Enhancing Prompt Understanding In Text-to-Image Diffusion Model

2024.05.03

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

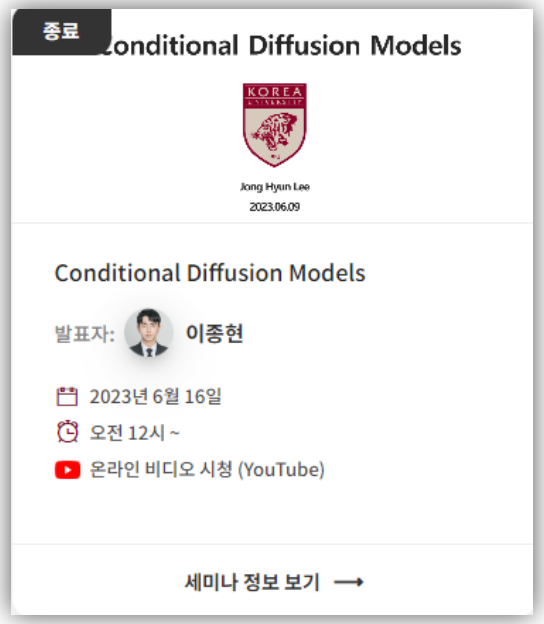
- dlwlsdn0225@korea.ac.kr

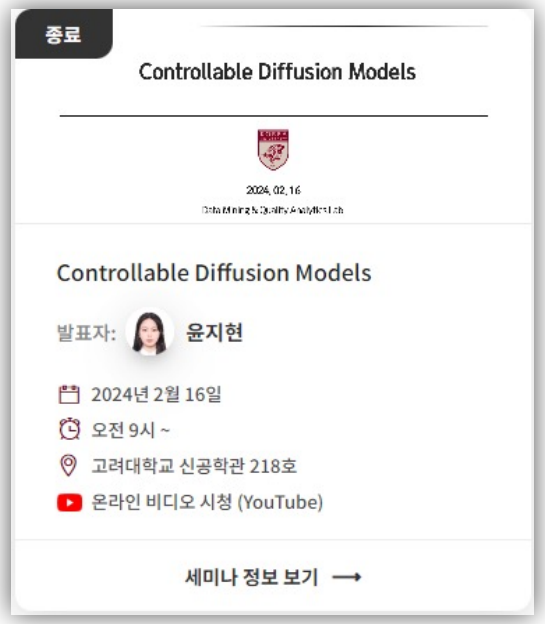
목차

Introduction

Text to Image Generation

- **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
- **Controllable Diffusion Models:** 텍스트 이외 입력 조건에 대해서도 컨디션을 부여하는 ControlNet, Composer, Uni-ControlNet 소개





Introduction
Background

❖ 이미지 생성 모델

- 상상을 현실로 그려낼 수 있는 흥미로운 분야
  - **Text to Image generation**: 사용자가 **프롬프트**를 입력하여 원하는 이미지를 생성하는 방식으로 작동
- 현재 디퓨젼 모델이 해당 분야에서 주를 이룸

올림픽 400m 접영 종목에서 수영하는 곰인형

초밥으로 만든 집에 살고있는 귀여운 코기

Imagen

Midjourney - '스페이스 오페라 극장

https://imagen.research.google/
https://www.midjourney.com/home/

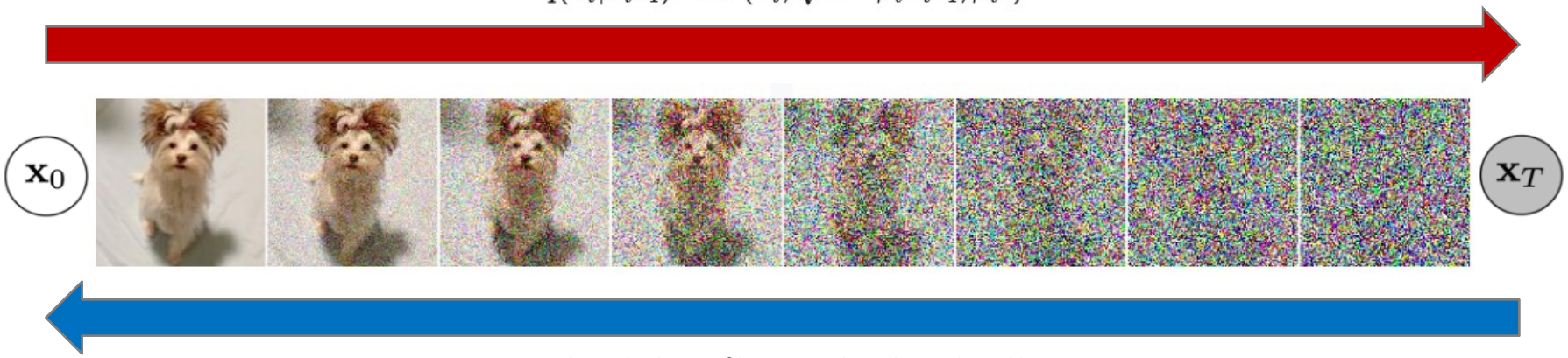
Introduction
DDPM

❖ Denoising Diffusion Probabilistic Models (DDPM)

- **Forward Process**
  - 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

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



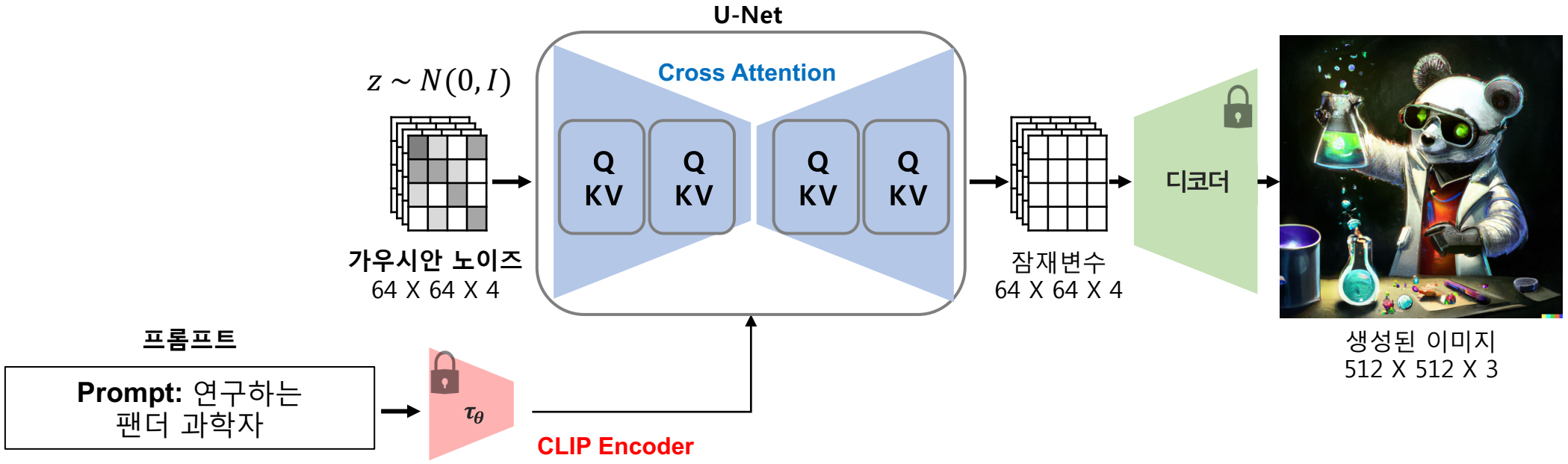
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

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

Introduction
Text to Image Generation

❖ 입력 조건에 맞는 이미지 생성: Latent Diffusion Model

- **Text to Image Generation** : LDM은 가우시안 노이즈와 프롬프트를 입력받아 원하는 이미지를 생성
  - **CLIP Encoder** : 텍스트와 이미지의 joint embeddding을 구하는 **텍스트 인코더**
  - **Cross-Attention**: 프롬프트로 주어진 **condition을 이미지에 반영**
- 많은 연구들이 Stable Diffusion을 기반으로 진행



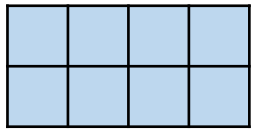
R. Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models" CVPR (2022)
https://www.filipcvetko.com/dalle2/

Introduction
Text to Image Generation

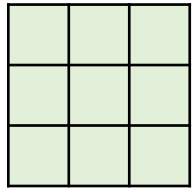
❖ Cross Attention in LDM

- 이미지와 프롬프트 의 관계를 반영
- **Query**: 이미지로부터 추출
- **Key, Value**: 프롬프트에서 추출

이미지

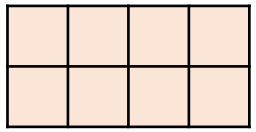


Query

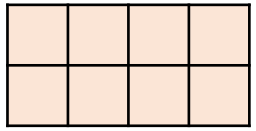


텍스트 프롬프트

"연구하는 팬더 과학자"

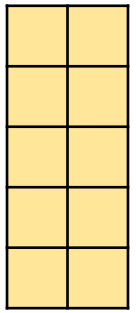


Key



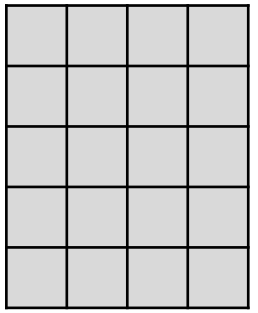
Value

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



이미지(Query)와 프롬프트내 단어(Key)의 관계

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



각 단어가 이미지의 어느 부분을 참조할지

팬더

연구

과학자

https://www.youtube.com/watch?v=Z8WWrilh1PU&t=3368s

https://v.daum.net/v/20230515160516680

Introduction

Text to Image Generation

그럼 디퓨전 모델은

사용자가 입력한 '**프롬프트**'를

완벽하게 반영하는가?

Introduction
Text to Image Generation

❖ **Text to Image 모델**

Prompt: 빨간 자동차와 하얀 양

Prompt: 말과 원숭이

Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).
Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-10.

Introduction
Text to Image Generation

❖ **Text to Image 모델에서 발생하는 문제점**

- 사용자가 입력하는 프롬프트를 완벽하게 반영하기 어려움
- **잘못된 속성**이 부여되거나(**attribute binding**), 특정 **객체가 생성되지 않는** (**missing objects**)등의 문제 발생

Prompt: 빨간 자동차와 **하얀 양**

< Attribute binding >

Prompt: 말과 **원숭이**

< Missing Object >

Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).
Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-10.

