Controllable Diffusion Models

2024. 02. 16

Data Mining & Quality Analytics Lab.

윤지현

발표자 소개

❖ 윤지현 (Jihyun Yun)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석사 과정 (2023. 09 ~ Present)

❖ Research Interest

- Generative Model
- Diffusion Model

❖ Contact

- whle56@korea.ac.kr

Contents

- Introduction
- Preliminary Study
- Controllable Diffusion Model
  - ControlNet
  - Composer
  - Uni-ControlNet
- Conclusions
- References

1. Introduction

Introduction

❖ Diffusion – Intuitive Understanding

- Diffusion : 확산
- 공간상에 모여 있던 분자들이 전 공간에 고르게 분포하게 되는 현상

시간이 지나면 잉크는 uniform 하게 분포하게 됨

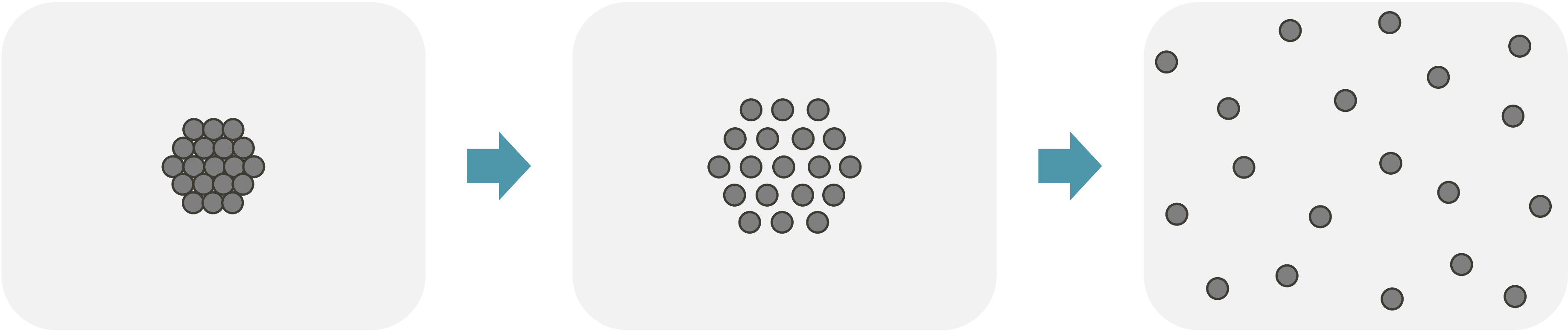
Introduction

❖ Diffusion – Intuitive Understanding

- Diffusion Process

  – 분자의 움직임 : Gaussian Distribution을 따름

  => 분자들의 다음 위치는 가우시안 분포 안에서 결정 됨



Introduction

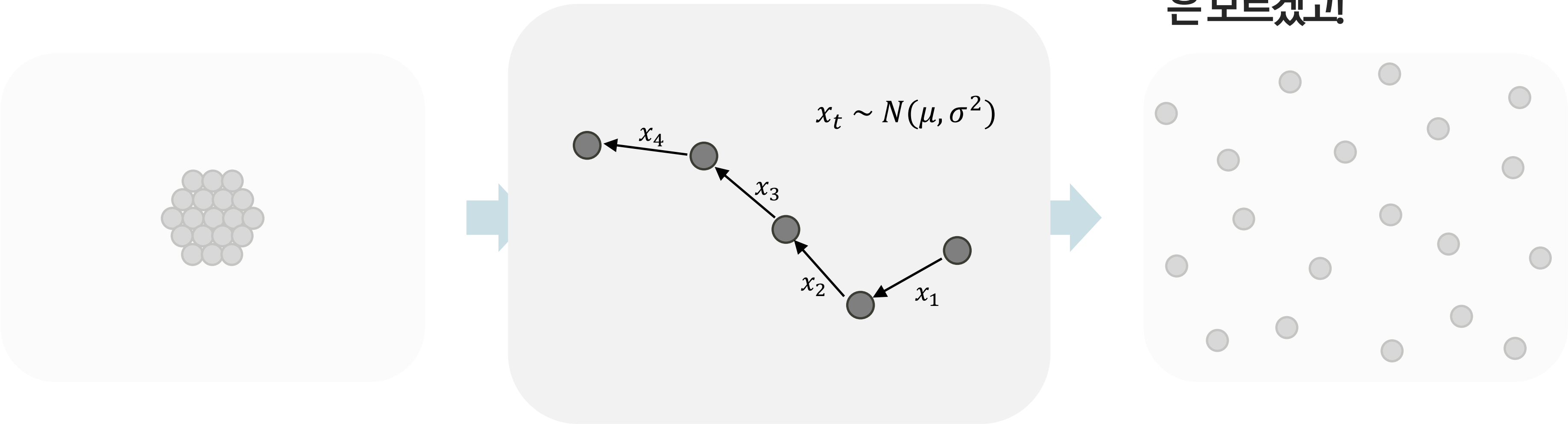
❖ Diffusion – Intuitive Understanding

- Reverse Process

  - 분자의 움직임 : Gaussian Distribution을 따름

균일하게 퍼진 잉크를 원래대로 돌릴 방법?

은 모르겠고!

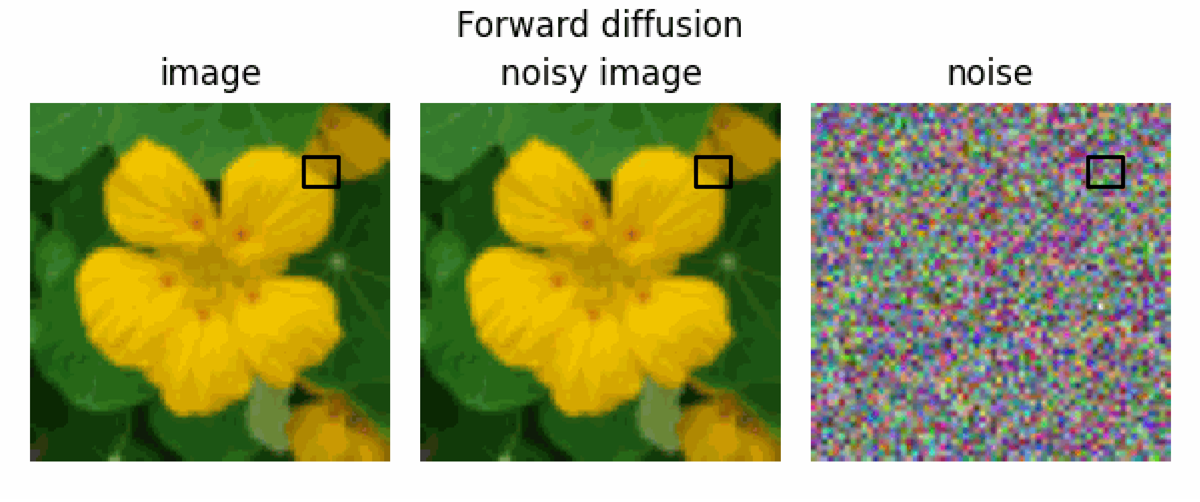


아주 짧은 시간 동안의 분자의 움직임을 안다면?

Introduction

❖ Diffusion – Generative Model

- 잉크가 물 안에서 서서히 퍼지는 것 = 이미지에 점점 노이즈가 더해져 완전한 노이즈가 되는 것
- Image의 pixel 값에 정규 분포를 따르는 noise를 추가한다고 생각!

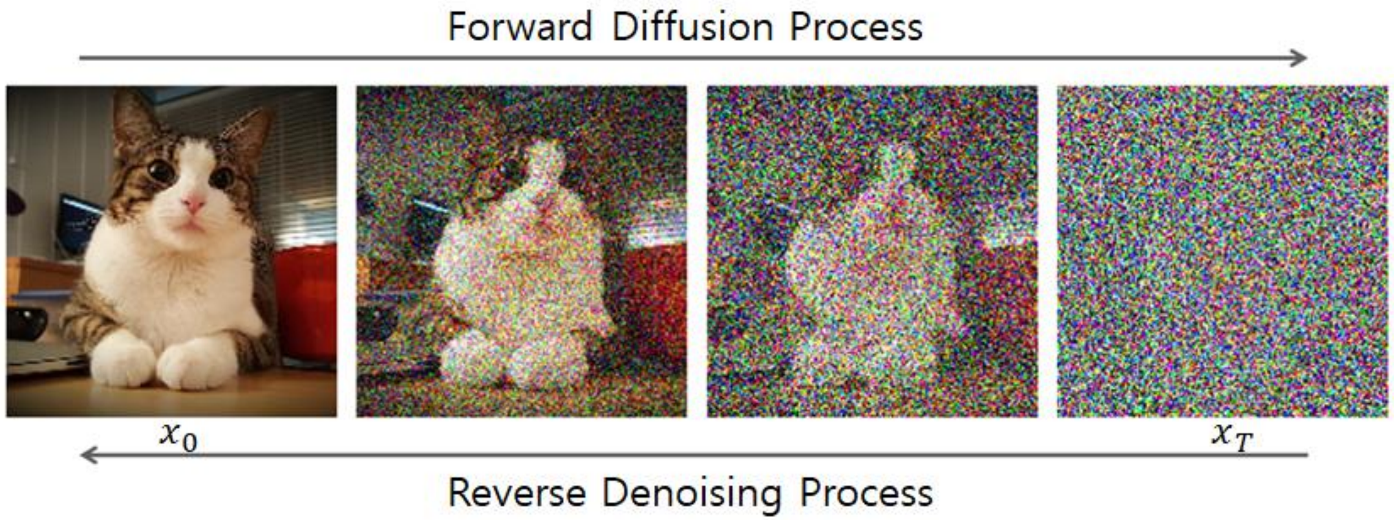


pixel noise(ε)

