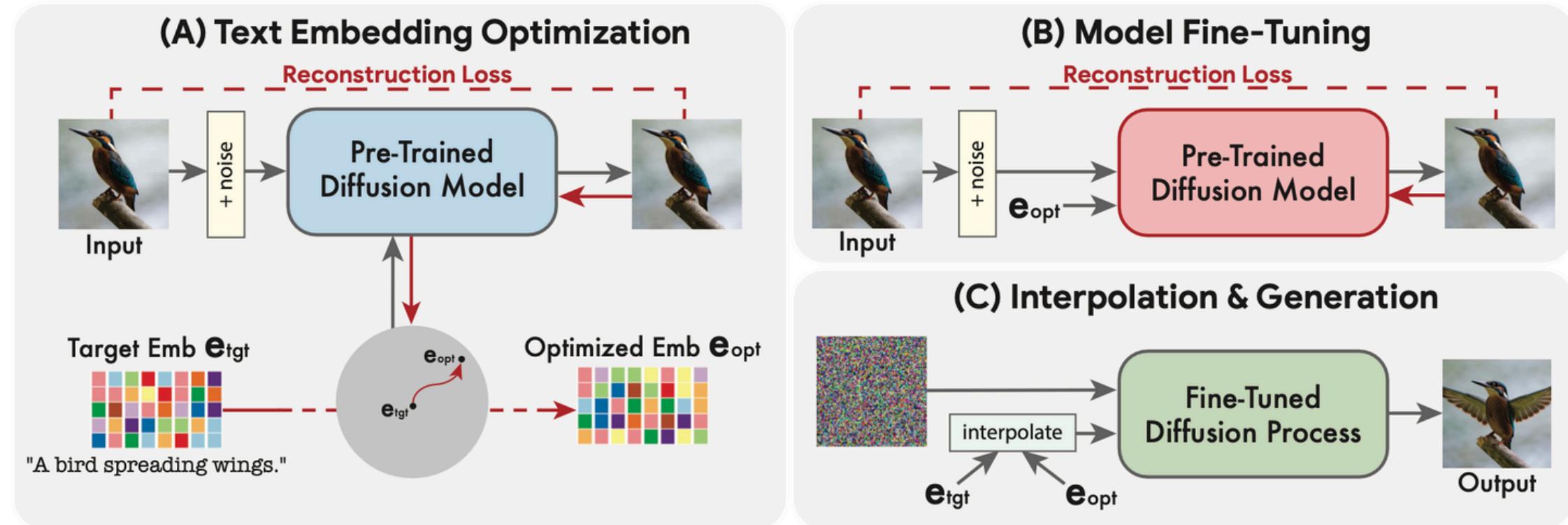
$$\mathbf{X}_t \xrightarrow{f_\theta(x_t, t, y)} \mathbf{X}_{t-1}$$

조건부 분포 학습

- 보조적인 입력 y 에 의존하여 노이즈 제거 네트워크를 조건부로 설정
- 조건부 입력 y : 원하는 이미지의 저해상도 버전, 클래스 레이블, 이미지를 설명하는 텍스트 시퀀스 등
- LLMs, 하이브리드 vision-language 모델에서 비롯된 지식 사용
 - 텍스트 프롬프트만으로 현실적인 고해상도 이미지를 생성할 수 있음