Enhancing Prompt Understanding

Structure Diffusion

TRAINING-FREE STRUCTURED DIFFUSION GUIDANCE FOR COMPOSITIONAL TEXT-TO-IMAGE SYNTHESIS

(ICLR 2023)

Attend-and-Excite

Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models

(ACM-TOG 2023)

Enhancing Prompt Understanding

Structure Diffusion

TRAINING-FREE STRUCTURED DIFFUSION GUIDANCE FOR COMPOSITIONAL TEXT-TO-IMAGE SYNTHESIS

(ICLR 2023)

Attend-and-Excite

Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models

(ACM-TOG 2023)

Enhancing Prompt Understanding

Structure Diffusion

❖ **TRAINING-FREE STRUCTURED DIFFUSION GUIDANCE FOR COMPOSITIONAL TEXT-TO-IMAGE SYNTHESIS (ICLR 2023)**

- **Motivation**: 객체에 잘못된 속성이 부여 - Attribute binding
- 원인: ① Training bias ② **프롬프트에 대한 이해 부족**

Prompt: 빨간 자동차와 **하얀 양**

Prompt: **하얀 건물** 앞에 있는 **갈색 벤치**

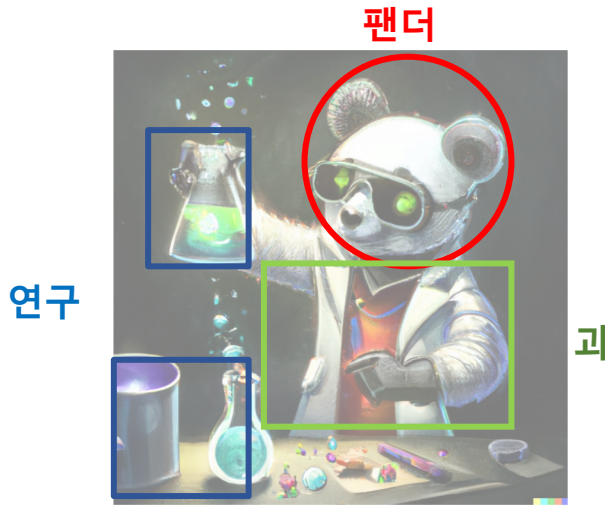
Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).

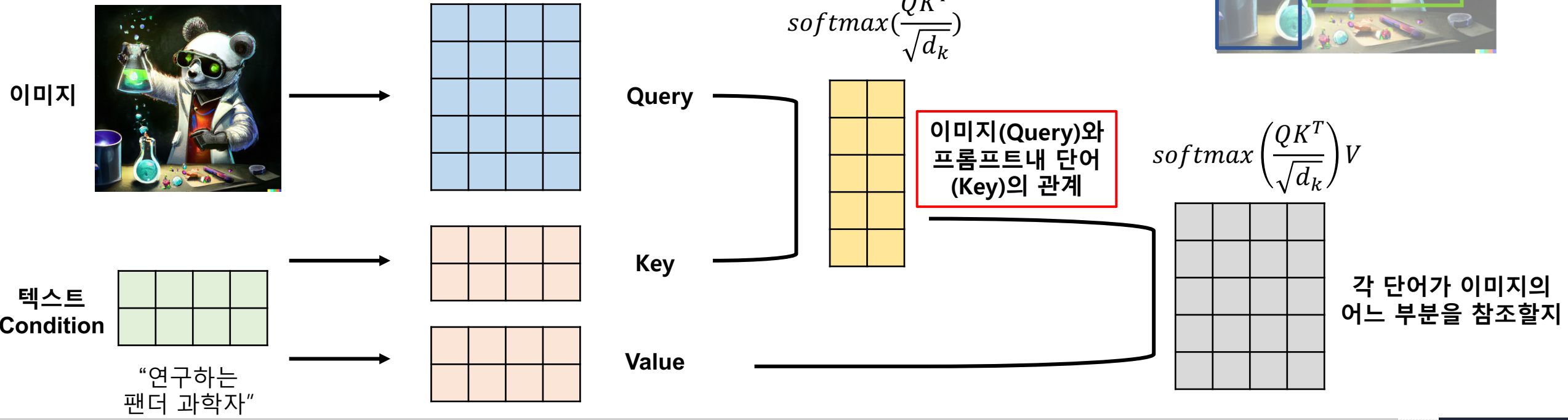
Enhancing Prompt Understanding

Structure Diffusion

❖ **Cross attention map**

- Cross-Attention: 프롬프트로 주어진 condition을 이미지에 반영
- **Cross attention map**은 각 프롬프트내 단어가 이미지 어느 부분에 영향을 주는지 나타냄





https://www.youtube.com/wtch?v=Z8WWrilh1PU&t=3368s
https://www.filipcvetko.com/dalle2/

Enhancing Prompt Understanding

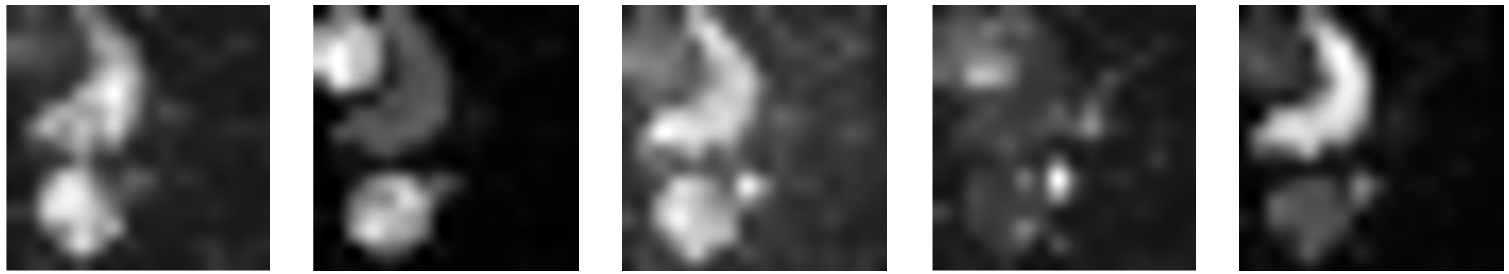
Structure Diffusion

❖ **Cross attention map**

- 각 단어가 이미지 어느 부분에 영향을 주는지 확인 (하얀색일수록 강한 영향력)
- **객체와 속성이 잘못 매칭되는 문제**
- **프롬프트에 대한 이해도를 높여 이를 해결**

Prompt: 노란 사과와 빨간 바나나

생성된 이미지



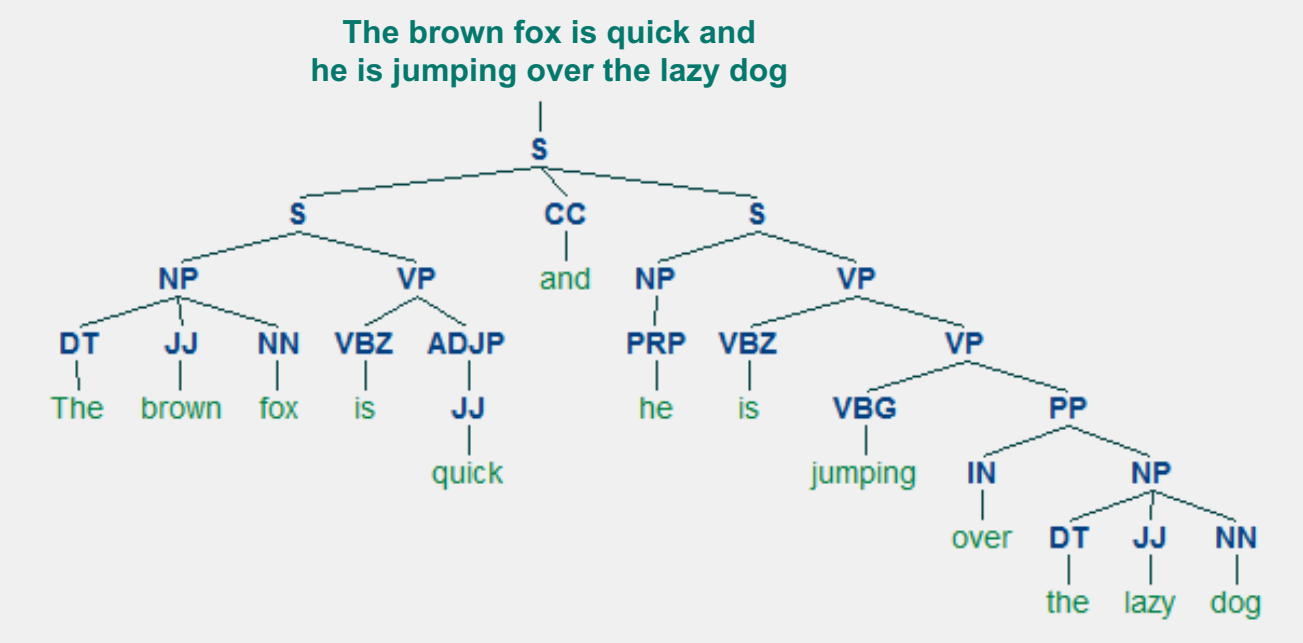
Cross attention map

Enhancing Prompt Understanding

Structure Diffusion

❖ 프롬프트 문법적 구성 파악: Constituency Tree

- 각 객체와 속성이 잘 매칭되도록 임베딩 방식에 언어 구조를 활용
- **Parsing**: 문장을 단어 조각들로 분해하고, 단어간의 관계를 파악하여 문법적 구성을 이해
- 명사구(Noun Phrase)는 객체와 속성을 포함



Constituency Tree

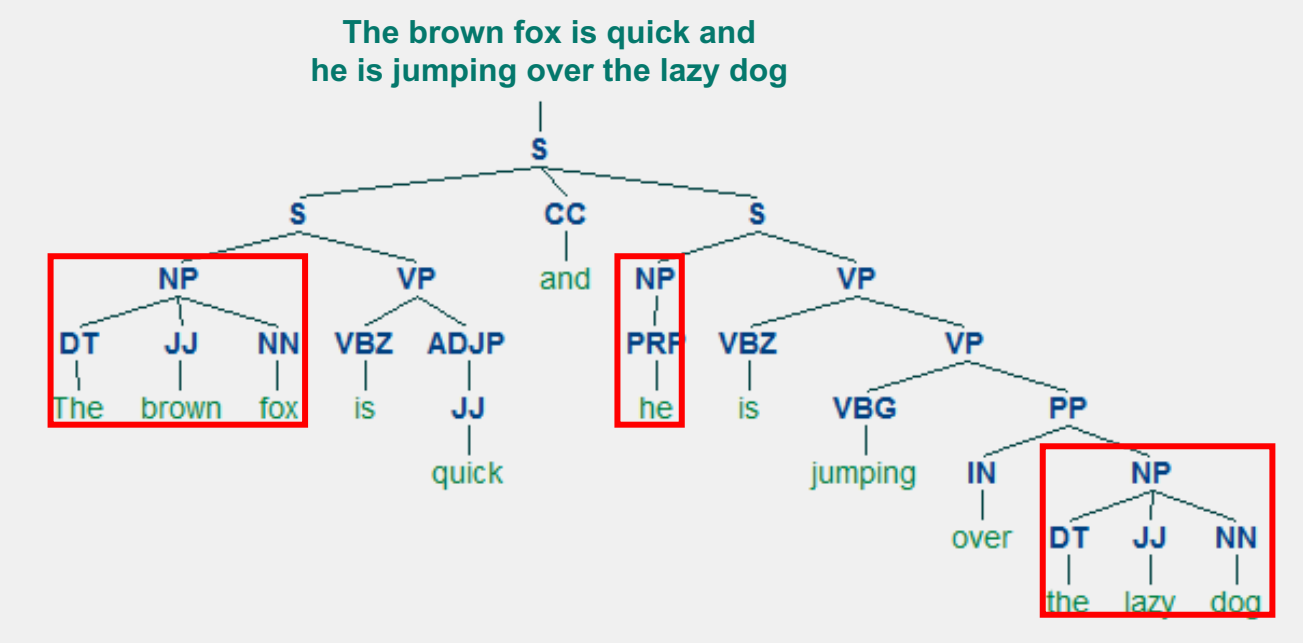
https://nltk.sourceforge.net/doc/en/ch07.html

