2024 겨울 세미나

Text-to-Image Diffusion Models and Image Editing

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
Min Seok Kang

Outline

- Background
  - Diffusion Models
  - Classifier & Classifier-free Guidance

- Saharia, Chitwan, et al. **"Photorealistic text-to-image diffusion models with deep language understanding."** Advances in Neural Information Processing Systems(NeurIPS), 2022

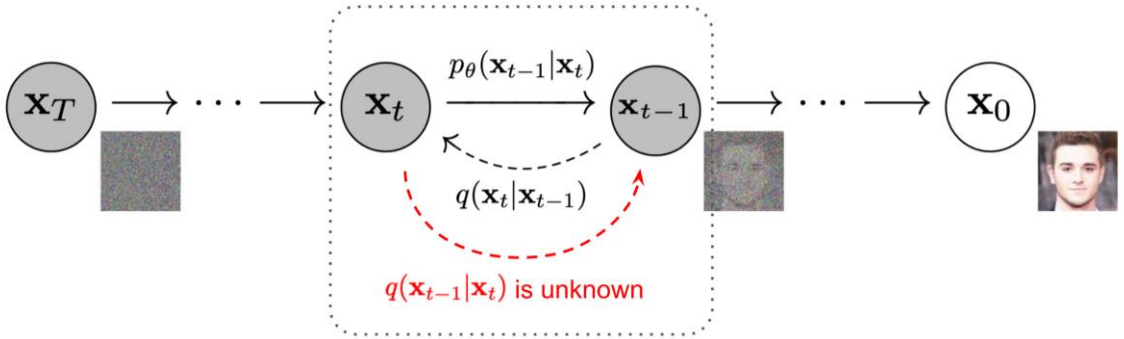
- Hertz, Amir, et al. **"Prompt-to-prompt image editing with cross attention control."** International Conference on Learning Representations(ICLR), 2022

Background

- Diffusion Models
  - Gaussian Noise를 반복적인 denoising 과정을 거쳐서 학습된 data의 분포(image)로 변환하는 생성 모델
  - Conditional diffusion models
    - Diffusion model을 class label/text/저해상도 image로 conditioning 가능
  - Diffusion model $\hat{x}_\theta$는 아래와 같은 denoising objective로 학습

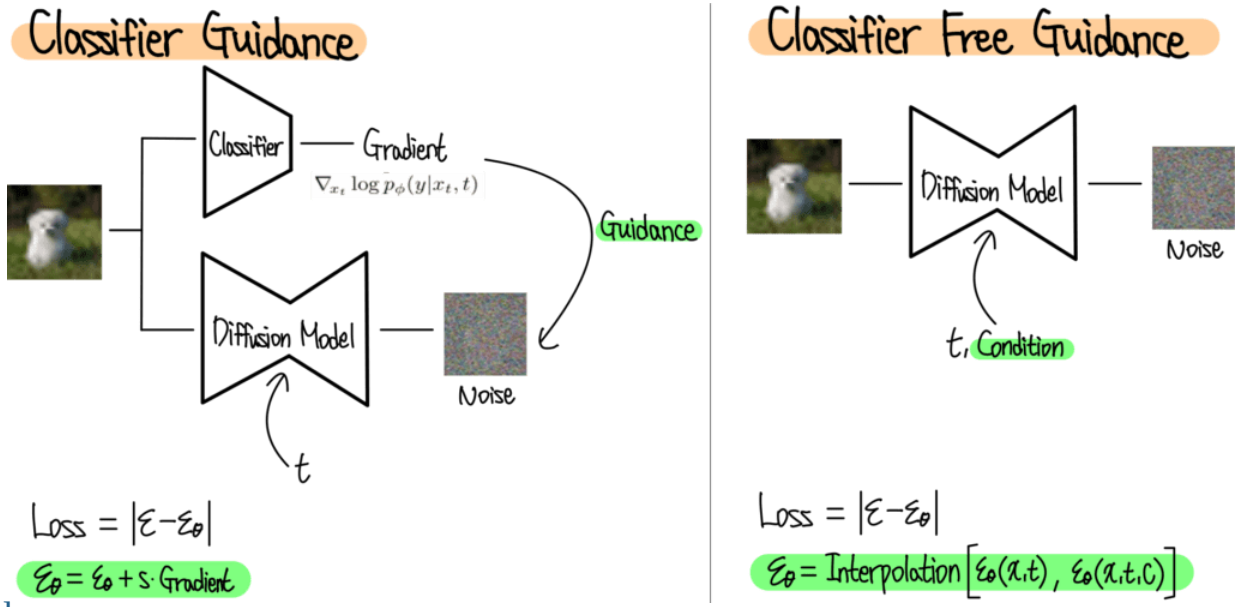
$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\left\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\right\|_2^2\right]$$

    - (x, c): data-condition pair, $t \sim U([0,1]), \epsilon \sim N(0, I)$(Gaussian Noise)
  - 직관적으로 diffusion model은 noise가 있는 $z_t := \alpha_t x + \sigma_t \epsilon$를 x로 denoising하는 것
    - 이 식을 reparameterization trick을 이용하여 ϵ-space에서 ϵ_θ에 대해 squared error loss를 적용
    - 원본 image 자체를 예측하는 문제에서 timestep t에서 t-1로 갈 때 제거할 noise를 예측하는 문제로 전환



Background

- Classifier Guidance vs. Classifier-free Guidance



- Classifier Guidance
  - Classifier model을 따로 두어서 class **y**에 대한 data sample의 gradient score를 뽑고, 이를 사용하여 Diffusion Model을 점점 class가 있는 쪽으로 유도해 생성하도록 함
  - Classifier를 추가로 훈련해야 하고, ImageNet 같이 class가 있는 task에서만 사용가능한 문제가 존재
- Classifier-free Guidance
  - Classifier Guidance의 수식을 Bayes Rule을 이용해 정리하면, conditional function $\epsilon_\theta(x_t, c)$과 unconditional function $\epsilon_\theta(x_t)$의 값을 섞어서 사용하는 Classifier-free Guidance 수식이 나옴
  - Condition을 timestep과 함께 직접 Denoising U-Net에 input으로 넣어줌

$$\tilde{\epsilon}_\theta(x_t, c) = w\epsilon_\theta(x_t, c) + (1 - w)\epsilon_\theta(x_t)$$

Photorealistic text-to-image diffusion models with deep language understanding

Abstract

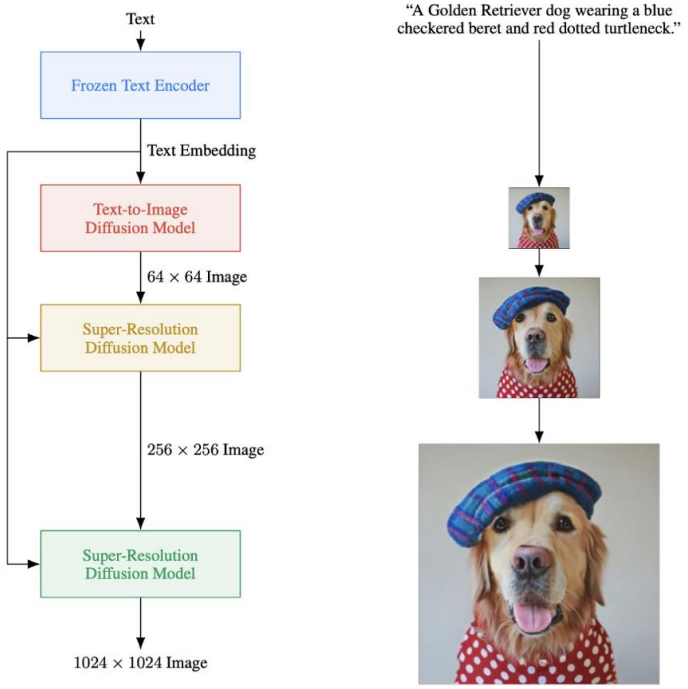
- Google에서 발표한 Text-to-Image Diffusion Model
- Imagen은 두 개의 강력한 모델로 구성됨
  - 뛰어난 언어 이해력을 가진 Transformer 기반 Large Language Models
  - 고품질의 image를 생성해 내는 Diffusion models
- image 없이 text 만으로 pre-trained 된 Language Model을 사용(e.g., T5)
  - image 생성을 위한 text encoding에서 뛰어난 효과를 보임
- COCO Dataset에 대한 FID score에서 7.27의 SOTA 달성
- DrawBench라는 벤치마크를 새로 소개

The Toronto skyline with Google brain logo written in fireworks.