03 Imagic: Diffusion-Based Real Image Editing

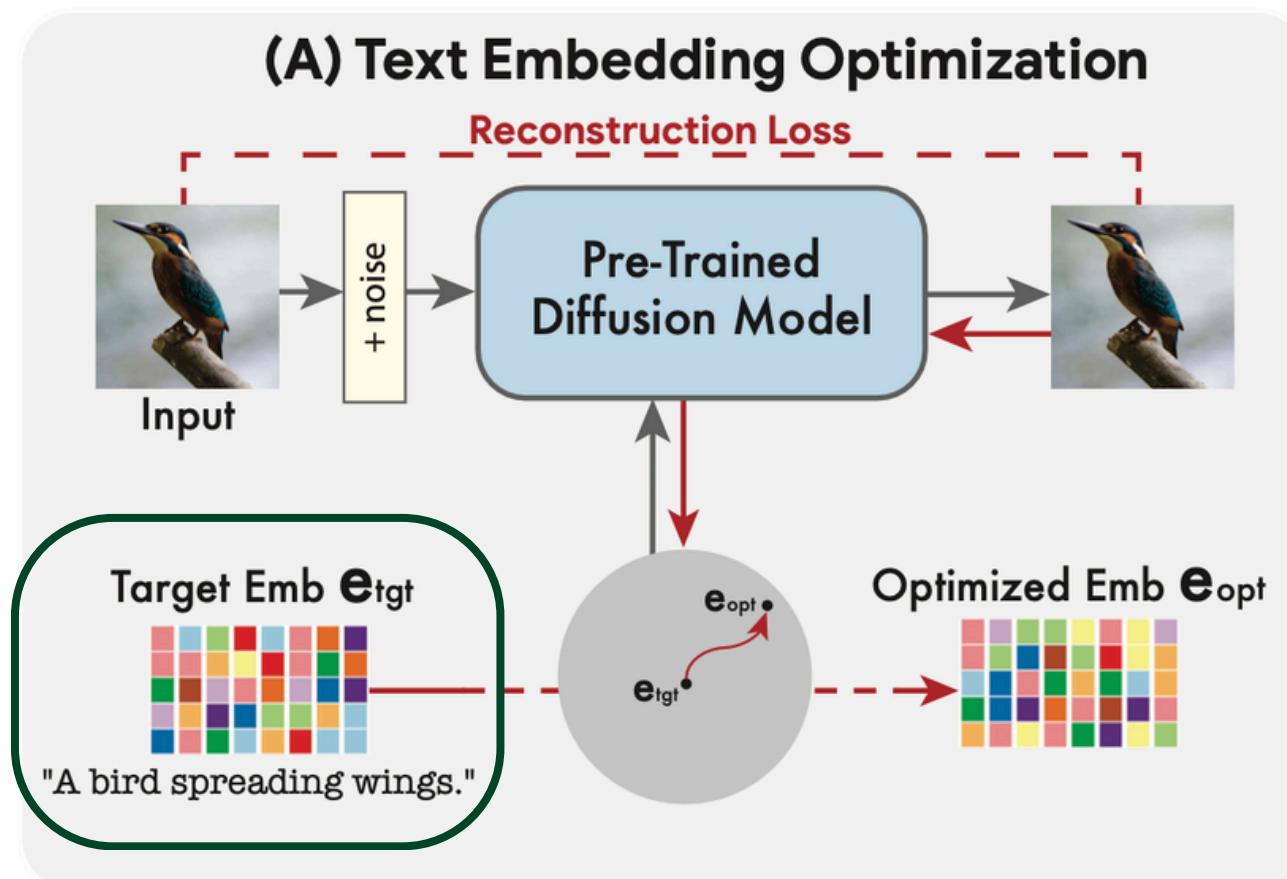
2. Method



- 입력 이미지 x 와 원하는 편집을 설명하는 Target text 가 주어짐
- 목표: 주어진 텍스트를 만족시키면서 입력 x 의 최대한 많은 세부 정보를 보존하는 이미지 편집
→ Diffusion 모델의 텍스트 임베딩 레이어를 활용하여 의미적 조작을 수행
- (A) Text embedding optimization → (B) Model fine-tuning → (C) Text embedding interpolation

03 Imagic: Diffusion-Based Real Image Editing

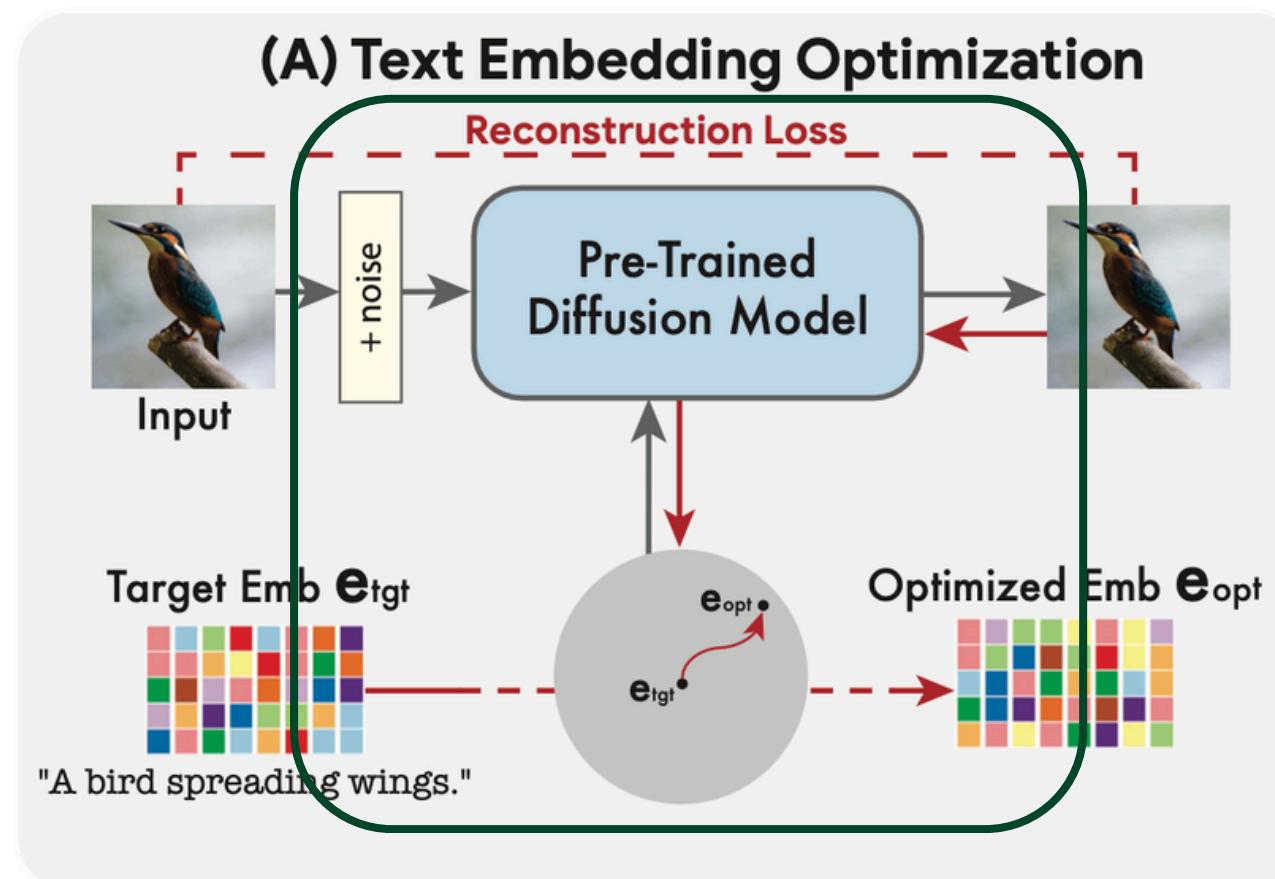
2. Method - (A) Text Embedding Optimization



Target text는 텍스트 인코더로 전달,
인코더는 해당하는 텍스트 임베딩 e_{tgt} 를 출력

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (A) Text Embedding Optimization



