Autoregressive Models in Vision: From Next-Token to Next-Scale Prediction

Open DMQA Seminar
2025.09.12

조한샘

고려대학교
KOREA UNIVERSITY

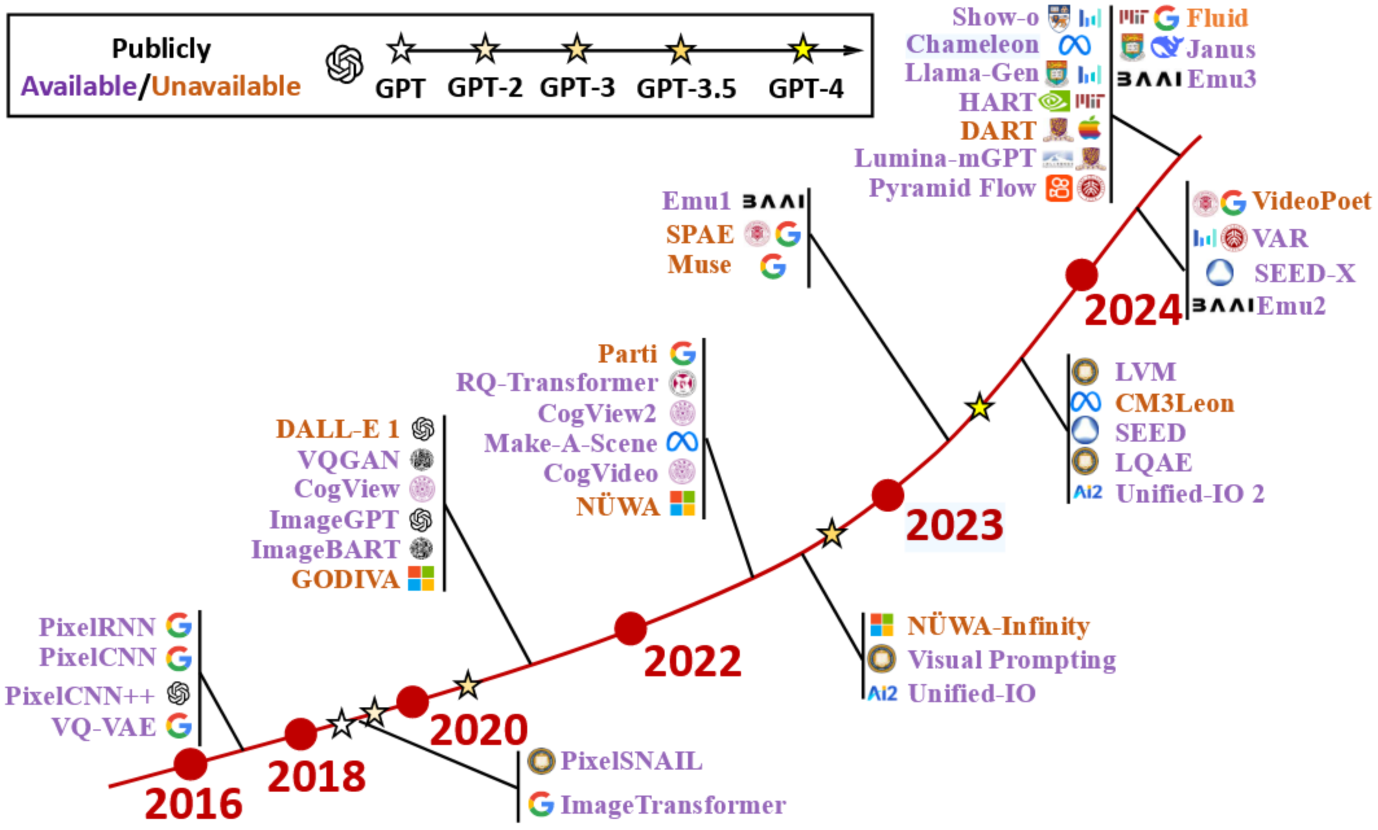
발표자 소개

- **조한샘**
  - ✓ Data Mining & Quality Analytics Lab
  - ✓ 석·박통합과정 (2020.09~)
- **관심 연구 분야**
  - ✓ Visual Generative Models
  - ✓ Controllable Generation
- **Contact**
  - ✓ chosam95@korea.ac.kr

Introduction

Autoregressive Models in Vision

- LLM의 성공으로 컴퓨터 비전 분야에서도 autoregressive model에 대한 연구 활발히 수행

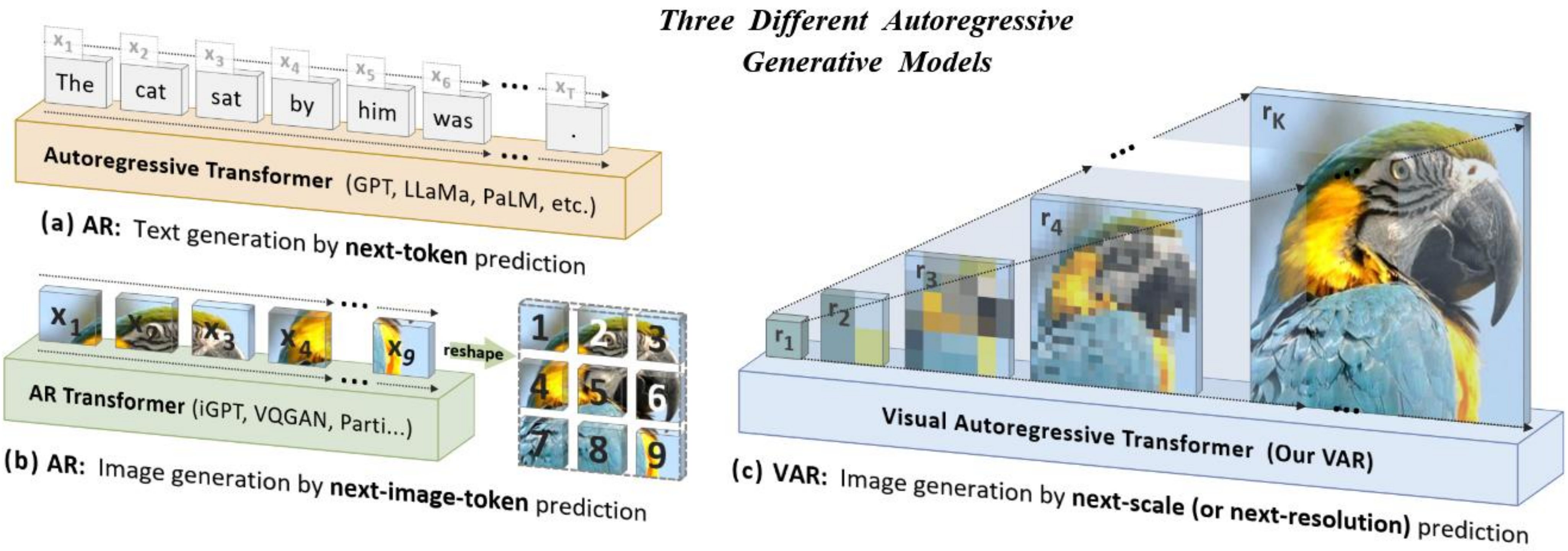


Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., ... & Wong, N. Autoregressive Models in Vision: A Survey. *Transactions on Machine Learning Research.*

Introduction

From Next-Token to Next-Scale Prediction

- 기존 autoregressive model은 LLM과 동일하게 next-token prediction을 기반으로 모델링
- 최근 next-scale prediction이라는 새로운 모델링 방식 제안