A blue jay standing on a large basket of rainbow macarons.

Introduction

- Text Encoder로 frozen된 T5-XXL를 사용
- 64x64 image를 생성하는 Text-to-Image Diffusion Model과 2개의 Super-Resolution Diffusion Model을 사용해 최종적으로 1024x1024 image 생성
- 모든 Diffusion models는 text embedding에 대해 condition되어 있고, classifier-free guidance를 사용



Key Contributions

1. Text만으로 학습한 Large frozen language models를 text encoder로 사용했을 때 image 생성에 매우 효과적임을 밝힘
2. 새로운 diffusion sampling 기법인 dynamic thresholding을 제안
3. COCO FID에서 7.27로 새로운 SOTA를 달성
4. Text-to-Image task의 새로운 평가 지표인 DrawBench를 소개하며, Imagen이 DALL-E 2 등 다른 모델들에 비해 더 우수함을 확인

Imagen – Pretrained text encoders

- Text-to-Image model
  - text input의 complexity와 compositionality[1)]를 파악하기 위해 문장의 의미를 이해할 수 있는 text encoder가 필요
- 이전의 연구
  - image-text paired data를 사용하여 text encoder를 학습하는 것이 표준(e.g., CLIP)
  - Text encoder를 통해 visually semantic한 표현을 encoding
- Imagen
  - text만으로 이루어진 corpus가 더욱 광범위한 text distribution을 가지고 있다고 주장
  - Large Language model이 text-to-image generation task를 위한 encoder로 활용될 수 있음(e.g., BERT, GPT, T5)
  - BERT, T5, CLIP에 대해서 모두 실험
  - Text encoder는 pre-trained model을 사용하여 추가 학습 없이 freeze함

1) Compositionality(합성성의 원리): 문장을 구성하는 요소들의 의미와 구성 요소들의 조합 방식에 따라 문장의 의미가 결정되는 원리

Imagen – Pretrained text encoders

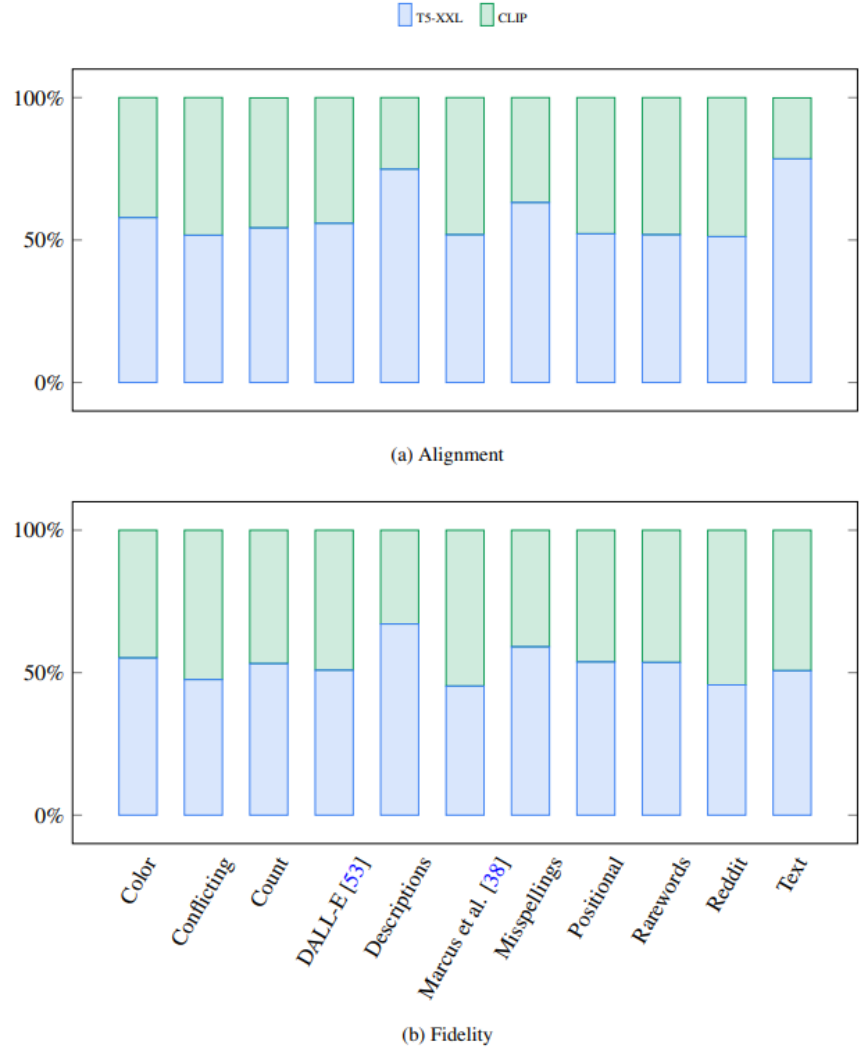
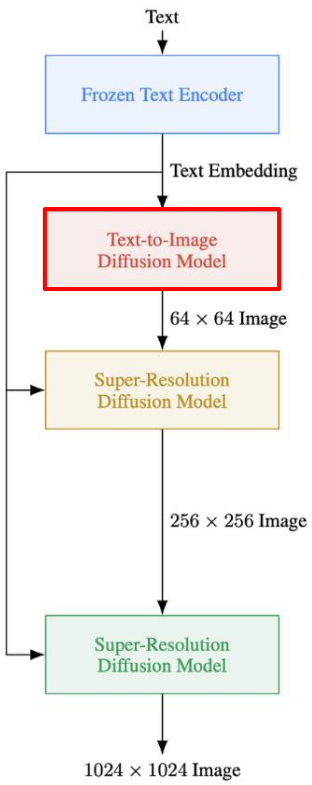


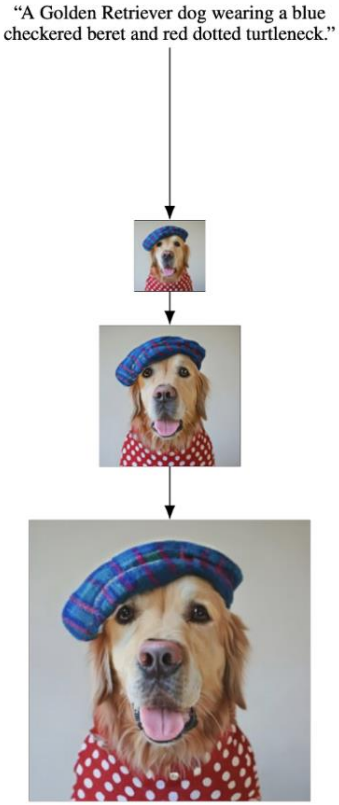
Figure A.7: T5-XXL vs. CLIP text encoder on DrawBench a) image-text alignment, and b) image fidelity.