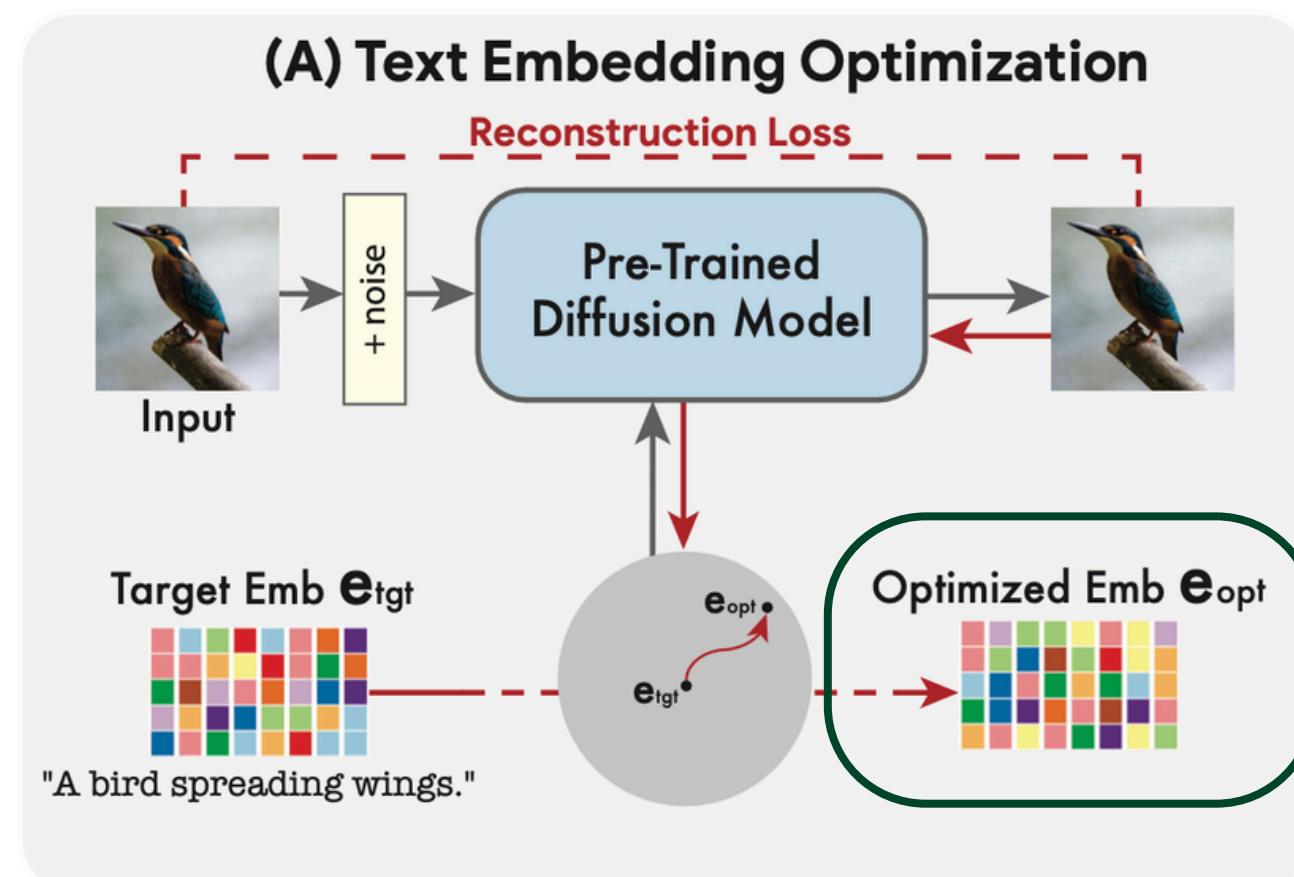
Diffusion 모델 f_{Θ} 의 매개변수를 고정,
denoising diffusion objective
 $\rightarrow e_{tgt}$ 최적화

$$\mathcal{L}(\mathbf{x}, \mathbf{e}, \theta) = \mathbb{E}_{t, \epsilon} \left[\|\epsilon - f_{\theta}(\mathbf{x}_t, t, \mathbf{e})\|_2^2 \right]$$

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (A) Text Embedding Optimization