Enhancing Prompt Understanding

Structure Diffusion

❖ 프롬프트 문법적 구성 파악: Constituency Tree

- 각 객체와 속성이 잘 매칭되도록 임베딩 방식에 언어 구조를 활용
- **Parsing**: 문장을 단어 조각들로 분해하고, 단어간의 관계를 파악하여 문법적 구성을 이해
- **명사구(Noun Phrase)**는 객체와 속성을 포함



Constituency Tree

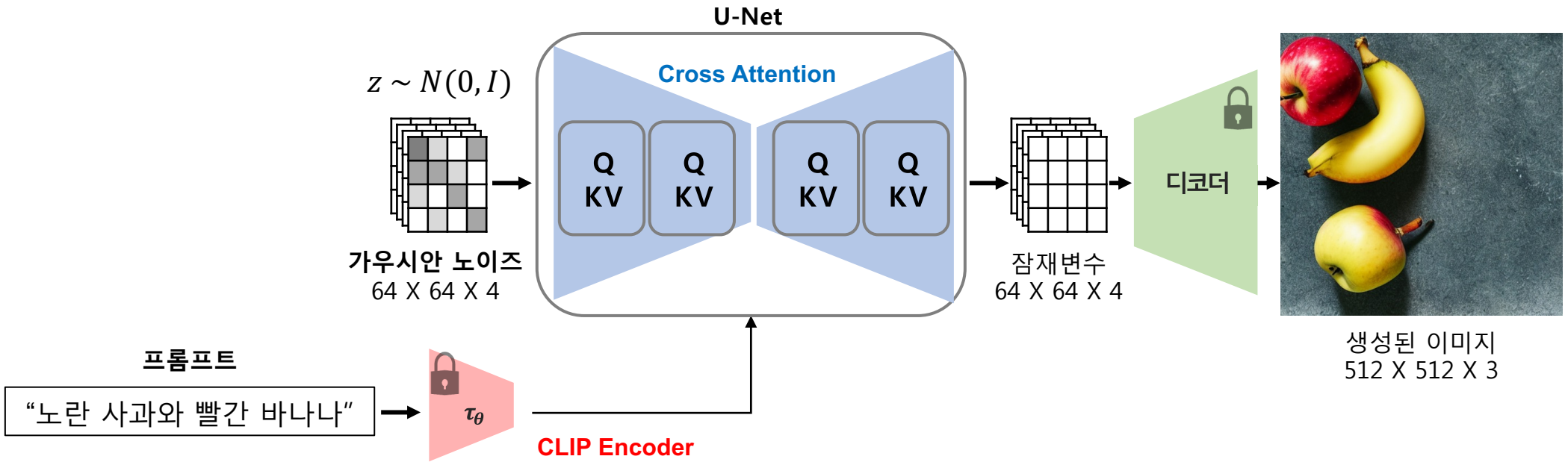
https://nltk.sourceforge.net/doc/en/ch07.html

Enhancing Prompt Understanding

Structure Diffusion

❖ LDM에서 프롬프트 임베딩 방식

- **CLIP Encoder** : 프롬프트를 임베딩해주는 텍스트 인코더
- Transformer 기반으로 텍스트와 이미지의 joint embeddding을 구하는 텍스트 인코더

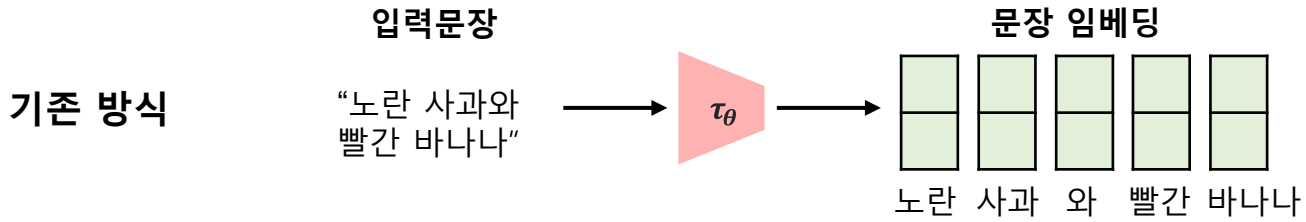


Enhancing Prompt Understanding

Structure Diffusion

❖ 임베딩 방식 차이 – 기존 방식

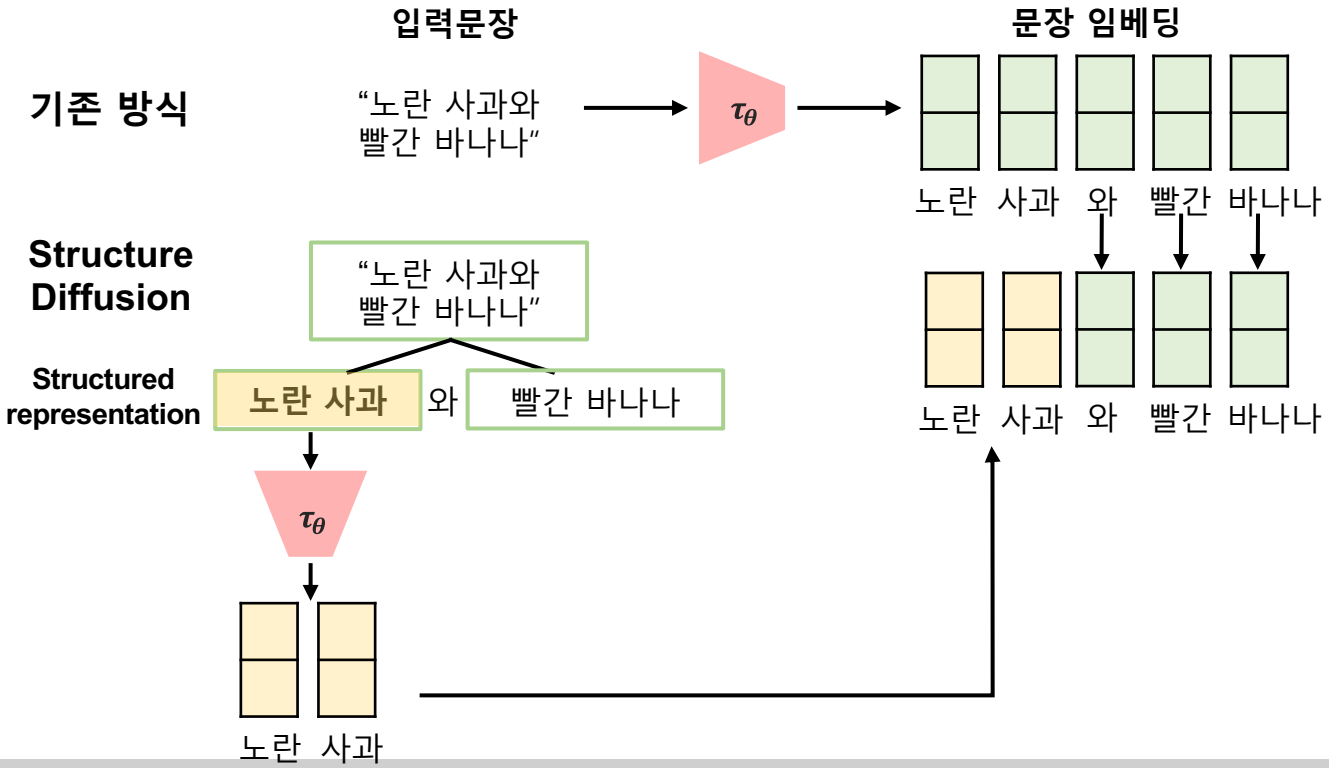
- **기존 방식**: 문장내 모든 단어들을 한번에 임베딩
- 문장의 **전체적인 의미를 담고 있는 임베딩** (단어와 단어의 관계 등)



Enhancing Prompt Understanding
Structure Diffusion

❖ 임베딩 방식 차이 – Structure Diffusion

- **Structure Diffusion:** 문장을 parsing한 후에 **명사구**(객체와 속성)는 **별도로 임베딩**
- **객체와 속성이 잘 매칭**되어 attribute binding 문제 완화