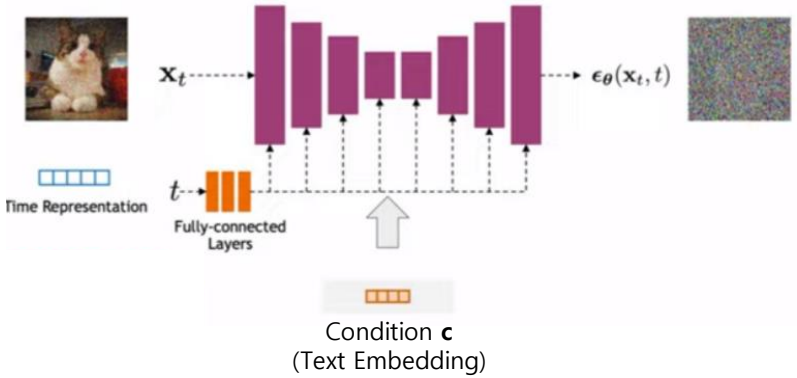
- T5-XXL encoder vs. CLIP encoder
  - MS-COCO와 같은 간단한 benchmark에서는 비슷한 성능
  - Challenging prompt들로 이루어진 DrawBench에서는 image-text alignment와 image fidelity 모두에서 T5-XXL가 우수

Imagen – Classifier-free Guidance

- Imagen에서는 Text-to-Image Diffusion Model로 Classifier-free Guidance Diffusion Model을 사용







Condition **c**
(Text Embedding)

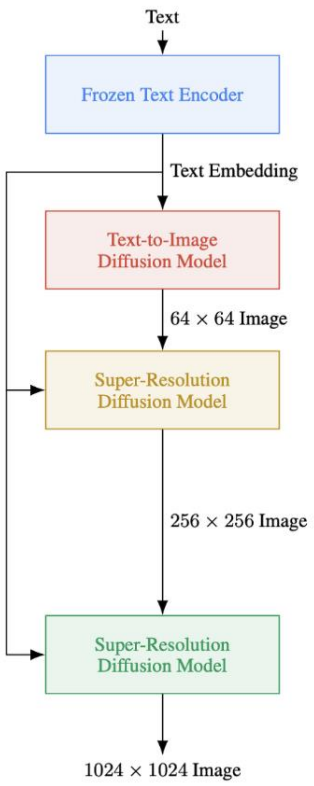
$$\tilde{\epsilon}_\theta(x_t, c) = w\epsilon_\theta(x_t, t, c) + (1 - w)\epsilon_\theta(x_t, t)$$

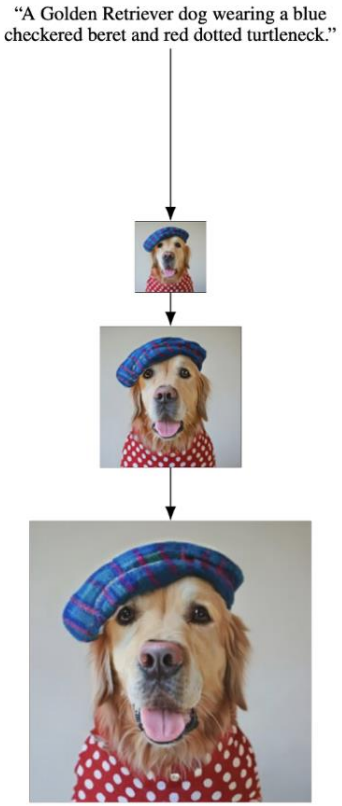
Imagen – Large Guidance Weight Samplers

- Classifier-free Guidance 방식으로 Diffusion Model을 학습했을 때, guidance weight **w**값이 일정 수준보다 커지면 pixel 값이 포화되는 문제가 발생
  - 생성한 image가 검정색으로 가득해짐
  - **x**-prediction인 $\hat{x}_0^t$의 값이 training data **x**의 범위인 [-1, 1]을 벗어나면서 발생
- Static Thresholding
  - **x**-prediction의 pixel 값의 범위를 [-1, 1]로 clipping
- Dynamic Thresholding
  - 각 sampling step마다 threshold s를 dynamic하게 정하고, 정한 [-s, s]의 범위에서 특정 비율을 정해 그 값을 넘는 pixel을 잘라줌

(a) No thresholding. (b) Static thresholding. (c) Dynamic thresholding.

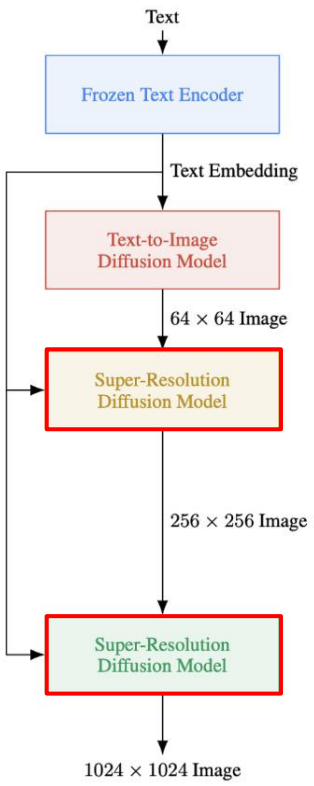
Imagen – Super Resolution Diffusion Model

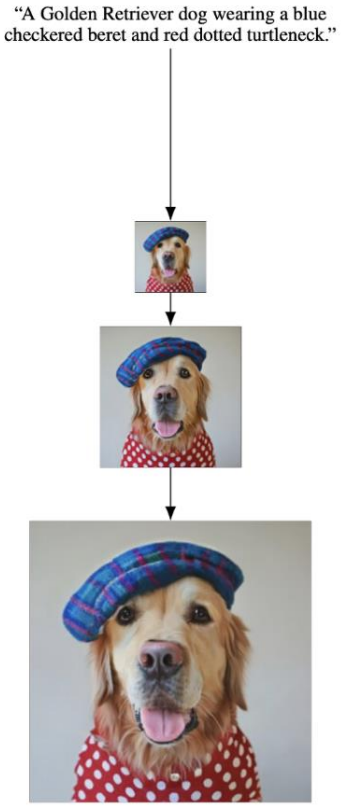




- Text-to-Image Diffusion Model로부터 생성된 image는 64x64의 low resolution image
- 이후 Upscaling을 위한 SR Diffusion Model을 이어 붙임
  - 64x64 → 256x256
  - 256x256 → 1024x1024
  - SR3 논문의 SR Diffusion Model구조를 사용

Imagen – Super Resolution Diffusion Model

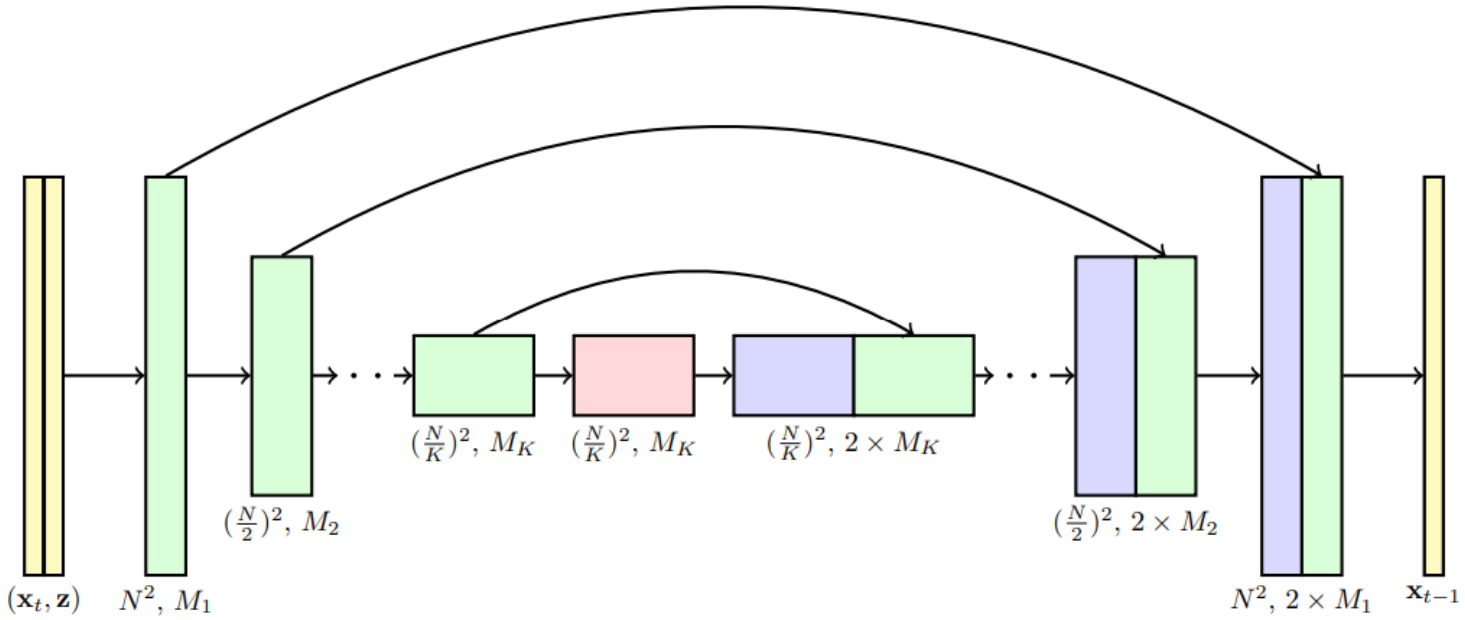




- Text-to-Image Diffusion Model로부터 생성된 image는 64x64의 low resolution image
- 이후 Upscaling을 위한 SR Diffusion Model을 이어 붙임
  - 64x64 → 256x256
  - 256x256 → 1024x1024
  - SR3 논문의 SR Diffusion Model구조를 사용