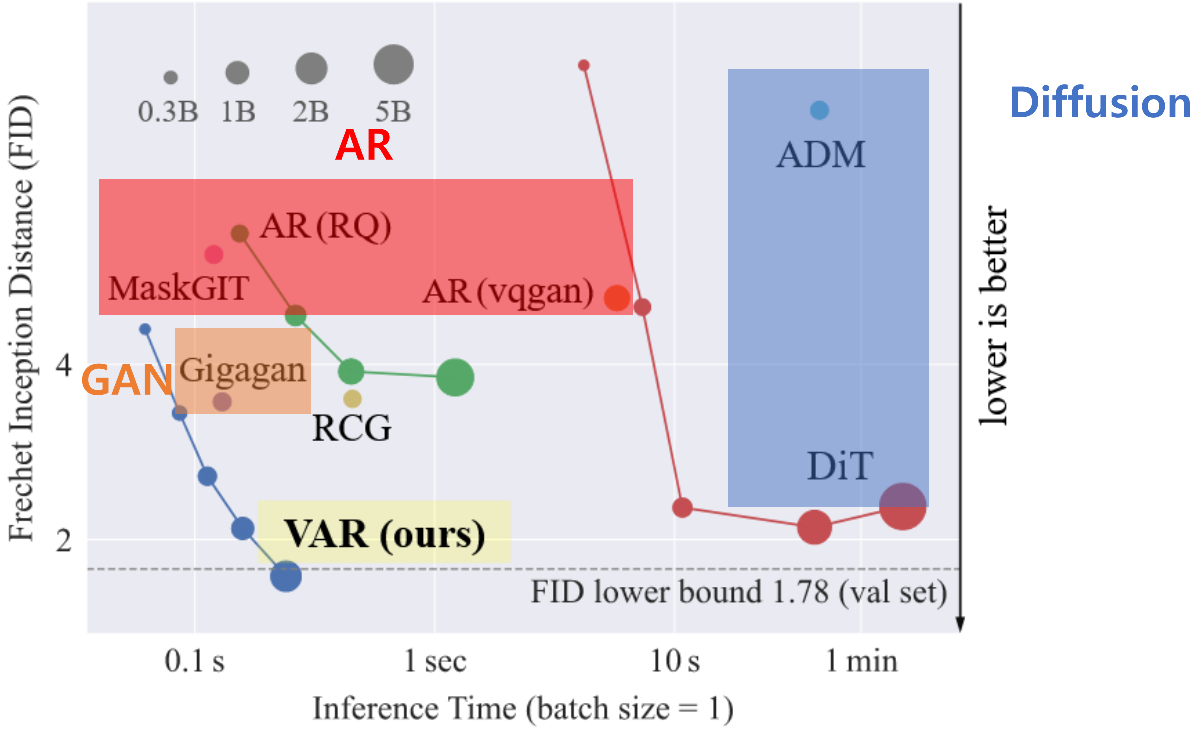
Tian, K., Jiang, Y., Yuan, Z., Peng, B., & Wang, L. (2024). Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems, 37, 84839-84865.

Introduction

From Next-Token to Next-Scale Prediction

- 기존 모델들보다 이미지 생성 속도와 퀄리티 측면에서 좋은 성능

이미지 생성 퀄리티



이미지 생성 속도

Tian, K., Jiang, Y., Yuan, Z., Peng, B., & Wang, L. (2024). Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems, 37, 84839-84865.

Introduction

From Next-Token to Next-Scale Prediction

- NeurIPS 2024 best paper

Oral

Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction

Keyu Tian · Yi Jiang · Zehuan Yuan · BINGYUE PENG · Liwei Wang

Best Paper

[Abstract] [Visit Oral Session 2D: Generative Models]

[OpenReview]

https://neurips.cc/virtual/2024/oral/97960

Autoregressive Models

Formulation

- Autoregressive model은 생성모델 → 데이터의 분포를 학습
- 데이터에 순서가 있다고 가정하고 모델링 진행

Step1
: Tokenization

Step2
: Modeling

$$p(x)$$

$$= p(x_1, x_2, \dots, x_T)$$

$$= \prod_{t=1}^{T} p(x_t | x_1, x_2 \dots, x_{t-1})$$

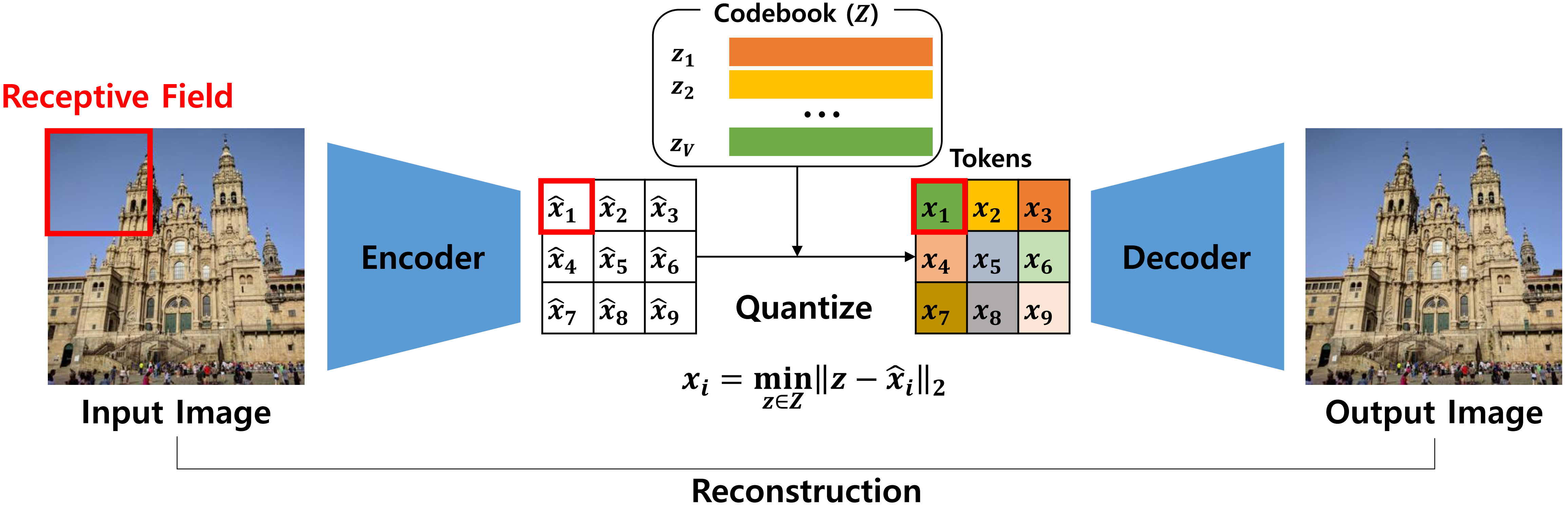
Chain Rule

$$x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_T$$

Next-Token Prediction

Step1: Tokenization

- **Codebook**: learnable vector들의 집합
- **Quantization**: continuous feature를 discreate token으로 변환하는 과정 (codebook에서 가장 가까운 벡터 추출)
- 패치단위 tokenization