Introduction

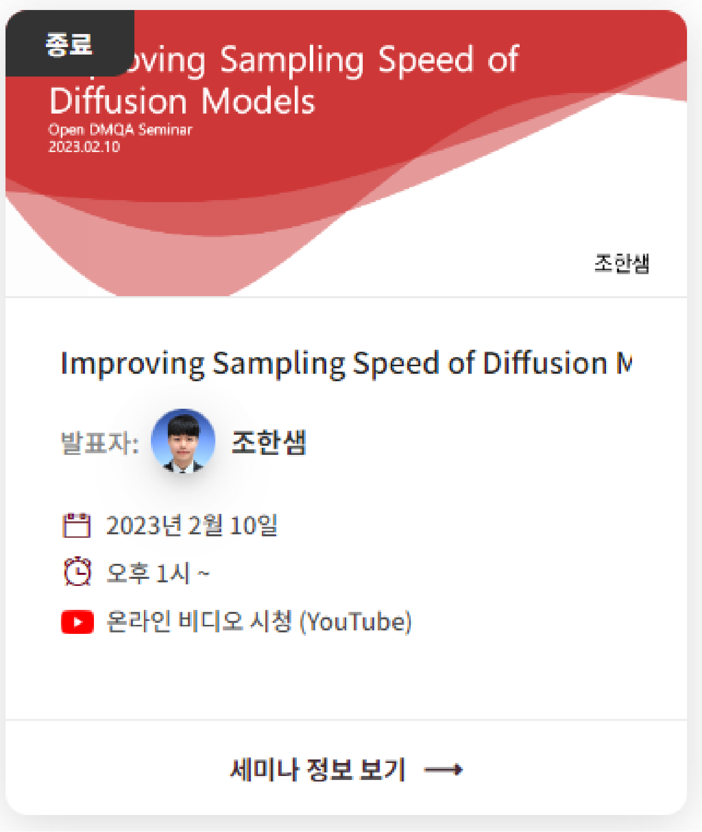
❖ Diffusion Model

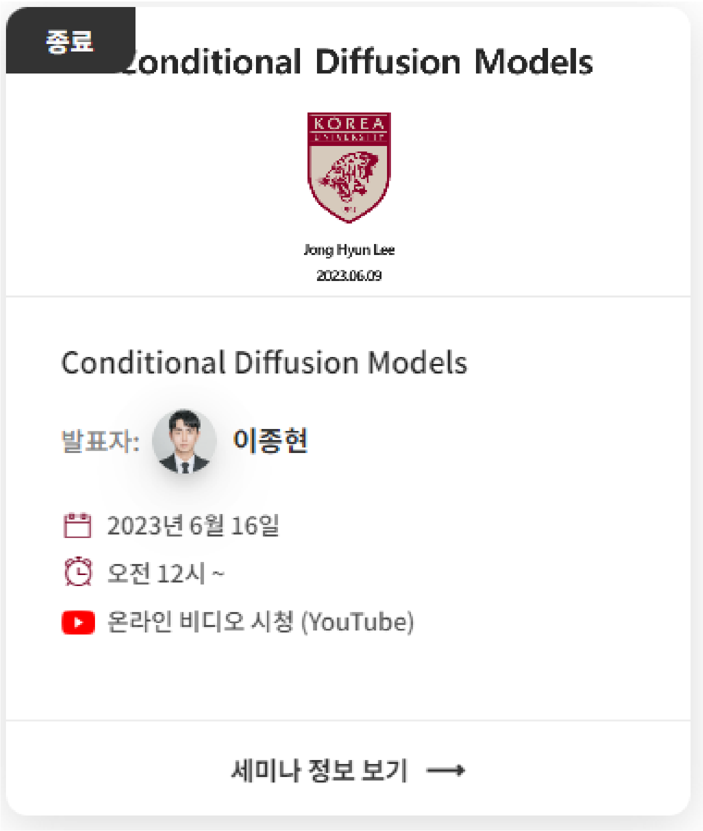
- 이미지 생성 모델 중 하나로써 입력 이미지와 유사한 확률 분포를 가진 결과 이미지를 생성하는 모델
- Forward Diffusion Process : 이미지에 고정된(fixed) gaussian noise를 더해 주는 과정
- Reverse Denoising Process : noise로부터 data를 복원하는 과정



Introduction

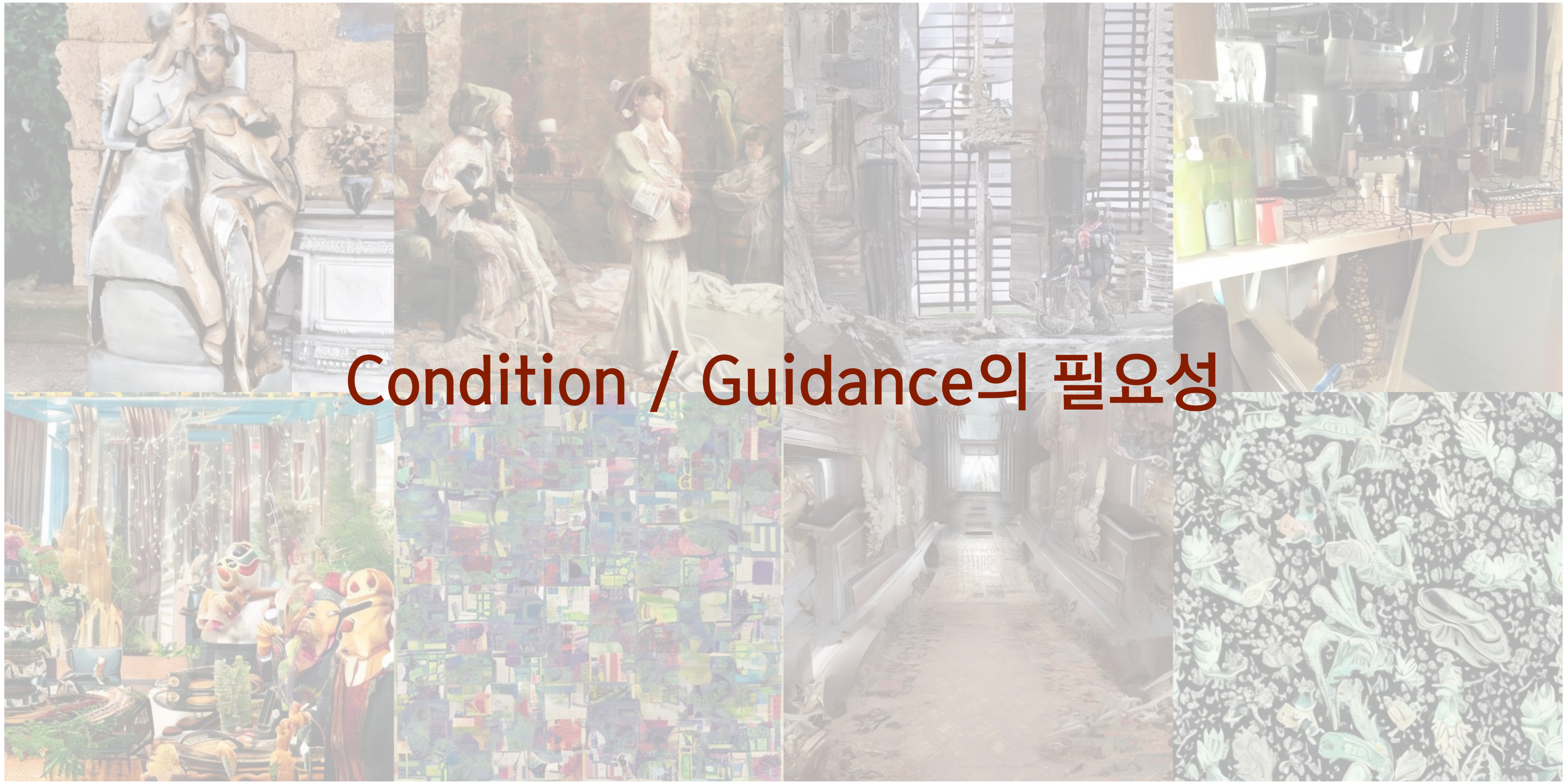
❖ Recommended Seminar





Introduction

❖ Diffusion can generate almost everything but..



Condition / Guidance의 필요성

2. Preliminary Study

Text-to-Image Diffusion

Conditional Diffusion Models

❖ GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

- 이미지를 설명하는 text(condition)을 condition으로 받음
- Classifier free guidance와 CLIP guidance를 활용하여 원하는 표현이 이미지에 잘 반영되도록 만들어줌

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

Alex Nichol * Prafulla Dhariwal * Aditya Ramesh * Pranav Shyam Pamela Mishkin Bob McGrew Ilya Sutskever Mark Chen

Abstract

Diffusion models have recently been shown to generate high-quality synthetic images, especially when paired with a guidance technique to trade off diversity for fidelity. We explore diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance. We find that the latter is preferred by human evaluators for both photorealism and caption similarity, and often produces photorealistic samples. Samples from a 3.5 billion parameter text-conditional diffusion model using classifier-free guidance are favored by human evaluators to those from DALL-E, even when the latter uses expensive CLIP reranking. Additionally, we find that our models can be fine-tuned to perform image inpainting, enabling powerful text-driven image editing. We train a smaller model on a filtered dataset and release the code and weights at https://github.com/openai/glide-text2im.

their corresponding text prompts.

On the other hand, unconditional image models can synthesize photorealistic images (Brock et al., 2018; Karras et al., 2019a;b; Razavi et al., 2019), sometimes with enough fidelity that humans can't distinguish them from real images (Zhou et al., 2019). Within this line of research, diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020b) have emerged as a promising family of generative models, achieving state-of-the-art sample quality on a number of image generation benchmarks (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho et al., 2021).

To achieve photorealism in the class-conditional setting, Dhariwal & Nichol (2021) augmented diffusion models with *classifier guidance*, a technique which allows diffusion models to condition on a classifier's labels. The classifier is first trained on noised images, and during the diffusion sampling process, gradients from the classifier are used to guide the sample towards the label. Ho & Salimans (2021) achieved similar results without a separately trained classifier through the use of *classifier-free guidance*, a form of guidance that interpolates between predictions from a

"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

[1] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., ... & Chen, M. (2022, June). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning* (pp. 16784-16804). PMLR.

Text to Image Diffusion

Conditional Diffusion Models

❖ Imagen : Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

- 높은 수준의 photorealism과 깊은 수준의 언어 이해를 갖춘 text-to-image 모델
- LLM의 텍스트 임베딩이 text-to-image 합성에 매우 효과적이라고 제시

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac,williamchan,mnorouzi}@google.com
{srbs,lala,jwhang,jonathanho,davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada

Abstract

We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g. T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, GLIDE and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment. See imagen.research.google for an overview of the results.

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

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

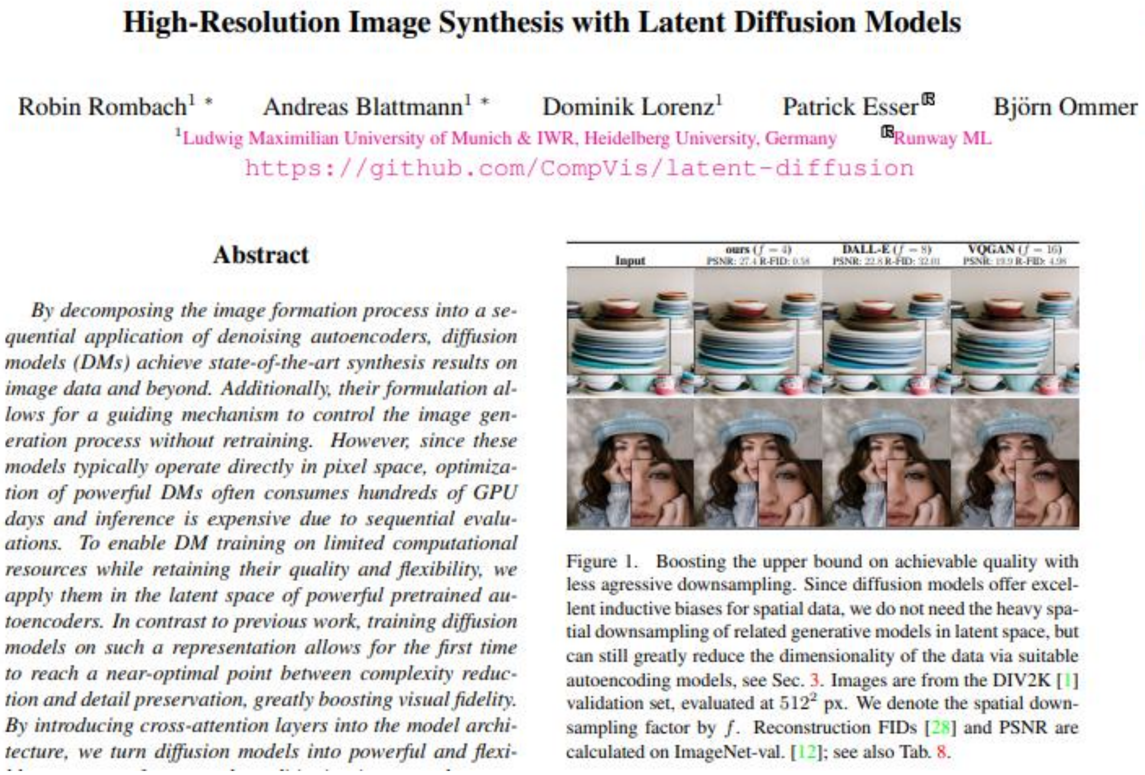
[2] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, *35*, 36479-36494.