Imagen – Super Resolution Diffusion Model

- x_t: High-resolution image에서 forward process를 통해 생성된 noisy image
- z: upsampled low-resolution image
  - high-resolution image x_t와 동일한 크기
  - augmentation level(Gaussian noise 강도)을 이용해 low-resolution image를 corrupt 시킴
    - 이러한 Noise conditioning augmentation은 high-fidelity image를 생성하는데 좋은 성능을 보임
- $M_1, M_2, \ldots, M_K$: channel multipliers
- Augmented z와 x_t 를 concatenate하여 text embedding과 함께 U-Net에 넣어줌



Evaluating Text-to-Image Models

- Text-to-Image Model은 COCO dataset의 검증 세트를 기준으로 FID & CLIP score로 성능 측정
  - FID: Image의 fidelity 측정
  - CLIP score: image-text alignment 측정
- COCO dataset의 한계를 보완하기 위해 DrawBench라는 새로운 벤치마크 제안

Category	Description	Examples
Colors	Ability to generate objects with specified colors.	"A blue colored dog." "A black apple and a green backpack."
Counting	Ability to generate specified number of objects.	"Three cats and one dog sitting on the grass." "Five cars on the street."
Conflicting	Ability to generate conflicting interactions b/w objects.	"A horse riding an astronaut." "A panda making latte art."
DALL-E [53]	Subset of challenging prompts from [53].	"A triangular purple flower pot." "A cross-section view of a brain."
Description	Ability to understand complex and long text prompts describing objects.	"A small vessel propelled on water by oars, sails, or an engine." "A mechanical or electrical device for measuring time."
Marcus et al. [38]	Set of challenging prompts from [38].	"A pear cut into seven pieces arranged in a ring." "Paying for a quarter-sized pizza with a pizza-sized quarter."
Misspellings	Ability to understand misspelled prompts.	"Rbefraigerator." "Tcennis rpacket."
Positional	Ability to generate objects with specified spatial positioning.	"A car on the left of a bus." "A stop sign on the right of a refrigerator."
Rare Words	Ability to understand rare words[3].	"Artophagous." "Octothorpe."
Reddit	Set of challenging prompts from DALLE-2 Reddit[4].	"A yellow and black bus cruising through the rainforest." "A medieval painting of the wifi not working."
Text	Ability to generate quoted text.	"A storefront with 'Deep Learning' written on it." "A sign that says 'Text to Image'."

Table A.1: Description and examples of the 11 categories in DrawBench.

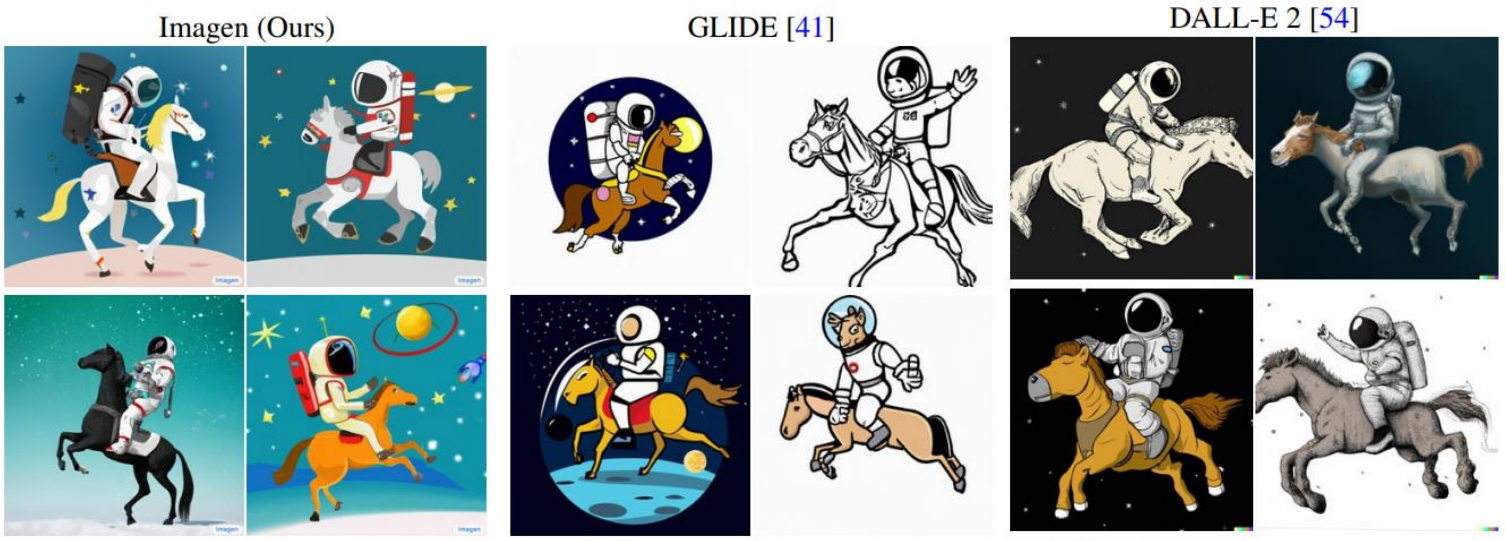
Experiments

New York Skyline with Hello World written with fireworks on the sky.

A storefront with Text to Image written on it.

Experiments

Experiments



Experiments

- Imagen과 COCO dataset의 image quality 및 alignment에 대한 인간 평가

Model	Photorealism ↑	Alignment ↑
Original		
Original	50.0%	91.9 ± 0.42
Imagen	$39.5 \pm 0.75\%$	91.4 ± 0.44
No people		
Original	50.0%	92.2 ± 0.54
Imagen	$43.9 \pm 1.01\%$	92.1 ± 0.55

- COCO validation set에 대한 FID score

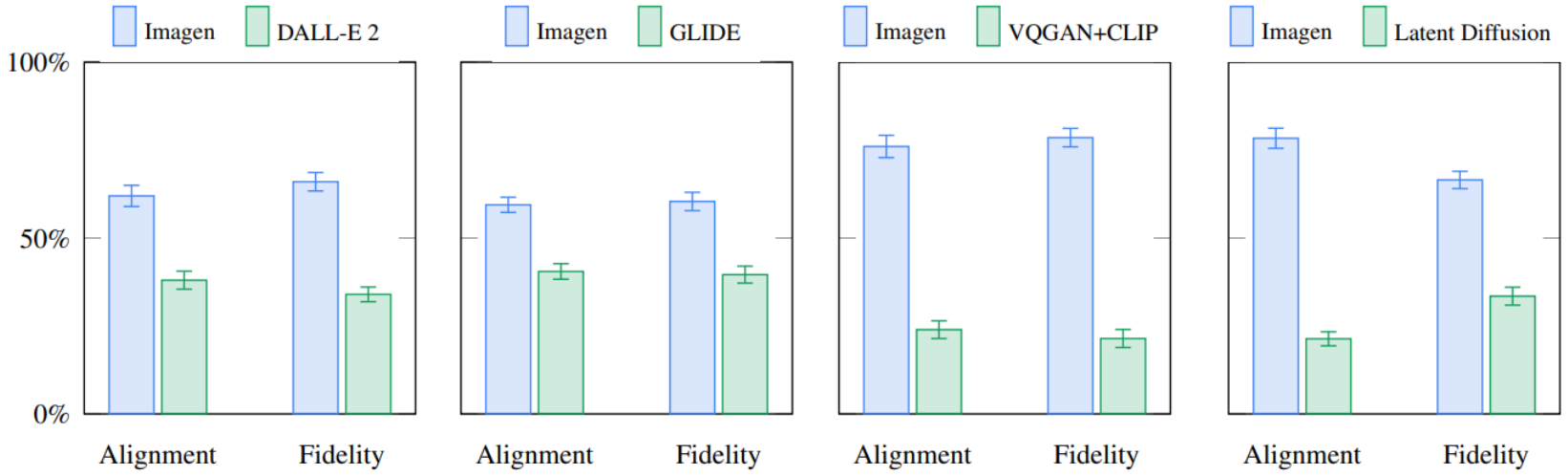
Model	FID-30K	Zero-shot FID-30K
AttnGAN [76]	35.49	
DM-GAN [83]	32.64	
DF-GAN [69]	21.42	
DM-GAN + CL [78]	20.79	
XMC-GAN [81]	9.33	
LAFITE [82]	8.12	
Make-A-Scene [22]	7.55	
DALL-E [53]		17.89
LAFITE [82]		26.94
GLIDE [41]		12.24
DALL-E 2 [54]		10.39
Imagen (Our Work)		**7.27**

