Image Editing with Diffusion Model

2023. 08. 25.

이진우

DMQA Open Seminar

발표자 소개

❖ 이진우 (Jin Woo Lee)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab (김성범 교수님)
- 석사과정 (2023.03 ~ Present)

❖ Research Interest

- Deep Generative Models
- Diffusion Models

❖ Contact

- dlwlsdn0225@korea.ac.kr

목차

Introduction
Background

❖ 이미지 생성 모델

- 상상을 현실로 그려낼 수 있는 흥미로운 분야
  - 주로 사용자가 **prompt**를 주면 원하는 이미지를 생성하는 방식으로 작동
- 사용자 의도를 반영한 고퀄리티 이미지 생성

Teddy bears swimming at the Olympics 400m Butter-fly event.

A cute corgi lives in a house made out of sushi.

Imagen

https://imagen.research.google/

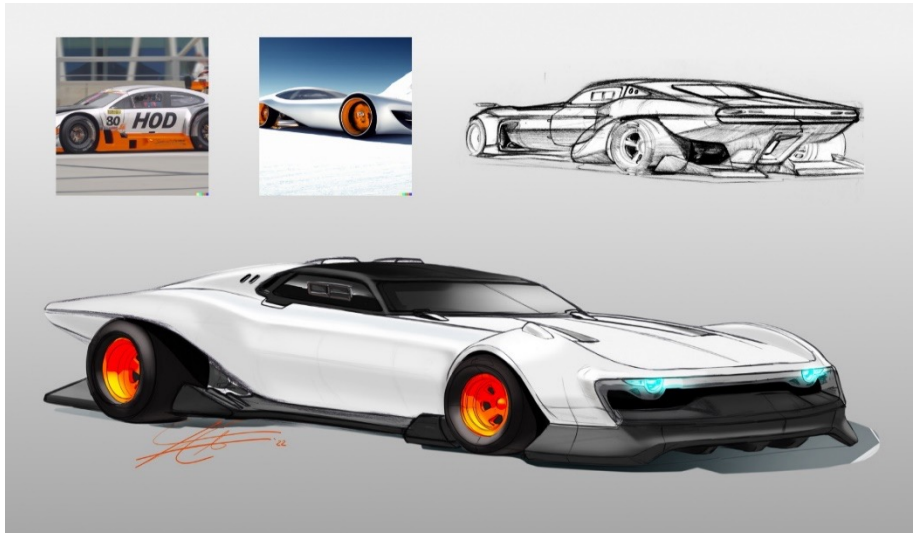
Midjourney - '스페이스 오페라 극장

https://www.midjourney.com/home/

Introduction
Background

❖ 이미지 생성 모델 활용

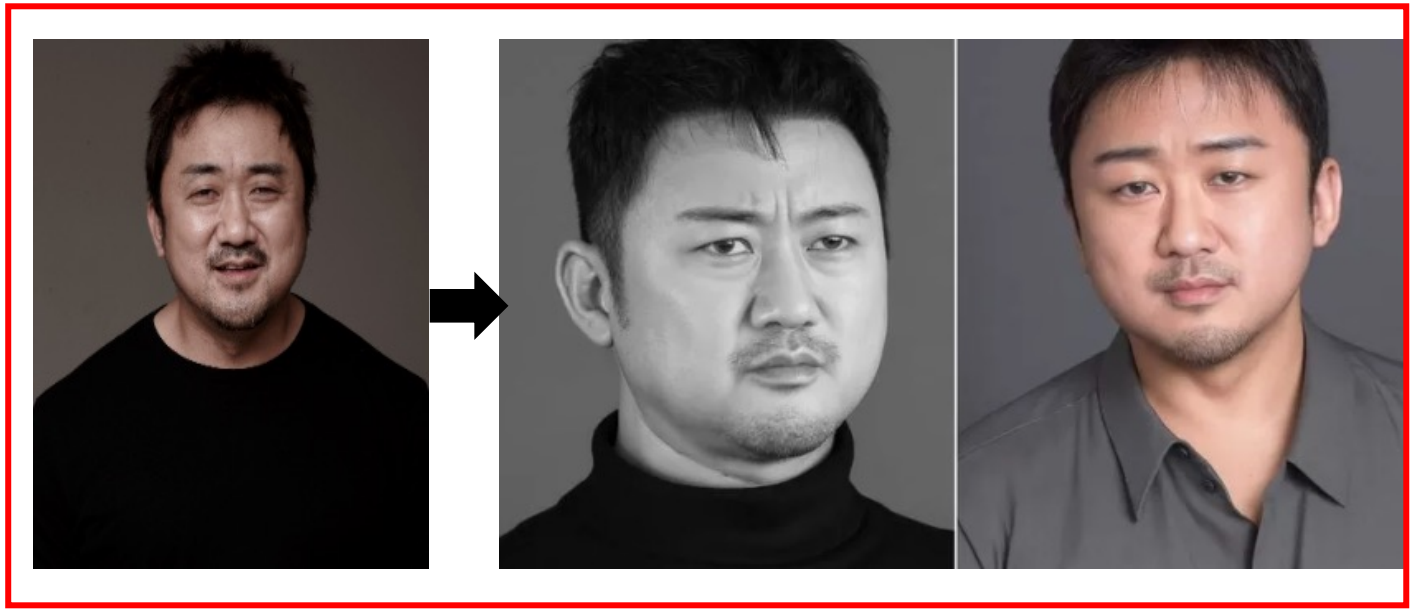
- 게임, 디지털아트, 뷰티, 마케팅등과 같이 다양한 분야에서 활용
  - **마텔** : 장난감 자동차 핫휠 디자인
  - **스노우** : AI filter



마텔 자동차 디자인

https://www.insight.co.kr/news/443295

이미지 편집



스노우 AI Filter

https://newsmoum.tistory.com/296
https://venturebeat.com/ai/dall-e-2-coming-to-microsofts-azure-ai-by-invitation/

Introduction
Background

❖ Image Editing

- 이미지 생성 분야에서 우수한 성능을 보인 **디퓨젼 모델 기반 editing 기법들이 많이 등장**
- 원본 이미지의 형태를 어느정도 유지하면서 , 객체, 스타일 등을 수정
- 주로 텍스트 입력 (prompt) 을 주고 이미지를 편집

" 의자에 앉아있는 모자 쓴 고양이"

원본 이미지

" 의자에 앉아있는 모자 쓴 코끼리"

생성된 이미지

Prompt-to-Prompt

" "

원본 이미지

" 금색 스파이더맨"

생성된 이미지

Plug-and-play

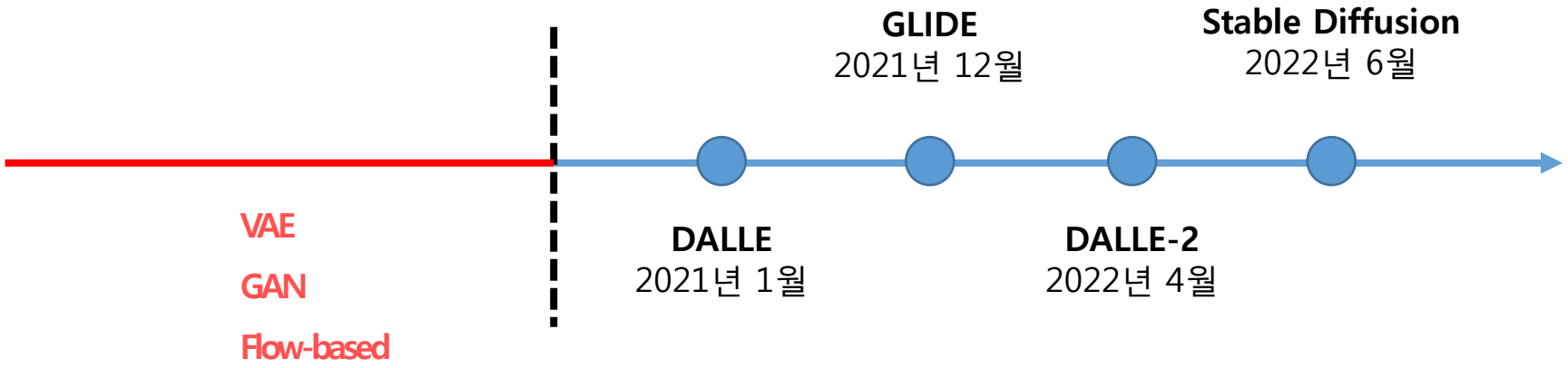
Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Introduction
Background

❖ 디퓨젼 모델

- 현재는 기존 생성 모델들에 비해 높은 퀄리티를 선보인 **디퓨젼 모델**이 각광
  - **GAN** :: mode collapse 와 학습 불안정
  - **VAE** : 낮은 품질의 생성 결과
  - **Flow-based** : 역함수가 존재하지 않으면 생성이 불가
- 2022년 6월 **Stable Diffusion**이 코드와 학습된 모델을 공개하면서 연구가 더욱 활발해짐



https://deepsense.ai/the-recent-rise-of-diffusion-based-models/

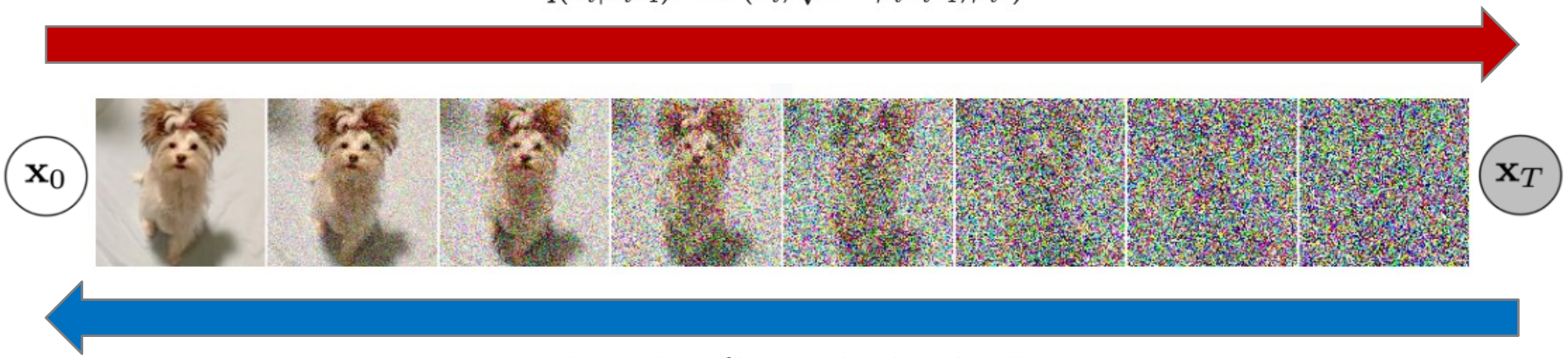
Diffusion Model
DDPM

❖ Denoising Diffusion Probabilistic Models (DDPM)

- **Forward Process** (고정)
  - 이미지 (x_0) 가 완전한 Gaussian Noise (x_T) 가 될 때까지 Gaussian Noise 를 점진적으로 추가하는 Markov Process
- **Reverse Process (학습)** :
  - Gaussian Noise (x_T) 에서 점진적으로 Gaussian Noise 를 제거하여 이미지 (x_0) 를 복원하는 Markov Process.

➢ **UNET을 학습하여** Reverse process에서 t 시점에서 제거할 **noise** $\epsilon_\theta(x_t, t)$를 예측

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$



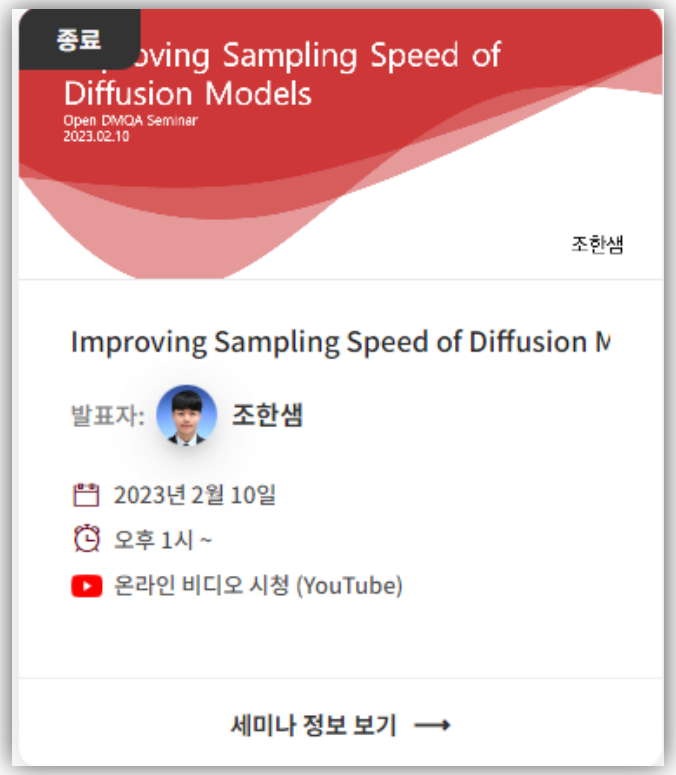
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

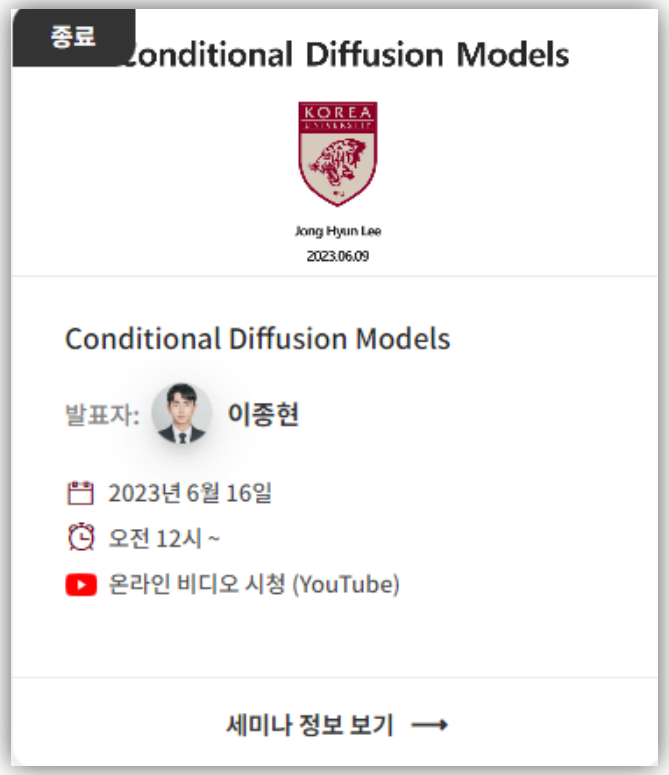
Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.