→ 이 과정을 통해 입력 이미지와 가장 일치하는 텍스트 임베딩을 얻을 수 있음

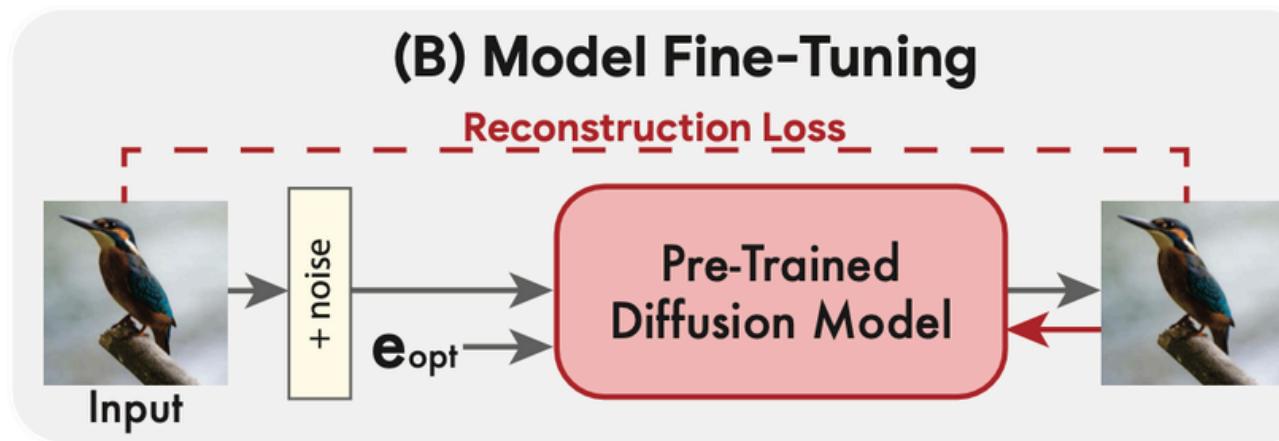


초기 target text 임베딩에 가깝게 유지된 e_{opt} 를 얻기 위해 비교적 적은 step으로 이 프로세스를 실행

이러한 근접성은 임베딩 공간에서 의미 있는 linear interpolation을 가능하게 함

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (B) Model Fine-Tuning



e_{opt} 에서 입력 이미지 x 에 맞게 모델을 이동

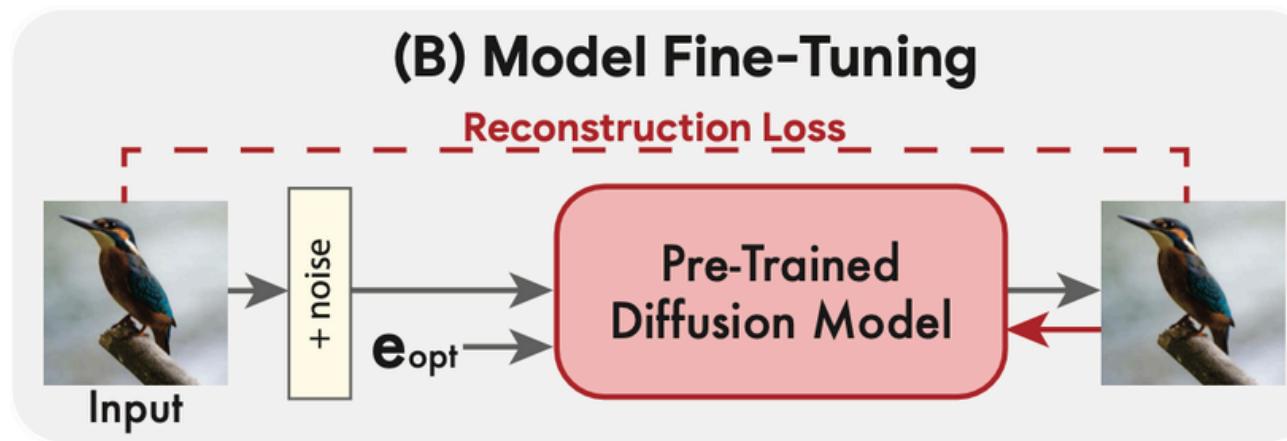
최적화된 임베딩 e_{opt} 가 생성 프로세스를 통과할 때 반드시 입력 이미지 x 로 이어지는 것은 아님

→ 최적화된 임베딩을 고정, 동일한 loss function을 사용하여 모델 파라미터 θ 를 최적화

→ 차이 해소

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (B) Model Fine-Tuning

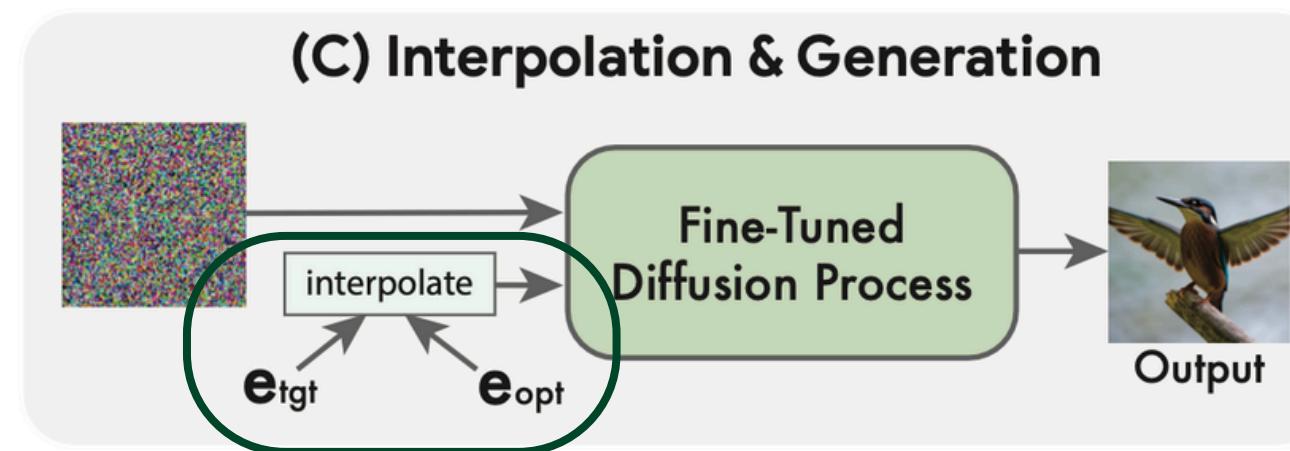


병렬적으로, super-resolution(SR) 모델과 같은 보조 diffusion model을 fine-tuning

- 동일하게 재구성된 loss로 fine-tuning: 편집된 이미지에서 작동하므로 e_{tgt} 가 조건
- 보조 모델의 최적화는 저해상도에 존재하지 않는 x 의 고주파수 디테일을 보존하는 것 보장
- inference 시에 e_{tgt} 를 보조 모델에 입력하는 것이 e_{opt} 를 사용하는 것보다 더 나은 성능

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (C) Text embedding interpolation



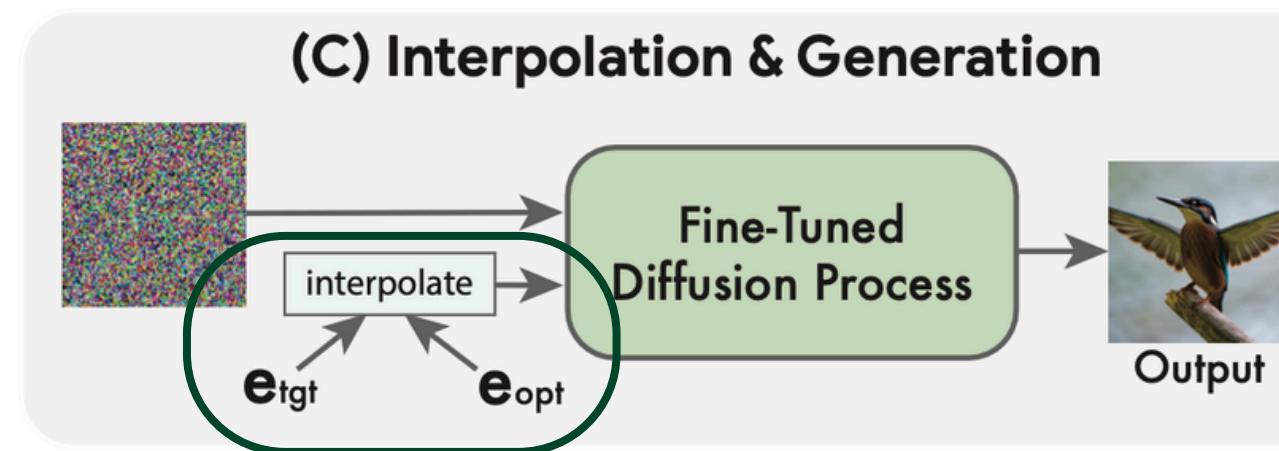
e_{tgt} 와 e_{opt} 사이의 Linear interpolation