Next-Token Prediction

Step2: Modeling

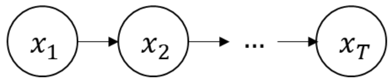
- Next-token prediction을 기반으로 학습 진행

$$p(x)$$

$$= p(x_1, x_2, \dots, x_T)$$

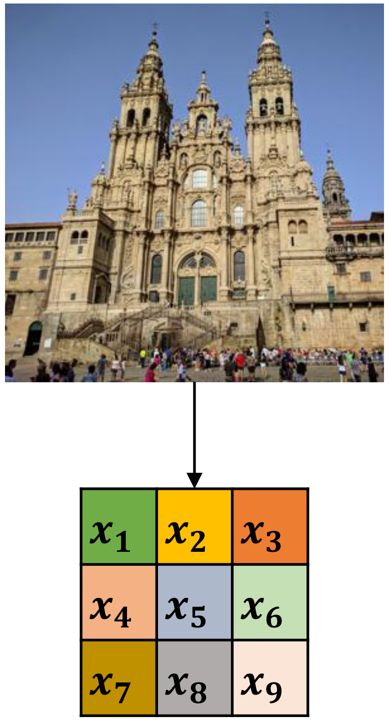
$$= \prod_{t=1}^{T} p(x_t | x_1, x_2 \dots, x_{t-1})$$

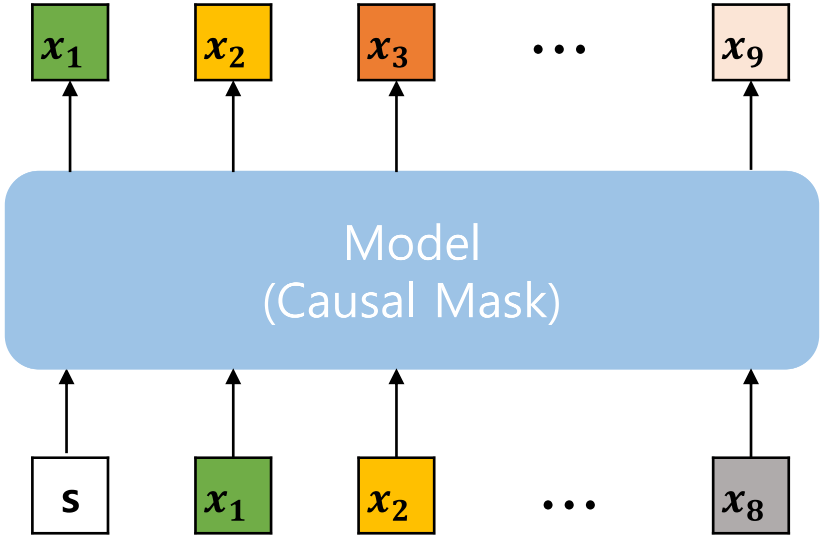
Chain Rule



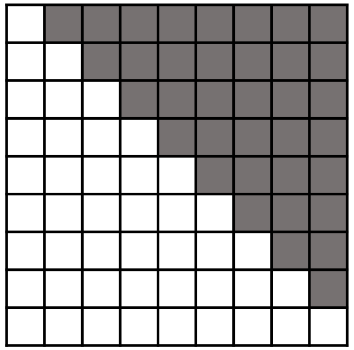
Image

Tokens





Causal Mask



Next-Token Prediction

Limitations

- 모델링을 위해서 토큰에 순서를 부여, but 이미지는 양방향 정보가 중요함

Image

Tokens

Next-Scale Prediction

From Next-Token to Next-Scale Prediction

- Reformulation

$$p(x)$$

$$= p(x_1, x_2, \ldots, x_T)$$

$$= \prod_{t=1}^{T} p(x_t | x_1, x_2 \ldots, x_{t-1})$$

<Next-Token Prediction>

$$p(x)$$

$$= p(r_1, r_2, \ldots, r_K)$$

$$= \prod_{k=1}^{K} p(r_k | r_1, r_2 \ldots, r_{k-1})$$

<Next-Scale Prediction>

Next-Scale Prediction

Step1: Tokenization

- **Next-token prediction**: patch 단위로 tokenize
- **Next-scale prediction**: scale 단위로 tokenize