Stable Diffusion

Conditional Diffusion Models

❖ High-Resolution Image Synthesis with Latent Diffusion Models

- Pixel 차원이 아닌 Latent Embedding을 학습하는 diffusion model
- Text가 아닌 condition도 입력으로 받을 수 있음



High-Resolution Image Synthesis with Latent Diffusion Models

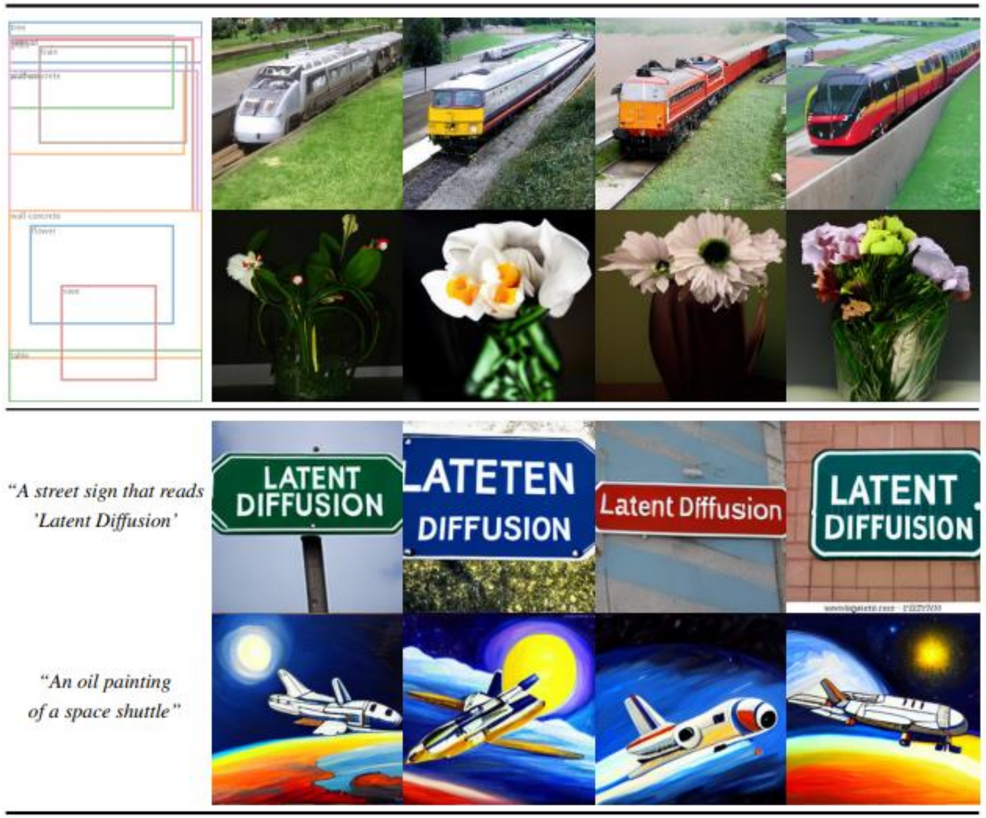
Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser Björn Ommer
[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany Runway ML
https://github.com/CompVis/latent-diffusion

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexi-

Input | ours ($f=4$) PSNR: 27.4 R-FID: 0.58 | DALL-E ($f=8$) PSNR: 22.8 R-FID: 32.01 | VQGAN ($f=16$) PSNR: 19.9 R-FID: 4.98

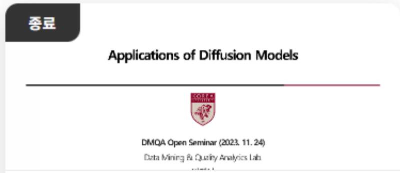
Figure 1. Boosting the upper bound on achievable quality with less agressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f. Reconstruction FIDs [28] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.



[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

Introduction

❖ Recommended DMQA Seminar



Applications of Diffusion Models

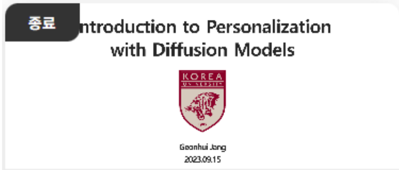
발표자: 박태남

2023년 11월 24일

오후 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶



Introduction to Personalization with Diffu

발표자: 장건희

2023년 9월 15일

오후 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Diffusion-based Anomaly Detection

발표자: 안시후

2024년 1월 12일

오후 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Accelerating Diffusion Models: Consisten

발표자: 조한샘

2023년 12월 15일

오후 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

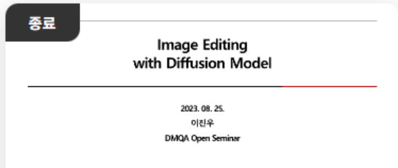


Image Editing with Diffusion Model

발표자: 이진우

2023년 8월 25일

오전 12시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Needs for Controllable Diffusion Model

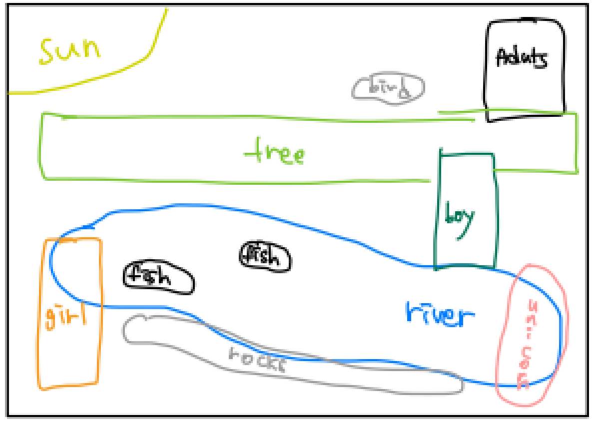
Conditional Diffusion Models

❖ Text prompt에 높은 의존

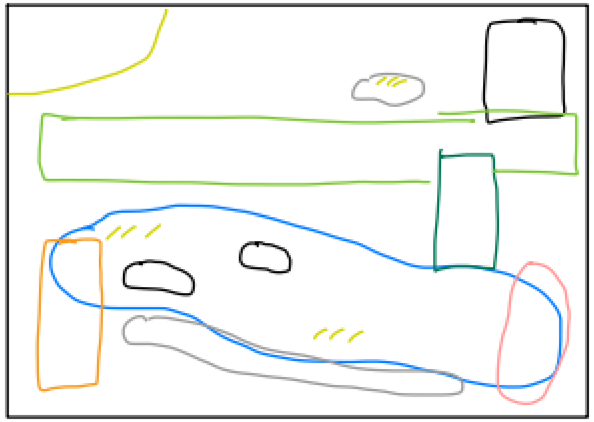
- Text만으로 원하는 이미지를 설명하는 것은 어려움
- Task specific한 도메인의 경우 일반적인 text-to-image 데이터 스케일만큼 크지 않음 (일반화 성능 하락)

물결이 흐르는 강, 앞에는 돌, 뒤에는 나무, 나무는 소나무와 은행 나무를 고르게 섞어서, 태양 빛은 강물에 반짝거리며 반사되고, 그 빛은 새들의 날개에도 은은하게 비치도록 그려줘. 강 안에는 물고기가 여러 마리 있고 그 중 몇 마리는 튀어 오르도록 그려줘. 여자 어린이는 왼쪽 남자는 오른쪽에 그 둘을 지켜보는 어른들도 그려줘. 여자 애는 만세하고 있고 남자애는 점프하고 있도록. 신비로운 분위기로. 아, 유니콘도 넣어줘.

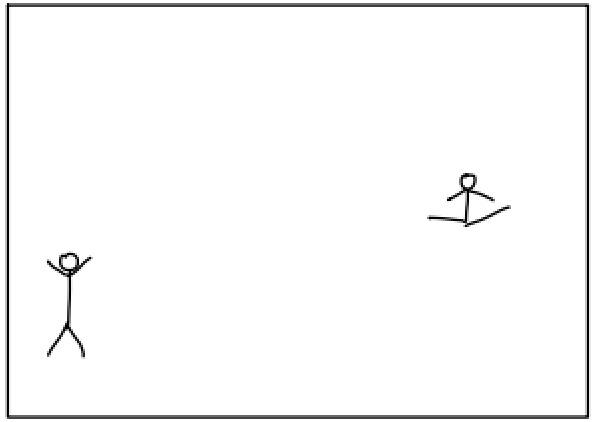
*정확한 비유 X 예시입니다.



Segmentation map



Colour pallete



Pose map

Depth map etc
…

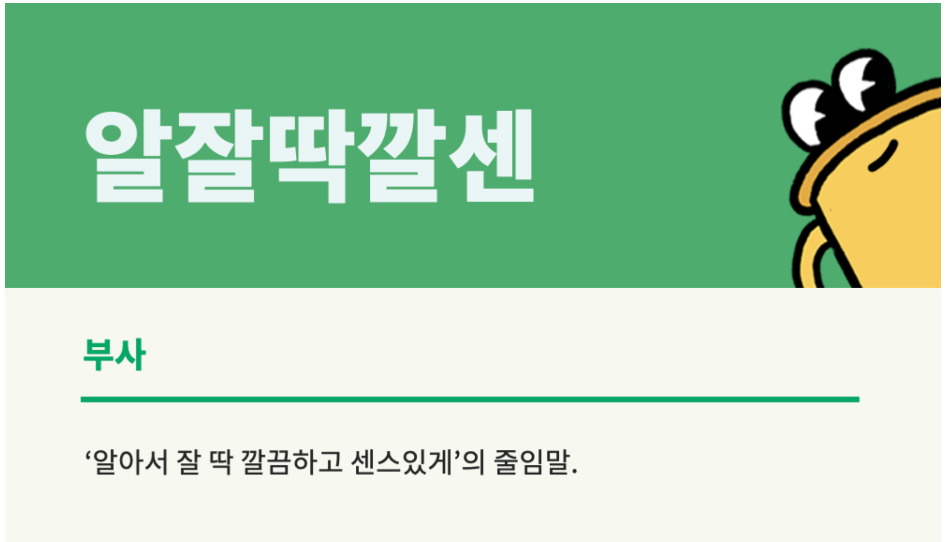
Needs for Controllable Diffusion Model

Conditional Diffusion Models

❖ Controllable Diffusion Model

- 우리가 의도한 결과가 생성되는 모델의 필요성

알아서 잘 딱 깔끔하고 센스 있게, 알지?



3. Controllable Diffusion Models

Controllable Diffusion Models

Control pretrained large diffusion models to support additional input conditions

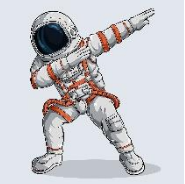
*dabbing

❖ Motivation

- Text prompt : “An astronaut dabbing, cartoon style”

Depth map

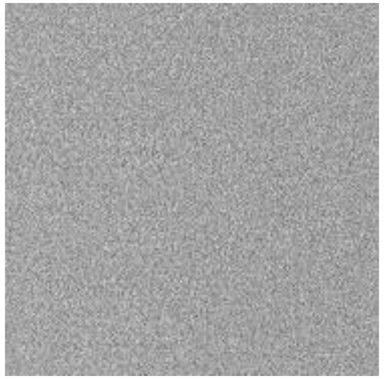
Image



…

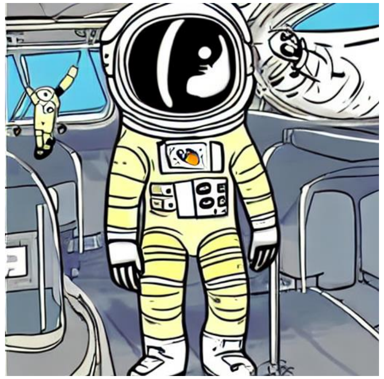
다른 condition을 추가 해보자!

Desired image



Stable Diffusion Model

Generated image



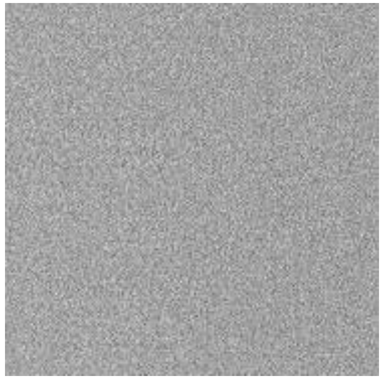
Controllable Diffusion Models

Control pretrained large diffusion models to support additional input conditions

❖ Motivation

- Okay, but how?
  - – training from scratch
  - – fine tuning light-weight adapters on frozen pretrained T2I diffusion models

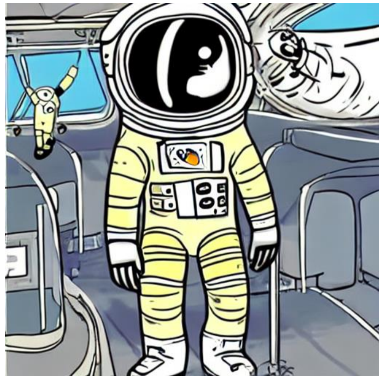
Desired image



다른 condition을 추가 해보자!