Diffusion model은 최적화된 임베딩 e_{opt} 에서 입력 이미지 x 를 완전히 재현하도록 학습됨

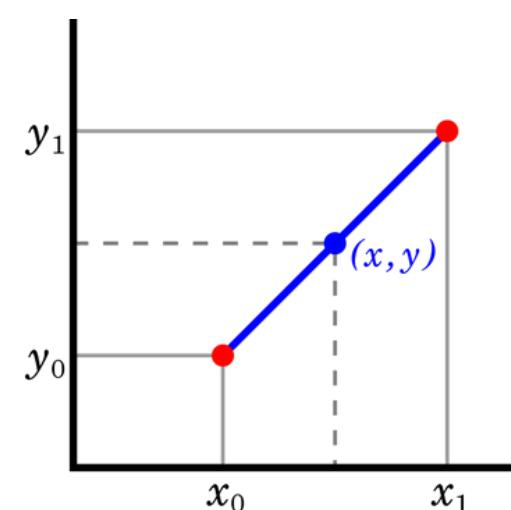
이를 사용하여 타겟 텍스트 임베딩 e_{tgt} 방향으로 진행하여 원하는 편집을 적용

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (C) Text embedding interpolation



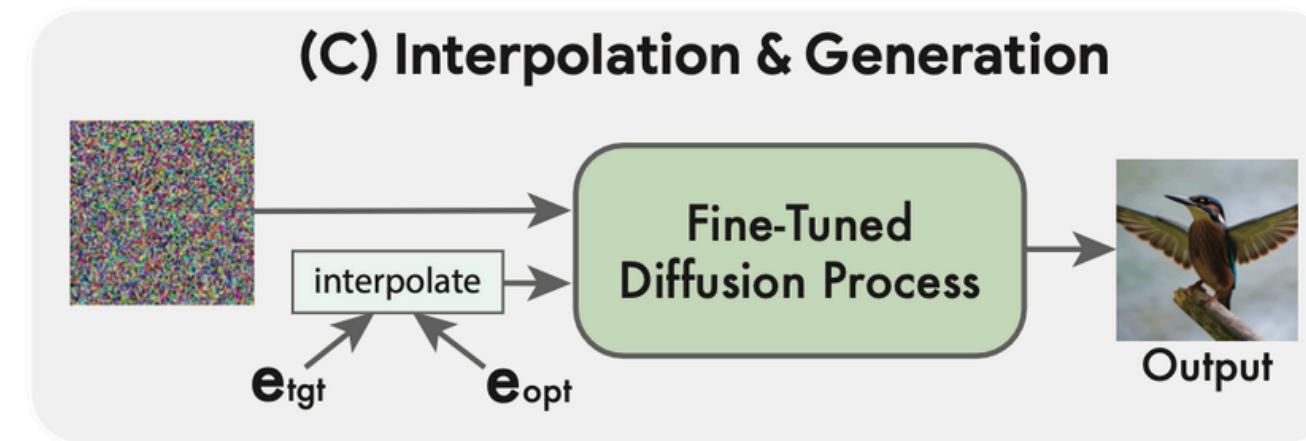
e_tgt와 e_opt 사이의 Linear interpolation



1차원에서 끝점의 값이 주어졌을 때
그 사이에 위치한 값을 추정하기 위하여
직선 거리에 따라 선형적으로 계산하는 방법

03 Imagic: Diffusion-Based Real Image Editing

2. Method - (C) Text embedding interpolation



고해상도의 최종적으로 편집된 이미지 x^- 출력

- 주어진 하이퍼파라미터 $\eta \in [0,1]$ 에 대해 원하는 편집된 이미지를 나타내는 임베딩 e^-
$$\bar{e} = \eta \cdot e_{tgt} + (1 - \eta) \cdot e_{opt},$$
- e^- 에 fine-tuning된 모델을 사용하여 이미지 생성 프로세스 적용
: 저해상도 편집 이미지 → target text에 맞춰 fine-tuning된 보조 모델로 super-resolution(SR) 이미지

03 Imagic: Diffusion-Based Real Image Editing

3. Implementation Details

: 두 가지 다른 최신의 diffusion 모델인 Imagen과 Stable Diffusion을 사용하여 이를 시연

Imagen

- optimizer: Adam (learning rate: 10^{-3})
- 텍스트 임베딩 최적화: 100 steps
- fine-tuning
 - 64×64: 1500 steps
 - 64×64 → 256×256: 1500 steps
 - 256×256 → 1024×1024: fine-tuning하지 않음
- 전체 프로세스: 2개의 TPUv4 칩에서 이미지당 8분 소요

Stable Diffusion

- optimizer: Adam
- 텍스트 임베딩 최적화: 1000 steps (learning rate: $2 * 10^{-3}$)
- fine-tuning: 1500 steps (learning rate: $5 * 10^{-7}$)
- 전체 프로세스: 1개의 Tesla A100 GPU에서 이미지당 7분 소요