Fréchet Inception Distance (FID)

$$\text{FID} = \|\mu_X - \mu_Y\|^2 + \text{Tr}\left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}\right)$$

Real and fake embeddings are two multivariate normal distributions

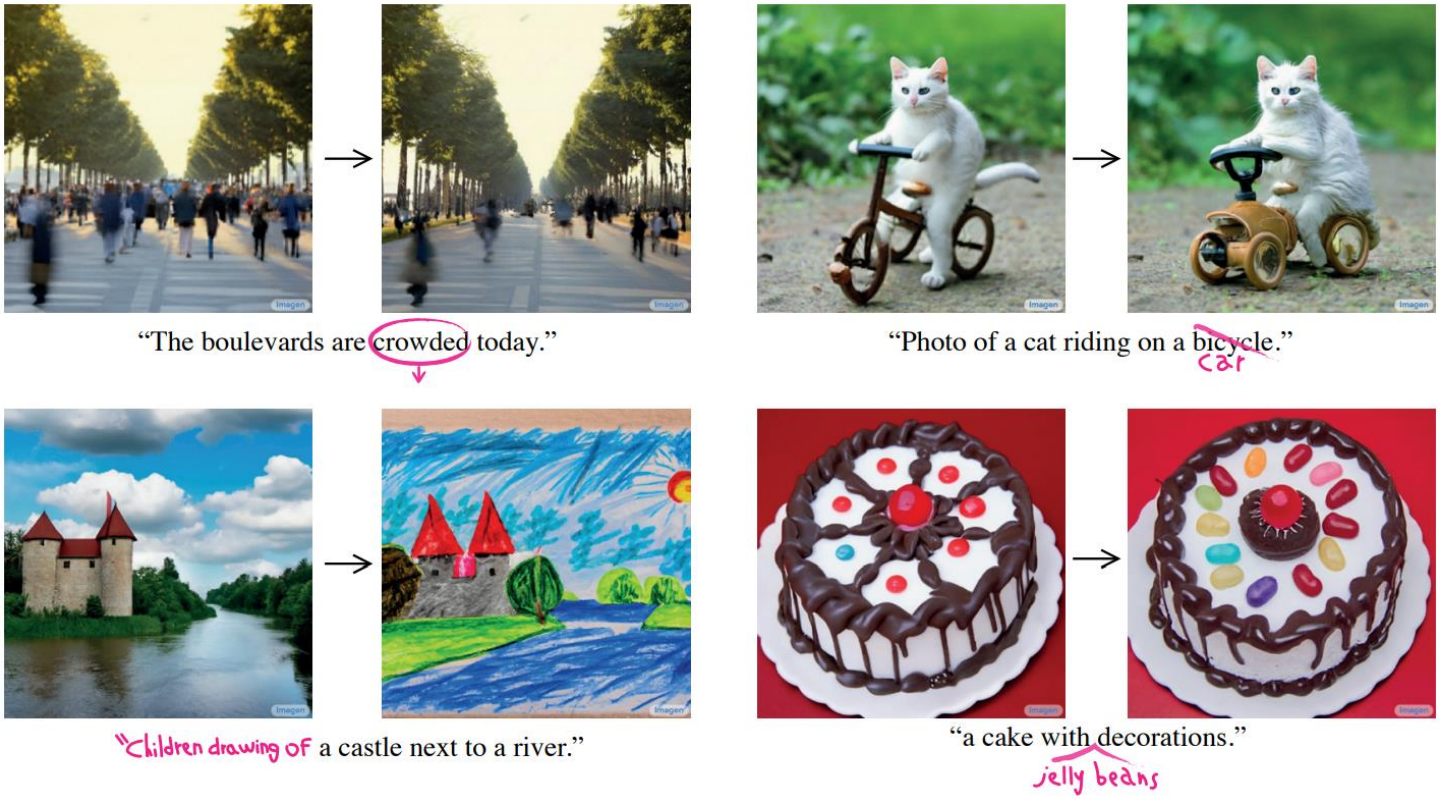
- DrawBench에 대한 fidelity 및 alignment 인간 평가



Prompt-to-prompt image editing with cross attention control

Abstract

- Text-based diffusion model은 image editing하기 어려움
  - Text prompt에 약간의 수정을 가해도 output이 완전히 달라짐
- Prompt-to-Prompt editing framework를 제시
  - Image의 공간적 위치와 prompt 각 단어 간의 relation을 cross-attention으로 조절



Introduction

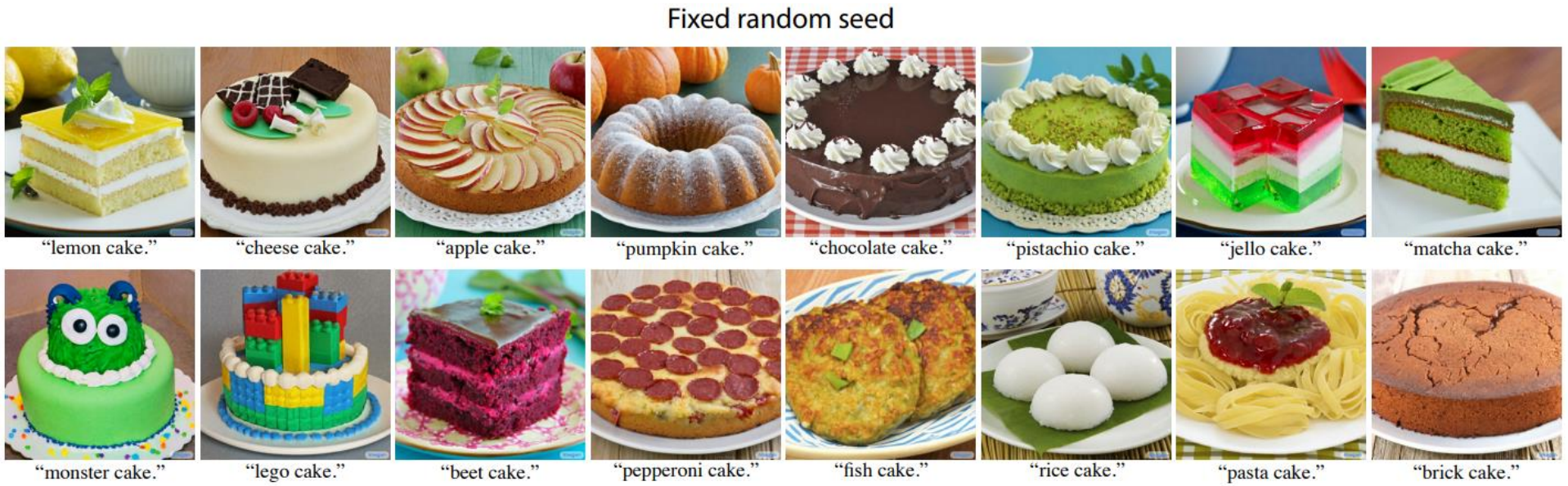
- Imagen, DALL-E 2 와 같은 Large-scale Language Image model은 높은 image generation 성능을 보임
  - 주어진 image의 특정 semantic region에 대한 control에는 취약함
- 기존에는 Image editing을 위해 masking 후 inpainting하는 방법 활용
  - Masking 절차가 번거로움
  - 빠르고 직관적인 text 기반 editing을 방해
  - Image 내의 구조적인 정보를 제거하는 문제 발생