Stable Diffusion Model

Generated image



ControlNet

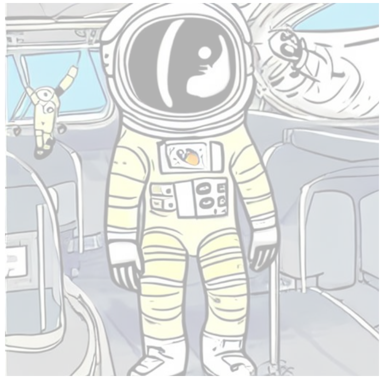
Control pretrained large diffusion models to support additional input conditions

❖ Scratch부터 다시 모델을 학습 시켜야 할까?

No! Stable Diffusion을 최대한 활용하자!

Desired image

Generated image



ControlNet

Control pretrained large diffusion models to support additional input conditions

❖ Adding Conditional Control to Text-to-Image Diffusion Models (2023, ICCV)

- Large diffusion model을 제어하여 task-specific한 입력 조건을 학습하는 end-to-end 구조의 ControlNet 제안
- Large diffusion model이 가지는 강력한 힘을 유지하고, 추가적인 input에 따라 바르게 모델을 build하는 방법

Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University
{lvmin, anyirao, maneesh}@cs.stanford.edu

Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc.*, to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

[4] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847).

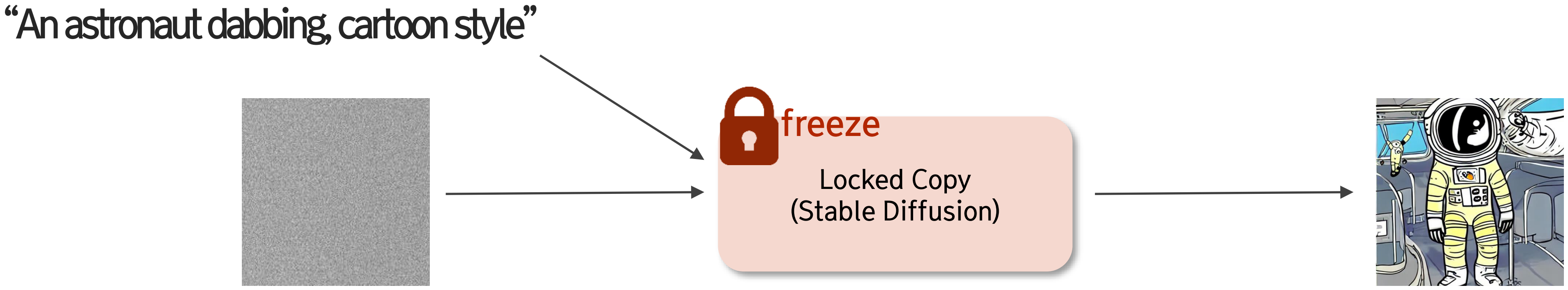
ControlNet

Control pretrained large diffusion models to support additional input conditions

❖ Adding Conditional Control to Text-to-Image Diffusion Models (2023, ICCV)

- Large diffusion model로부터 trainable copy와 locked copy를 복제함

Task-specific dataset에서 conditional control을 학습 대용량 이미지 데이터에서 학습된 weight를 보존



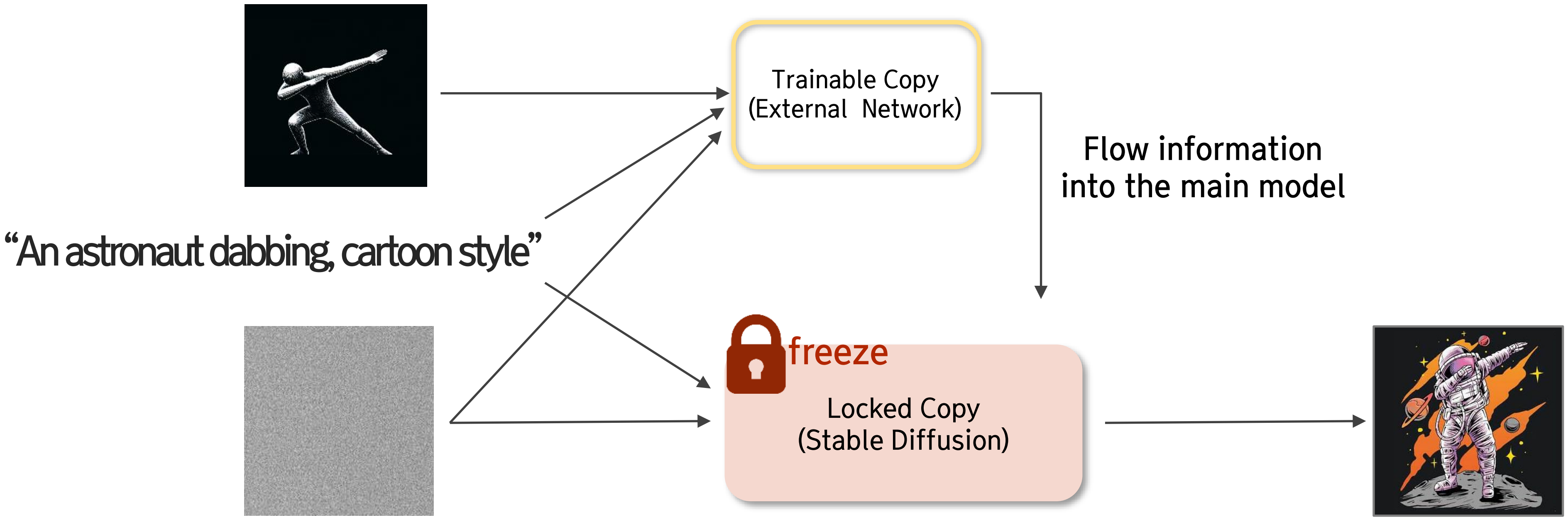
ControlNet

Control pretrained large diffusion models to support additional input conditions

❖ Adding Conditional Control to Text-to-Image Diffusion Models (2023, ICCV)

- Large diffusion model로부터 trainable copy와 locked copy를 복제함

Task-specific dataset에서 conditional control을 학습 대용량 이미지 데이터에서 학습된 weight를 보존



ControlNet

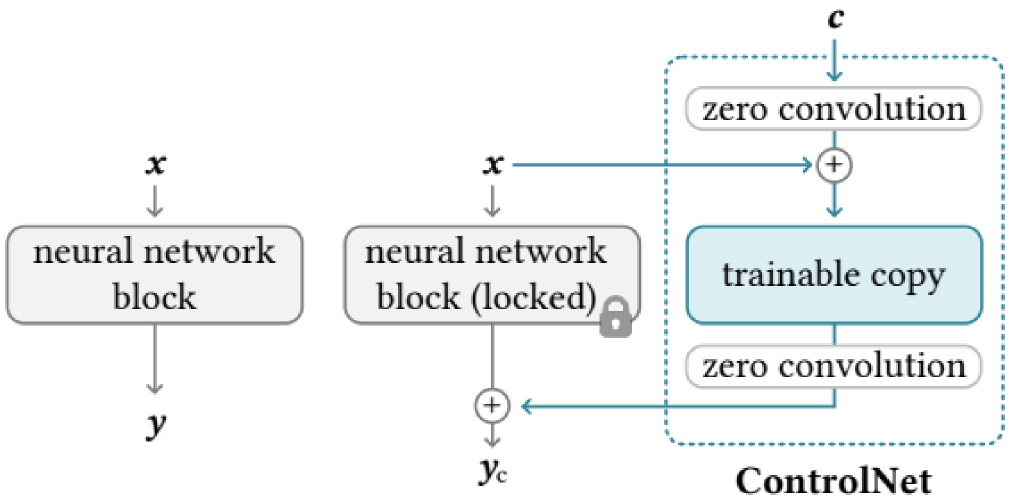
Control pretrained large diffusion models to support additional input conditions

❖ Technical Method

- Zero Convolution layer

  - 가중치와 바이어스가 모두 0으로 초기화된 1 * 1 convolution layer

  - 기존의 학습된 가중치를 유지하며 새로운 조건에 맞게 모델을 조정하는 역할



(a) Before (b) After

ControlNet 구조

$$y_c = \mathcal{F}(x;\theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c;\theta_{z1});\theta_c);\theta_{z2})$$

신경망 구조의 출력

$\mathcal{Z}(\cdot;\cdot)$: *Zero Convolution*

파라미터의 두 인스턴스 $\{\theta_{z1}, \theta_{z2}\}$

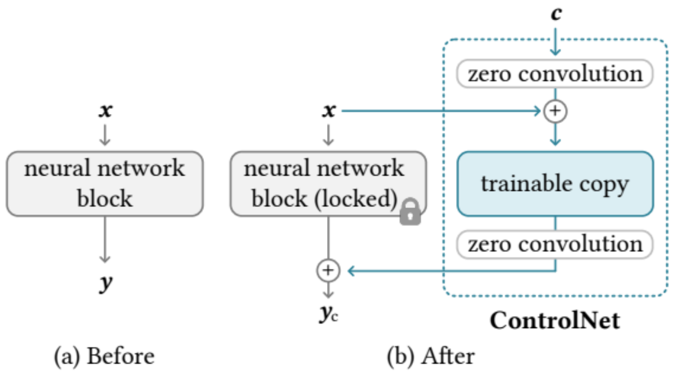
θ_c : original block 파라미터의 trainable copy

첫 번째 학습 step에서 y_c값은 0으로 초기화 되어

최적화 전의 zero convolution layer는 feature에 영향을 미치지 않음

ControlNet

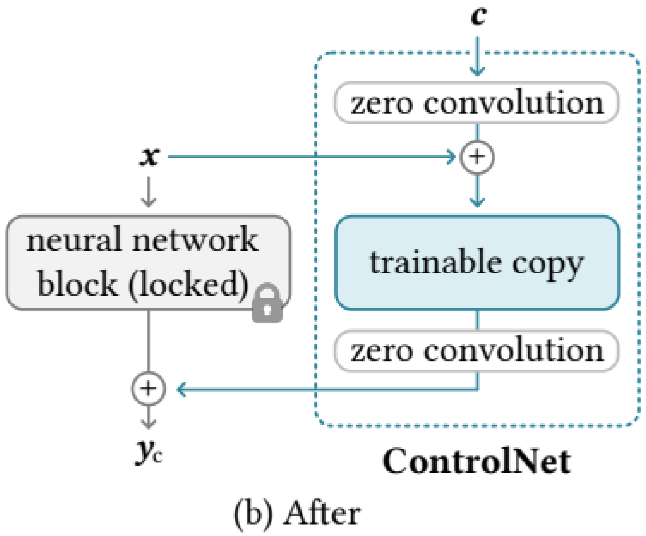
Control pretrained large diffusion models to support additional input conditions



❖ Technical Method

- Zero Convolution layer

  - weight와 bias를 0으로 설정하면 gradient가 흐르지 않는 것 아닌가?



$$y_c = \mathcal{F}(x;\theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c;\theta_{z1});\theta_c);\theta_{z2})$$

$$\mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i} = \boldsymbol{B}_i + \sum_j^c \boldsymbol{I}_{p,i}\boldsymbol{W}_{i,j}$$

$$\begin{cases} \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{B}_i} = 1 \\ \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{I}_{p,i}} = \sum_j^c \boldsymbol{W}_{i,j} = 0 \\ \dfrac{\partial \mathcal{Z}(\boldsymbol{I};\{\boldsymbol{W},\boldsymbol{B}\})_{p,i}}{\partial \boldsymbol{W}_{i,j}} = \boldsymbol{I}_{p,i} \neq \boldsymbol{0} \end{cases}$$

$$W^* = W - \beta_{\mathrm{lr}} \cdot \frac{\partial \mathcal{L}}{\partial \mathcal{Z}(I;\{W,B\})} \odot \frac{\partial \mathcal{Z}(I;\{W,B\})}{\partial W} \neq 0$$