Experiments



04 Experiments

1. Qualitative Evaluation

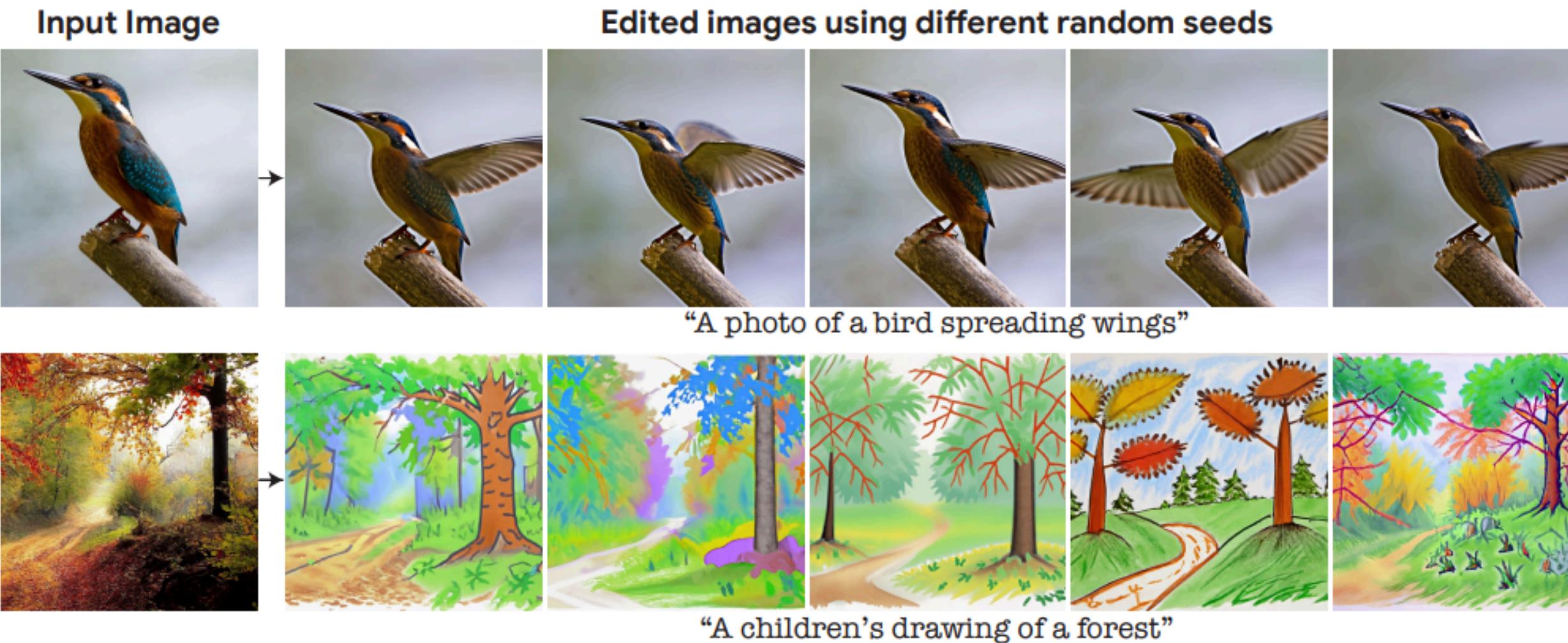
- 여러 타겟 텍스트를 동일한 이미지에 적용한 결과



04 Experiments

1. Qualitative Evaluation

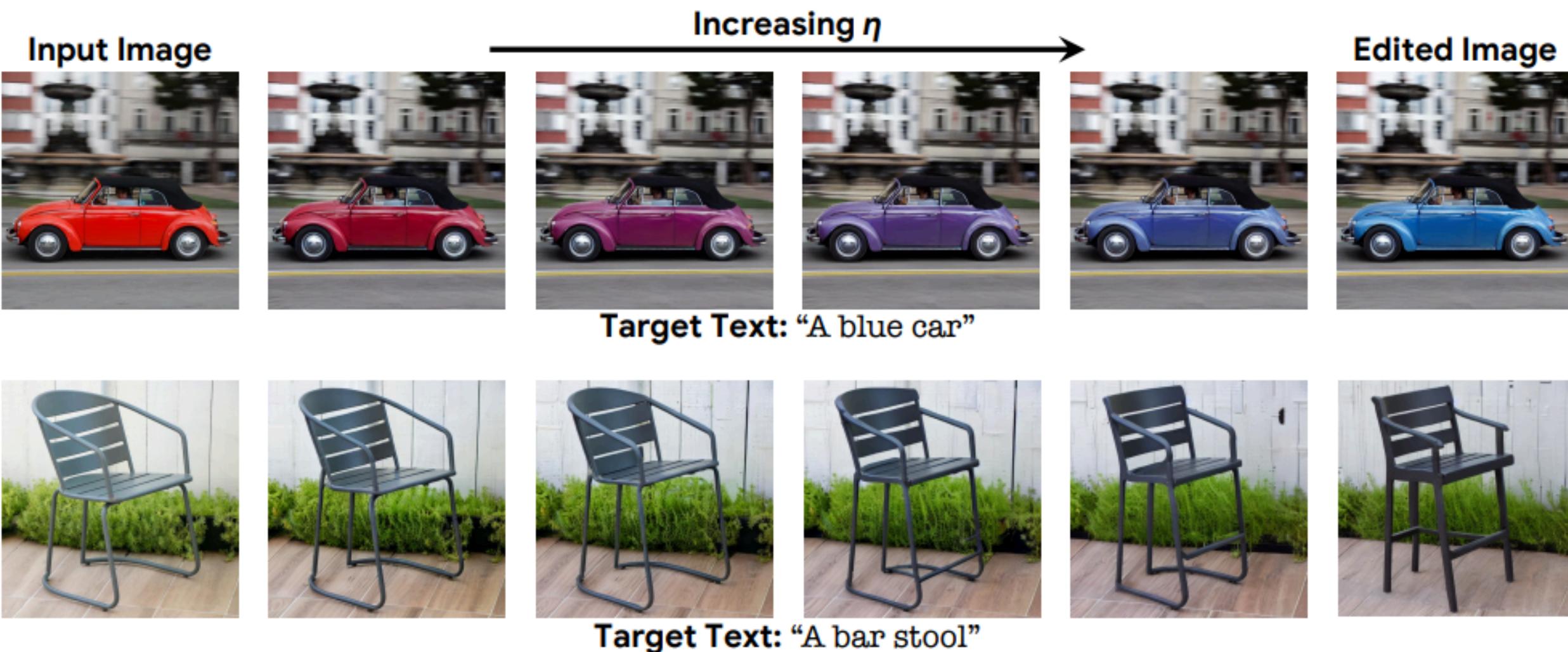
- 여러 랜덤시드로 편집한 결과
: Diffusion 모델이 확률적이기 때문에, 하나의 image-text 쌍에 대해 다양한 결과를 생성할 수 있음