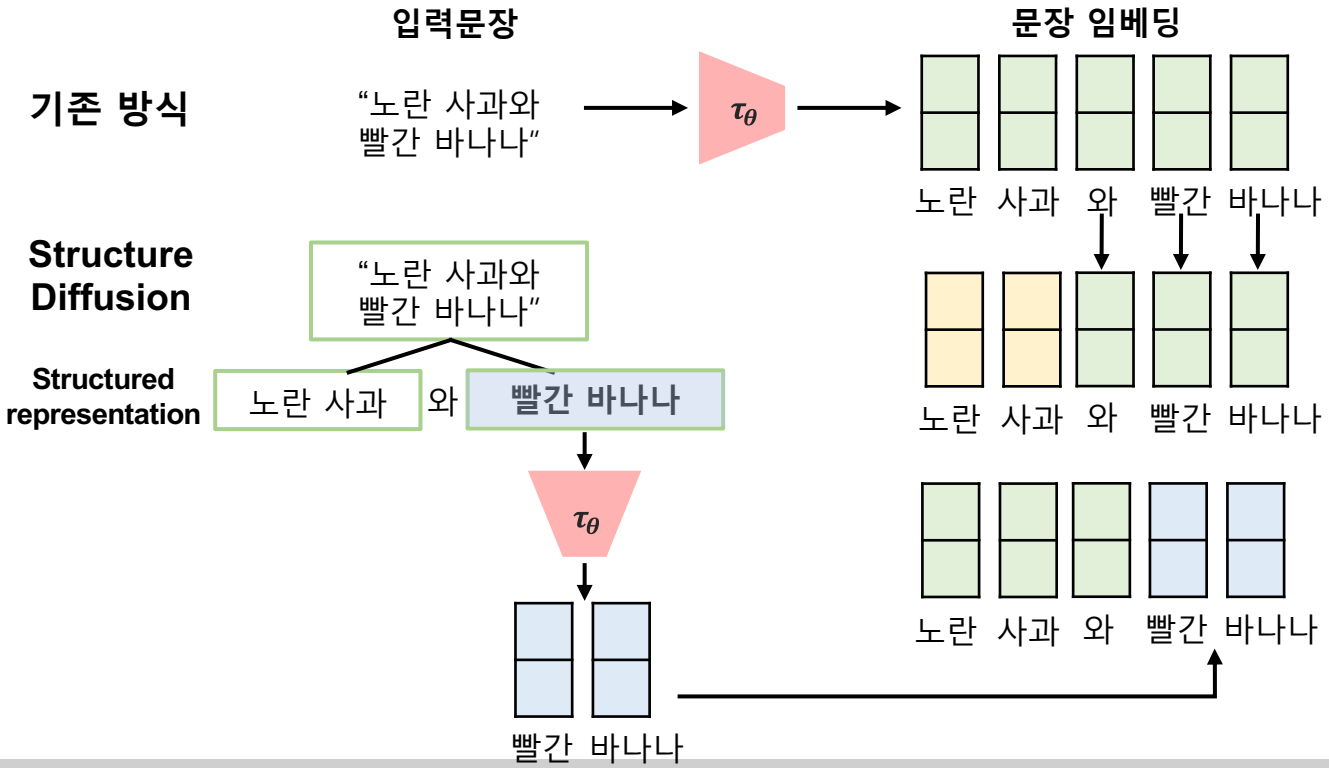


Enhancing Prompt Understanding

Structure Diffusion

❖ 임베딩 방식 차이 – Structure Diffusion

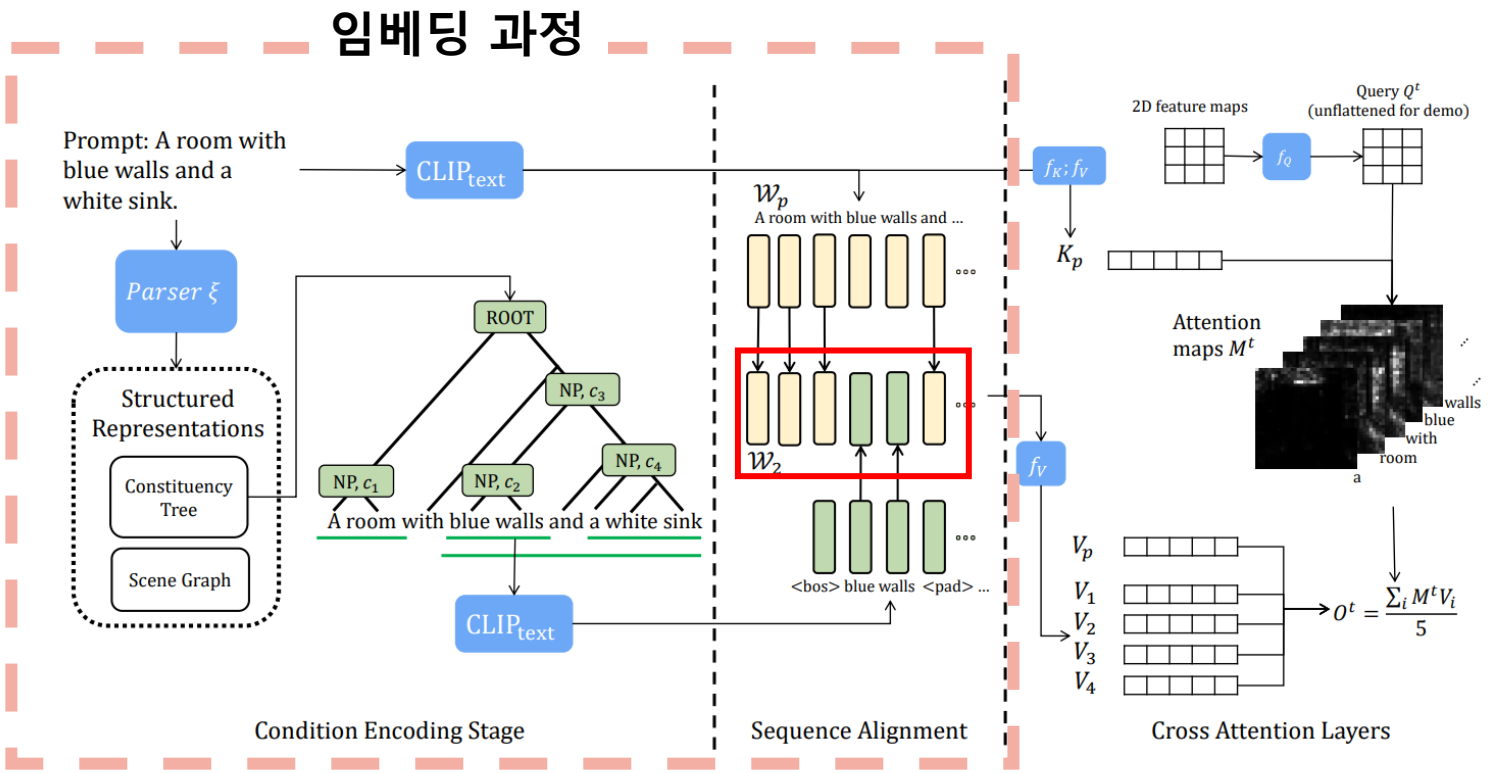
- **Structure Diffusion:** 문장을 parsing한 후에 **명사구**(객체와 속성)는 **별도로 임베딩**
- **객체와 속성이 잘 매칭**되어 attribute binding 문제 완화



Enhancing Prompt Understanding
Structure Diffusion

❖ Structure Diffusion 임베딩 과정

- **명사구**에 대한 임베딩은 별도로 진행하여 전체 문장 임베딩에 대체
- **새로운 임베딩**은 객체과 속성의 관계를 잘 반영



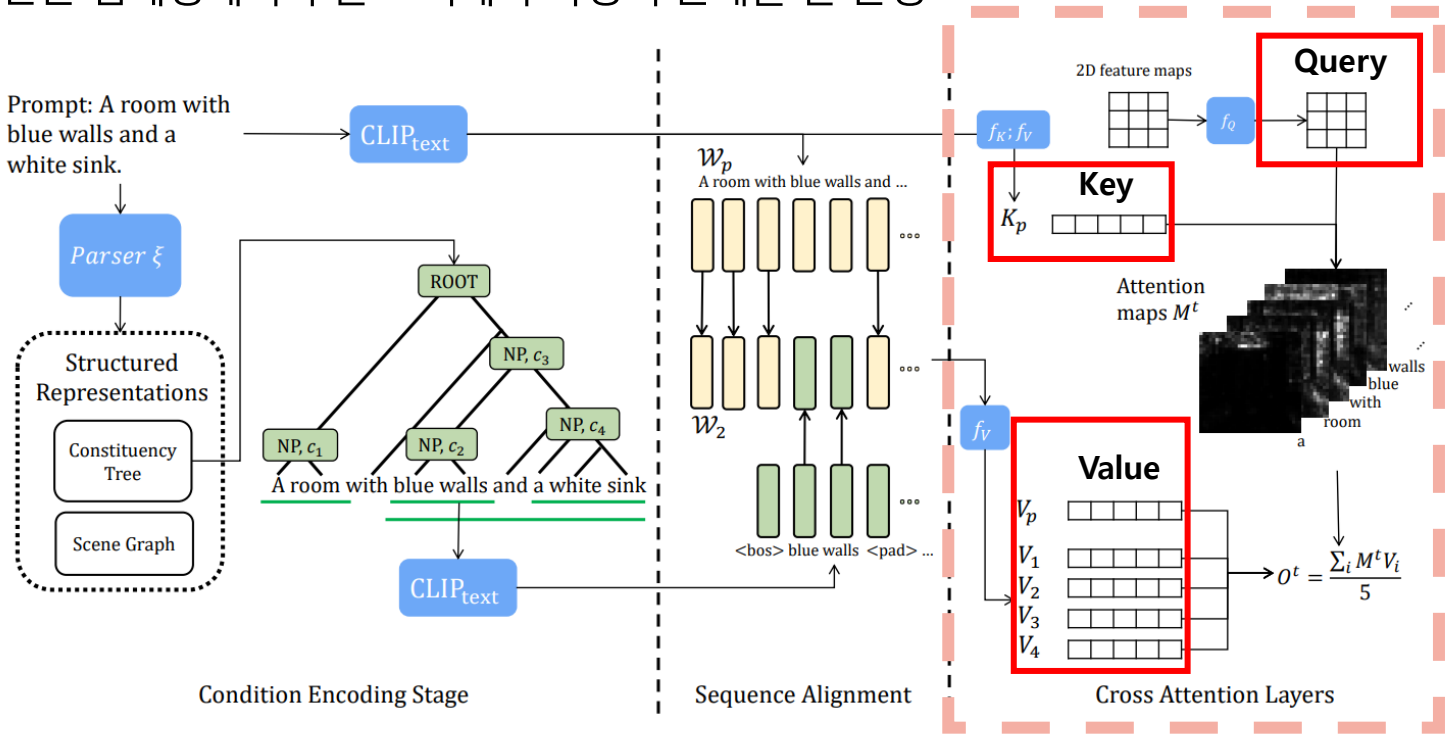
Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).

Enhancing Prompt Understanding

Structure Diffusion

❖ Cross Attention Layers

- **Query:** 이미지로부터 추출
- **Key:** 전체 프롬프트 임베딩에서 추출 → 문장의 전체적인 정보(layout과 composition)
- **Value:** 새로 만든 임베딩에서 추출 →객체와 특성의 관계를 잘 반영



Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).

Enhancing Prompt Understanding
Structure Diffusion

❖ 실험결과

- Structure Diffusion 활용 여부에 따라 생성된 이미지를 비교
- Attribute Binding, Missing Object 문제 해결

Stable Diffusion | **With Structure Diffusion**

Prompt: 빨간 자동차와 하얀 양

< Attribute binding >

Stable Diffusion | **With Structure Diffusion**

Prompt: 파란색 가방과 갈색 코끼리

< Missing Object >

Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).