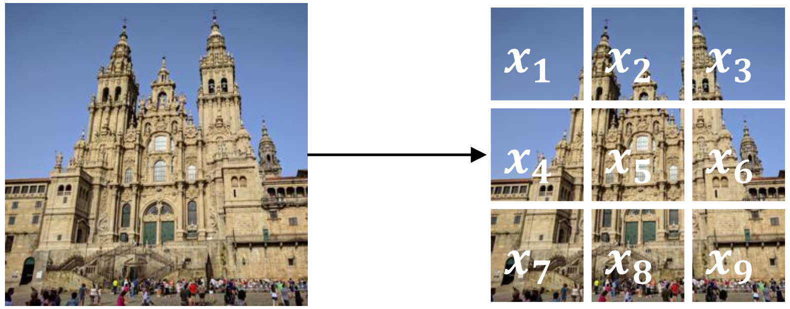


Image **Tokens**

<Patch-wise Tokenization>

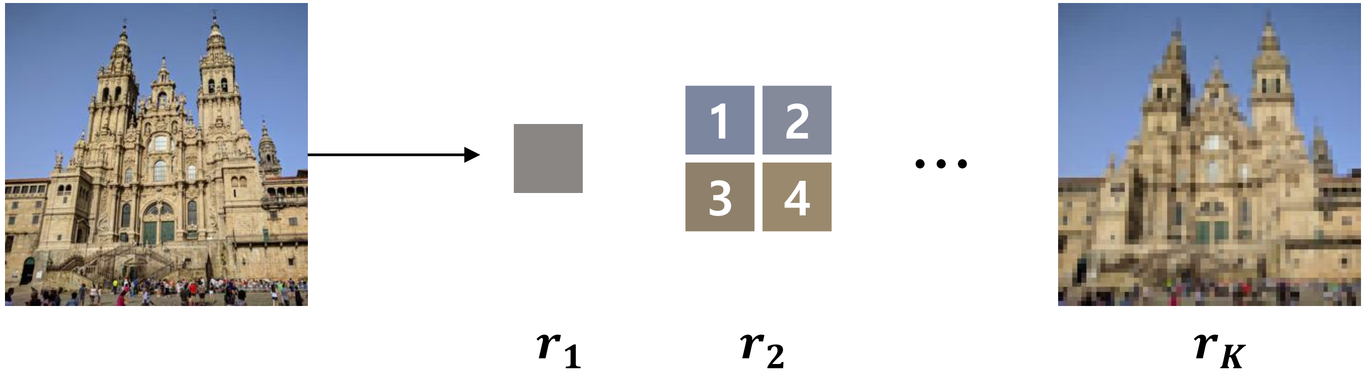


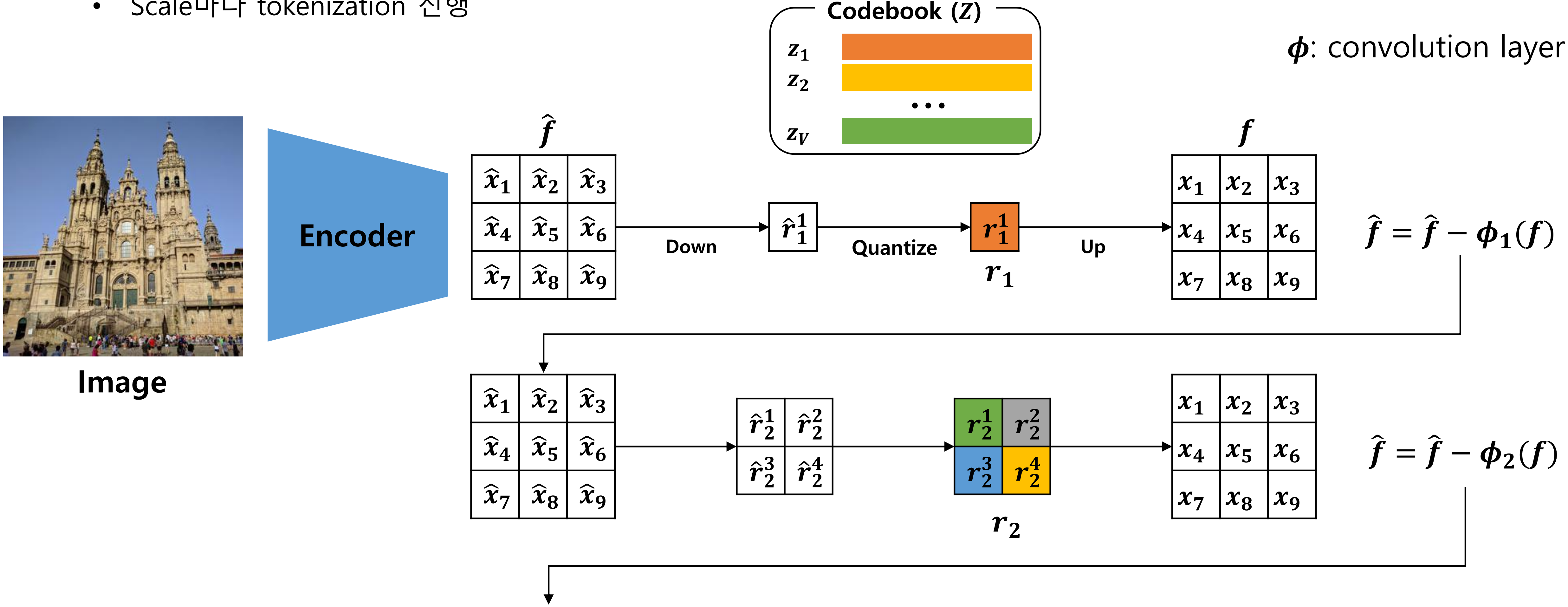
Image **Tokens**

<Scale-wise Tokenization>

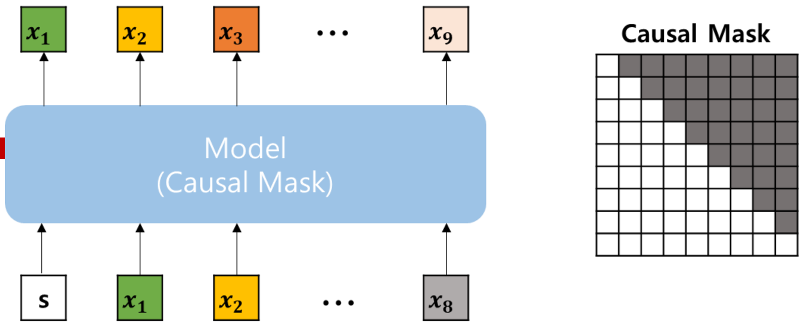
Next-Scale Prediction

Step1: Tokenization

- Scale마다 tokenization 진행

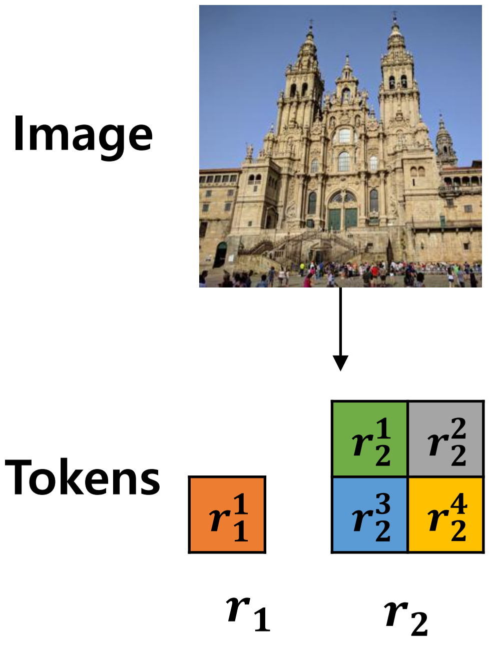


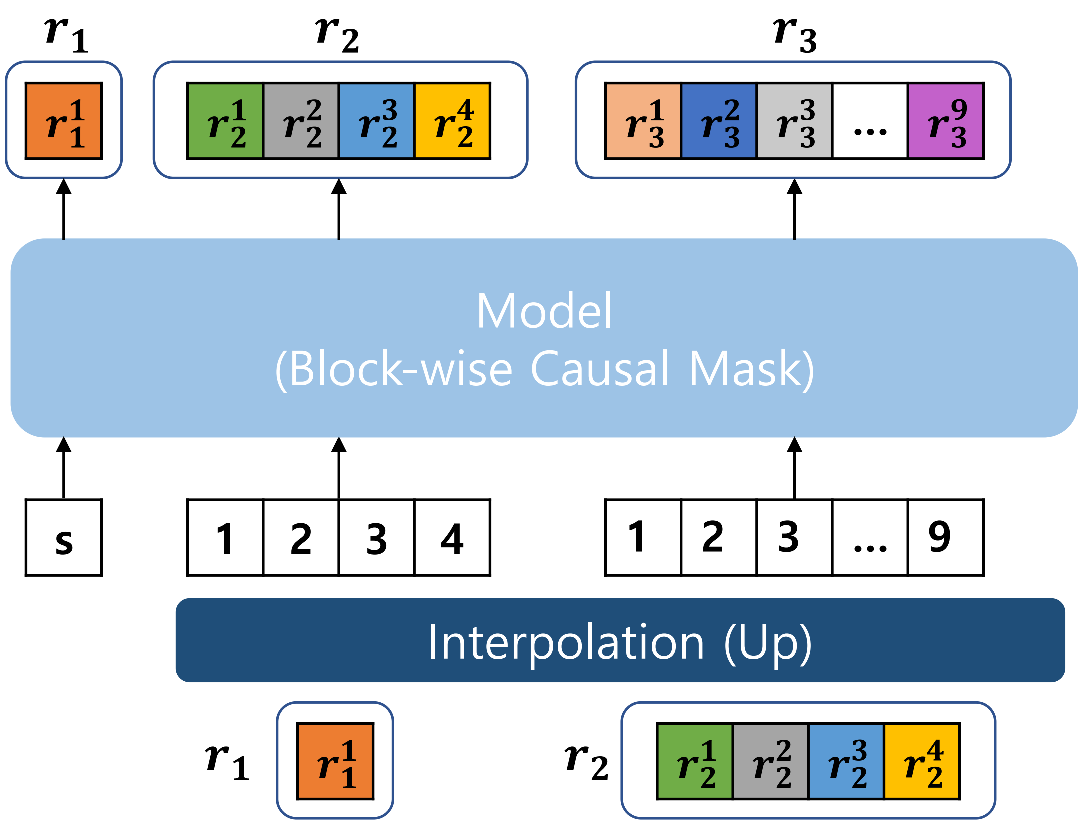
Next-Scale Prediction

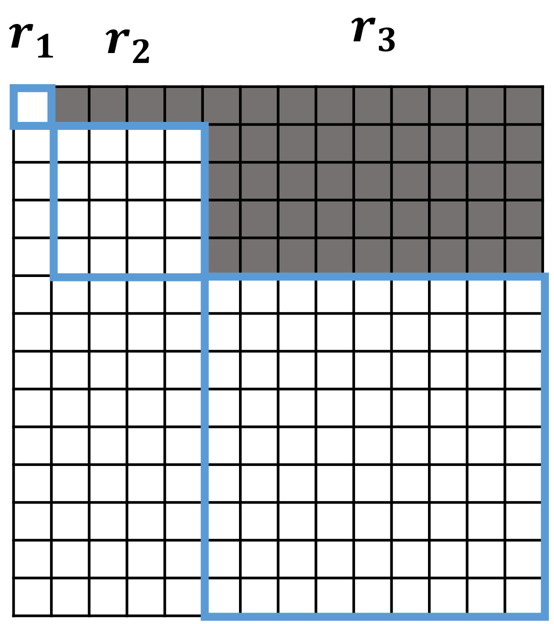


Step2: Modeling

- 이전 scale을 입력 받아 next-scale에 대한 예측을 진행





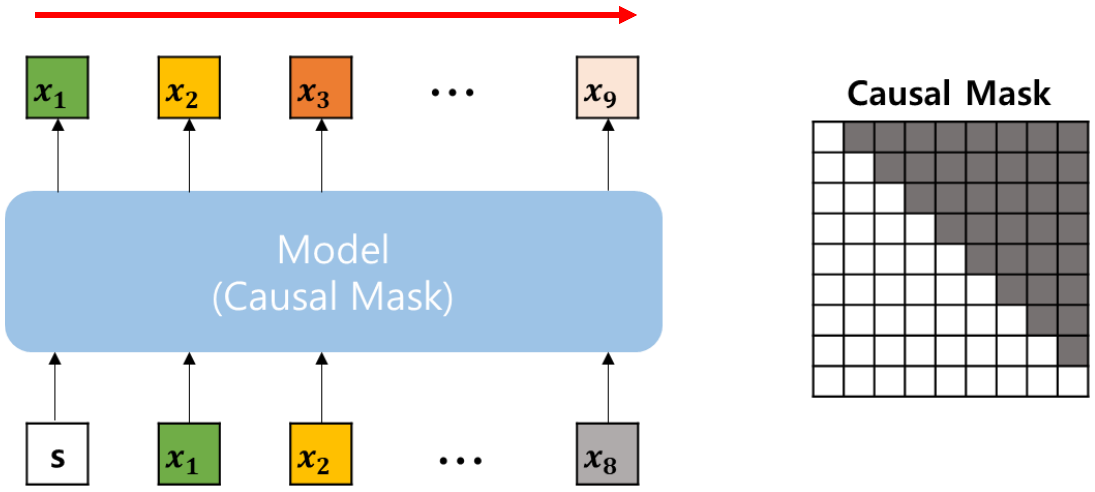