04 Experiments

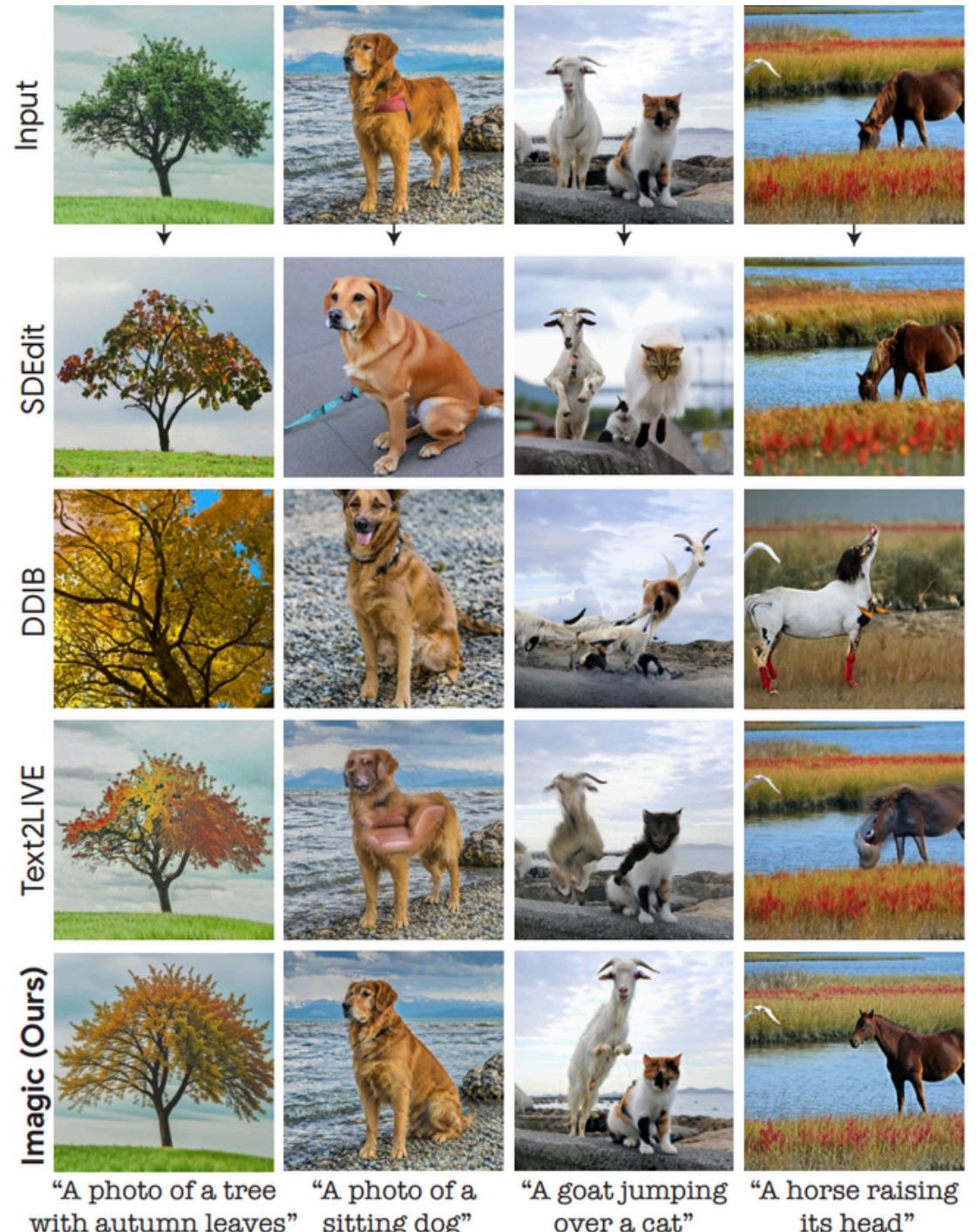
1. Qualitative Evaluation

- 임베딩 interpolation 결과



04 Experiments

2. Comparisons



Imagic을 Text2LIVE, DDIB, SDEdit과 비교한 결과

→ Imagic: 입력 이미지에 대한 높은 충실도를 유지하면서 원하는 편집을 적절하게 수행.

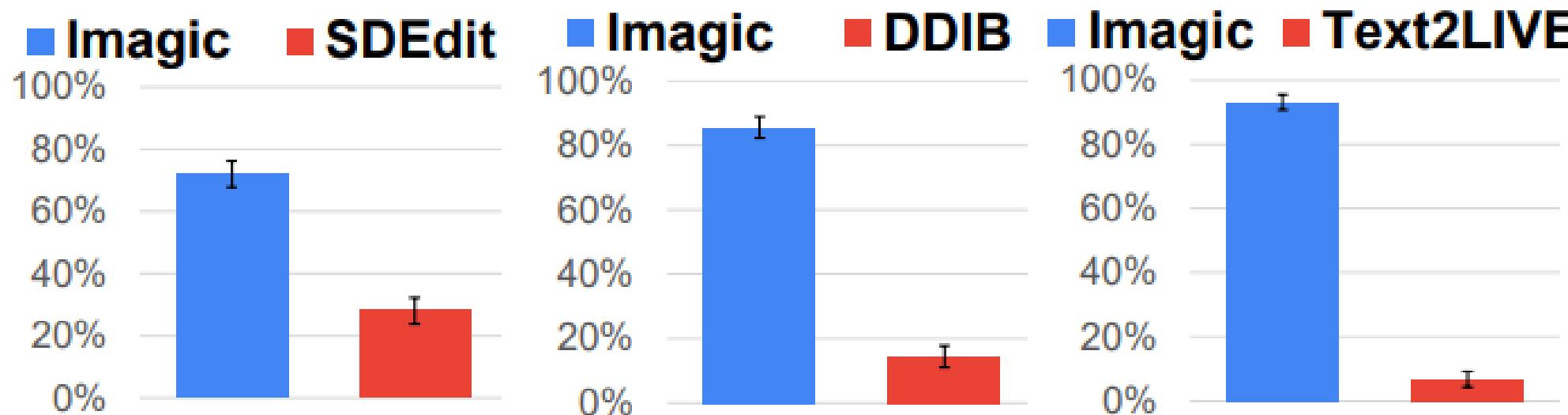
04 Experiments

3. TEdBench and User Study

- TEdBench(텍스트 편집 벤치마크)