ControlNet

Control pretrained large diffusion models to support additional input conditions

❖ Results

- 학습된 condition을 바탕으로 만들어진 이미지
- Default prompt : “a high-quality, detailed, and professional image”

Input Canny edge | Default | “masterpiece of fairy tale, giant deer, golden antlers” | “..., quaint city Galic”

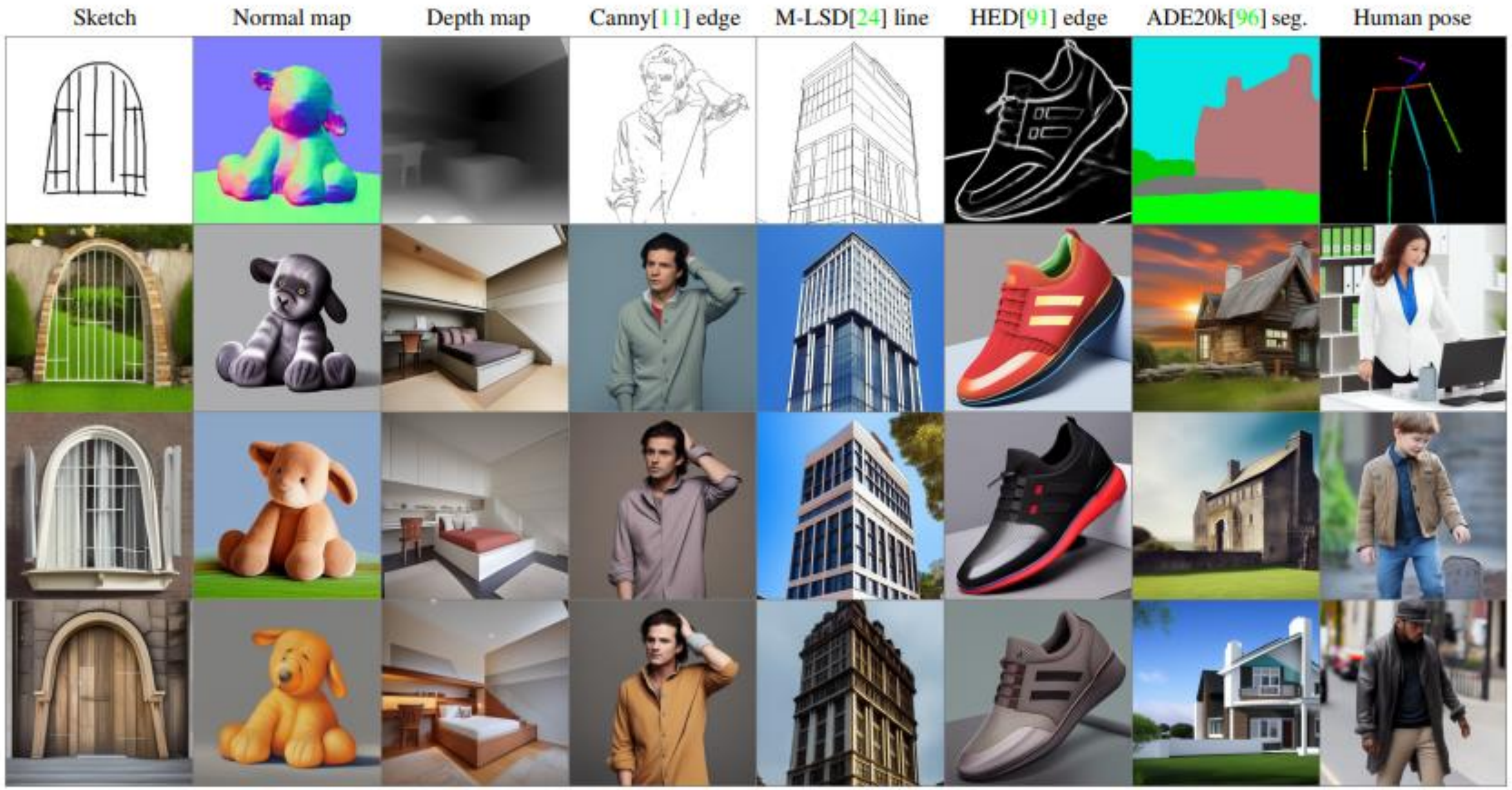
Input human pose | Default | “chef in kitchen” | “Lincoln statue”

ControlNet

Control pretrained large diffusion models to support additional input conditions

❖ Results

- Controlling Stable diffusion with various conditions **without prompts**



ControlNet

Control pretrained large diffusion models to support additional input conditions

❖ Conclusion

- 원본 모델을 고정하고 각 condition에 대한 adapter를 fine tuning하는 방법
- Trainable copy를 통해 데이터셋이 작을 때도 overfitting방지 가능
- Training cost의 현저한 감소

❖ Limitation

- 각 단일 조건에 대해 하나의 독립적인 adapter를 필요로 함
- 제어 조건의 수가 증가함에 따라 fine-tuning 비용과 모델 크기가 비례하게 증가

Composer

Creative and Controllable Inage Synthesis with Composable Conditions

❖ Composer : Creative and Controllable Image Synthesis with Composable Conditions (2023, ICML)

- 시각적 구성 요소들을 재결합하여 새로운 이미지를 생성하는 multi conditional diffusion model
- 분해 (decompose) 단계와 recompose(합성) 단계를 통해 합성 가능한 생성형 모델 제시

Composer: Creative and Controllable Image Synthesis with Composable Conditions

Lianghua Huang [1] Di Chen [1] Yu Liu [1] Yujun Shen [2] Deli Zhao [1] Jingren Zhou [1]

Abstract

Recent large-scale generative models learned on big data are capable of synthesizing incredible images yet suffer from limited controllability. This work offers a new generation paradigm that allows flexible control of the output image, such as spatial layout and palette, while maintaining the synthesis quality and model creativity. With *compositionality* as the core idea, we first decompose an image into representative factors, and then train a diffusion model with all these factors as the conditions to recompose the input. At the inference stage, the rich intermediate representations work as composable elements, leading to a huge design space (*i.e.*, exponentially proportional to the number of decomposed factors) for customizable content

produce photorealistic and diverse images (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2021; Yu et al., 2022; Chang et al., 2023). To further achieve customized generation, many recent works extend the text-to-image models by introducing conditions such as segmentation maps (Rombach et al., 2021; Wang et al., 2022b; Couairon et al., 2022), scene graphs (Yang et al., 2022), sketches (Voynov et al., 2022), depthmaps (stability.ai, 2022), and inpainting masks (Xie et al., 2022; Wang et al., 2022a), or by finetuning the pretrained models on a few subject-specific data (Gal et al., 2022; Mokady et al., 2022; Ruiz et al., 2022). Nevertheless, these models still provide only a limited degree of controllability for designers when it comes to using them for practical applications. For example, generative models often struggle to accurately produce images with specifications for semantics, shape, style, and color all at once, which is common in real-world

[5] Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., & Zhou, J. (2023). Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778.*

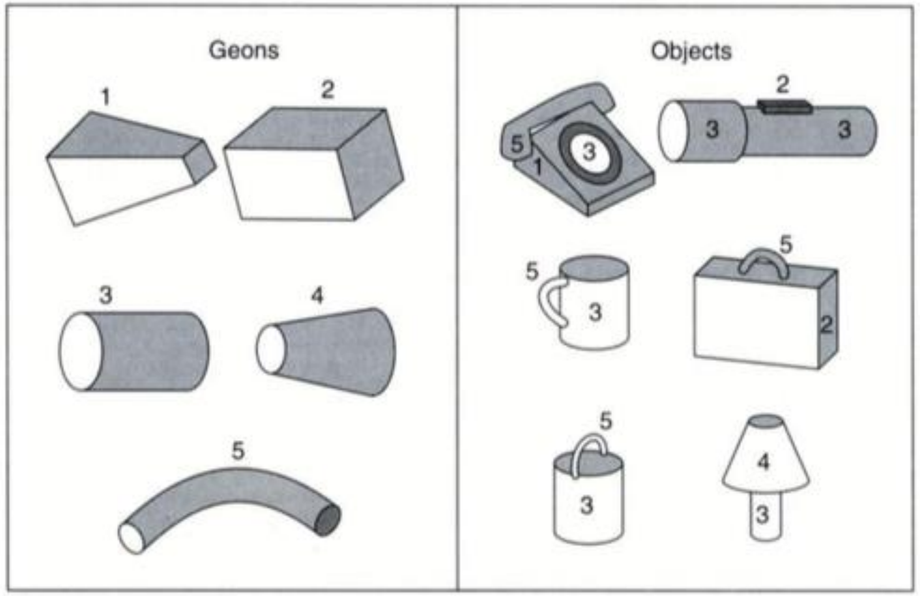
Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Compositionality (합성성)

- Controllable diffusion model의 핵심을 'Compositionality'로 정의
- 기본 요소(primitive elements)의 결합을 통해 새로운 표현이 구성될 수 있음

Sample Complexity 감소 / Deeper Generalization 가능




```python
class UNet(nn.Module):
    def __init__(self):
        super(UNet, self).__init__()

        def CBR2d(in_channels, out_channels, kernel_size=3, stride=1, padding=1, bias=True):
            layers = []
            #convolution layer
            layers += [nn.Conv2d(in_channels=in_channels, out_channels=out_channels,
                                 kernel_size=kernel_size, stride=stride, padding=padding,
                                 bias=bias)]
            #BN layer (convolution layer의 out_channels만큼)
            layers += [nn.BatchNorm2d(num_features=out_channels)]
            layers += [nn.ReLU()]

            cbr = nn.Sequential(*layers)

            return cbr
```


[7] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. Behavioral and Brain Sciences, 40, 2017.
[8] Biederman, I. (1987). Recognition by Components: A Theory of Human Image Understanding. *Psychological Review*, *94*(2), 115–147.

Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Compositionality (합성성)

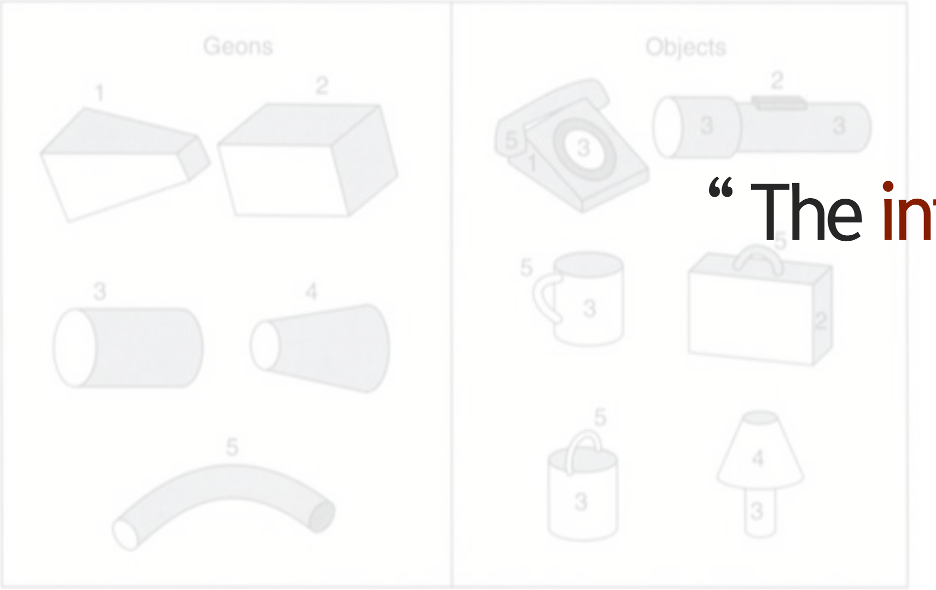
- 이미지 생성의 핵심은 합성 !
- 유한한 개수의 요소를 무한하게 활용할 수 있다

1. Introduction

"The infinite use of finite means."