Introduction

- Prompt-to-prompt 조작을 통해서 직관적이고 강력한 textual editing 방법을 소개
  - Pixel ↔ Text prompt token 사이의 cross-attention maps를 사용
- 본 논문에서는 3가지 editing methods를 제시
  1. 생성된 image에서 특정 단어에 대한 **semantic effect를 확대하거나 축소**
  2. **prompt의 single token**'s value를 바꿈
  3. **prompt에 새로운 단어를 추가**하여 새로운 attention flow를 제공

Method

- Text-guided diffusion model에서 random seed만을 고정하고 수정된 text prompt로 re-generate 시도
  - “lemon cake” prompt에서 lemon만 다른 재료들로 수정하려 했지만 실패
  - 이미지의 구조와 모습이 유지되지 않음



Method

- 해당 실험을 통한 key observation
  - 생성된 image의 구조와 모습에는 random seed 뿐만 아니라, **pixel과 text embedding 간의 interaction** 또한 영향을 줌

Fixed attention maps and random seed

Cross-Attention in Text-Conditioned Diffusion Model

- Imagen 논문의 model을 backbone으로 사용
- 64x64-Resolution Text-to-Image Diffusion model에만 적용
  - Image의 구성과 모습이 64x64 resolution의 diffusion process에서 결정됨
  - Imagen의 SR model은 그대로 사용

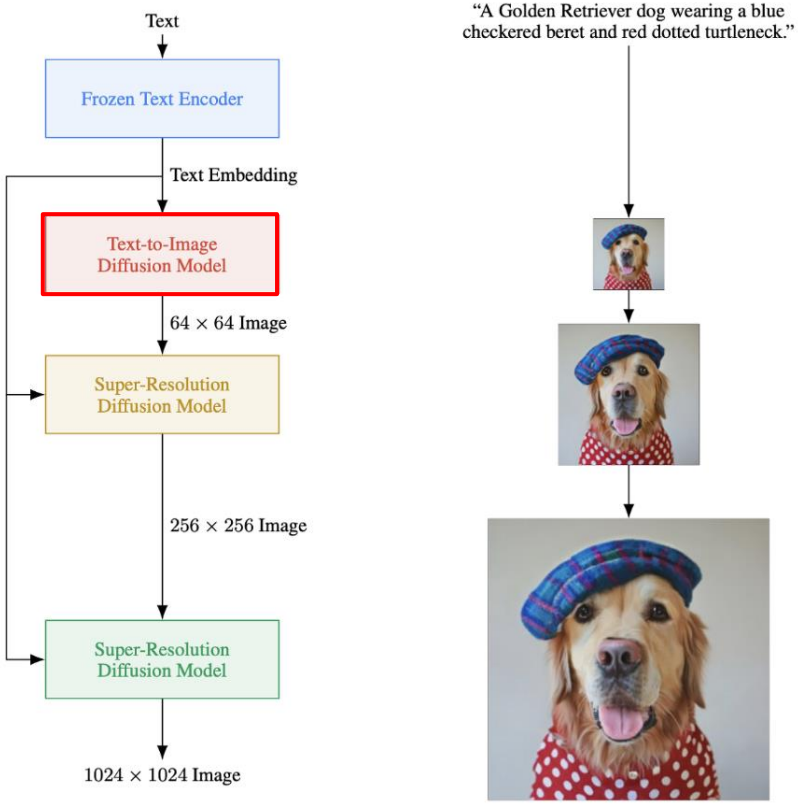
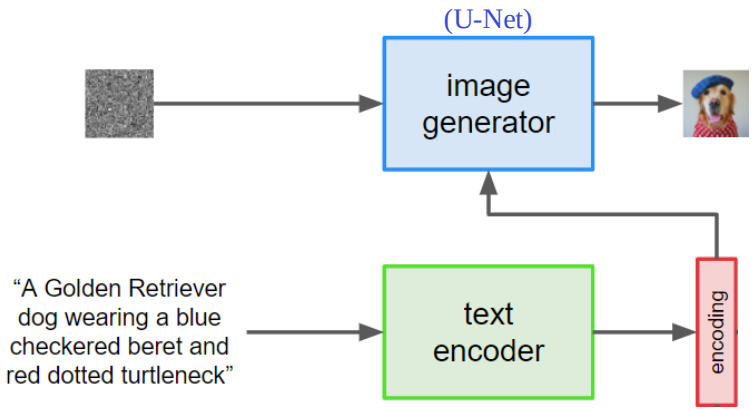
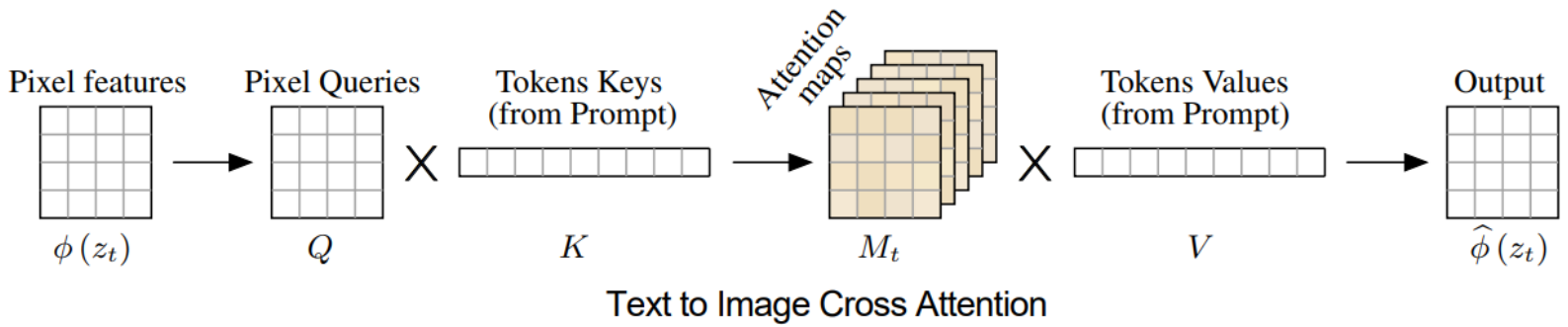


Imagen architecture

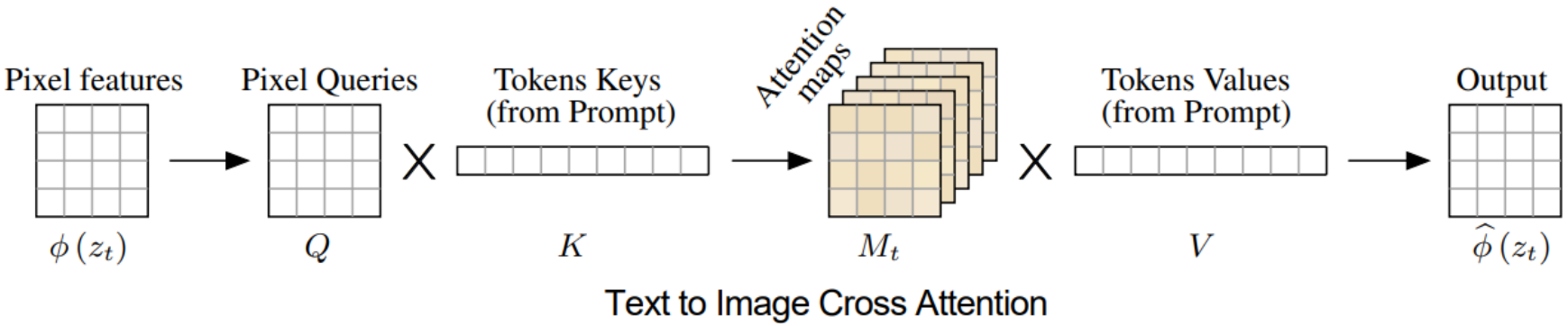
Cross-Attention in Text-Conditioned Diffusion Model