Diffusion Model

Latent Diffusion Model

- **Improving Sampling Speed of Diffusion Model :** 디퓨젼 모델의 기초인 DDPM과 DDIM과 이미지 생성 속도를 높이는 방법론 소개
- **Conditional Diffusion Model :** 디퓨젼 모델에 컨디션을 부여하는 Classifier Guidance, Classifier Free Guidance 그리고 Latent Diffusion 소개

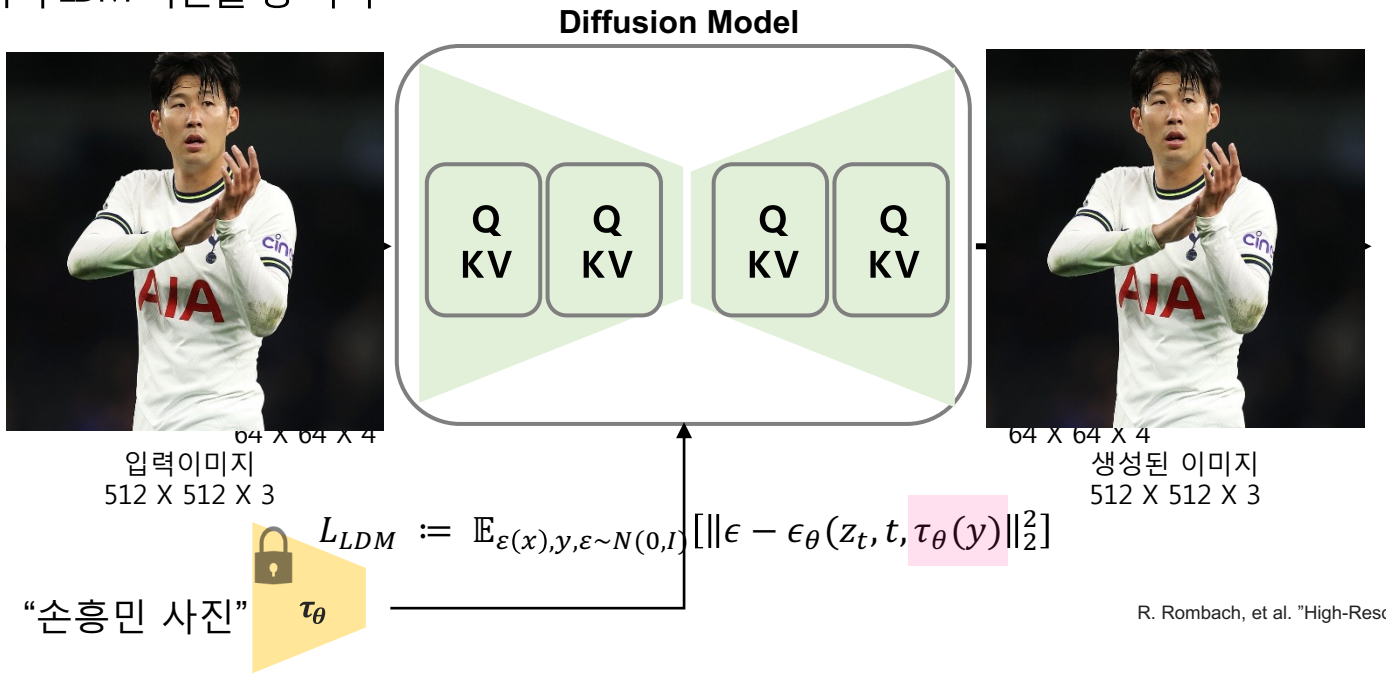




Diffusion Model
Latent Diffusion Model

❖ **"High-Resolution Image Synthesis with Latent Diffusion Model (LDM)"**

- 많은 디퓨젼 기반 editing 모델들이 LDM을 사용
- Pixel space 상에서 작동하는 디퓨젼 모델은 연산량이 높음 → 오토인코더(**VQ-VAE**)를 사용하여 **latent space**에서 작동, 연산량을 줄임
- Image inpainting, class-conditioned image synthesis 그리고 **text-to-image synthesis task** 수행
- Stable Diffusion 역시 LDM 버전들 중 하나



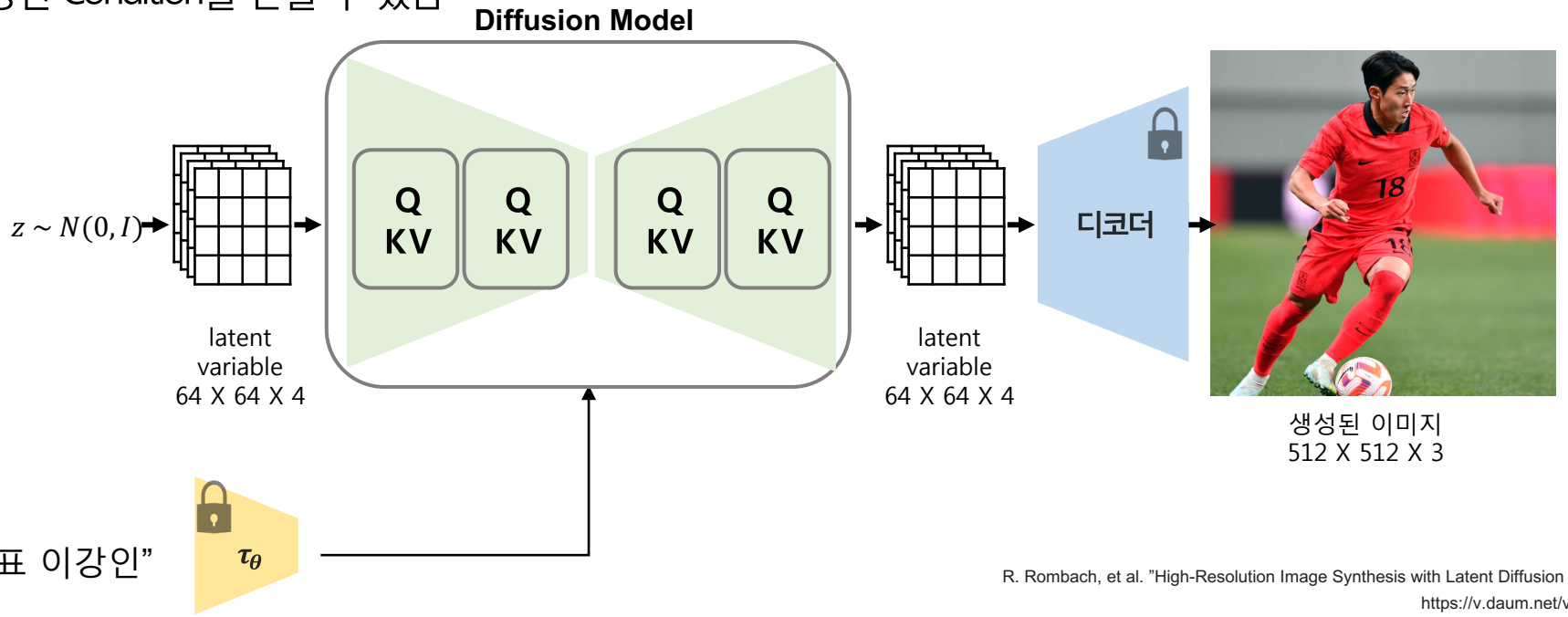
R. Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models" CVPR (2022)

https://www.spotvnews.co.kr/news/articleView.html?idxno=558962

Diffusion Model
Latent Diffusion Model

❖ Sampling

- **Text to Image Generation** : LDM은 Gaussian Noise와 Condition을 입력받아 원하는 이미지를 생성
  - **CLIP** : text 와 image의 joint embeddding
  - U-Net에 **Cross-Attention**을 추가하여 Condition 반영
- Text 이외에도 다양한 Condition을 받을 수 있음

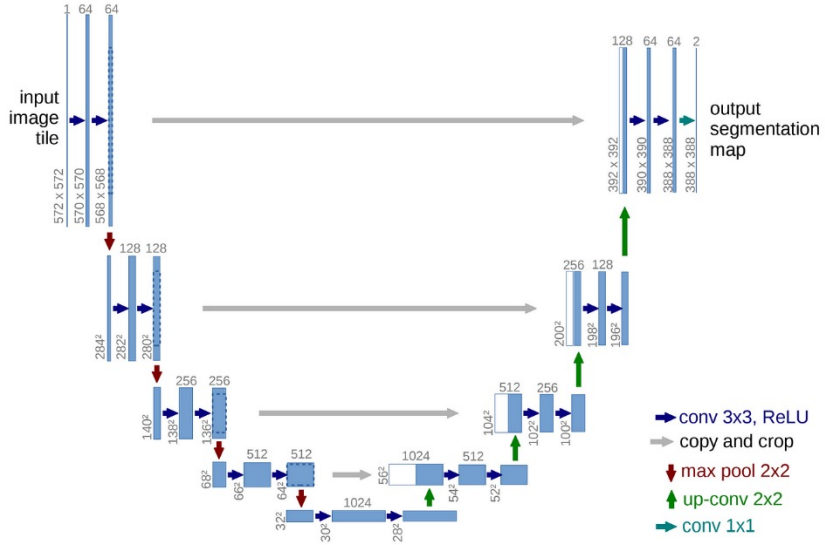


R. Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models" CVPR (2022)
https://v.daum.net/v/20230515160516680

Diffusion Model
Latent Diffusion Model

❖ Inside Diffusion Model

- 모델 구조 : U-Net (Denoising U-Net)
  - Transformer block과 Residual block으로 구성
  - Transformer block : Self Attention , Cross attention
  - Residual block : convolution 연산

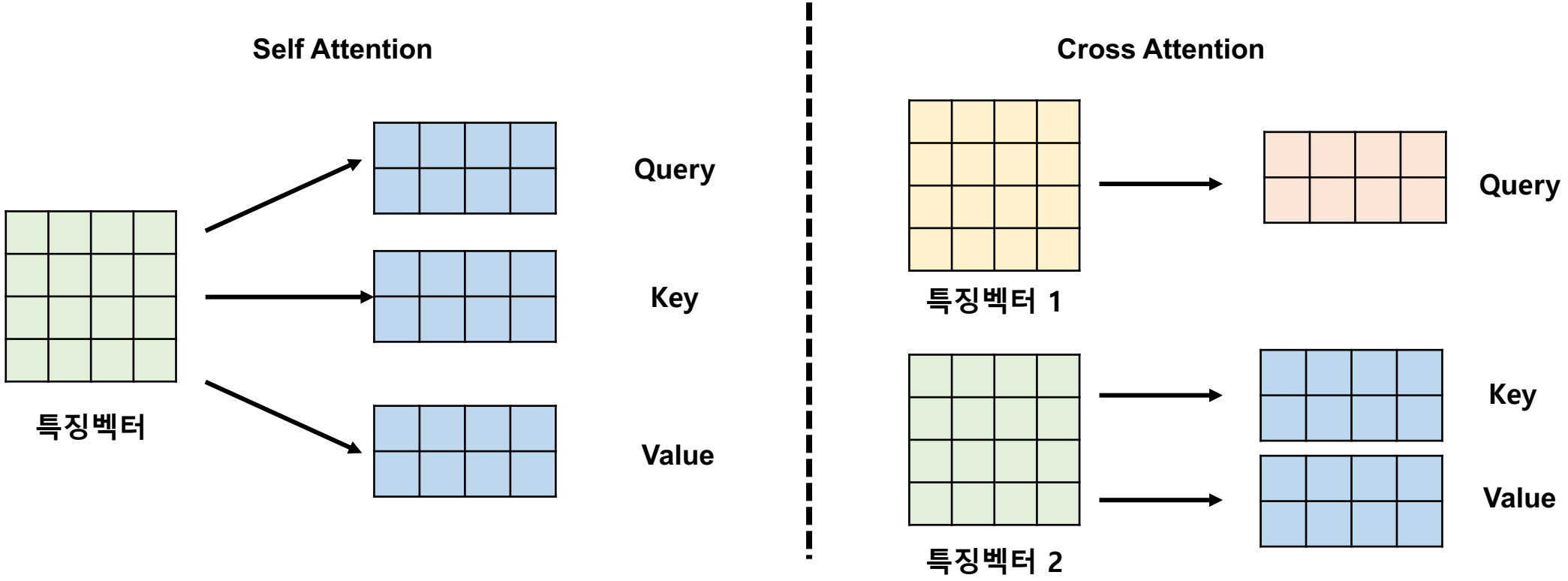


Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

Diffusion Model
Conditioning

❖ **Self Attention VS Cross Attention**

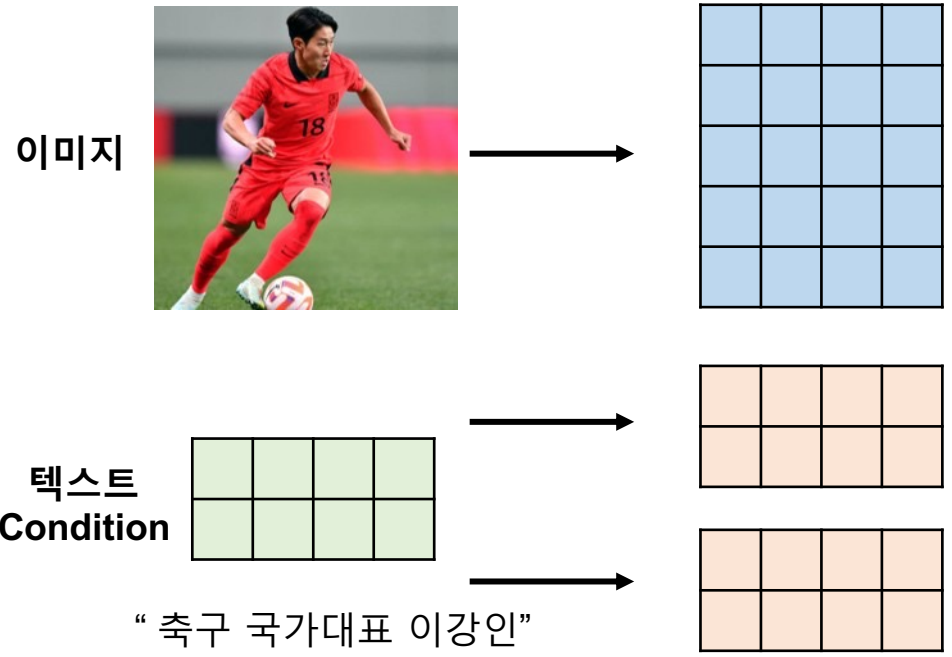
- **Attention** : 전체를 보고 필요한 것에만 집중하자
- **Self Attention** : 동일한 특징벡터로부터 Query, Key, Value가 나옴 (**이미지 특징들간의** 관계)
- **Cross Attention** : Query와 Key, Value를 추출한 특징벡터가 다름 (**이미지와 Condition**의 관계)



Diffusion Model
Conditioning

❖ Cross Attention in LDM

- 이미지와 Condition 의 관계를 반영
- **Query :** 이미지로부터 추출
- **Key, Value** : Condition에서 추출



$$softmax(\frac{QK^T}{\sqrt{d_k}})$$

Query

Key

Value

Query(이미지와 Key(Condition) 의 관계

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

각 단어가 이미지의 어느 부분을 참조할지

Diffusion Model

Conditioning

디퓨젼 모델은 내부적으로 복잡한 구조를 갖는데,

Residual block, self attention 그리고 cross attention의

feature 들을 조작한다면 어떨까?