Block-wise Causal Mask

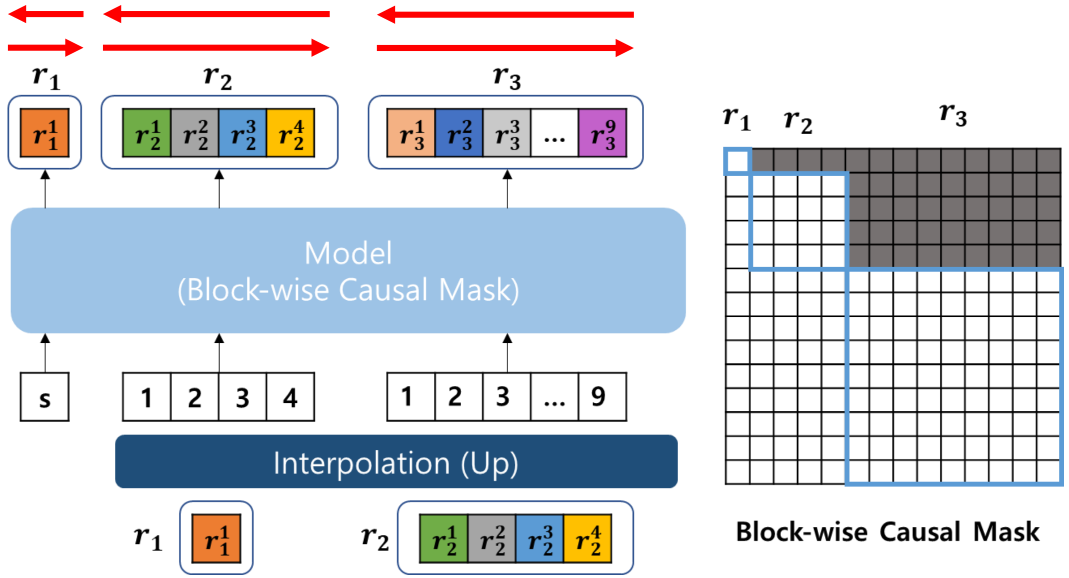
Next-Scale Prediction

Next-Token vs Next-Scale

- Unidirectional vs Bidirectional



<Next-Token Prediction>

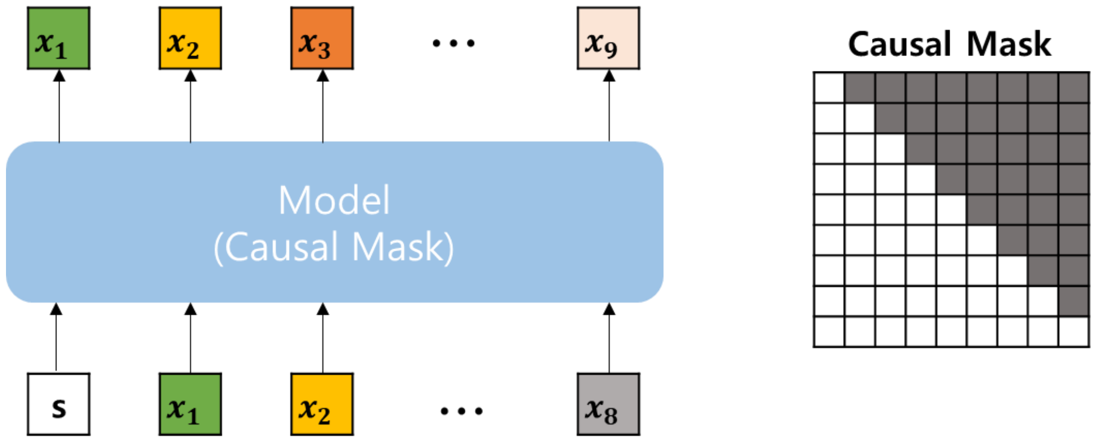


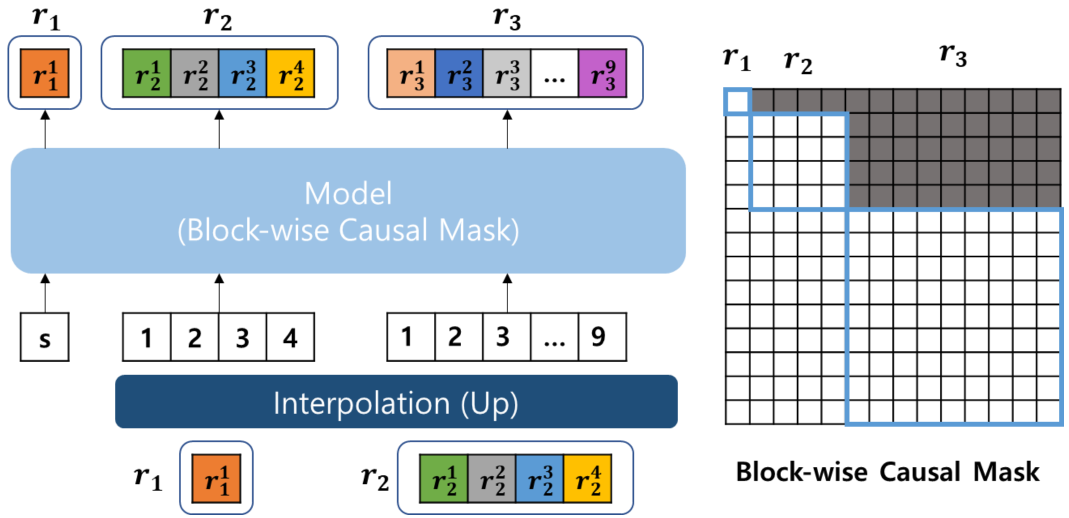
<Next-Scale Prediction>

Next-Scale Prediction

Next-Token vs Next-Scale

- Efficiency (Appendix B*)





$$1^2 + 2^2 + \cdots + 9^2 = 285$$

$$O(n^6)$$

$$1^2 + 5^2 + 14^2 = 222$$

$$O(n^4)$$

*Tian, K., Jiang, Y., Yuan, Z., Peng, B., & Wang, L. (2024). Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems, 37, 84839-84865.