: 원하는 복잡한 비구조적 편집을 설명하는 100쌍의 입력 이미지와 대상 텍스트로 구성된 새로운 데이터셋

- Imagic의 성능을 TEdBench에서의 유저스터디를 통해 정량적으로 평가



Imagic vs SDEdit, DDIB, Text2LIVE
→ 평가자들은 모든 편집 결과에 대해 Imagic을 강력히 선호

04 Experiments

4. Ablation Study

Fine-tuning and optimization



임베딩 interpolation 시 사전 학습된 모델(위)과 fine-tuning된 모델(아래)의 결과를 동일한 랜덤시드에서 비교한 결과

04 Experiments

4. Ablation Study

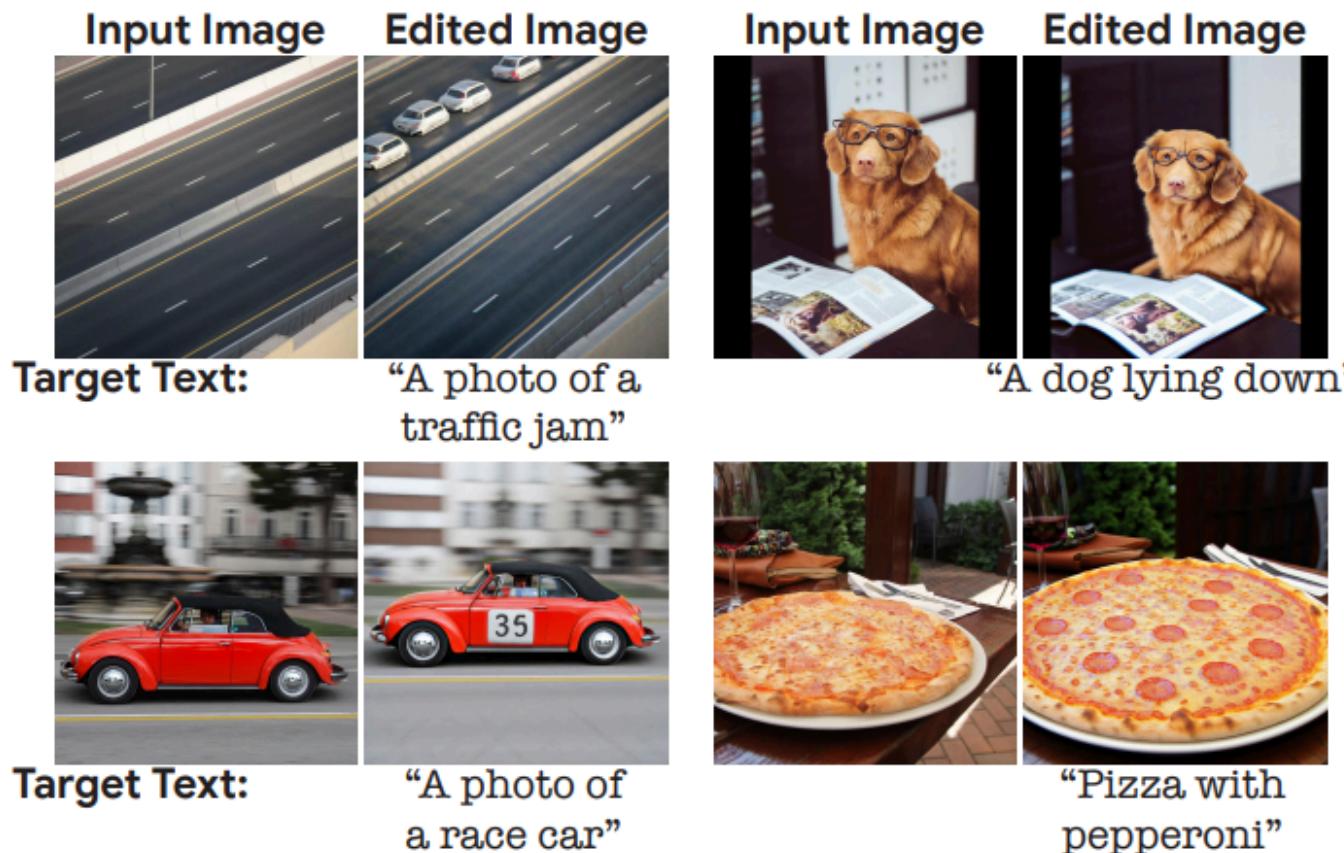
Fine-tuning and optimization

- Fine-tuning x: 스키마는 $\eta = 0$ 에서 원래 이미지를 완전히 재구성하지 못하고,
 η 가 증가함에 따라 이미지의 세부 사항을 유지하지 못함
 - Fine-tuning: 최적화된 임베딩 이상의 입력 이미지 세부 사항을 부과,
스키마가 중간 η 값에 대해 디테일을 유지하도록 함,
의미론적으로 Linear interpolation을 가능하게 함
- 따라서, 모델 파인튜닝이 Imagic의 성공에 중요함

04 Experiments

5. Limitations

Imagic의 두 가지 주요 실패 사례



- 원하는 편집이 매우 미묘하게 적용되어 타겟 텍스트와 잘 정렬되지 않음
- 편집 내용이 잘 적용되지만 줌이나 카메라 각도와 같은 외부 이미지 디테일에 영향을 미침

04 Experiments

5. Limitations

- Imagic은 사전 학습된 text-to-image diffusion model에 의존
→ 모델의 생성적 한계와 편향을 상속:
기본 모델의 실패 사례 생성이 포함되면 원치 않는 아티팩트가 생성됨
- Imagic에 필요한 최적화 속도가 느림: 사용자 대상 애플리케이션에 직접 배포하는 데 방해

Conclusions



05 Conclusions and Future Work

Future Work

- 방법의 입력 이미지에 대한 충실도 및 식별 보존, 무작위 시드 및 interpolation 매개 변수 η 에 대한 민감성을 더욱 향상시키는 데 초점을 맞출 수 있음
- 각 요청된 편집에 대해 η 를 선택하는 자동화된 방법의 개발

05 Conclusions and Future Work

□ Societal Impact

- 타겟 편집에 대한 텍스트 설명을 사용하여 실제 세계 이미지의 복잡한 편집을 가능하게 하는 것이 목표
 - 기본 텍스트 기반 생성 모델의 사회적 편향에 노출될 수 있음
- 악의적인 사용자들에 의해 가짜 이미지를 합성할 수 있음
 - 합성으로 편집된, 생성된 콘텐츠의 식별에 대한 추가 연구가 필요

THANK YOU