Image Editing with Diffusion Model

Image Editing

- 사용자가 원하는 형태로 이미지를 편집하면서도, 원본 이미지의 형태를 유지하는게 중요
- 주로 사용자가 Text로 원하는 바를 입력하는 **Prompt** 기반 편집
  - 예 : "노란색 털을 가진 고양이 " → " 노란색 털을 가진 강아지
- **사전학습된 LDM을 사용하면서, denoising 과정에서 attention을 조정**

" 의자에 앉아있는 모자 쓴 고양이"

원본 이미지

" 의자에 앉아있는 모자 쓴 코끼리"

생성된 이미지

Prompt-to-Prompt

" "

원본 이미지

" 금색 스파이더맨"

생성된 이미지

Plug-and-play

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

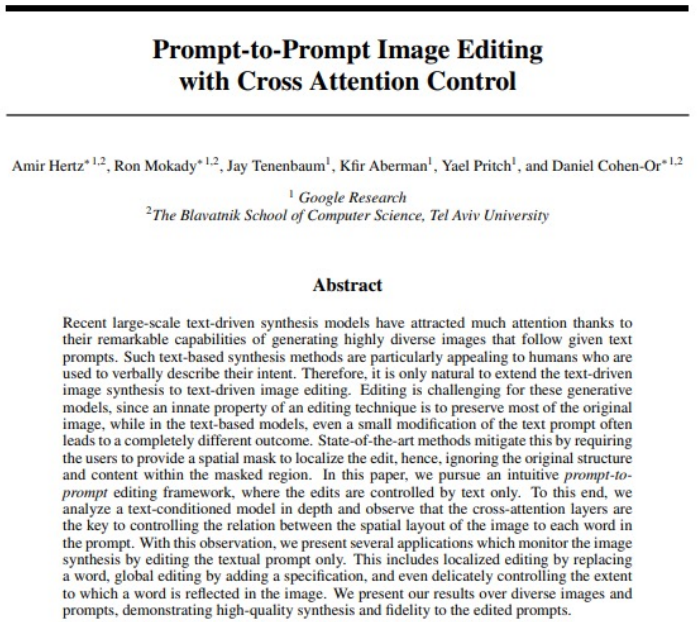
Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Prompt-to-Prompt Image Editing with Cross Attention Control

- **Prompt to Prompt Editing** : 텍스트 prompt 내 일부분을 수정하여 이미지의 편집을 수행
- 원본 이미지의 **attention map** 주입하여 수정하고자 하는 이미지의 attention map을 조절
- Finetuning 과 학습할 필요가 없는 editing 기법



Prompt-to-Prompt Image Editing with Cross Attention Control

Amir Hertz*[1,2], Ron Mokady*[1,2], Jay Tenenbaum[1], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or*[1,2]

[1] *Google Research*
[2] *The Blavatnik School of Computer Science, Tel Aviv University*

Abstract

Recent large-scale text-driven synthesis models have attracted much attention thanks to their remarkable capabilities of generating highly diverse images that follow given text prompts. Such text-based synthesis methods are particularly appealing to humans who are used to verbally describe their intent. Therefore, it is only natural to extend the text-driven image synthesis to text-driven image editing. Editing is challenging for these generative models, since an innate property of an editing technique is to preserve most of the original image, while in the text-based models, even a small modification of the text prompt often leads to a completely different outcome. State-of-the-art methods mitigate this by requiring the users to provide a spatial mask to localize the edit, hence, ignoring the original structure and content within the masked region. In this paper, we pursue an intuitive *prompt-to-prompt* editing framework, where the edits are controlled by text only. To this end, we analyze a text-conditioned model in depth and observe that the cross-attention layers are the key to controlling the relation between the spatial layout of the image to each word in the prompt. With this observation, we present several applications which monitor the image synthesis by editing the textual prompt only. This includes localized editing by replacing a word, global editing by adding a specification, and even delicately controlling the extent to which a word is reflected in the image. We present our results over diverse images and prompts, demonstrating high-quality synthesis and fidelity to the edited prompts.

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

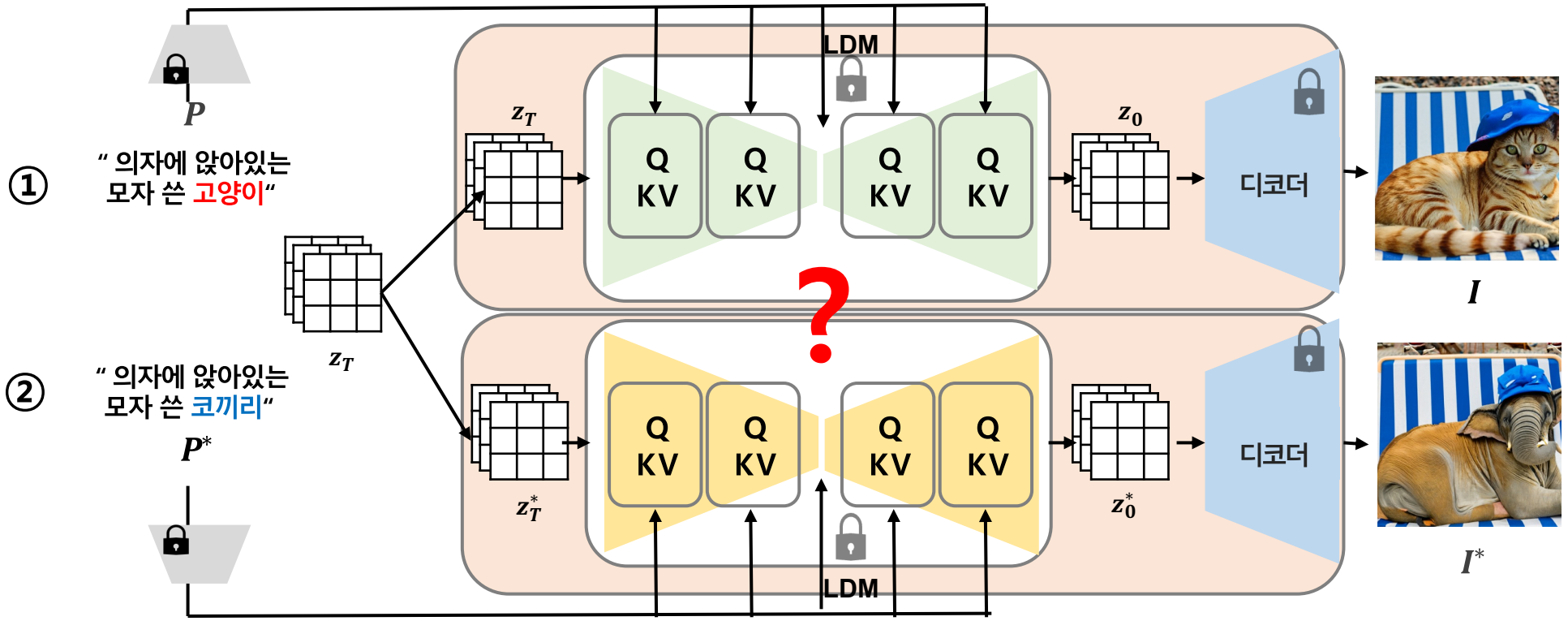
Image Editing with Diffusion Model

Prompt-to-Prompt

❖ 문제상황

① 원본 Prompt P 를 바탕으로 이미지 I 를 생성하는 디퓨전 모델 (**원본 이미지 생성과정**)

② 이미지 I 가 타겟 Prompt P^* 를 반영하는 이미지 I^* 로 edit 하는 디퓨전 모델 (**원본 이미지의 편집 과정**)



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt

❖ 문제상황

- 기존 연구들은 사용자가 직접 mask를 제공해야하는 번거로움 → **기존 연구들과 달리 mask 없이 edit 하는 방식**
- 원본 이미지의 모습은 유지하는 editing 방법

① 디퓨젼 모델은 Gaussian Noise에서 시작하기 때문에, seed 값에 영향을 받음

② 디퓨젼 모델은 생성과정에서 이미지와 condition의 상호작용

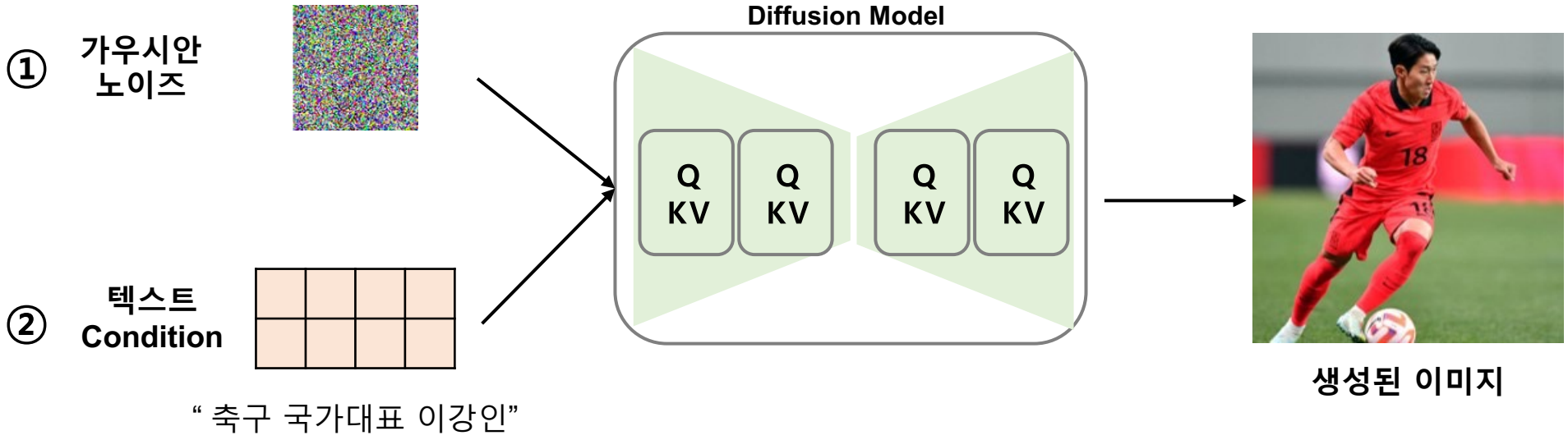
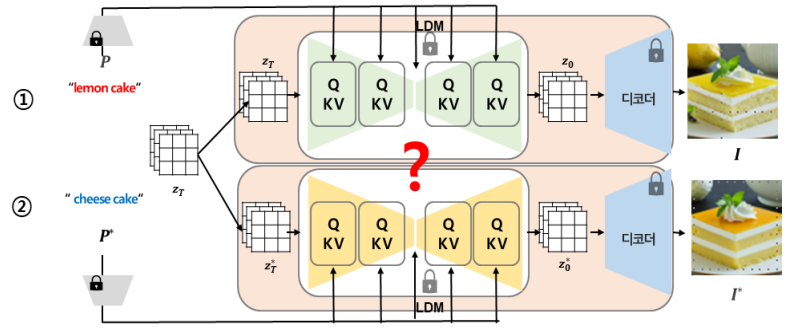


Image Editing with Diffusion Model

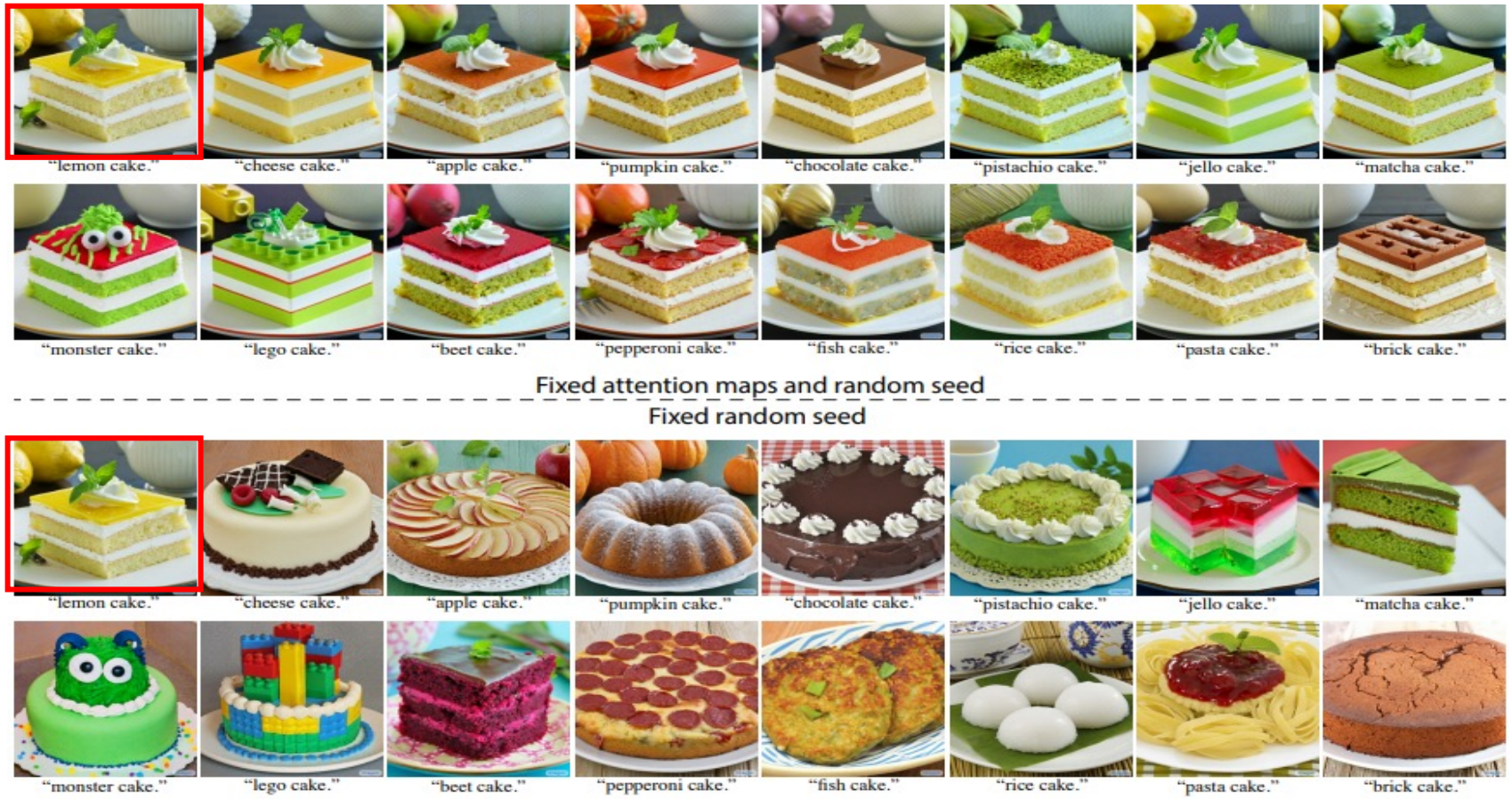
Prompt-to-Prompt



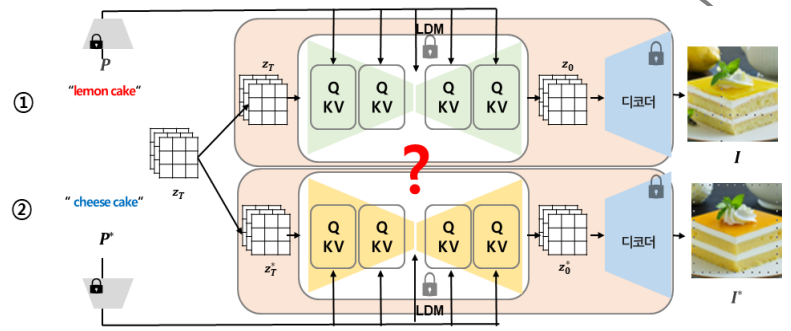
❖ Seed와 Condition에 변형을 주어 이미지를 생성하는 실험 (lemon cake이 원본