– Noam Chomsky (Chomsky, 1965)

Sample Complexity 감소 / Deeper Generalization 가능



" The infinite use of finite means"

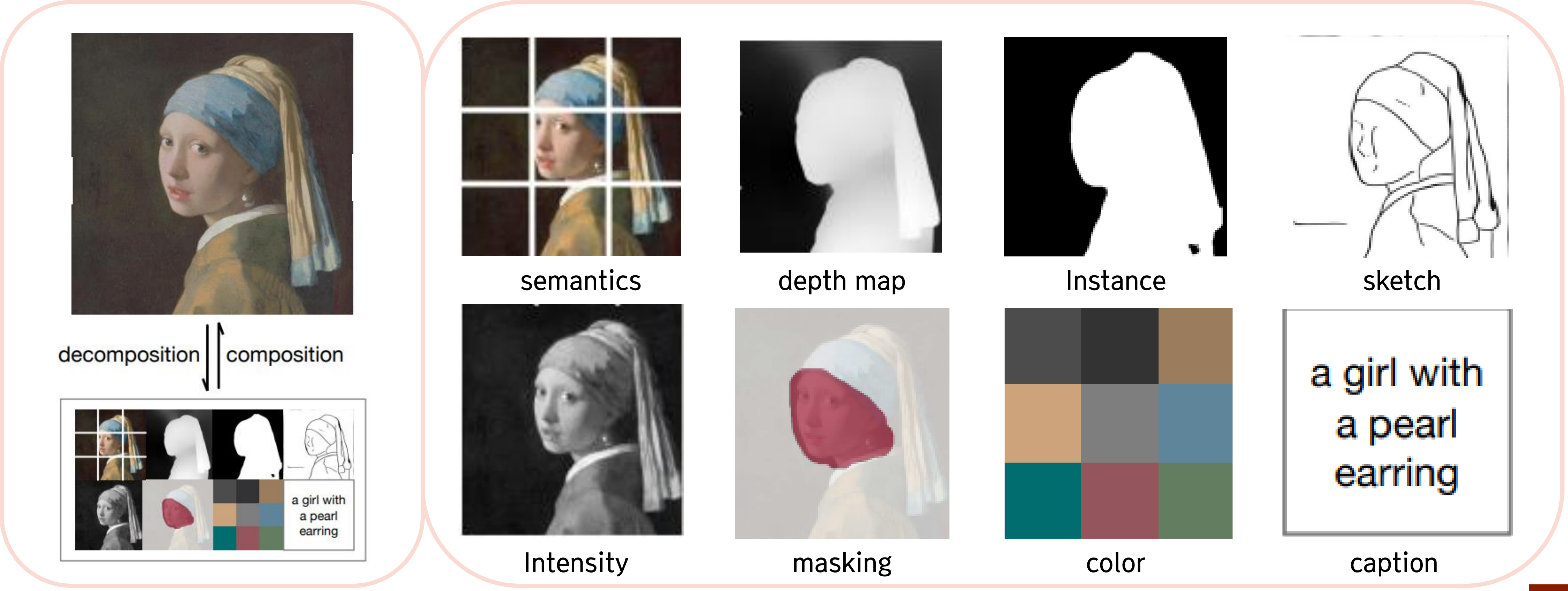
Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Method

- Decomposition

  – 이미지를 8개의 representation으로 학습 중 즉석에서 분해



Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Method

- Composition

  - Representation의 집합으로부터 이미지를 recompose(재구성) 하는 단계

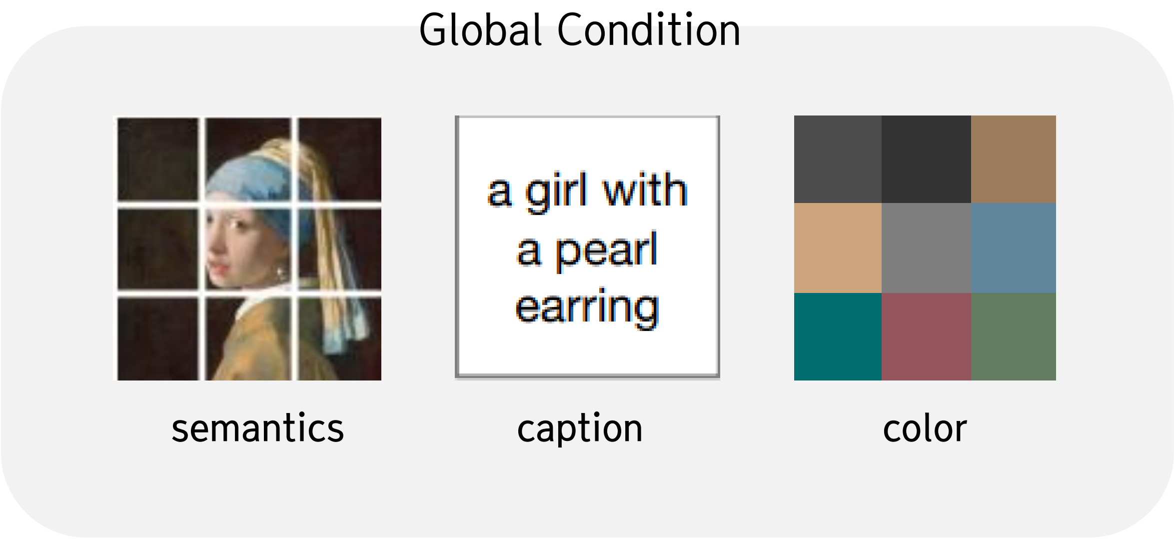
  1) Global conditioning : 이미지 전체에 대한 조건
  2) Localized conditioning : 이미지 내 특정 영역이나 구성 요소에 대한 조건
  3) Joint training strategy

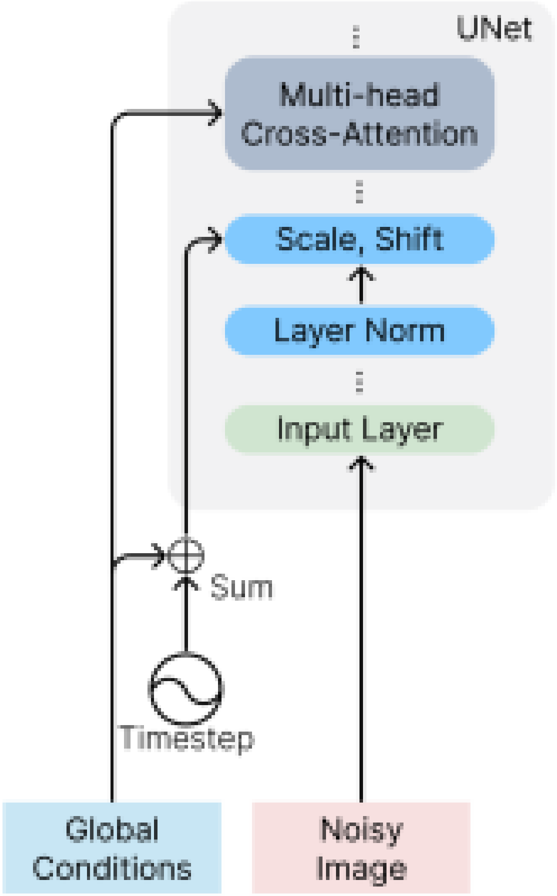
Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Method

- Composition : Representation의 집합으로부터 이미지를 recompose(재구성) 하는 단계
  1) **Global conditioning** : 이미지 전체에 대한 조건
  2) Localized conditioning : 이미지 내 특정 영역이나 구성 요소에 대한 조건
  3) Joint training strategy



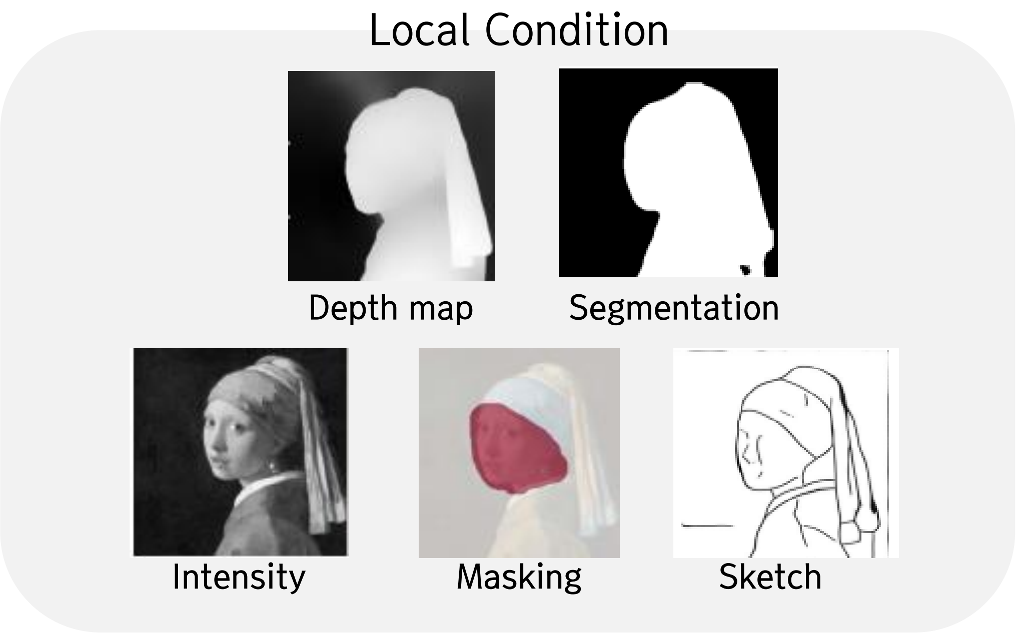


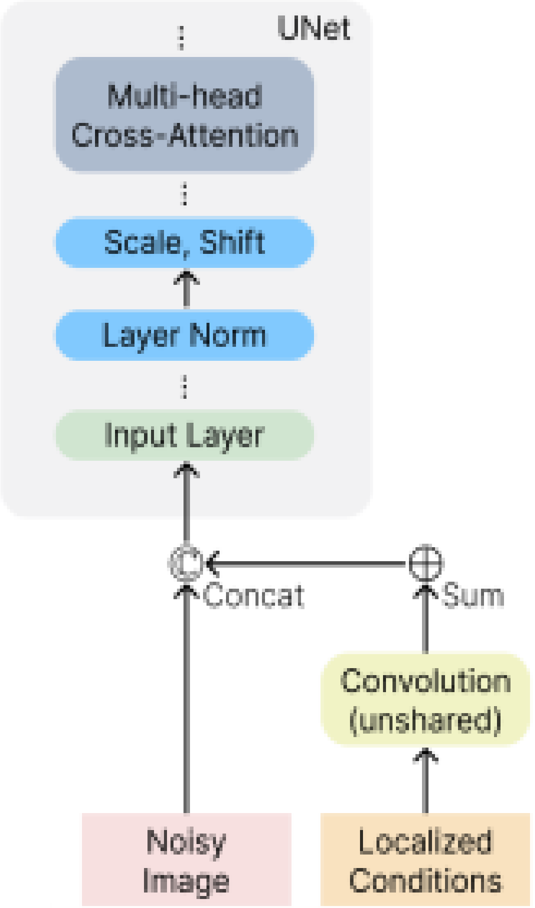
Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Method

- Composition : Representation의 집합으로 부터 이미지를 recompose(재구성) 하는 단계
  1) Global conditioning : 이미지 전체에 대한 조건
  2) **Localized conditioning** : 이미지 내 특정 영역이나 구성 요소에 대한 조건
  3) Joint training strategy