Next-Scale Prediction

Experiments - Overall

- 짧은 시간에 고품질의 이미지 생성

Type	Model	FID↓	IS↑	Pre↑	Rec↑	#Para	#Step	Time
GAN	BigGAN [13]	6.95	224.5	**0.89**	0.38	112M	1	–
GAN	GigaGAN [42]	3.45	225.5	0.84	**0.61**	569M	1	–
GAN	StyleGan-XL [74]	2.30	265.1	0.78	0.53	166M	1	0.3 [74]
Diff.	ADM [26]	10.94	101.0	0.69	0.63	554M	250	168 [74]
Diff.	CDM [36]	4.88	158.7	–	–	–	8100	–
Diff.	LDM-4-G [70]	3.60	247.7	–	–	400M	250	–
Diff.	DiT-L/2 [63]	5.02	167.2	0.75	0.57	458M	250	31
Diff.	DiT-XL/2 [63]	2.27	278.2	0.83	0.57	675M	250	45
Diff.	L-DiT-3B [3]	2.10	304.4	0.82	0.60	3.0B	250	>45
Diff.	L-DiT-7B [3]	2.28	316.2	0.83	0.58	7.0B	250	>45
Mask.	MaskGIT [17]	6.18	182.1	0.80	0.51	227M	8	0.5 [17]
Mask.	RCG (cond.) [51]	3.49	215.5	–	–	502M	20	1.9 [51]
AR	VQVAE-2† [68]	31.11	~45	0.36	0.57	13.5B	5120	–
AR	VQGAN† [30]	18.65	80.4	0.78	0.26	227M	256	19 [17]
AR	VQGAN [30]	15.78	74.3	–	–	1.4B	256	24
AR	VQGAN-re [30]	5.20	280.3	–	–	1.4B	256	24
AR	ViTVQ [92]	4.17	175.1	–	–	1.7B	1024	>24
AR	ViTVQ-re [92]	3.04	227.4	–	–	1.7B	1024	>24
AR	RQTran. [50]	7.55	134.0	–	–	3.8B	68	21
AR	RQTran.-re [50]	3.80	323.7	–	–	3.8B	68	21
VAR	VAR-*d*16	3.30	274.4	0.84	0.51	310M	10	0.4
VAR	VAR-*d*20	2.57	302.6	0.83	0.56	600M	10	0.5
VAR	VAR-*d*24	2.09	312.9	0.82	0.59	1.0B	10	0.6
VAR	VAR-*d*30	1.92	323.1	0.82	0.59	2.0B	10	1
VAR	VAR-*d*30-re	1.73	**350.2**	0.82	0.60	2.0B	10	1

Next-Scale Prediction

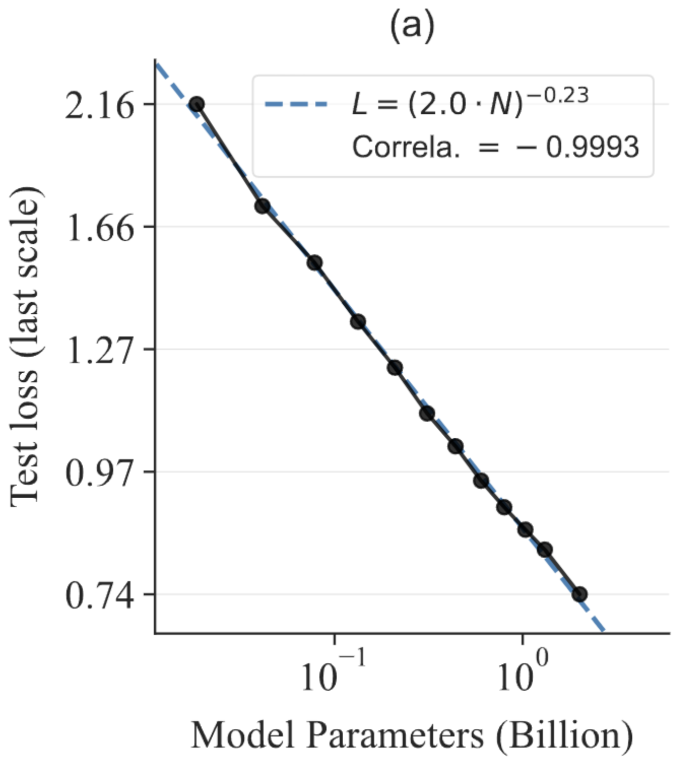
Experiments

- LLM의 성공 요인은 **scalability**와 **generalizability**
- **Scalability**: 모델의 크기가 증가할수록 모델의 성능도 증가함
- **Generalizability**: 다양한 task에 zero-shot 혹은 few-shot으로도 좋은 성능을 보임

Next-Scale Prediction

Experiments - Scalability

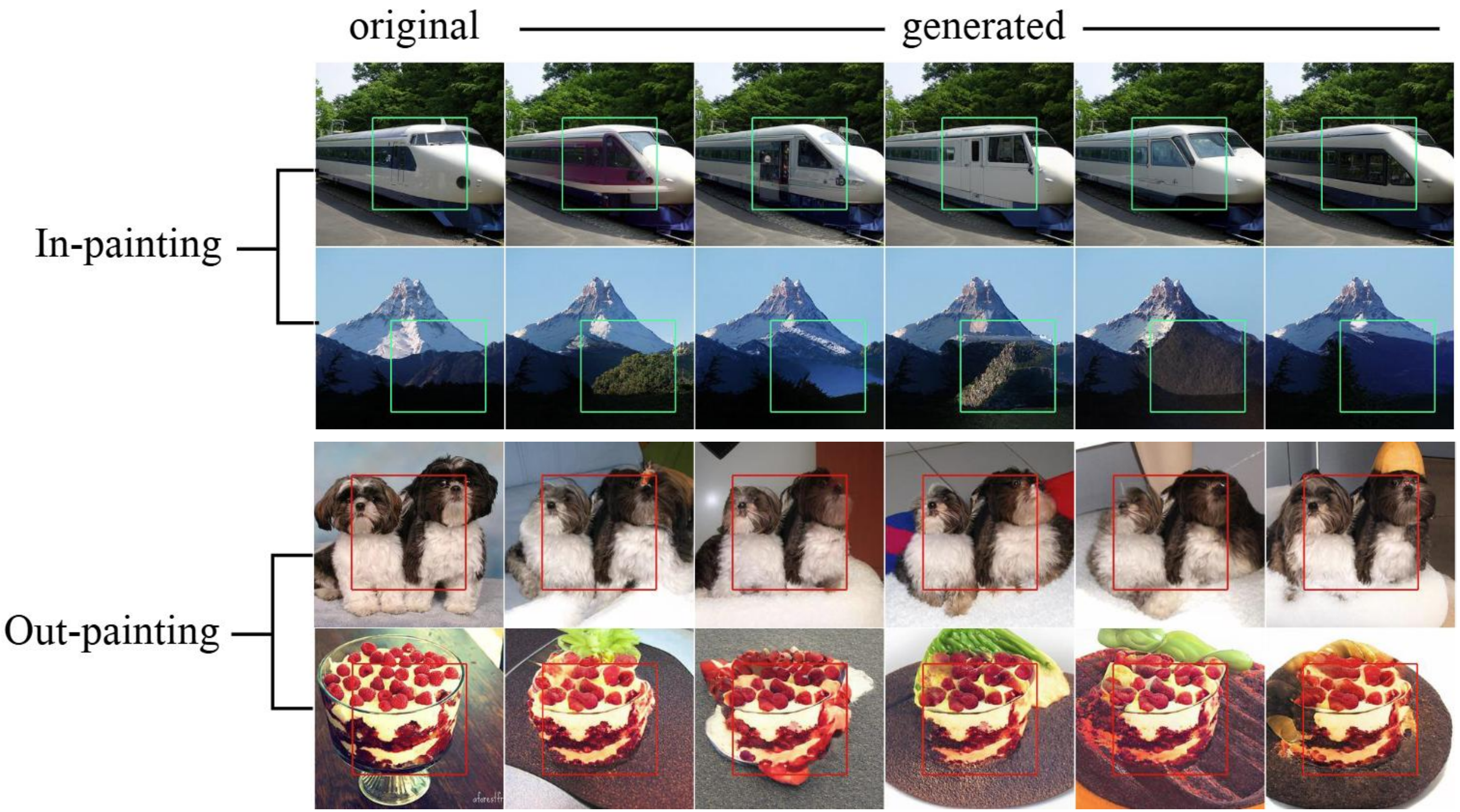
- 모델의 depth 증가에 따라서 모델 파라미터 증가 (12개의 모델)
- 모델 파라미터가 증가할수록 test loss가 감소
- 작은 모델로 다양한 실험 → 큰 모델로 확장 가능, 자원의 효율적인 배분 가능