고정된
Attention map
Random Seed

고정된
Random Seed



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022



이미지의 형태와 모습이 단순히 random seed에만 의존하는게 아니라,

Cross attention에도 영향을 받음

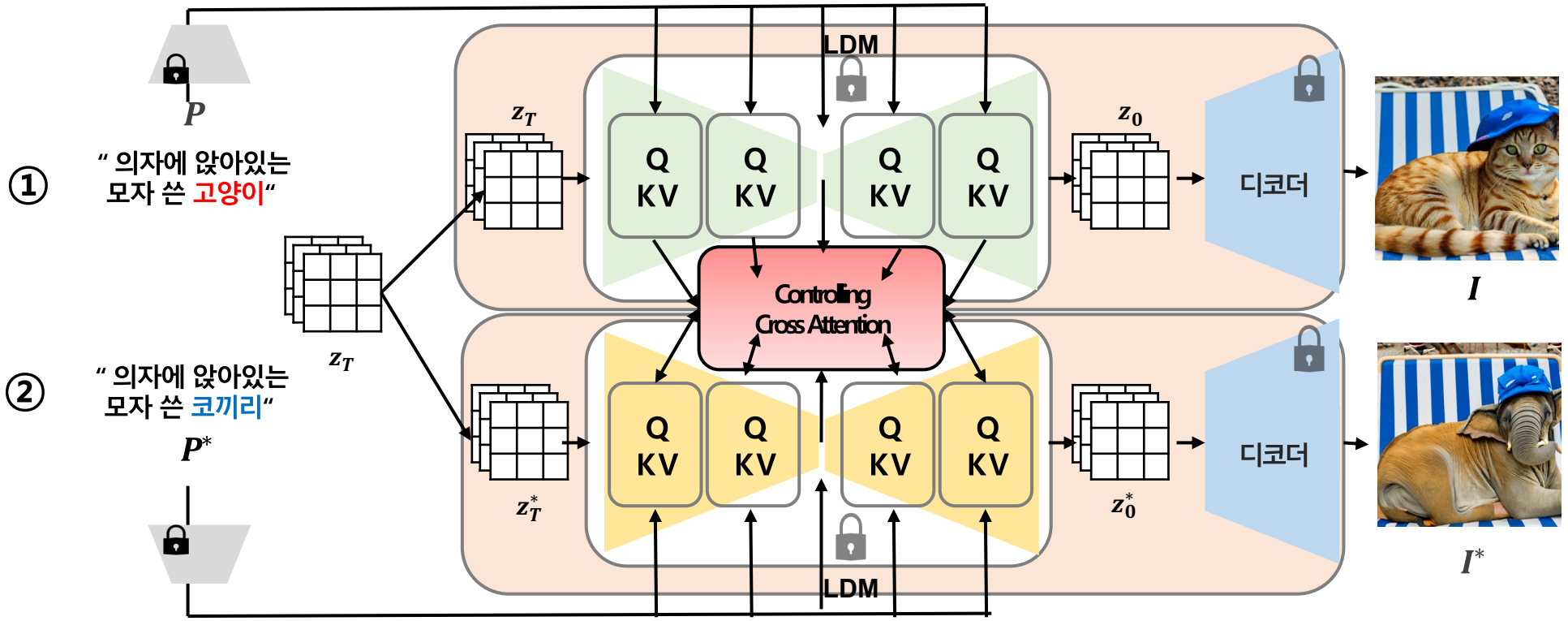
→ 타겟 프롬프트를 입력 받아 이미지를 생성해내는 과정에서 원본 cross attention map을 주입하면 원본 이미지의 형태를 유지하면서 edit 가능

Image Editing with Diffusion Model

Prompt-to-Prompt

❖ 문제상황

- **Prompt to Prompt Editing** : 텍스트 prompt 내 일부분을 수정하여 이미지의 편집을 수행
- 원본 Prompt P 를 바탕으로 하는 디퓨젼 과정과 타겟 Prompt P^*의 디퓨젼 과정에서 Cross Attention을 조작하자



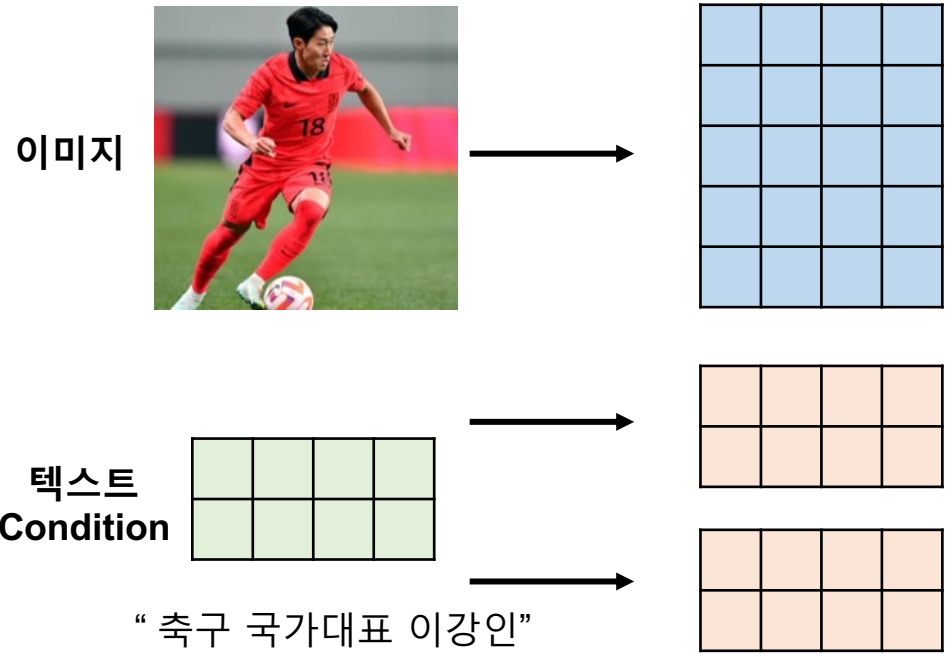
Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Controling Cross Attention

- 이미지와 Condition 의 관계를 반영
- **Query :** 이미지로부터 추출
- **Key, Value** : Condition에서 추출



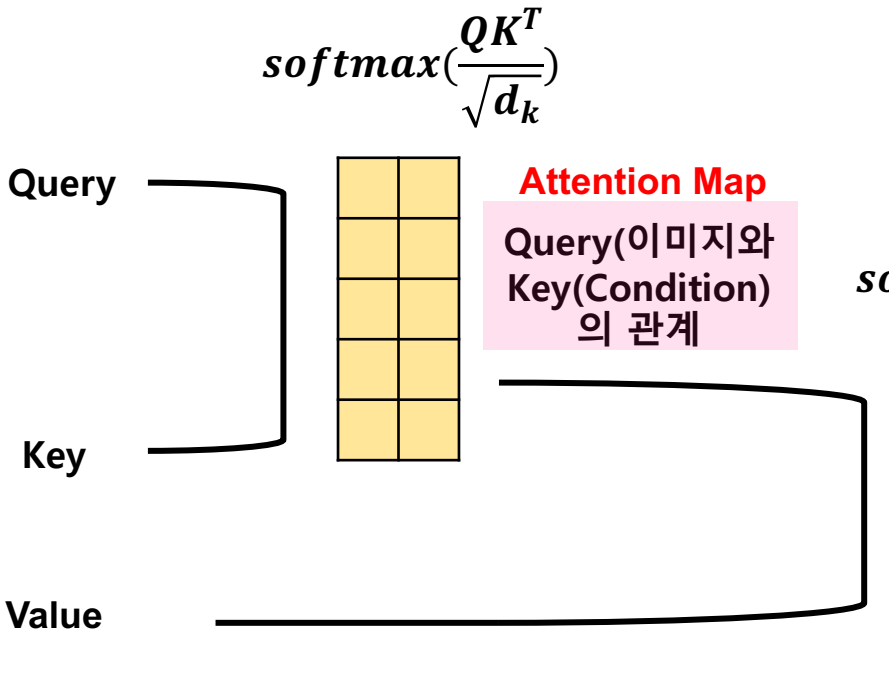
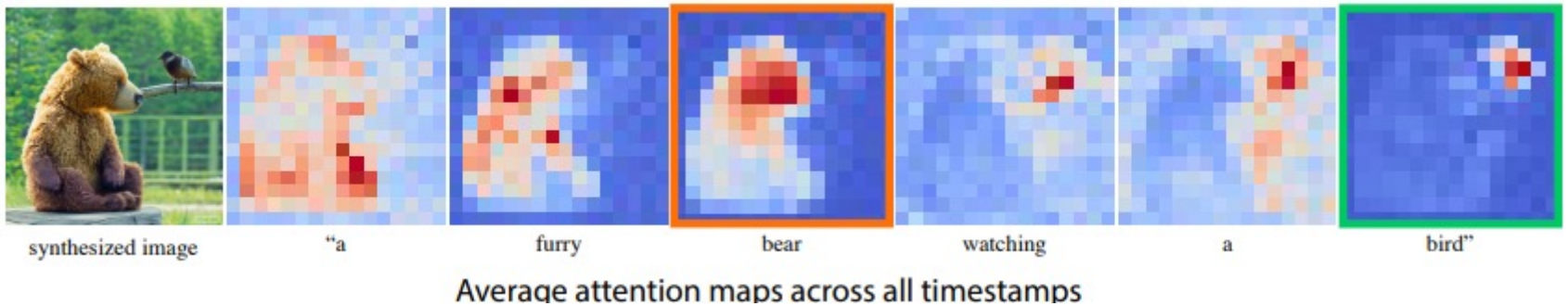


Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Controlling Cross Attention : Attention Map (Cross Attention Map)

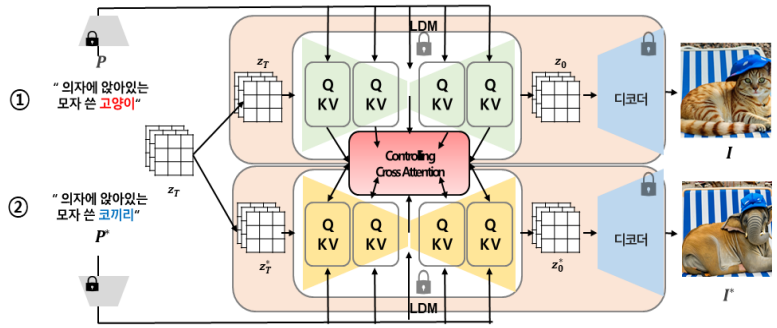
- 이미지(Query)와 Condition(Key) 의 관계를 반영
- 각 단어에 해당하는 attention map을 timestamp에 대해 평균
- " bear " : 곰의 형체, "watches" : 눈 주변



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt



❖ **Controlling Cross Attention : Attention Map (Cross Attention Map)**

① 원본 Prompt P 와 타겟 Prompt P^* 의 디퓨젼 과정에서 attention map M_t, M_t^* 추출 (Seed 고정)

② attention map M_t, M_t^* 에 대해 cross attention control (Edit) 을 하여 attention map $\widehat{M_t}$ 를 생성

③ 타겟 Prompt P^* 와 수정한 attention map $\widehat{M_t}$ 로 이미지 생성

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^*)

① ② ③

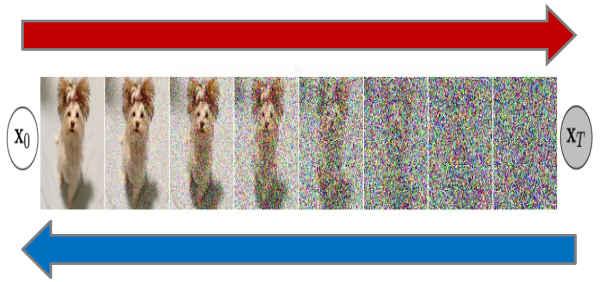
Prompt 의 변화에 따라
Attention Map Edit 방식이 다름

1. Word-swap
2. Adding a new phrase
3. Attention Re-weighting

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt



❖ Word Swap

- 원본 Prompt P 와 새로운 Prompt P^* 내 **특정 단어만** 바뀐 경우
- Soft attention constrain : 특정 time step τ 이후부터 수정한 attention map 을 사용
- 이미지의 형태는 초기 denoising step에서 결정

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

I

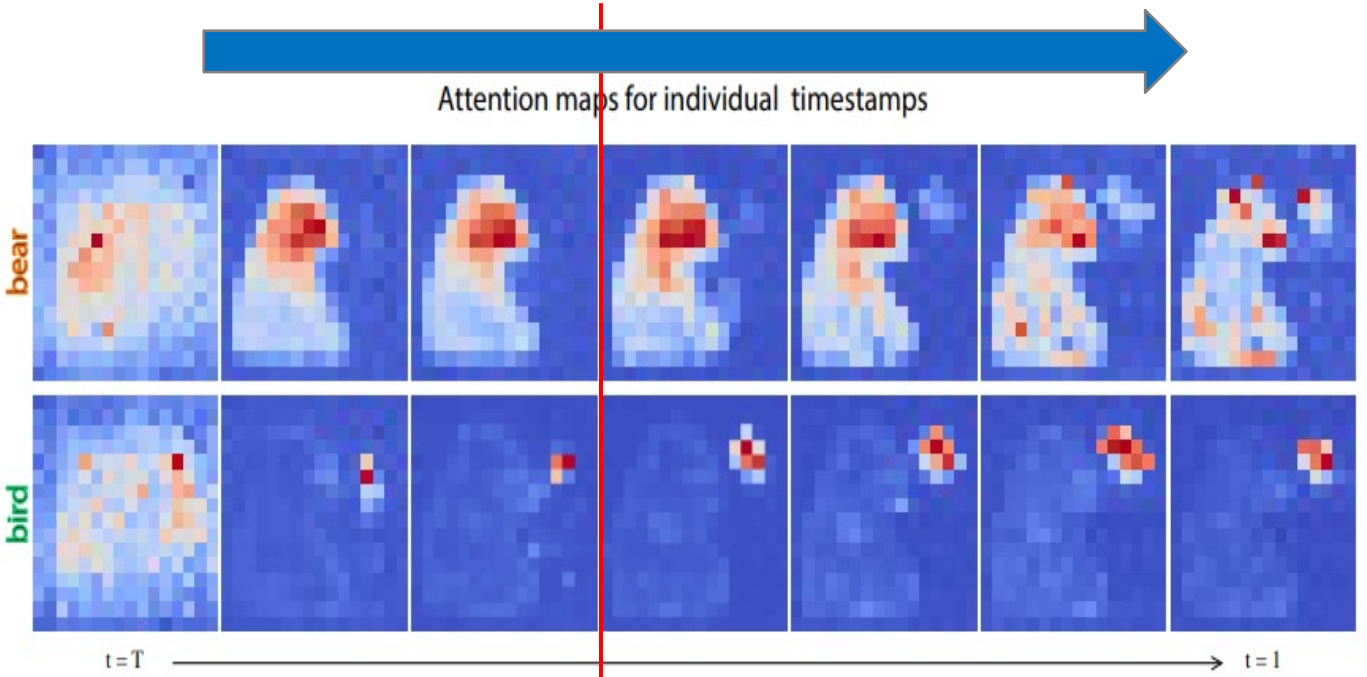
"모자 쓴 고양이"

P

I^*

" 모자 쓴 강아지"