Enhancing Prompt Understanding

Structure Diffusion

❖ 실험결과

- 두개의 객체 그리고 각 객체의 색상을 명시한 프롬프트로 생성된 이미지를 비교
- Stable Diffusion은 물론 Composable Diffusion 보다 우수한 성능

Stable Diffusion | **Composable Diffusion** | **With Structure Diffusion**

Prompt: 빨간 새와 초록 사과

< Attribute binding >

Methods	CC-500 (Prompt format: "*a* [*colorA*] [*objectA*] and *a* [*colorB*] [*objectB*]")					
	Human Annotations			GLIP		Human-GLIP Consistency
	Zero/One obj. (↓)	Two obj.	Two obj. w/ correct colors	Zero/One obj. (↓)	Two obj.	
Stable Diffusion	65.5	34.5	19.2	69.0	31.0	46.4
Composable Diffusion	69.7	30.3	20.6	74.2	25.8	48.9
StructureDiffusion (Ours)	**62.0**	**38.0**	**22.7**	**68.8**	**31.2**	47.6

< 정량적 평가 >

Feng, Weixi, et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis." *arXiv preprint arXiv:2212.05032* (2022).

Enhancing Prompt Understanding

Structured Guidance	Attend-and-Excite
TRAINING-FREE STRUCTURED DIFFUSION GUIDANCE FOR COMPOSITIONAL TEXT-TO-IMAGE SYNTHESIS (ICLR 2023)	Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models (ACM-TOG 2023)

Enhancing Prompt Understanding

Attend-and-Excite

❖ **Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models (ACM-TOG 2023)**

- **Motivation:** 객체들이 생성되지 않는 문제 - missing object
- 프롬프트 내 단어들의 영향력이 작아서 그런거 아닐까?

Prompt: 말과 **강아지**

Prompt: **왕관** 쓴 사자

Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-10.

Enhancing Prompt Understanding

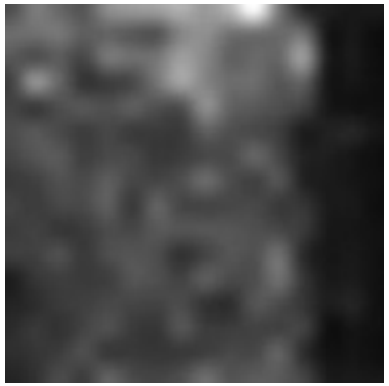
Attend-and-Excite

❖ Cross attention map

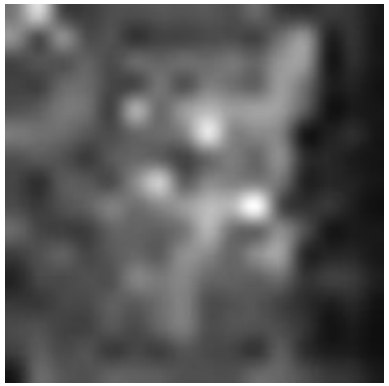
- 각 단어가 이미지 어느 부분에 영향을 주는지 확인 (하얀색일수록 강한 영향력)
- 특정 단어의 **영향력이 약하게** 나타나 해당 객체가 생성되지 않는 문제
- 그렇다면 영향력이 약한 단어의 cross attention map을 변형해 영향력을 높이자

Prompt: **왕관** 쓴 사자

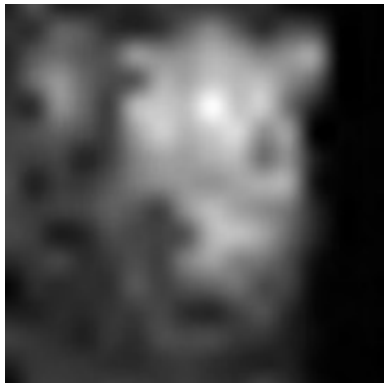
생성된 이미지



<왕관>



<쓴>



<사자>

Cross attention map

Enhancing Prompt Understanding

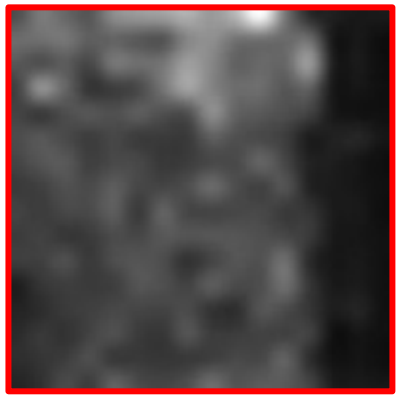
Attend-and-Excite

❖ **Cross attention map**

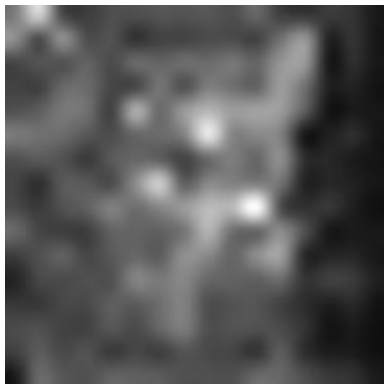
- 각 단어가 이미지 어느 부분에 영향을 주는지 확인 (하얀색일수록 강한 영향력)
- 특정 단어의 **영향력이 약하게** 나타나 해당 객체가 생성되지 않는 문제
- **그렇다면 영향력이 약한 단어의 cross attention map을 변형해 영향력을 높이자**