- Diffusion model은 U-Net을 사용해 noisy image z_t와 text embedding $\psi(P)$으로부터 noise ϵ를 predict
  - 마지막 step에서 image $I = z_0$를 생성
  - Language와 Visual 간의 interaction은 이 noise prediction 과정에서 일어남
  - 이 과정에서 textual tokens에 대한 spatial attention map을 만들어 내는 cross-attention layer를 활용





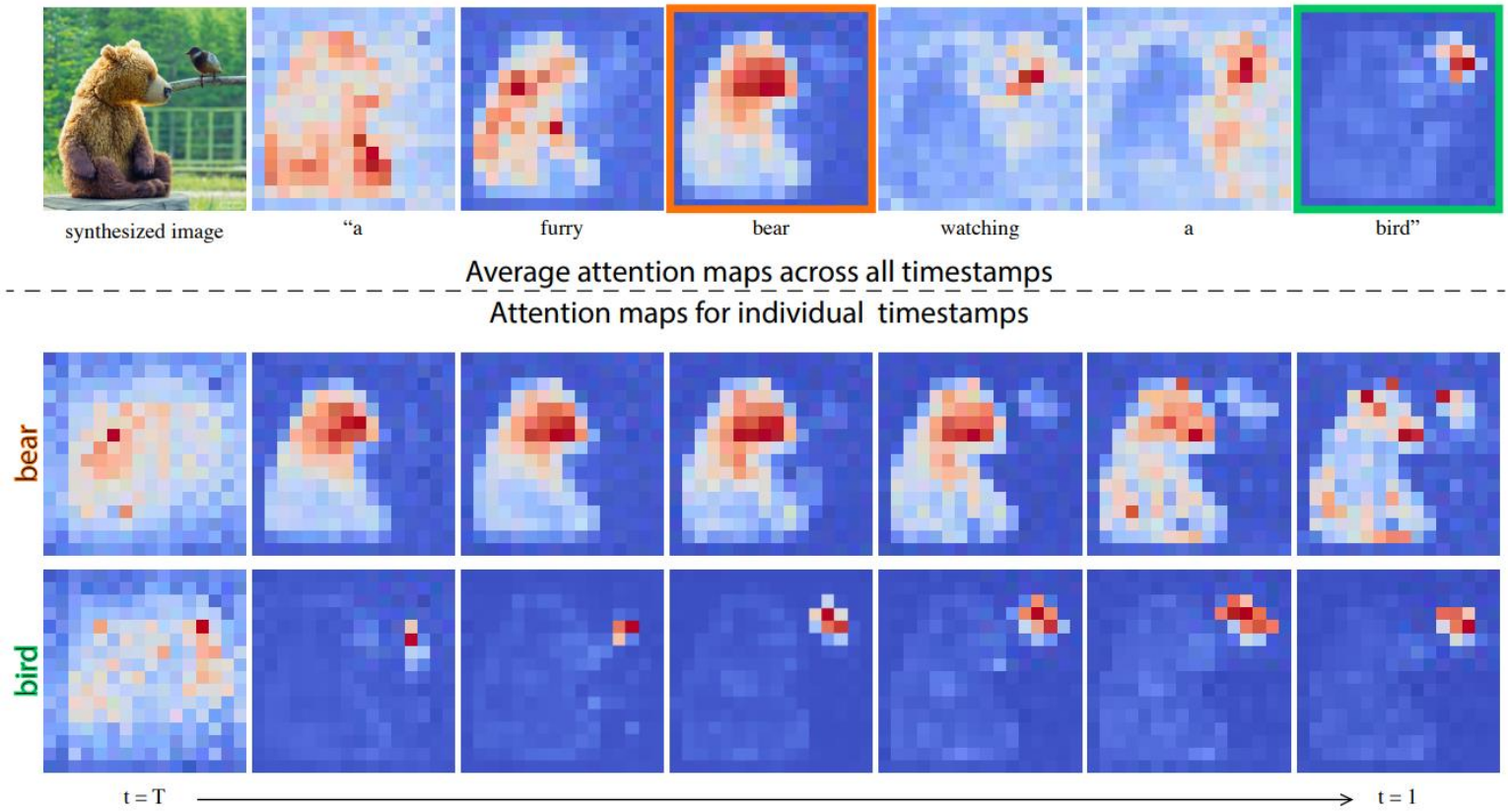
Cross-Attention in Text-Conditioned Diffusion Model



- Noisy image $\phi(z_t)$의 spatial feature → Query Matrix $Q = l_Q(\phi(z_t))$
- Text Embedding $\psi(P)$ → Key Matrix $K = l_K(\psi(P))$, Value Matrix $V = l_V(\psi(P))$
- Attention map $M = softmax(\frac{QK^T}{\sqrt{d}})$
  - Cell M_{ij}는 i번째 pixel과 j번째 token에 대한 weight를 정의
  - d는 key와 query의 latent projection dimension
- 최종 cross attention output $\hat{\phi}(z_t) = MV$로 정의
  - 이것을 pixel feature $\phi(z_t)$를 업데이트하는데 사용

Controlling the Cross-Attention

- 각 token에 대한 attention maps와 timestamp에 따른 attention map의 변화를 시각화
  - Token "bear"의 경우 pixel들이 곰에 더 연결되어 있음
  - Token "bird"의 경우 pixel들이 새에 더 연결되어 있음
- 이미지의 구조가 diffusion process 초반에 결정되는 것을 확인 가능



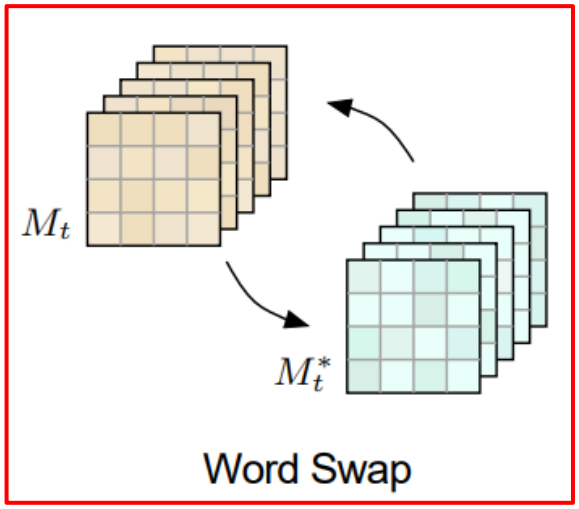
Controlling the Cross-Attention

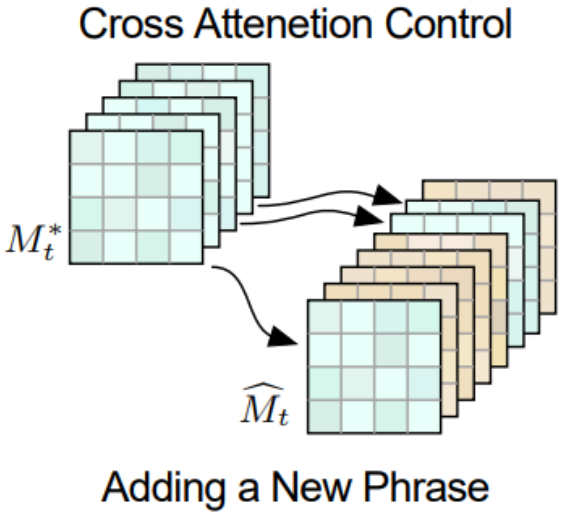
Algorithm 1: Prompt-to-Prompt image editing

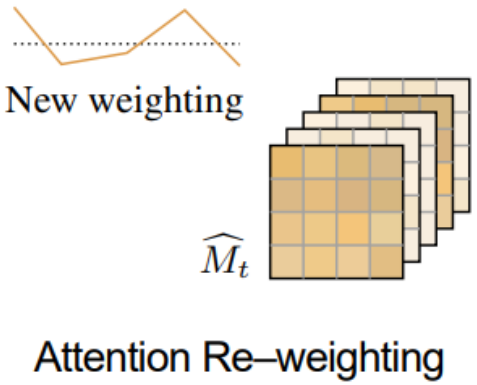
Input: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$, and a random seed s.
Output: A source image x_{src} and an edited image x_{dst}.
$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed s;
$z_T^* \leftarrow z_T$;
for $t = T, T-1, \ldots, 1$ **do**
  $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$;
  $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$;
  $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$;
  $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$;
end
Return (z_0, z_0^*)