P^*



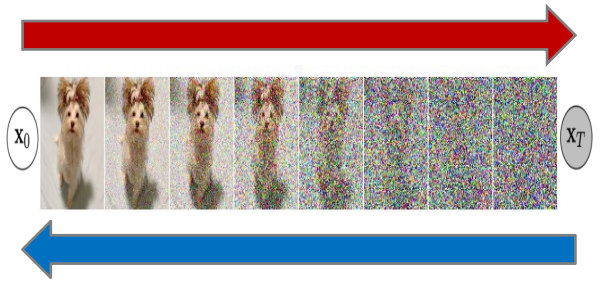
M_t : 원본 prompt를 반영한 attention map

M_t^* : 새로운 prompt를 반영한 attention map

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt

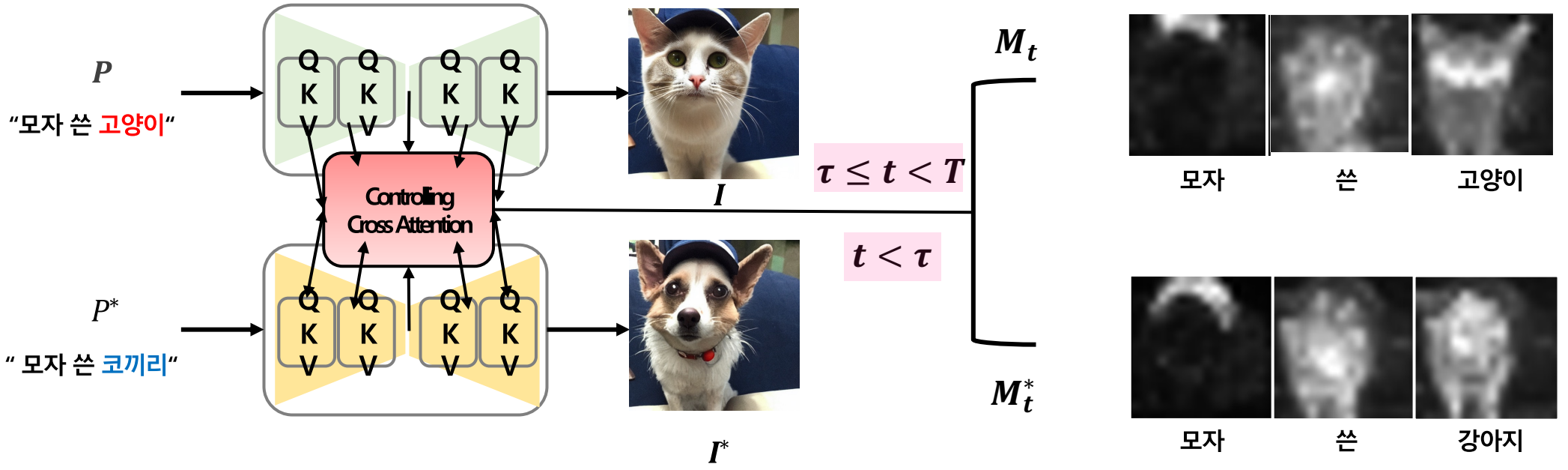


❖ Word Swap

- 원본 Prompt P 와 새로운 Prompt P^* 내 **특정 단어만** 바뀐 경우

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

- Soft attention constrain : 특정 time step τ 이후부터 수정한 attention map 을 사용
- 이미지의 구조는 초기 denoising step에서 결정



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

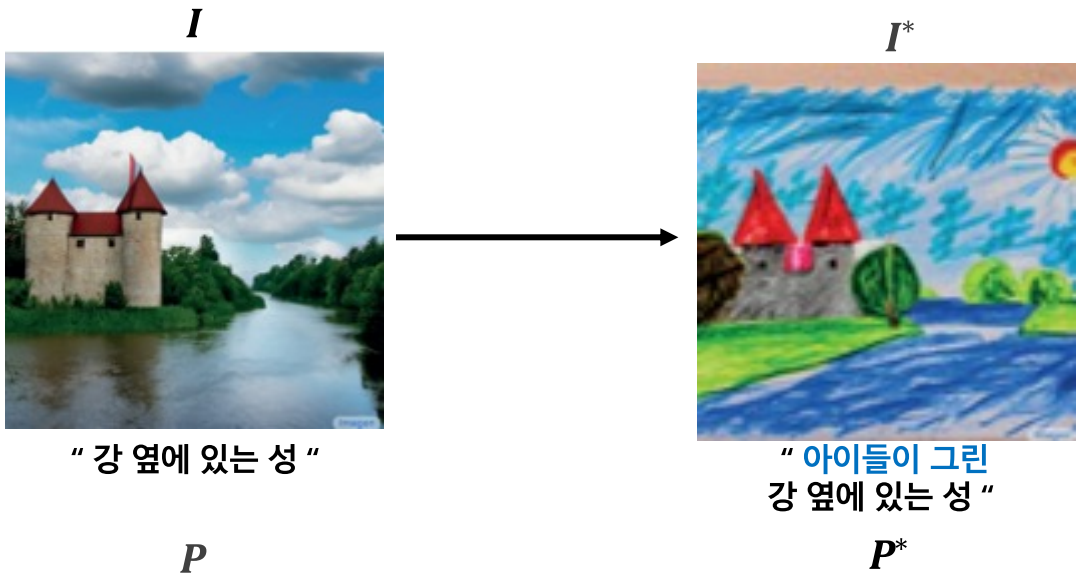
Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Adding a new phrase

- 원본 Prompt P 에 **새로운 문장이** 추가된 Prompt P^*
- 겹치는 문장일 경우 원본 이미지 M_t 사용, 아닐 경우 수정한 attention map M_t^* 사용
- $A(j)$: Alignment 함수 . 두 prompt 의 겹치는 부분을 찾아줌 ($A(j) = None$ 일 경우 겹치는 부분이 아님)

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



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

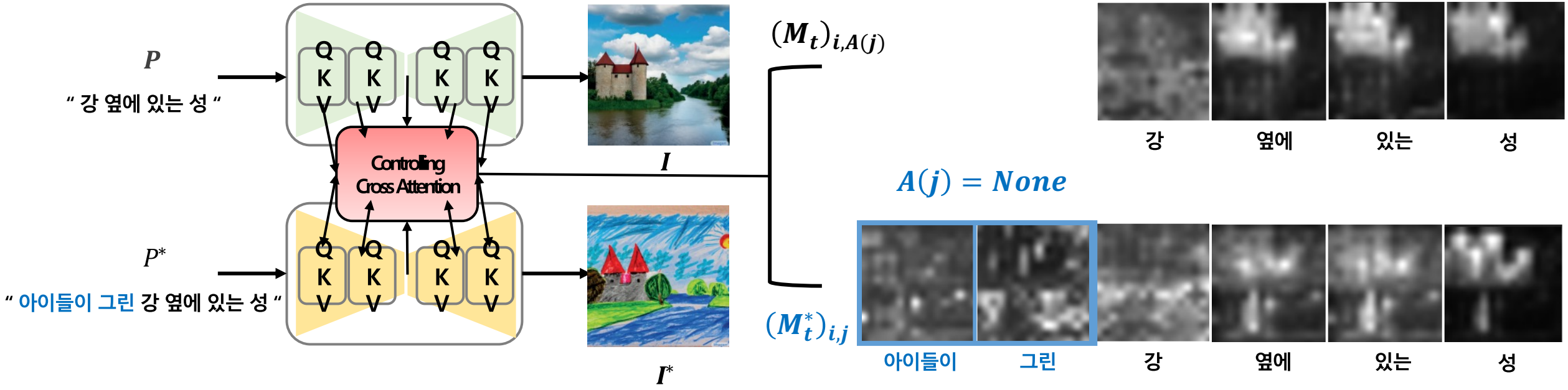
Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Adding a new phrase

- 원본 Prompt P 에 **새로운 문장이** 추가된 Prompt P^*
- 겹치는 문장일 경우 원본 이미지 M_t 사용, 아닐 경우 수정한 attention map M_t^* 사용
- $A(j)$: Alignment 함수 . 두 prompt 의 겹치는 부분을 찾아줌 ($A(j) = None$ 일 경우 겹치는 부분이 아님)

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



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Attention Re-weighting

- Prompt 내 **특정 단어의 영향력을** 조절
- 특정 단어 토큰 j 에 대해 가중치를 조절
- c : 특정 단어의 가중치 ($c \in [-2, 2]$)

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

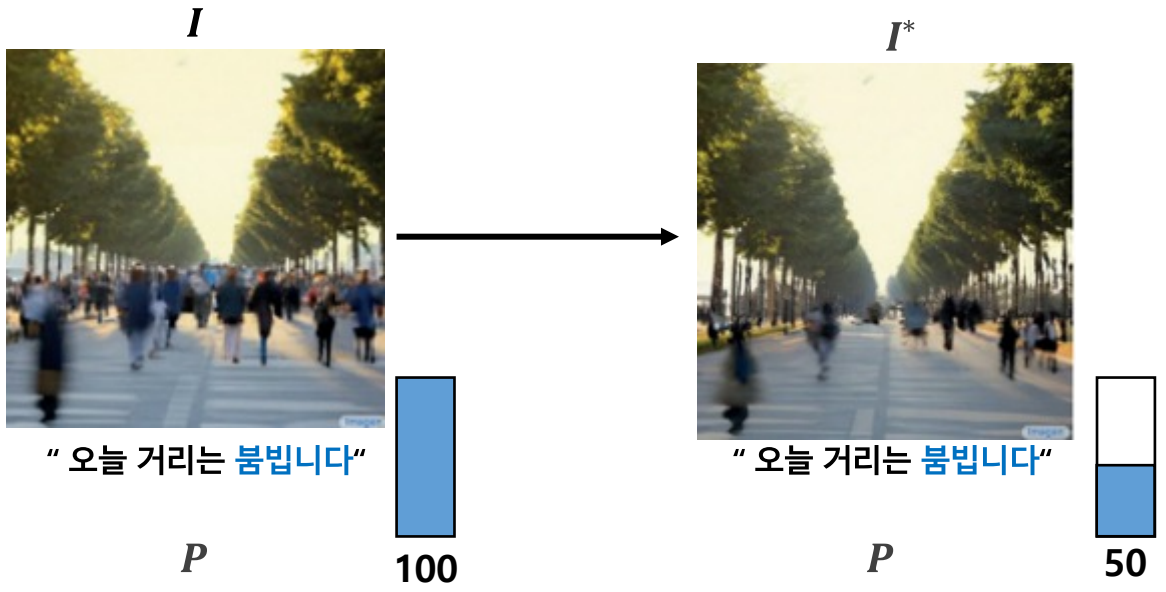


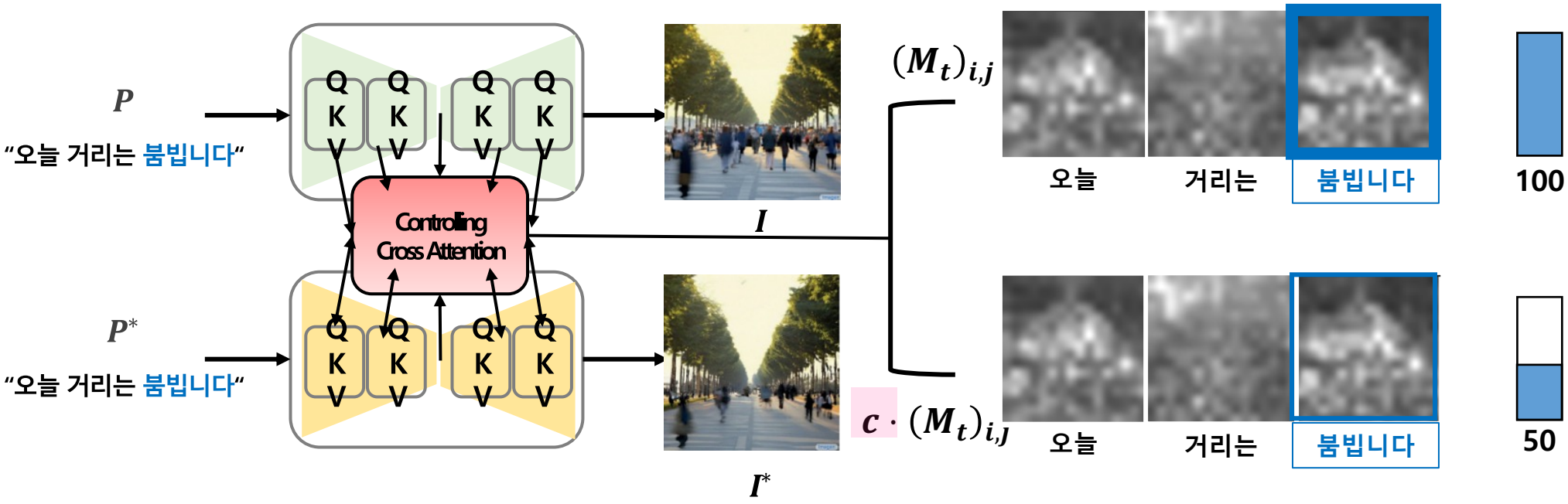
Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Attention Re-weighting

- Prompt 내 **특정 단어의 영향력을** 조절
- 특정 단어 토큰 j 에 대해 가중치를 조절
- c : 특정 단어의 가중치 ($c \in [-2, 2]$)

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



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Image Editing with Diffusion Model

Prompt-to-Prompt

❖ Results

- Prompt-to-Promt Editing 방식을 썼을 때, 원본 이미지 형태를 유지
- Prompt-to-Promt Editing 의 단점 :
  - Attention map 이 저화질이라 정교한 editing이 어려움
  - 이미지내에서 객체의 위치를 이동시킬 수 없음

" 햄버거 먹는 다람쥐"

" 햄버거 먹는 사자"

Prompt-to-prompt 없이

" 햄버거 먹는 다람쥐" → " 햄버거 먹는 사자"

Prompt-to-prompt

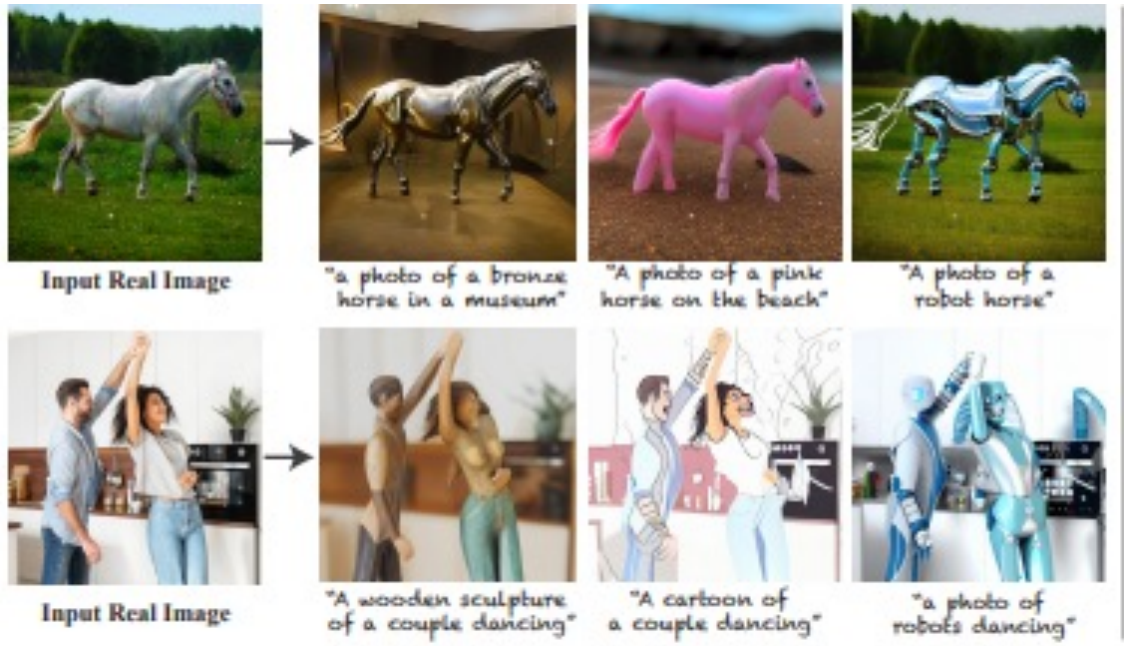
Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