Prompt: **왕관** 쓴 사자

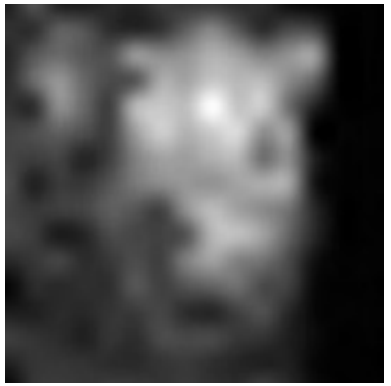
생성된 이미지



<왕관>



<쓴>



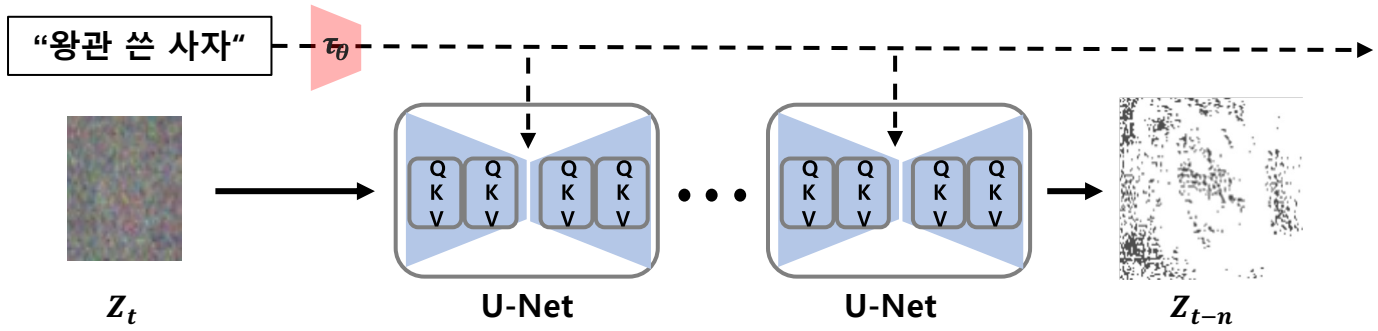
<사자>

Cross attention map

Enhancing Prompt Understanding

Attend-and-Excite

❖ 디퓨젼 모델 생성과정

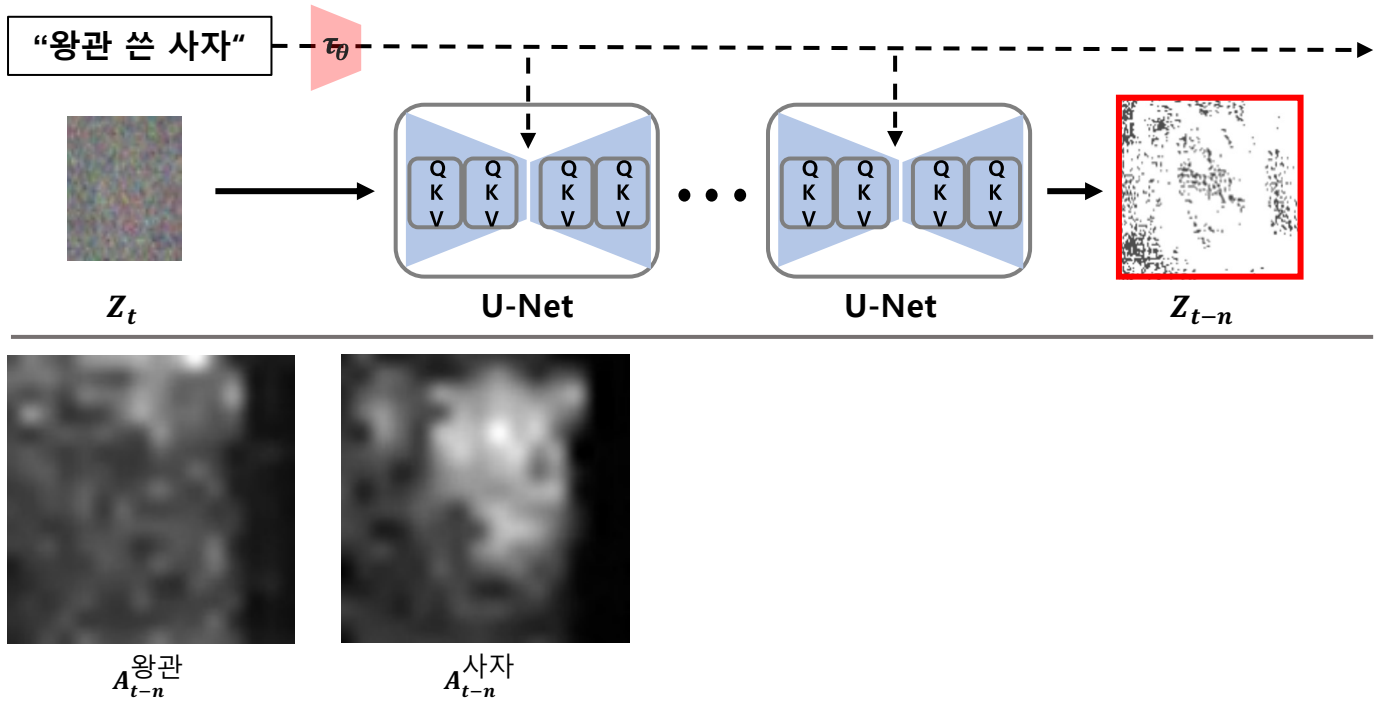


Enhancing Prompt Understanding

Attend-and-Excite

❖ **1단계: Cross attention map 추출**

- 이미지 생성과정에서 각 **명사의** cross attention map 추출
- **'왕관'**에 대한 영향력이 약하게 나타남

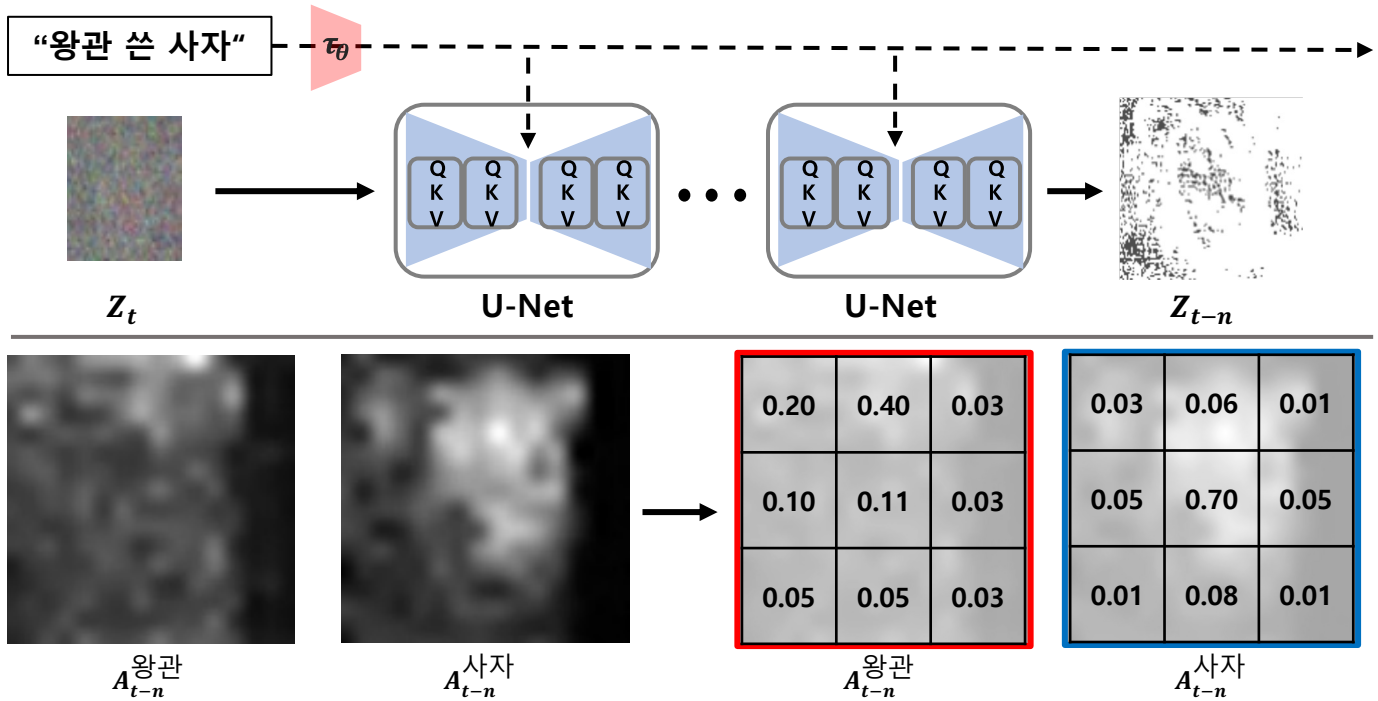


Enhancing Prompt Understanding

Attend-and-Excite

❖ **2단계: Cross attention map max값 비교**

- 어떤 객체가 생성되지 않는지 판별
- 최종적으로는 생성되지 않는 객체의 영향력을 높이기 위함

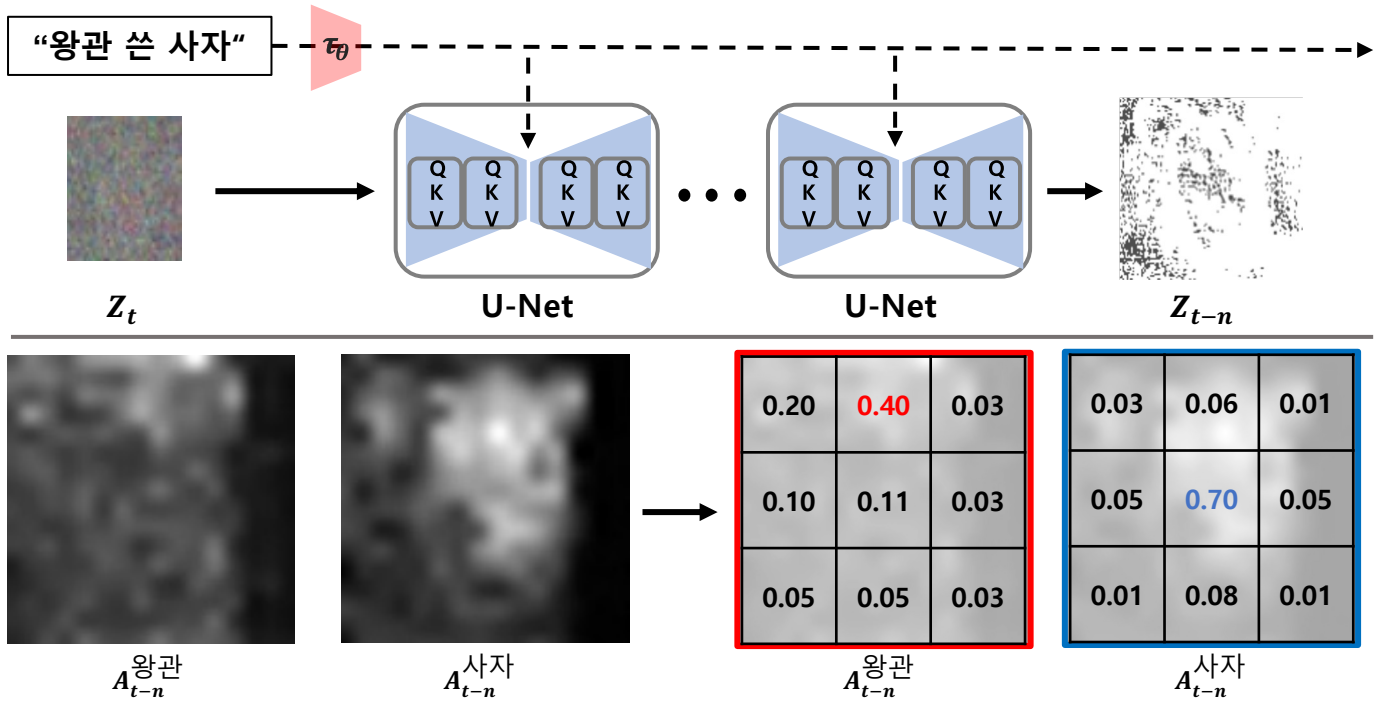


Enhancing Prompt Understanding

Attend-and-Excite

❖ 2단계: Cross attention map max값 비교

- Cross attention map의 **max** 값을 활용해 각 명사의 loss 산출
- Max 값이 작은 명사, 즉 영향력이 낮은 명사가 선택됨

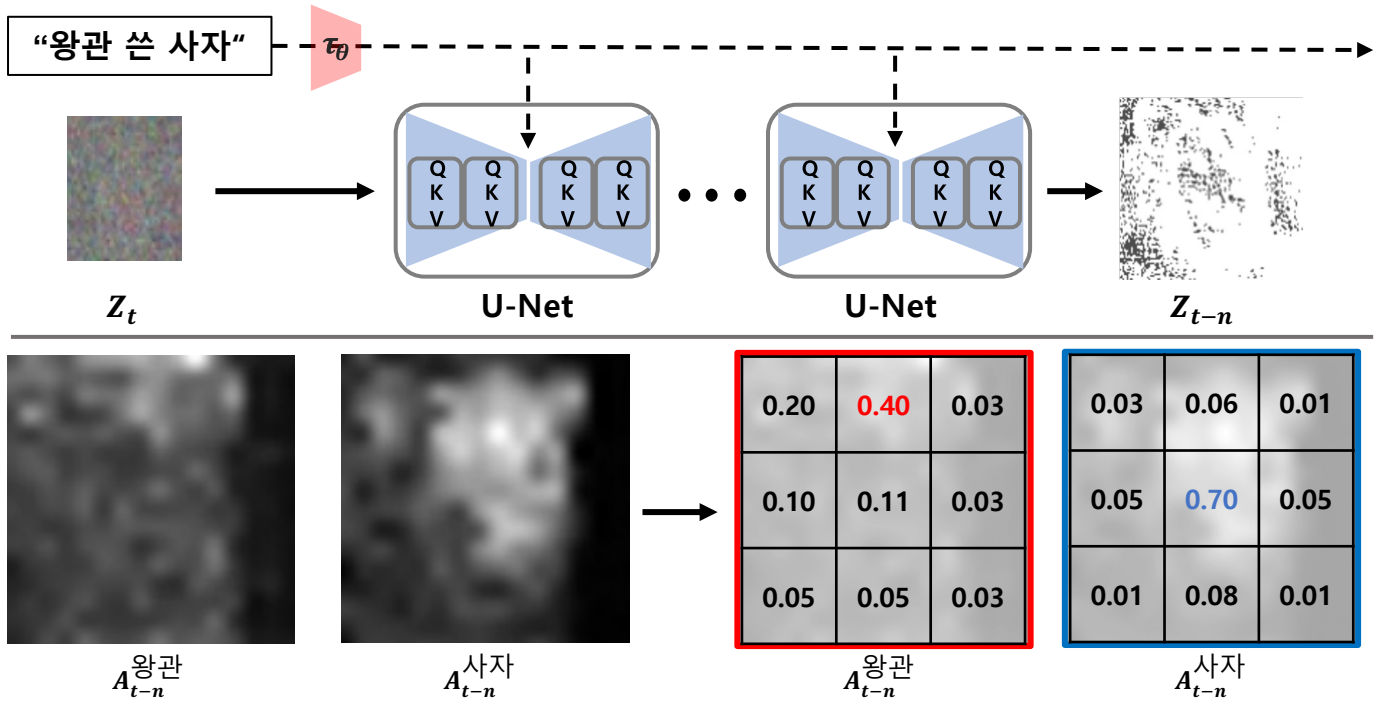


Enhancing Prompt Understanding

Attend-and-Excite

❖ 2단계: Cross attention map max값 비교

- Cross attention map의 **max** 값을 활용해 각 명사의 loss 산출
- Max 값이 작은 명사, 즉 영향력이 낮은 명사가 선택됨