Composer

Creative and Controllable Image Synthesis with Composable Conditions

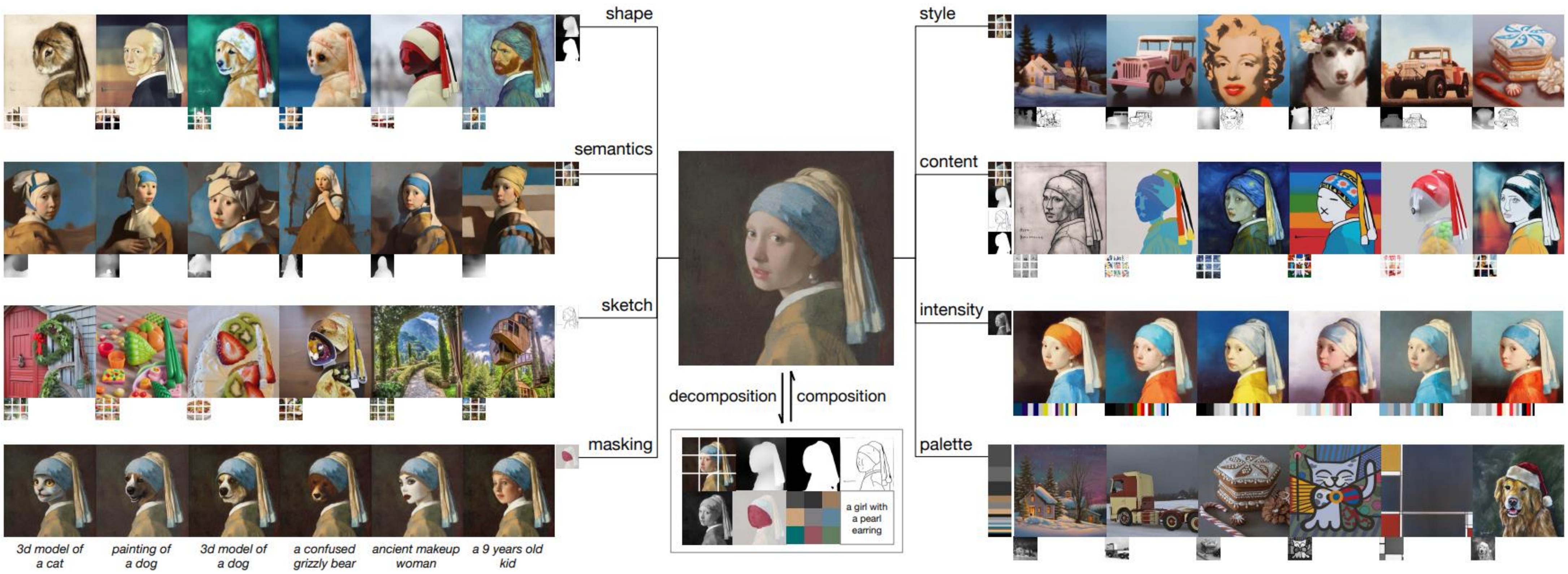
❖ Method

- Composition : Representation의 집합으로 부터 이미지를 recompose(재구성) 하는 단계
  1) Global conditioning : 이미지 전체에 대한 조건
  2) Localized conditioning : 이미지 내 특정 영역이나 구성 요소에 대한 조건
  3) **Joint training strategy**

Composer

Creative and Controllable Image Synthesis with Composable Conditions

- ❖ Results



Composer

Creative and Controllable Image Synthesis with Composable Conditions

- Results
  - Image generation

(a) Palette-based colorization.

(b) Style transfer.

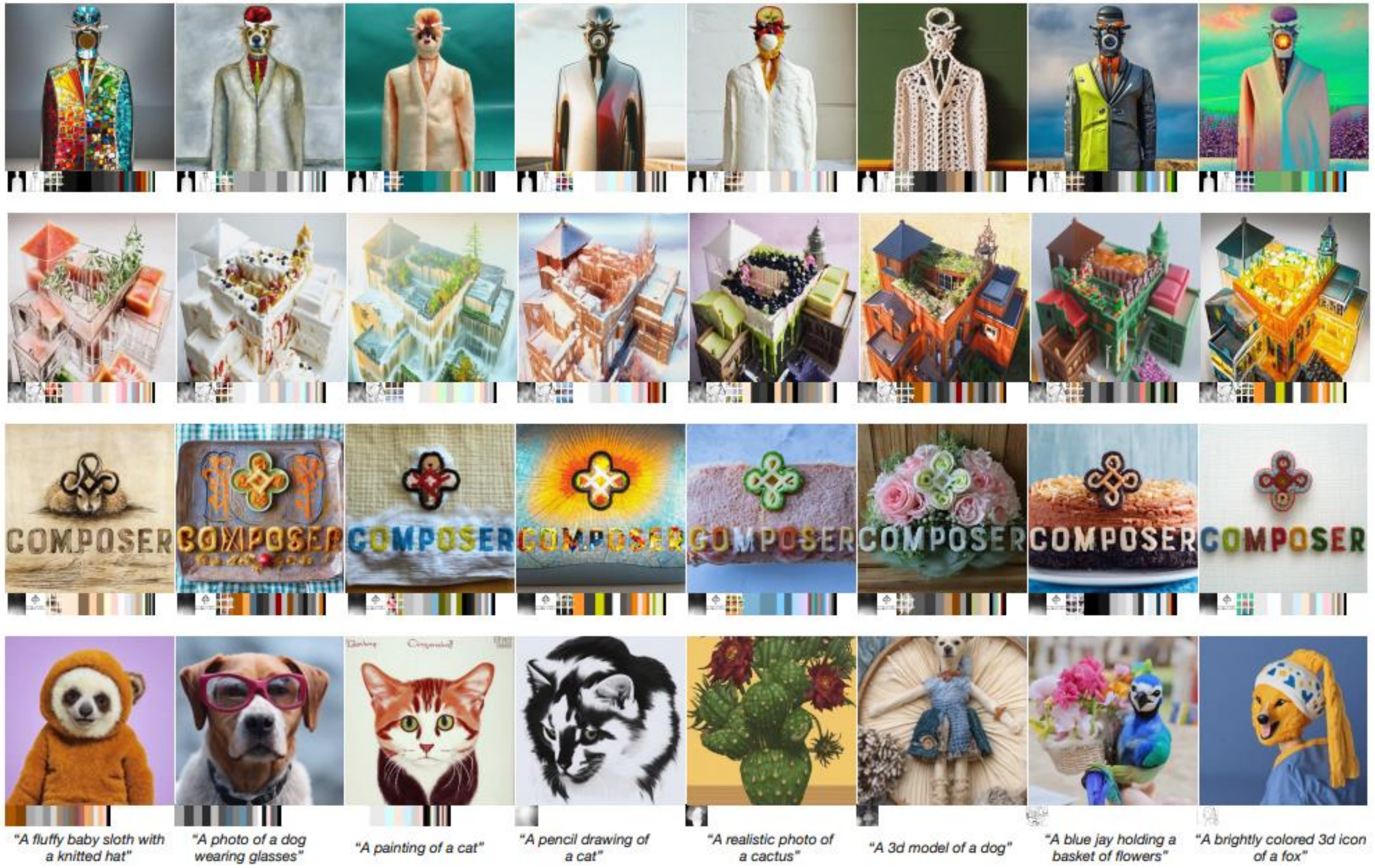
(c) Image translation.

Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Results

- Compositional한 이미지 생성 task



Composer

Creative and Controllable Image Synthesis with Composable Conditions

❖ Conclusion

- Scratch부터 큰 diffusion model을 새로 학습 시킴으로써 여러 condition에 맞게 compositional한 이미지 생성을 가능하게 함
- 단일 condition과 다중 condition 모두 높은 퀄리티로 task 수행
- 이미지 생성을 분해/합성 단계로 구분하여 높은 퀄리티의 이미지 생성 가능

❖ Limitation

- 매우 높은 training cost

Uni-ControlNet

All-in-One Control to Text-to-Image Diffusion Models

❖ Uni-ControlNet : All-in-One Control to Text-to-Image Diffusion Models (2023, NeurIPS)

- 하나의 모델 내에서 다양한 local control과 global control의 동시 사용이 가능하게 하는 controllable diffusion model

Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models

Shihao Zhao[†]
The University of Hong Kong
shzhao@cs.hku.hk

Dongdong Chen[*]
Microsoft
cddlyf@gmail.com

Yen-Chun Chen
Microsoft
yen-chun.chen@microsoft.com

Jianmin Bao
Microsoft
jianmin.bao@microsoft.com

Shaozhe Hao
The University of Hong Kong
szhao@cs.hku.hk

Lu Yuan
Microsoft
luyuan@microsoft.com

Kwan-Yee K. Wong[*]
The University of Hong Kong
kykwong@cs.hku.hk

Abstract

Text-to-Image diffusion models have made tremendous progress over the past two years, enabling the generation of highly realistic images based on open-domain text descriptions. However, despite their success, text descriptions often struggle to adequately convey detailed controls, even when composed of long and complex texts. Moreover, recent studies have also shown that these models face challenges in understanding such complex texts and generating the corresponding images. Therefore, there is a growing need to enable more control modes beyond text description. In this paper, we introduce Uni-ControlNet, a unified framework that allows for the simultaneous utilization of different local controls (e.g., edge maps, depth map, segmentation masks) and global controls (e.g., CLIP image embeddings) in a flexible and composable manner within one single model. Unlike existing

[6] Zhao, S., Chen, D., Chen, Y. C., Bao, J., Hao, S., Yuan, L., & Wong, K. Y. K. (2023). Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2305.16322.*

Uni-ControlNet

All-in-One Control to Text-to-Image Diffusion Models

❖ Comparisons of different controllable diffusion models

- pre-trained model 고정, 2개의 추가적인 adapter를 fine-tuning하는 과정만 요구됨
- Condition을 local과 global로 나눔
- 다양한 condition에 대한 훌륭한 composable control이 가능

	Fine-tuning	Composable Control	Fine-tuning Cost	Adapter Number
Composer	✗	✓	-	-
ControlNet	✓	✓	N	N
GLIGEN	✓	✗	N	N
T2I-Adapter	✓	✓	$N(+1)$	$N(+1)$
Uni-ControlNet (Ours)	✓	✓	2	2

Uni-ControlNet

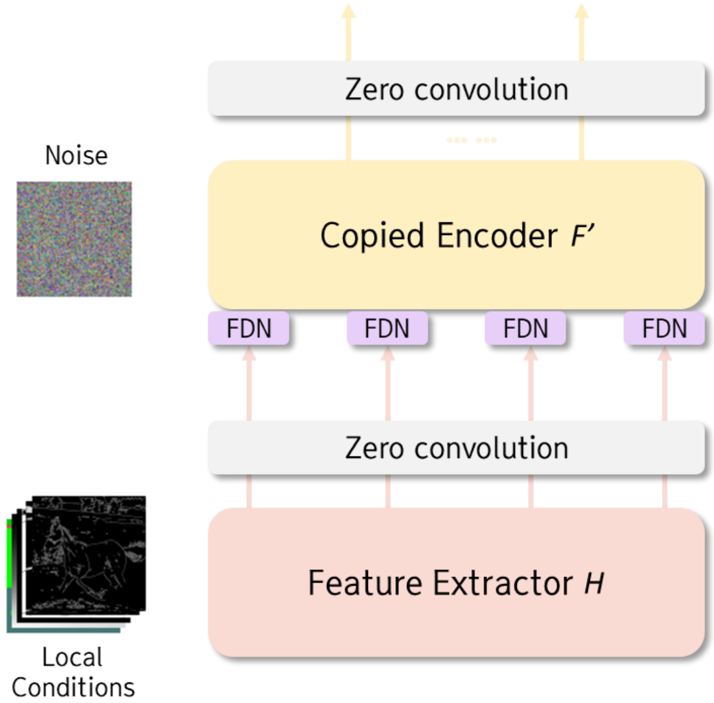
All-in-One Control to Text-to-Image Diffusion Models

❖ Method

- Local Control Adapter

  - 7개의 local condition 사용 (Canny edge, MLSD edge, HED boundary, Sketch, Openpose, Midas depth, Segmentation mask