- $DM(z_t, P, t, s)$: diffusion process의 single step t
  - Output: noisy image z_{t-1} + attention map M_t
- $DM(z_t, P, t, s)\{\text{M} \leftarrow \widehat{M}\}$: attention map M을 $\widehat{M}$으로 바꾼 diffusion step
- M_t^*: edited prompt P*로부터 생성된 attention map
- $Edit(M_t, M_t^*, t)$: general edit function, 본 논문에서는 세 가지 종류의 editing을 제공
  - Word Swap, Adding a new phrase, Attention Re-Weighting
  - 원본 image의 attention map M_t와, editing image의 attention map M_t^*를 모두 사용하여 새로운 map $\widehat{M}$를 생성
    - 원본 image의 구조를 적절히 유지하면서 editing을 진행함을 의미

Controlling the Cross-Attention





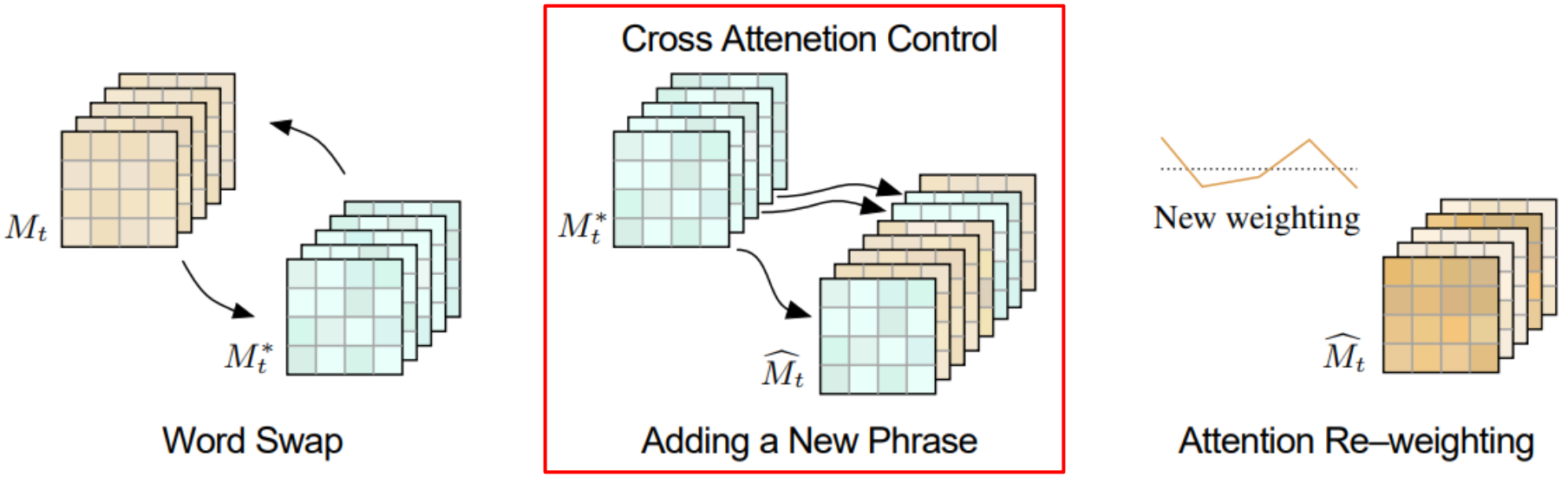


- **Word Swap**: original prompt의 token을 다른 것으로 교환
  - Ex. P = "a big red **bicycle**", P* = "a big red **car**"

$$Edit(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise.} \end{cases}$$

  - τ: attention map 교체가 적용되는 step을 결정하는 timestep 파라미터

Controlling the Cross-Attention

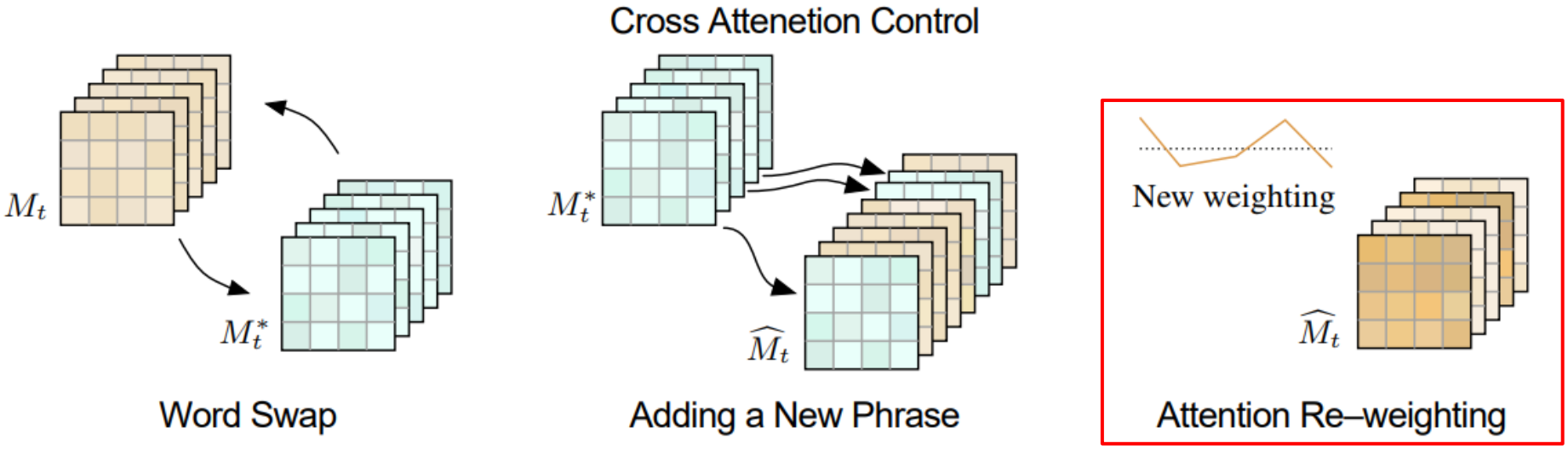


- **Adding a New Phrase**: prompt에 새로운 token을 추가
  - Ex) P = "a castle next to a river.", P* = "**children drawing of** a castle next to a river."

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = None \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

- 일반적인 detail을 유지하기 위해서, original prompt의 attention maps에다가 새로운 token에 해당되는 attention map들만 주입
  - Alignment function A(j): P*의 j번째 token에 매칭되는 P의 token index 또는 None을 반환
    - A(j)가 None일 경우, P*의 j번째 token이 새로운 token임을 의미

Controlling the Cross-Attention



- **Attention Re-weighting**: 각 token이 결과 image에 영향을 미치는 정도를 강화/약화
  - Ex) P = "a **fluffy** red ball", 공을 더 fluffy 또는 덜 fluffy 하도록 변경
  - $c \in [-2, 2]$: 강화/약화하도록 지정된 특정 token(j*)의 weighting parameter

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$

Experiments: Word Swap

$$Edit(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise.} \end{cases}$$

Source image and prompt:

"photo of a cat riding on a bicycle."

bicycle → motorcycle

bicycle → car

bicycle → airplane

bicycle → train

W.O. attention injection → Full attention injection

Experiments: Adding a New Phrase

"A photo of a butterfly on a flower."

source image

"...on a spikey flower."

"...wither flower."

"...origami flower."

"...flower made out from candies."

"...wooden flower."

Experiments: Adding a New Phrase

Experiments: Attention Re-weighting

“A tiger is sleeping(↑) in a field.”

“A smiling(↑) teddy bear.”

“Photo of a cubic(↓) sushi.”

Experiments: Attention Re-weighting

“The modern(↓) city.”

“My colorful(↓) bedroom.”

“Photo of a field of poppies at night(↓).”

Any Questions?