$$L_{왕관} = 1 - \max(A_{t-n}^{왕관}) = 1 - 0.40 = 0.60$$

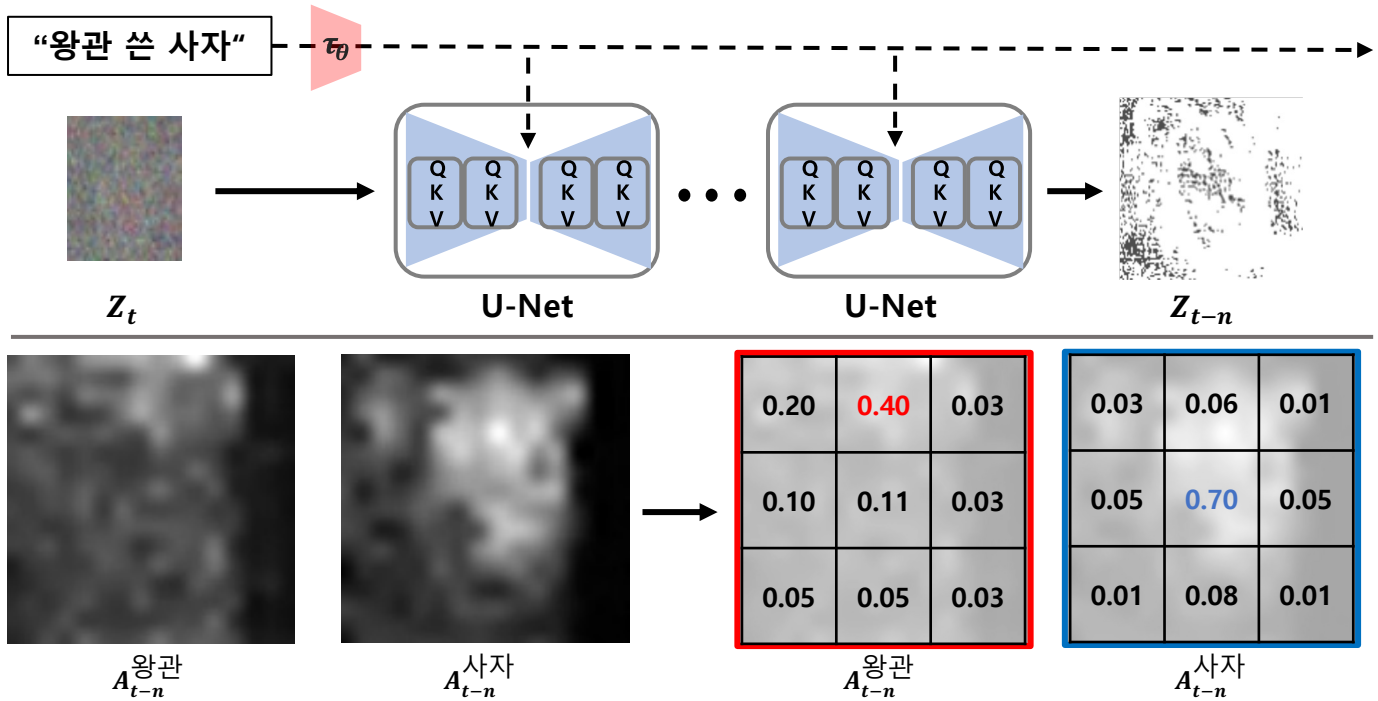
$$L_{사자} = 1 - \max(A_{t-n}^{사자}) = 1 - 0.70 = 0.30$$

Enhancing Prompt Understanding

Attend-and-Excite

❖ 2단계: Cross attention map max값 비교

- Cross attention map의 **max** 값을 활용해 각 명사의 loss 산출
- Max 값이 작은 명사, 즉 **영향력이 낮은 명사**가 선택됨



$$L_{왕관} = 1 - \max(A_{t-n}^{왕관}) = 1 - 0.40 = 0.60$$

$$L_{사자} = 1 - \max(A_{t-n}^{사자}) = 1 - 0.70 = 0.30$$

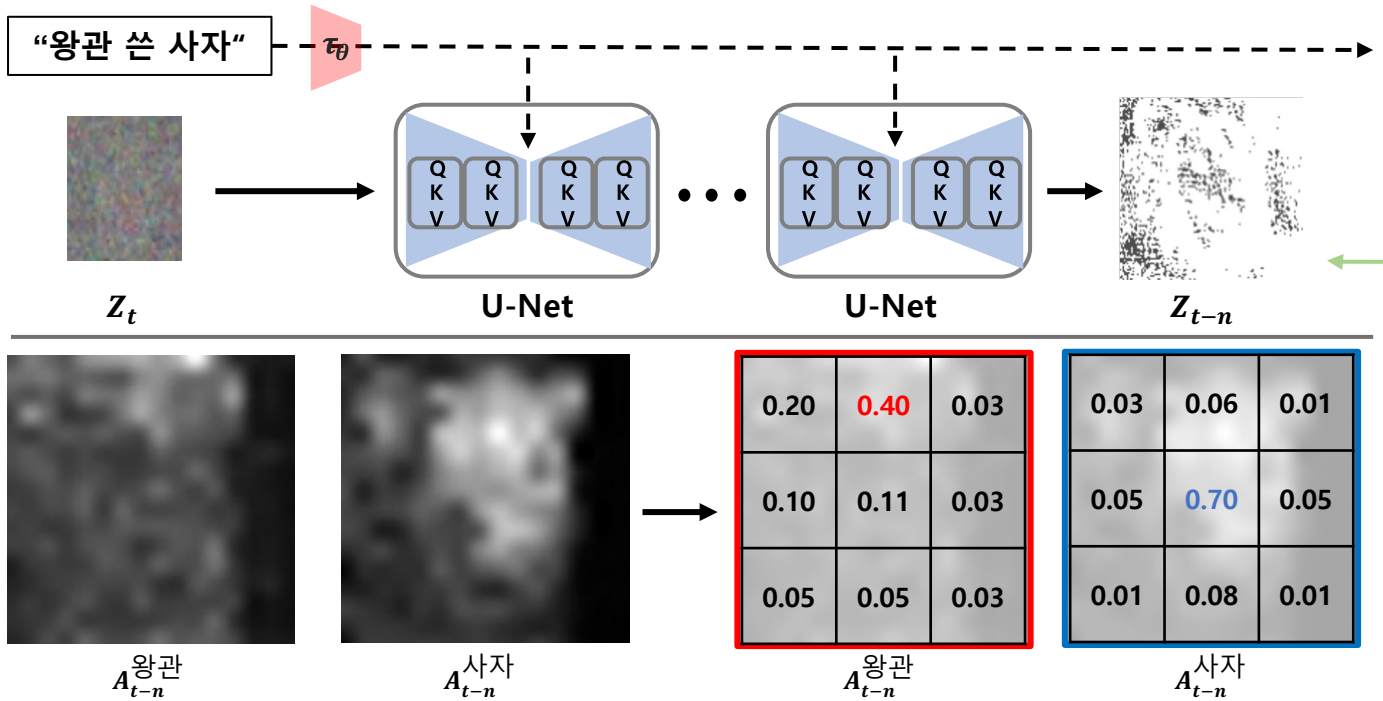
$$L = \max(L_{왕관}, L_{사자}) = L_{왕관} = 0.6$$

Enhancing Prompt Understanding

Attend-and-Excite

❖ 3단계: Latent 업데이트

- 영향력이 적은 명사가 생성되도록 latent를 업데이트
- **'왕관'**에 대한 영향력이 증가
- **그러나 이러한 방식은 하나의 patch에 대해서만 업데이트**



$$L_{왕관} = 1 - \max(A^{왕관}_{t-n}) = 1 - 0.40 = 0.60$$

$$L_{사자} = 1 - \max(A^{사자}_{t-n}) = 1 - 0.70 = 0.30$$

$$L = \max(L_{왕관}, L_{사자}) = 0.6$$

$$update: Z'_{t-n} = Z_{t-n} - \alpha \nabla_t L$$

Enhancing Prompt Understanding

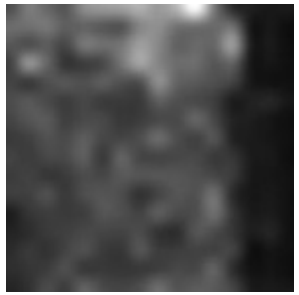
Attend-and-Excite

❖ 가우시안 필터 효과

- 업데이트 시 인접한 패치들에도 영향
- Attention value가 인접한 패치와의 **선형 조합**으로 생성

$update: Z'_{t-n} = Z_{t-n} - \alpha \nabla_t L$

가우시안 필터 사용 X



$A^{왕관}_{t-n}$

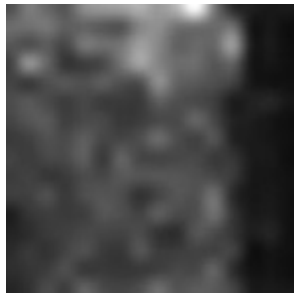
0.20	0.40	0.03
0.10	0.11	0.03
0.05	0.05	0.03