Next-Scale Prediction

Experiments - Generalizability

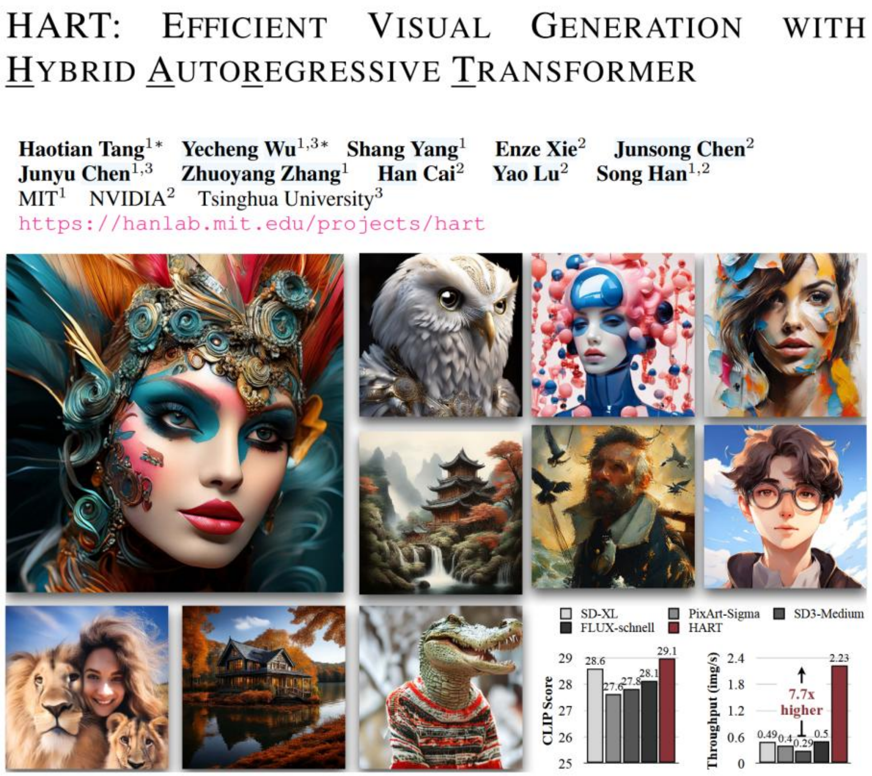
- Zero-shot으로 in-painting, out-painting이 가능하다는 것을 보여줌



Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

- ICLR 2025 / MIT, NVIDIA, 칭화대
- 60회 인용 (2025년 9월 12일)

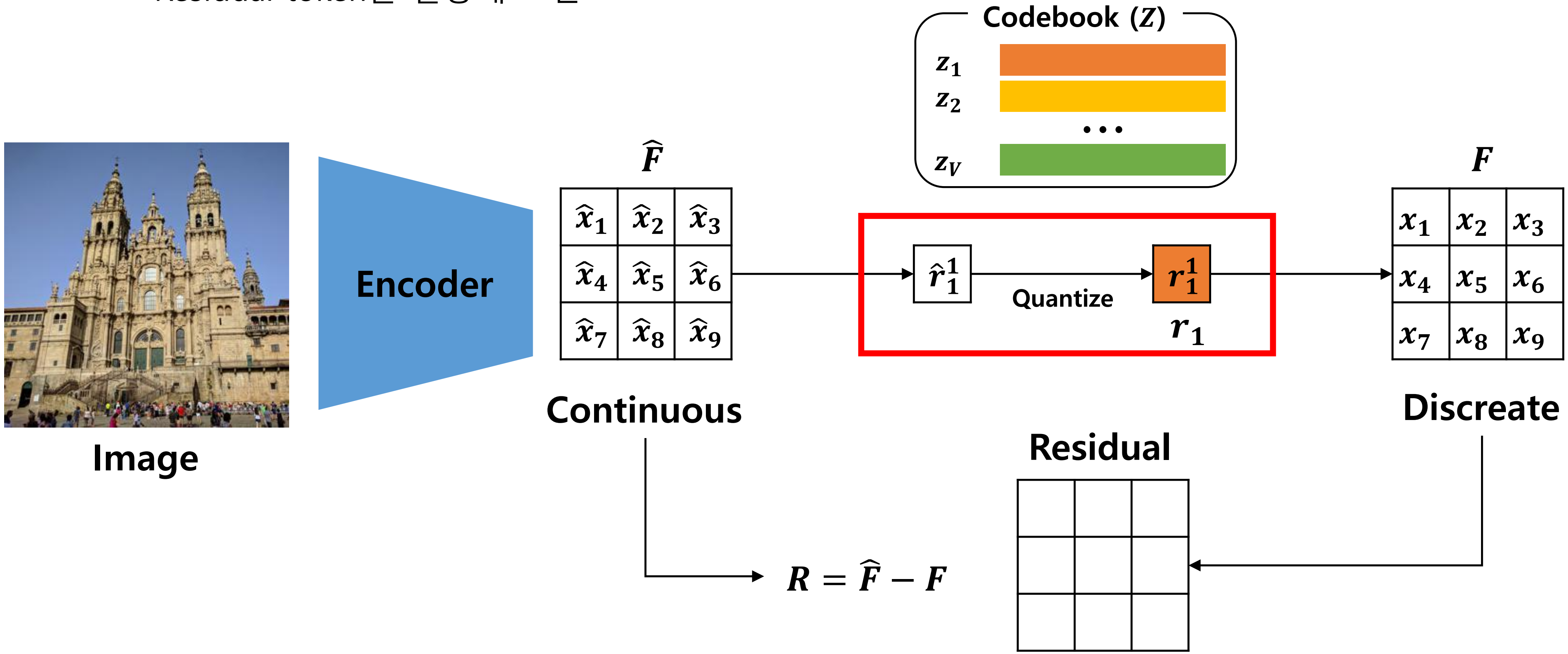


Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., ... & Han, S. HART: Efficient Visual Generation with Hybrid Autoregressive Transformer. In *The Thirteenth International Conference on Learning Representations.*

Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

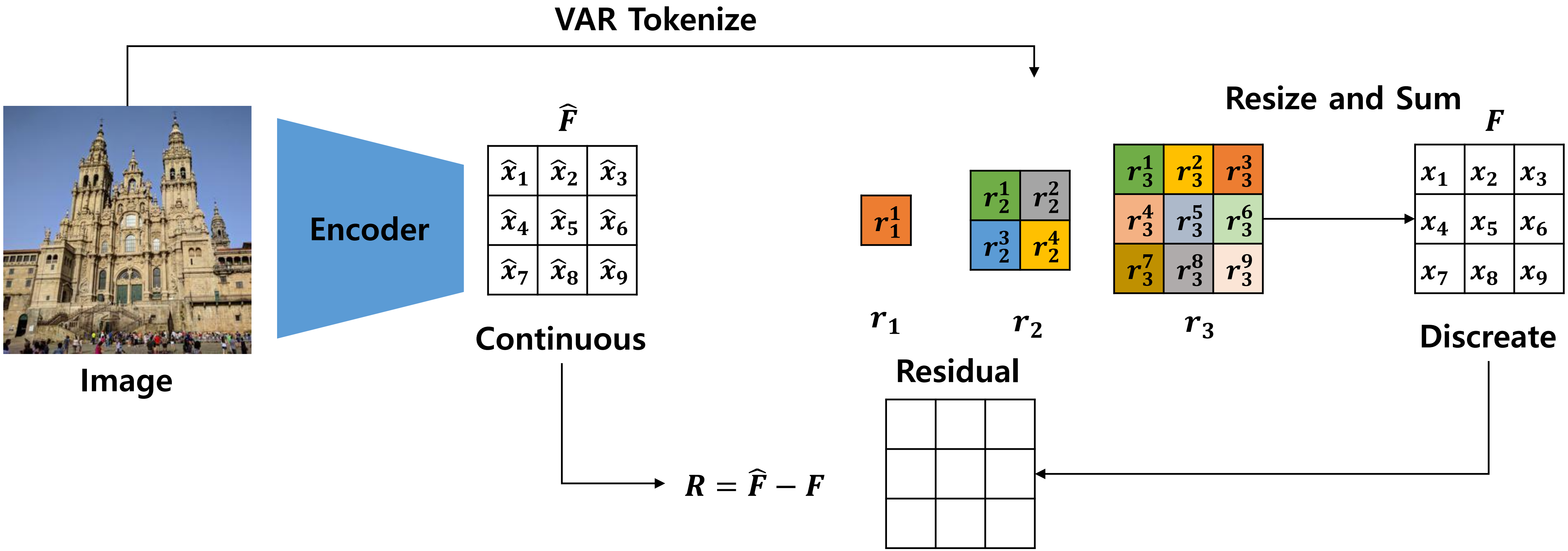
- Continuous → discreate token이 되는 과정에서 정보 손실
- Residual token을 활용해 보완



Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

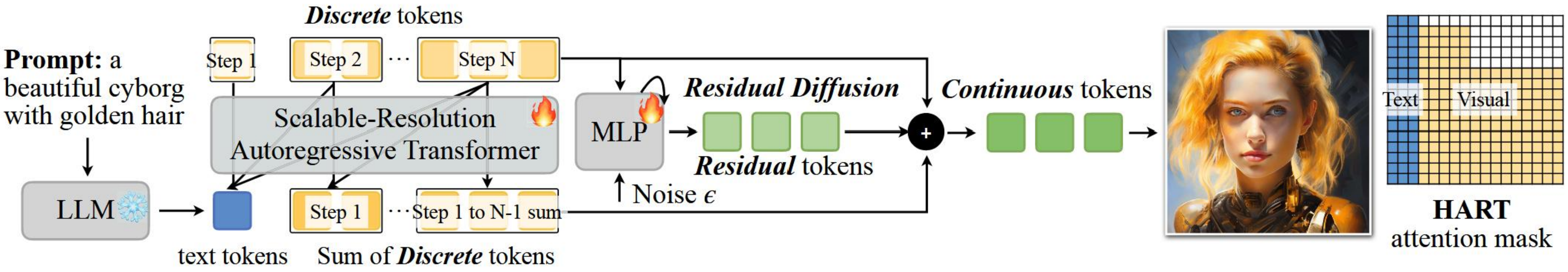
- Token들의 합을 discreate token, feature map을 continuous token으로 정의
- Continuous token과 discreate token으로 residual token 정의



Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

- Text에 대한 정보를 첫번째 토큰으로 사용
- Residual token을 diffusion 모델을 활용해 생성
- Continuous = Discreate + Residual



Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

- HART의 tokenizer가 VAR에 비해 우수한 성능

Reference (1k×1k) VAR (discrete) **HART** (hybrid)

Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

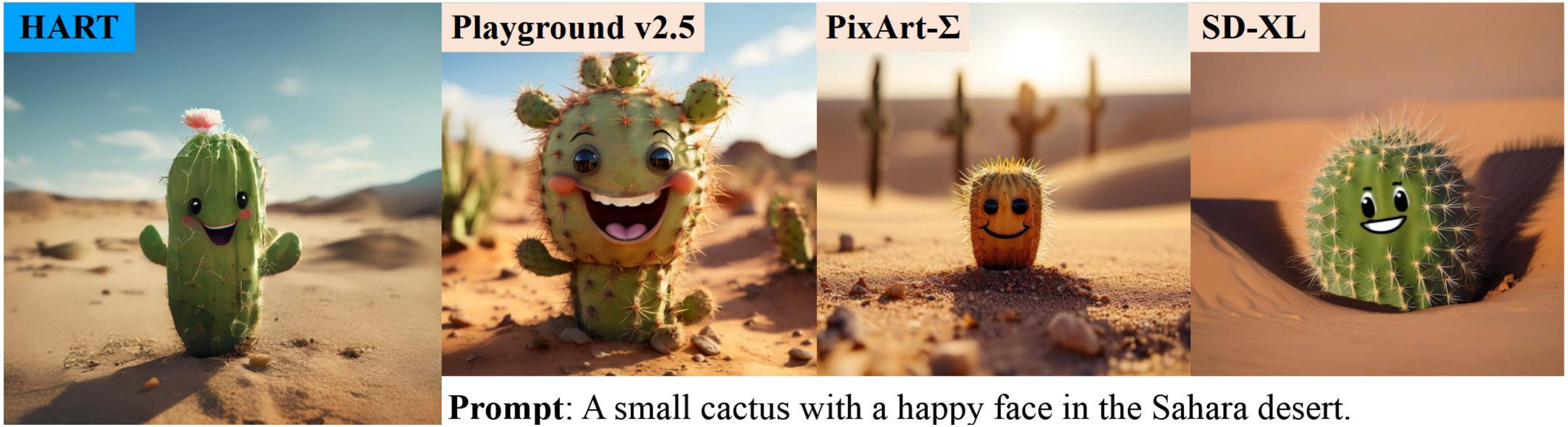
- Residual token은 이미지의 디테일을 담당

Prompt: A close up of a helmet on a person, digital art, victorian armor.

Text-to-Image Generation

HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

- VAR을 활용하여 SOTA인 diffusion 기반 모델들과 유사한 이미지 생성 퀄리티



Prompt: A small cactus with a happy face in the Sahara desert.

Conclusion

From Next-Token to Next-Scale Prediction

- LLM의 성공으로 컴퓨터 비전 분야에서도 autoregressive model에 대한 연구 수행
- AR은 unidirectional, computational cost ↑
- VAR은 bidirectional, computational cost ↓
- HART는 residual token을 활용해 text-to-image generation으로 확장