  - Multi scale condition 주입 전략을 사용



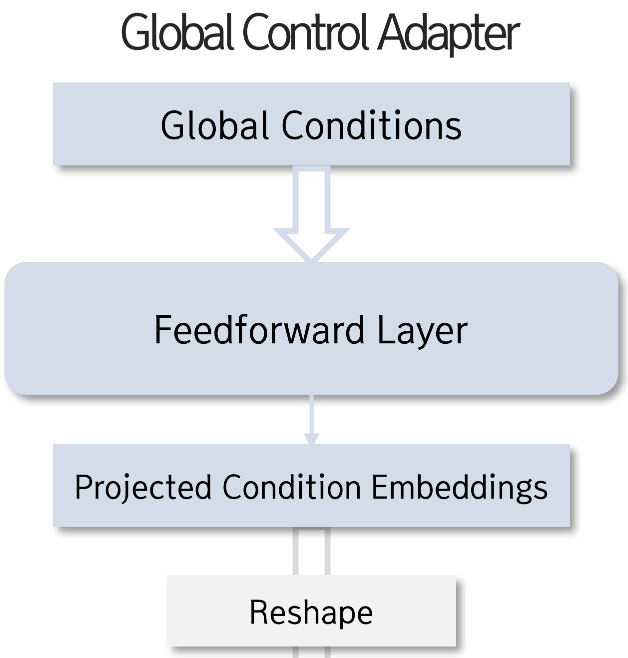
Local Control Adapter

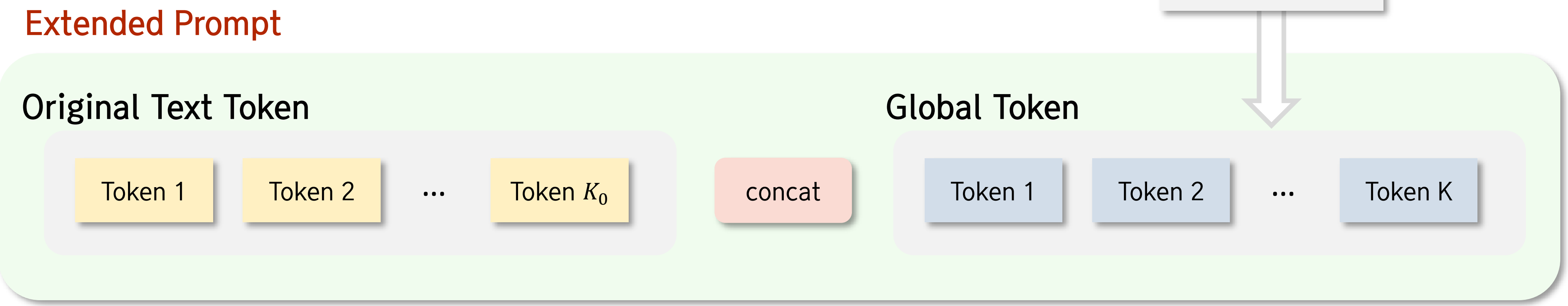
Uni-ControlNet

All-in-One Control to Text-to-Image Diffusion Models

❖ Method

- Global Control Adapter
  - CLIP 이미지 인코더로부터 추출된 global image embedding
  - original text token과 global token을 concatenate하여 extended prompt 생성
  - extended prompt : main model과 control adapter 모두의 cross attention input



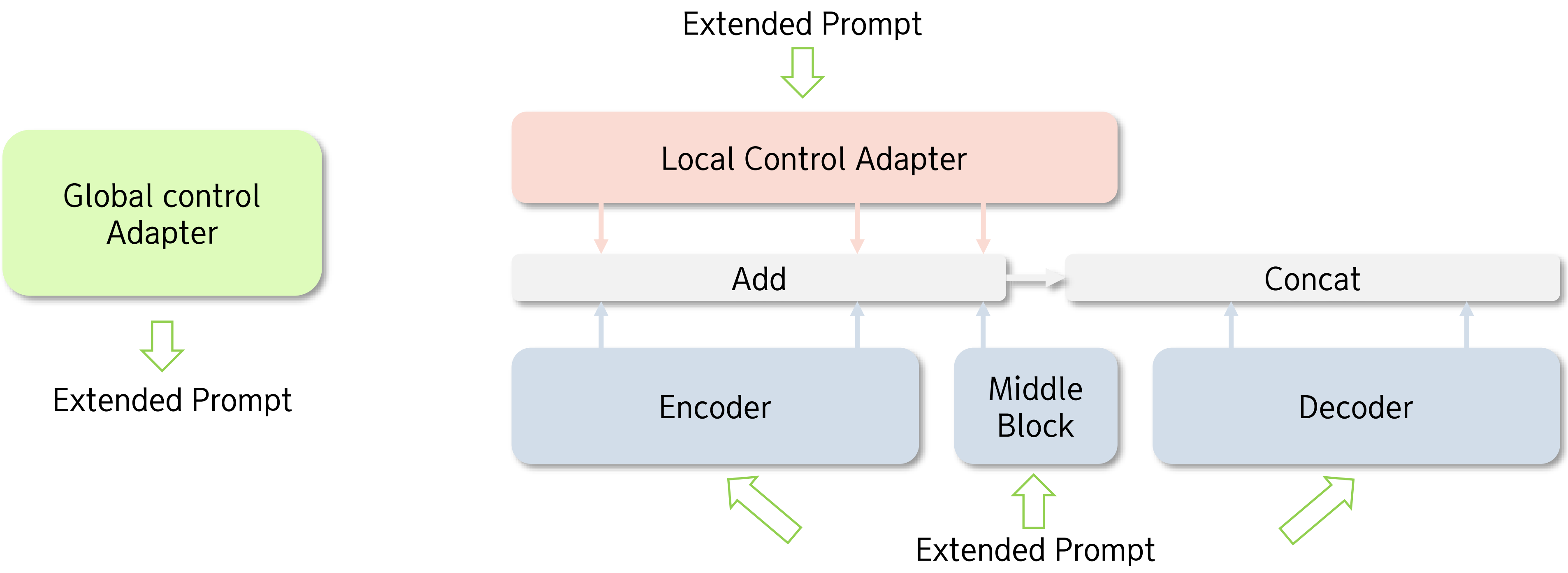


Uni-ControlNet

All-in-One Control to Text-to-Image Diffusion Models

❖ Method – framework

- Local control과 global control을 각각을 분리하여 fine tuning



Uni-ControlNet

All-in-One Control to Text-to-Image Diffusion Models

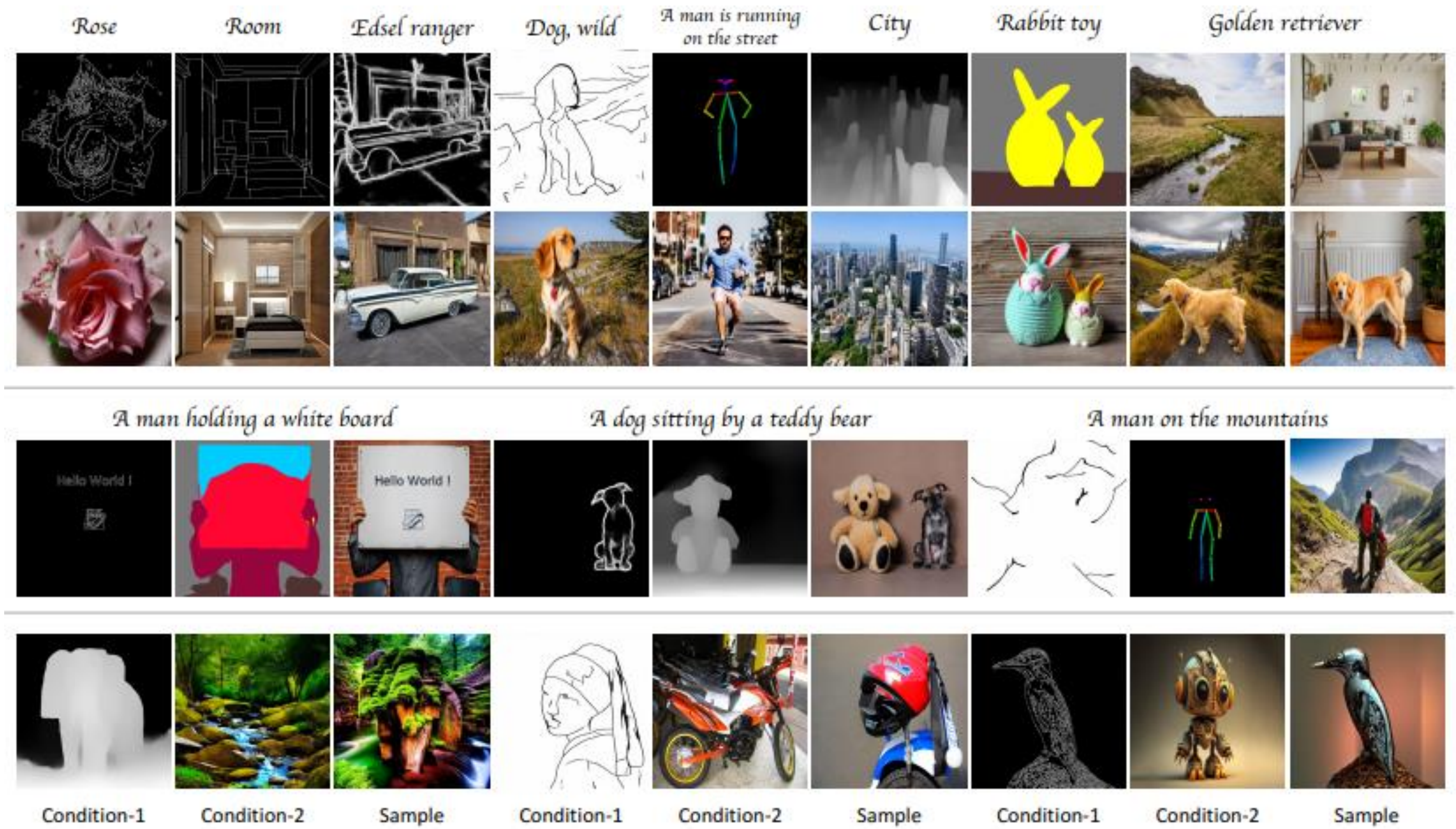
❖ Results

Single condition

2 Local conditions

1 Local condition

+ 1 Global condition



Uni-ControlNet

All-in-One Control to Text-to-Image Diffusion Models

❖ Results

- FID score와 다양한 정량적 평가 지표에서 높은 성능

Table 2: FID on different controllable diffusion models. The best results are in **bold**.

	Canny	MLSD	HED	Sketch	Pose	Depth	Segmentation	Style\Content
ControlNet	18.90	31.36	26.59	22.19	27.84	21.25	**23.08**	31.17
GLIGEN	24.74	-	28.57	-	**24.57**	21.46	27.39	25.12
T2I-Adapter	18.98	-	-	**18.83**	29.57	21.35	23.84	28.86
Ours	**17.79**	**26.18**	**17.86**	20.11	26.61	**21.20**	23.40	**23.98**

Table 3: Quantitative evaluation of the controllability. The best results are in **bold**.

	Canny (SSIM)	MLSD (SSIM)	HED (SSIM)	Sketch (SSIM)	Pose (mAP)	Depth (MSE)	Segmentation (mIoU)	Style\Content (CLIP Score)
ControlNet	0.4828	**0.7455**	0.4719	0.3657	0.4359	**87.57**	**0.4431**	0.6765
GLIGEN	0.4226	-	0.4015	-	0.1677	88.22	0.2557	0.7458
T2I-Adapter	0.4422	-	-	0.5148	**0.5283**	89.82	0.2406	0.7078
Ours	**0.4911**	0.6773	**0.5197**	**0.5923**	0.2164	91.05	0.3160	**0.7753**

5. Conclusion

Conclusions

Controllable Diffusion Model

❖ Controllable Diffusion Model

- ControlNet
- Composer
- Uni ControlNet

❖ Future works

- 이미지 내 매우 세밀한 영역의 control
- 조건이 충돌하는 경우의 condition을 조절할 방법
- etc

6. Reference

References

[1] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., ... & Chen, M. (2022, June). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In International Conference on Machine Learning (pp. 16784-16804). PMLR.
[2] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35, 36479-36494.
[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
[4] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3836-3847).
[5] Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., & Zhou, J. (2023). Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778.
[6] Zhao, S., Chen, D., Chen, Y. C., Bao, J., Hao, S., Yuan, L., & Wong, K. Y. K. (2023). Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. arXiv preprint arXiv:2305.16322.
[7] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. Behavioral and Brain Sciences, 40, 2017.
[8] Biederman, I. (1987). Recognition by Components: A Theory of Human Image Understanding. Psychological Review, 94(2), 115–147.

Thank You