$A^{왕관}_{t-n}$

X_{t-n}

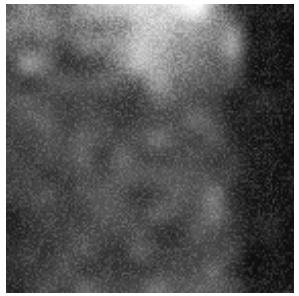
X'_{t-n}

가우시안 필터 사용



$A^{왕관}_{t-n}$

가우시안 필터



$G(A^{왕관}_{t-n})$

0.130	0.050	0.001
0.012	0.010	0.040
0.003	0.006	0.001

$G(A^{왕관}_{t-n})$

X_{t-n}

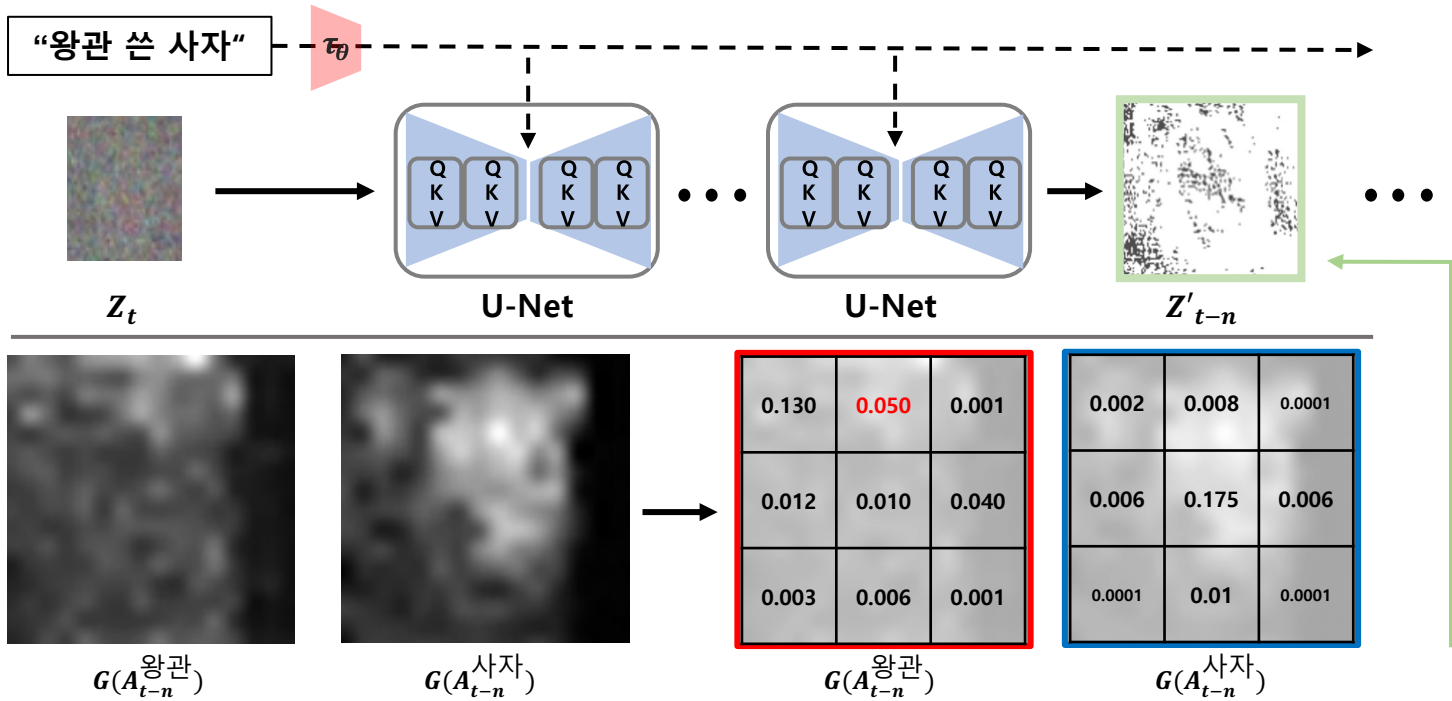
X'_{t-n}

Enhancing Prompt Understanding

Attend-and-Excite

❖ Attend-and-Excite

- **Attend:** 프롬프트 내 모든 명사들이 이미지에 생성
- **Excite:** 영향력이 작은 명사들에 대한 영향력을 높아짐
- Stable diffusion에 합쳐서 사용 가능하며, 생성 과정동안 업데이트 반복



Stable Diffusion

With Attend-and-Excite

$$L_{왕관} = 1 - \max(G(A_{t-n}^{왕관})) = 1 - 0.4 = 0.6$$

$$L_{사자} = 1 - \max(G(A_{t-n}^{사자})) = 1 - 0.7 = 0.3$$

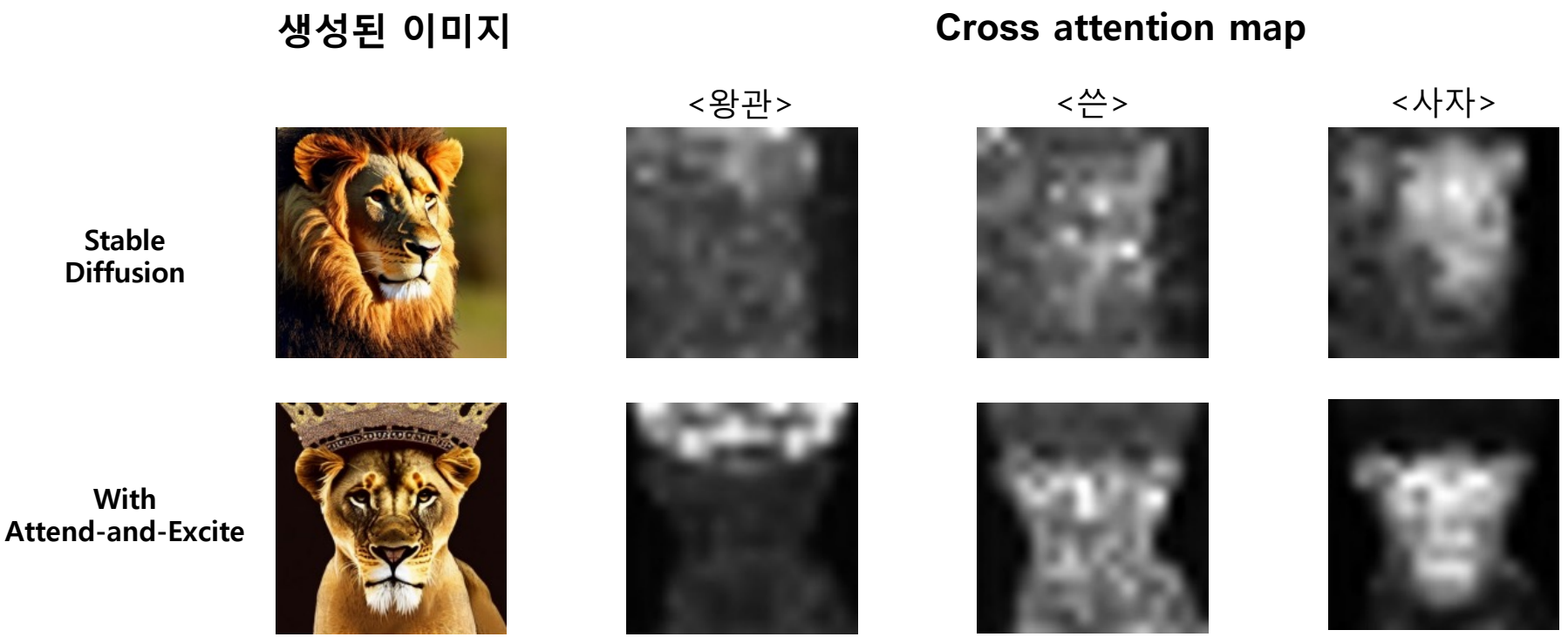
$$L = \max(L_{왕관}, L_{사자}) = 0.6$$

$$update: Z'_{t-n} = Z_{t-n} - \alpha \nabla_t L$$

Enhancing Prompt Understanding
Attend-and-Excite

❖ 실험 결과

- 매 timestep마다 cross attention map에 변형이 가해져 최종적으로 생성된 이미지는 기존 이미지와 다른 형태를 보임
- Attend-and-Excite를 사용하였을 때 '**왕관**'에 대한 영향력이 뚜렷해짐



Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-10.

Enhancing Prompt Understanding
Attend-and-Excite

❖ 실험 결과

- Missing object는 물론 attribute binding 문제도 완화

Stable Diffusion | **With Attend-and-Excite**

Prompt: 노란 그릇과 파란 고양이


< Missing object >

Stable Diffusion | **With Attend-and-Excite**

Prompt: 노란 리본과 갈색 벤치

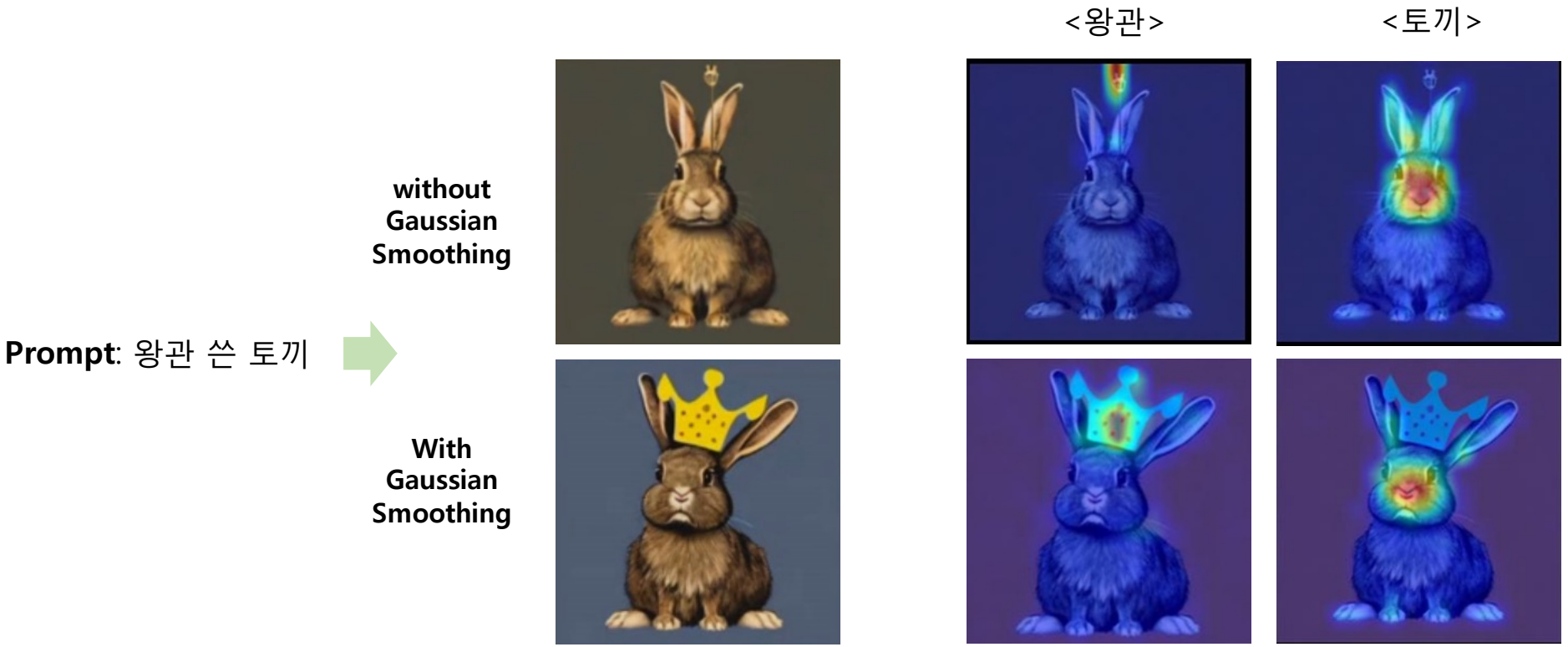

< Attribute binding >

Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-10.

Enhancing Prompt Understanding
Attend-and-Excite

❖ 실험 결과

- Gaussian Filter를 사용하였을 때 객체가 더 강한 영향력을 가짐



Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-10.

Conclusion

❖ **Stable Diffusion의 문제점**

- 프롬프트를 완벽하게 이해하기는 어려움
- 이에 따라 attribute binding 그리고 missing object 문제 발생

❖ **Structure Guidance**

- 객체와 속성이 잘못 매칭되는 **Attribute binding** 문제에 집중
- **언어 구조**를 활용해 디퓨젼 모델의 임베딩과 cross attention을 변형

❖ **Attend-and-Excite**

- 객체가 생성되지 않는 **Missing object** 문제를 해결하기 위해 고안된 방법론
- 영향력이 약한 명사의 **attention value를 최대화**하는 방향으로 latent 업데이트

고맙습니다