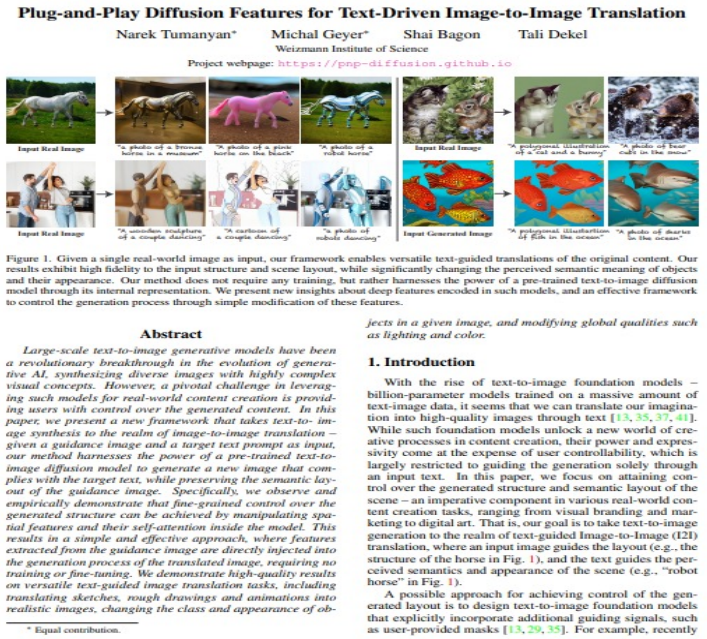
Image Editing with Diffusion Model

Plug-and-Play

❖ **Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation**

- **Image to image Translation**: 이미지를 다른 양상으로 mapping
- 원본 이미지의 **spatial feature 와 self attention의 Query와 Key를 활용하여** 이미지를 Edit
- Finetuning 과 학습할 필요가 없는 editing 기법





Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

Narek Tumanyan* Michal Geyer* Shai Bagon Tali Dekel

Weizmann Institute of Science

Project webpage: https://pnp-diffusion.github.io

Figure 1. Given a single real-world image as input, our framework enables versatile text-guided translations of the original content. Our results exhibit high fidelity to the input structure and scene layout, while significantly changing the perceived semantic meaning of objects and their appearance. Our method does not require any training, but rather harnesses the power of a pre-trained text-to-image diffusion model through its internal representation. We present new insights about deep features encoded in such models, and an effective framework to control the generation process through simple modification of these features.

Abstract

Large-scale text-to-image generative models have been a revolutionary breakthrough in the evolution of generative AI, synthesizing diverse images with highly complex visual concepts. However, a pivotal challenge in leveraging such models for real-world content creation is providing users with control over the generated content. In this paper, we present a new framework that takes text-to- image synthesis to the realm of image-to-image translation – given a guidance image and a target text prompt as input, our method harnesses the power of a pre-trained text-to-image diffusion model to generate a new image that complies with the target text, while preserving the semantic layout of the guidance image. Specifically, we observe and empirically demonstrate that fine-grained control over the generated structure can be achieved by manipulating spatial features and their self-attention inside the model. This results in a simple and effective approach, where features extracted from the guidance image are directly injected into the generation process of the translated image, requiring no training or fine-tuning. We demonstrate high-quality results on versatile text-guided image translation tasks, including translating sketches, rough drawings and animations into realistic images, changing the class and appearance of objects in a given image, and modifying global qualities such as lighting and color.

1. Introduction

With the rise of text-to-image foundation models – billion-parameter models trained on a massive amount of text-image data, it seems that we can translate our imagination into high-quality images through text [13, 35, 37, 41]. While such foundation models unlock a new world of creative processes in content creation, their power and expressivity come at the expense of user controllability, which is largely restricted to guiding the generation solely through an input text. In this paper, we focus on attaining control over the generated structure and semantic layout of the scene – an imperative component in various real-world content creation tasks, ranging from visual branding and marketing to digital art. That is, our goal is to take text-to-image generation to the realm of text-guided Image-to-Image (I2I) translation, where an input image guides the layout (e.g., the structure of the horse in Fig. 1), and the text guides the perceived semantics and appearance of the scene (e.g., "robot horse" in Fig. 1).

A possible approach for achieving control of the generated layout is to design text-to-image foundation models that explicitly incorporate additional guiding signals, such as user-provided masks [13, 29, 35]. For example, recently

Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

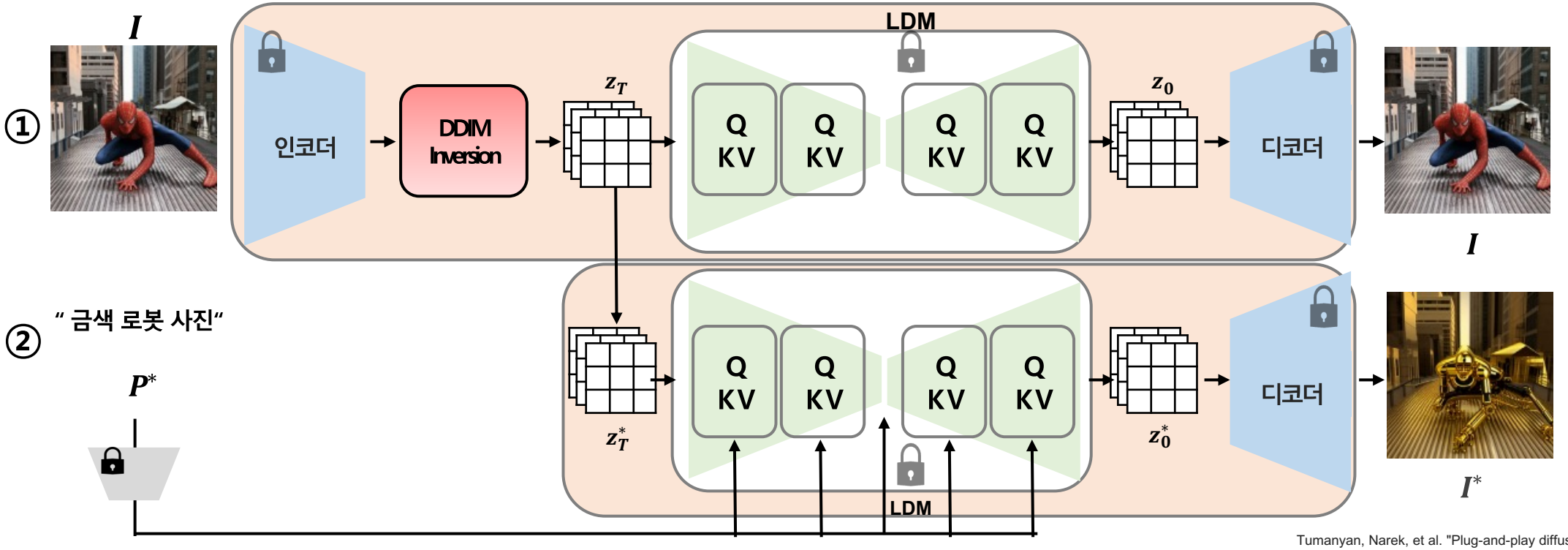
Image Editing with Diffusion Model

Plug-and-Play

❖ 문제상황

① 원본 이미지 I 를 Gaussian Noise z_t 로 만들고, 다시 복원하는 디퓨젼 모델 **(원본 이미지 생성과정)**

② 원본 이미지 I 의 타겟 Prompt P^* 를 반영하는 이미지 I^* 로 edit 하는 디퓨젼 모델 **(원본 이미지의 편집 과정)**



Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023.

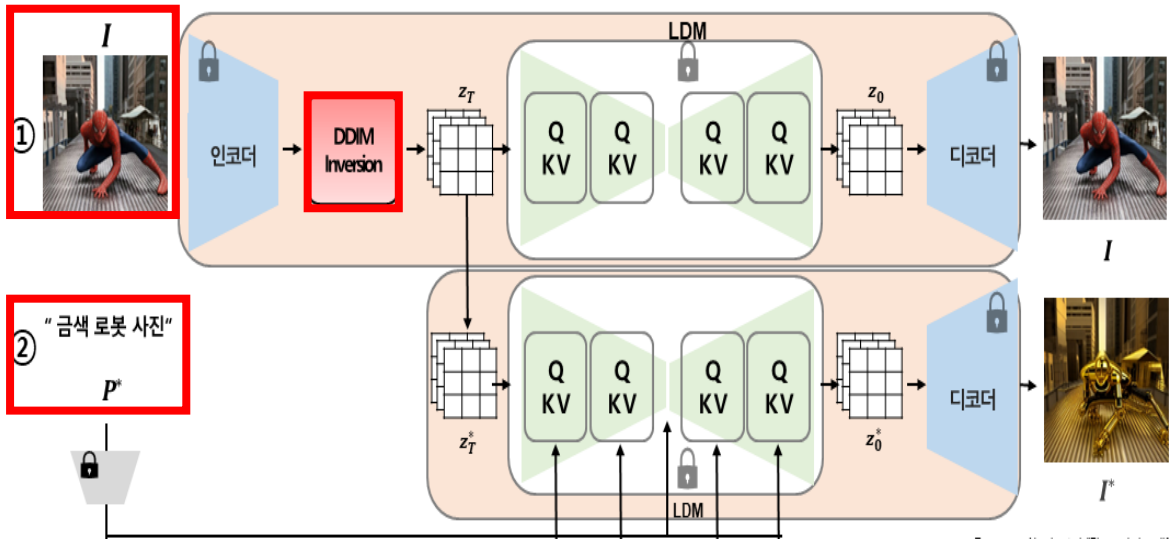
Image Editing with Diffusion Model

Plug-and-Play

❖ **문제상황 (Plug-and-Play 와 Prompt-to-Prompt 비교)**

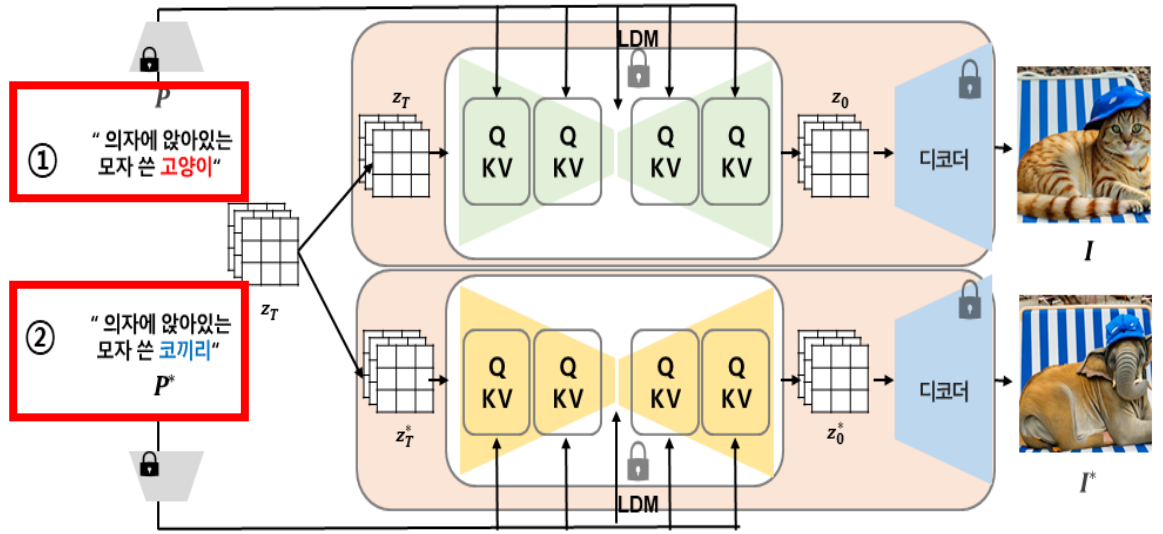
- **Plug-and-Play** : 원본 이미지 I 와 타겟 프롬프트 P^* 가 주어지고, **Inversion (원본 이미지를 복원하는 initial noise 찾는 과정)** 필요
- **Prompt-to-Prompt :** 원본 프롬프트 P 와 타겟 프롬프트 P^* 가 주어짐.

Plug-and-Play



Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Prompt-to-Prompt



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

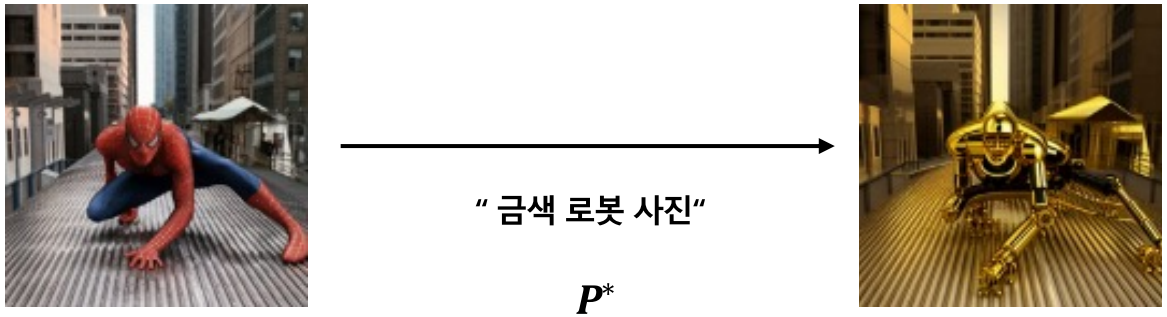
Image Editing with Diffusion Model

Plug-and-Play

❖ 문제상황

- 원본 이미지 I 가 존재할 때 사용자가 타겟 프롬프트 P^* 를 입력하면, **원본 이미지 I 의 구조와 semantic layout을 보존**하면서 해당 프롬프트를 반영하는 이미지 I^* 를 생성

➢ 원본 이미지를 복원해내는 디퓨전 모델 **디코더 레이어** 의 ① **Residual Block내 spatial feature** 와 ② **Self Attention**을 이용하면 원본 이미지의 형태를 유지할 수 있을 것이다.



Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

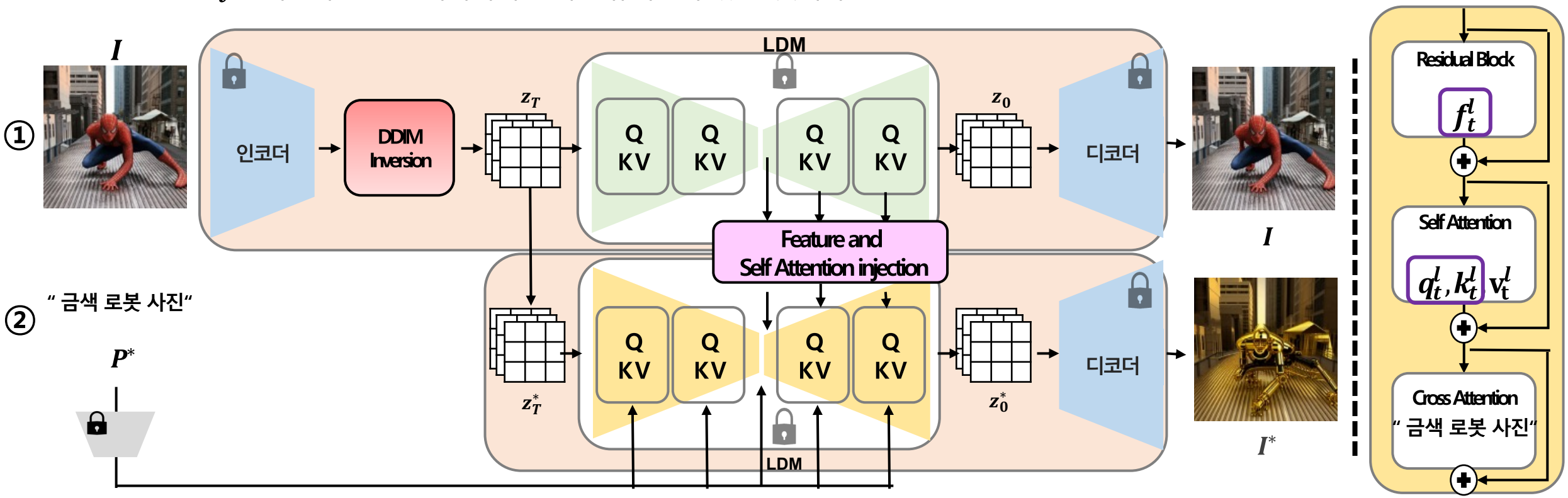
Image Editing with Diffusion Model

Plug-and-Play

❖ 문제상황

➢ 원본 이미지를 복원해내는 디퓨전 모델 **디코더 레이어**의 ① **Residual Block내 spatial feature** f_t^l와 ② **Self Attention affinities** A_t^l을 이용하면 원본 이미지의 형태를 유지할 수 있을 것이다.

$$A_t^l = softmax(q_t^l, k_t^{l^T})$$



Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Image Editing with Diffusion Model

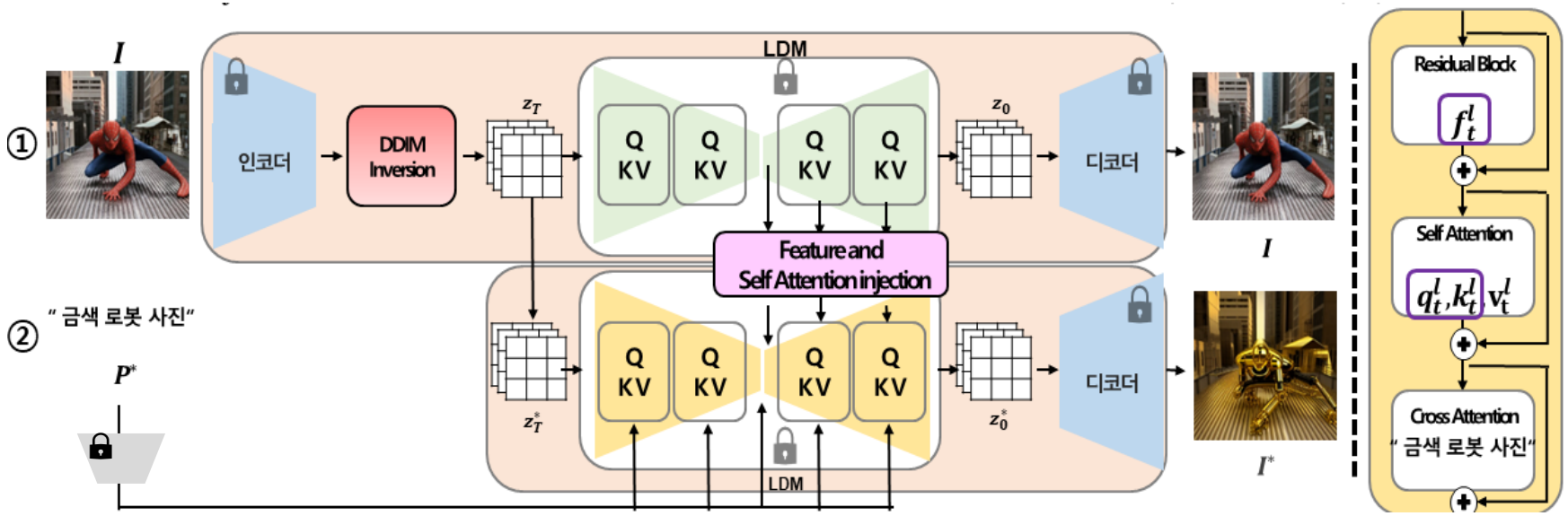
Plug-and-Play

❖ Plug-and-Play 과정

① DDIM inversion 을 통해 원본 이미지 I 에 대한 initial noise z_T 를 구함

② 원본 이미지의 initial noise z_T 와 동일한 z_T^* 를 바탕으로 타겟 프롬프트 P^* 를 반영한 이미지 생성 **(이미지 편집 과정)**

➢ 디코더 몇번 째 **레이어 l** 부터 **spatial feature f_t^l**와 **Self Attention affinities A_t^l**을 주입해야하는가?



$$A_t^l = softmax(q_t^l, {k_t^l}^T)$$

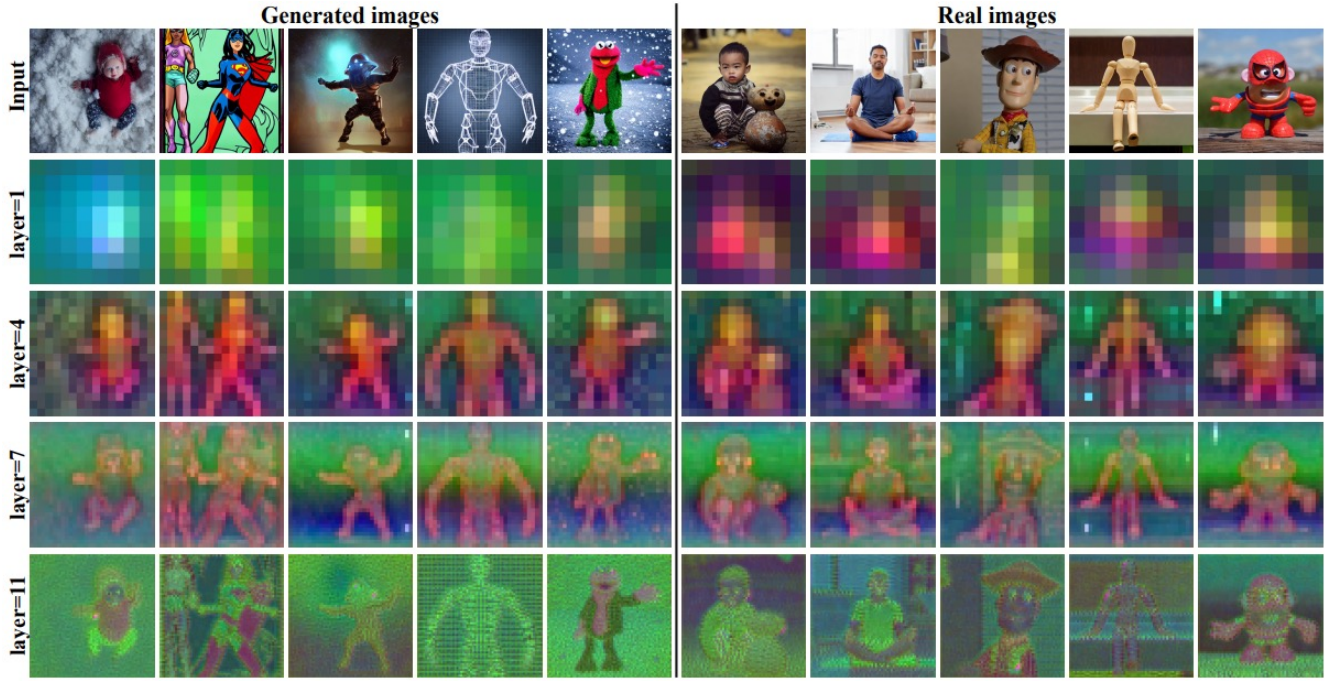
Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Image Editing with Diffusion Model

Plug-and-Play

❖ Spatial Features 실험 1

- 휴머노이드 이미지를 복원하는 과정에서 디퓨젼 모델 디코더 레이어에서 spatial feature를 추출하고 PCA로 시각화 (t = 540)
- 낮은 레이어에서는 배경과 객체와 분리, **디코더 레이어 = 4** 일 때부터 어느정도 휴머노이드의 윤곽이 나타남
- 같은 부위는 동일한 색상이 나타남 (머리 : 노란색, 다리 : 빨강)



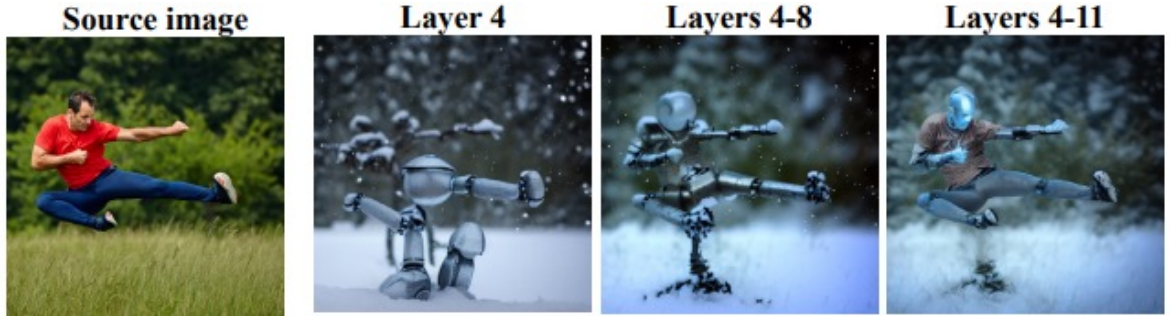
Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Image Editing with Diffusion Model
Plug-and-Play

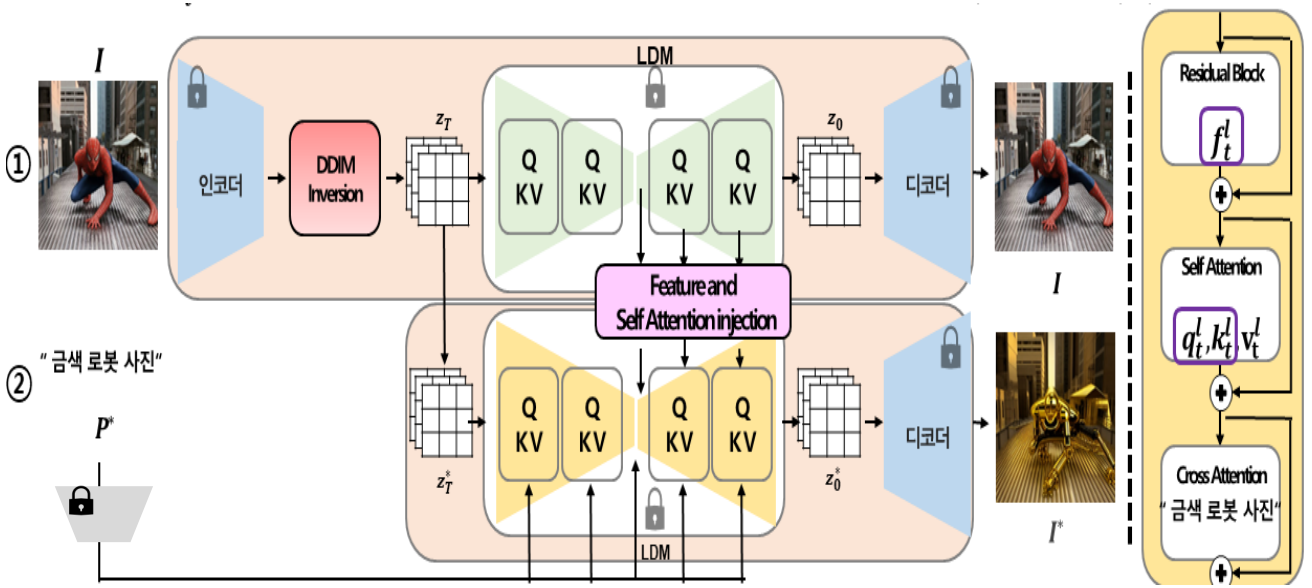
❖ Spatial Features 실험 2

- 이미지를 edit 하는 과정에서 다양한 디코더 레이어에 원본 이미지의 spatial feature f_t^l 를 주입하는 실험
  - **레이어 4** : 원본 이미지 형태 유지 불가
  - **레이어 4 ~ 11** : 원본 이미지 형태가 너무 많이 나타남

$$z_{t-1}^* = \hat{\epsilon}_\theta(z_t^*, P, t; \{f_t^l\})$$



P : " 눈 속에 있는 로봇 사진"



Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Image Editing with Diffusion Model

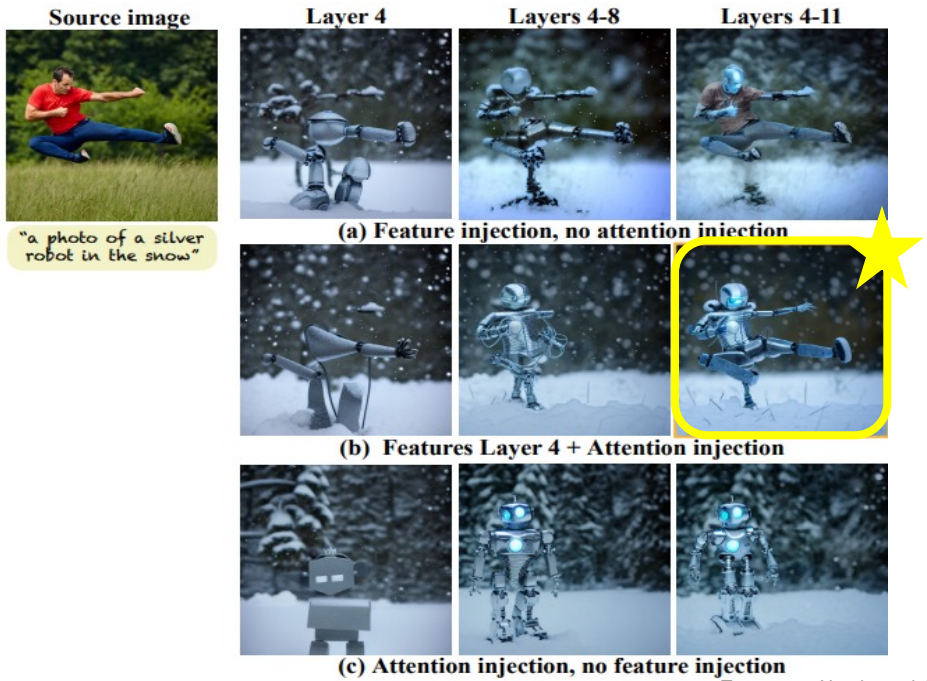
Plug-and-Play

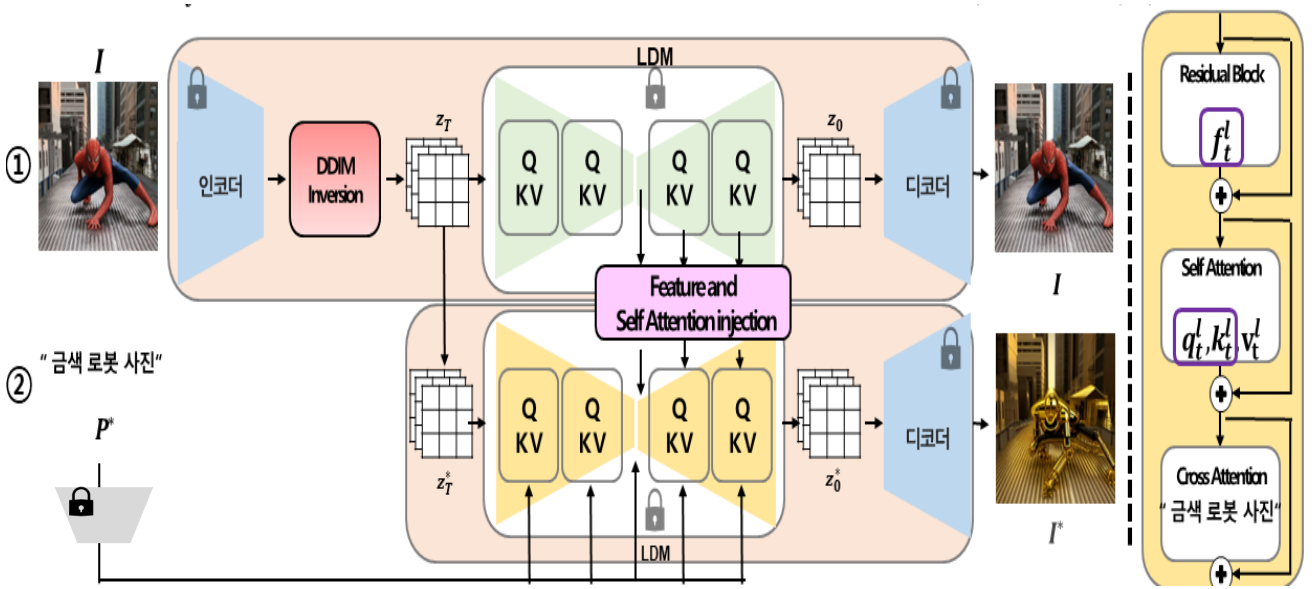
❖ Spatial Feature 와 Self Attention 실험

- 이미지를 edit 하는 과정에서 다양한 디코더 레이어에 원본 이미지의 spatial feature f_t^l 와 self attention A_t^l 를 주입하는 실험
  - **spatial feature f_t^l 는 레이어 4 , self attention A_t^l 은 레이어 4 ~11** 이 가장 좋음

$$z_{t-1}^* = \hat{\epsilon}_\theta(z_t^*, P, f_t^4 ; \{A_t^l\})$$

$$A_t^l = softmax(q_t^l, k_t^{l^T})$$





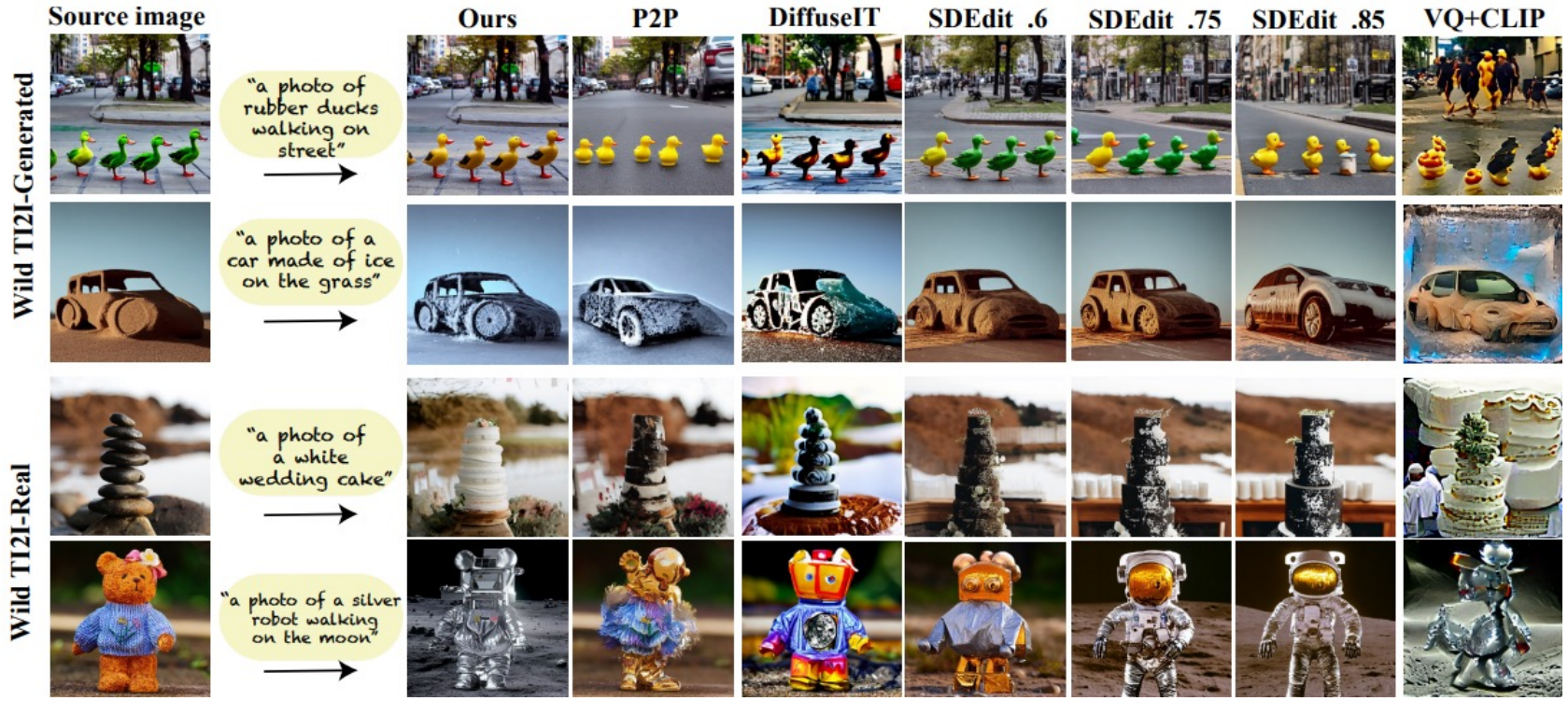
Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Image Editing with Diffusion Model

Plug-and-Play

❖ Results

- **원본 이미지 생성 과정의 디코어에서 spatial feature f_t^l 는 레이어 4 , :self attention A_t^l은 레이어 4 ~11** 이 가장 좋음
- 원본 이미지의 형태를 잘 유지하면서 Image-to-Image Translation 수행



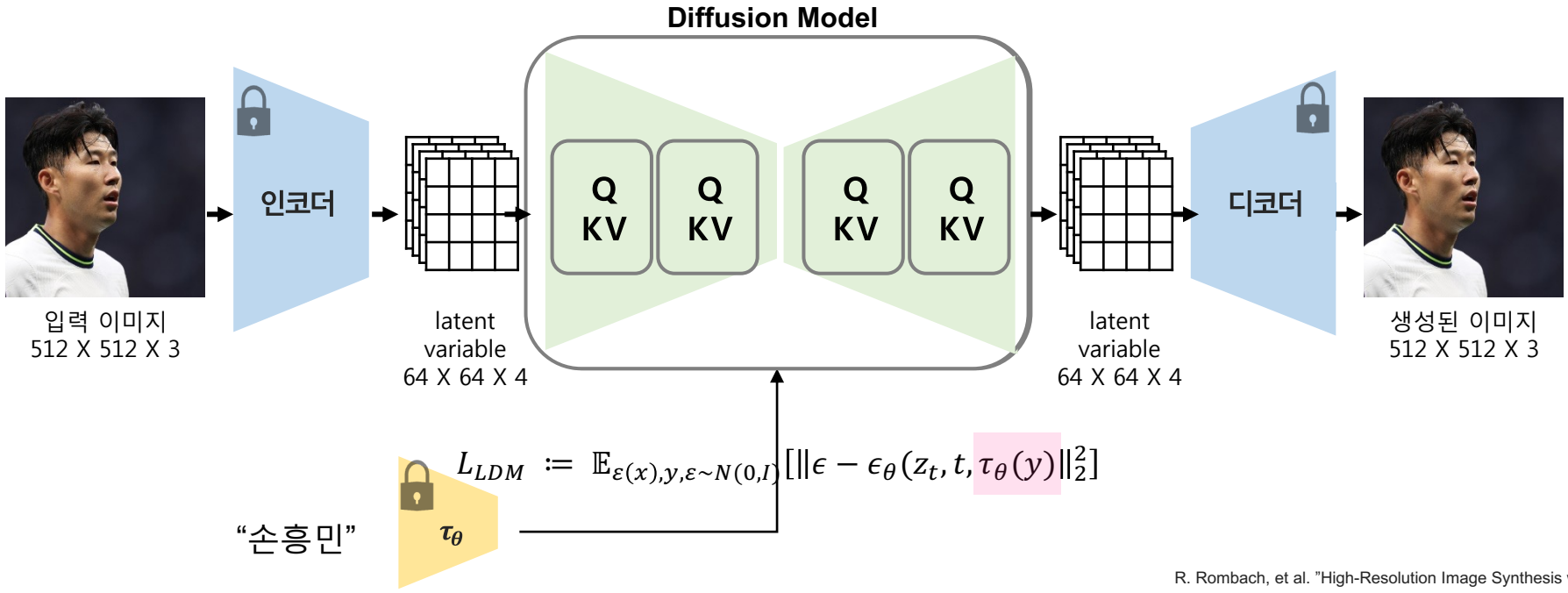
Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Conclusion

Conclusion

❖ Image Editing with Diffusion Models

- 디퓨젼 기반 생성모델은 단순한 Text to Image Generation을 뛰어넘어 현재는 image editing 까지 가능한 수준
- 이때 대부분의 디퓨젼 기반 image editing 모델은 LDM 을 backbone으로 작동
- **LDM** : 오토인코더**(VQ-VAE)**를 사용하여 **latent space**에서 작동, 연산량을 줄임



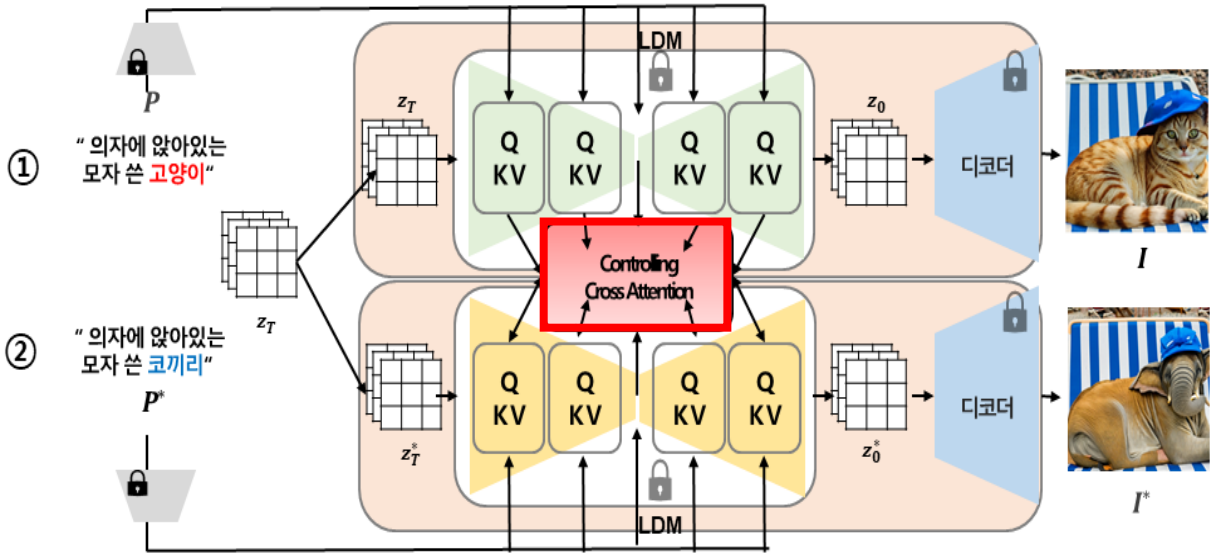
R. Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models" CVPR (2022)

Conclusion

❖ Image Editing with Diffusion Models

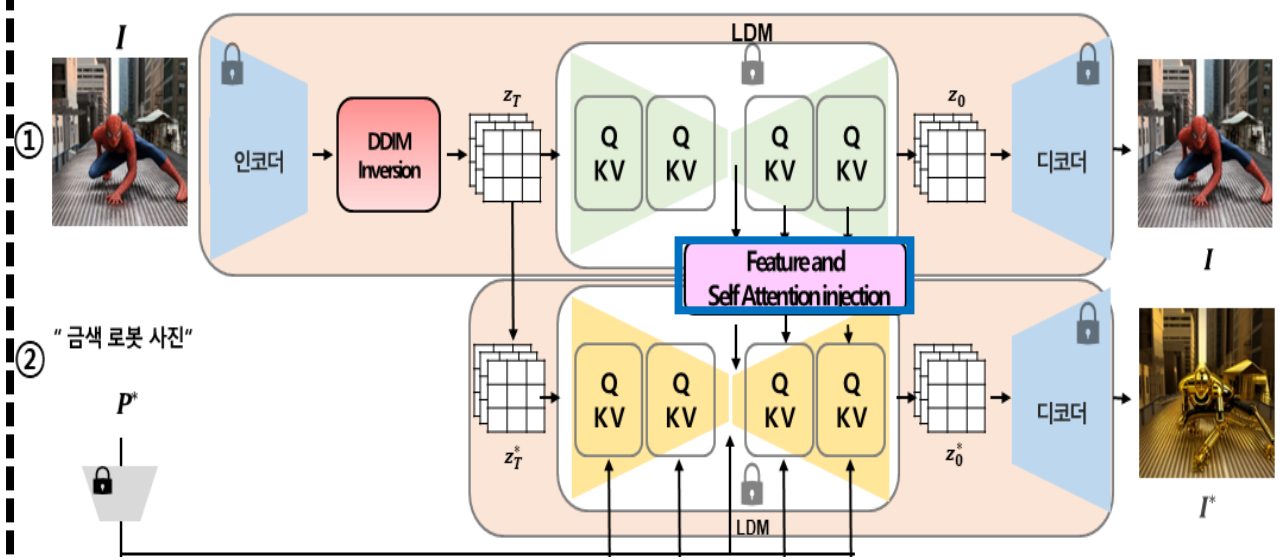
- 이러한 image editing 모델들은 원본 이미지의 형태를 유지하면서 이미지를 수정하는 것이 관건
- 따라서 원본 이미지를 생성하는 과정에서의 residual block spatial feature, self attention, cross attention을 활용하여 edit
- Prompt-to-prompt : cross attention map, plug-and-play : residual block spatial feature 와 self attention

Prompt-to-Prompt



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022

Plug-and-Play



Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

References

[1] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851

[2] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." *arXiv preprint arXiv:2010.02502* (2020).

[3] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." arXiv preprint arXiv:2207.12598 (2022)

[4] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022

[5] Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." *arXiv preprint arXiv:2208.01626* (2022).

[6] Tumanyan, Narek, et al. "Plug-and-play diffusion features for text-driven image-to-image translation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[7] 김준호 - NAVER AI LAB. "생성모델부터 Diffusion까지 3회 I 모두의연구소 모두팝." *YouTube*, YouTube, 3 Jan. 2023, www.youtube.com/watch?v=Z8WWriIh1PU&t=3370s.

고맙습니다