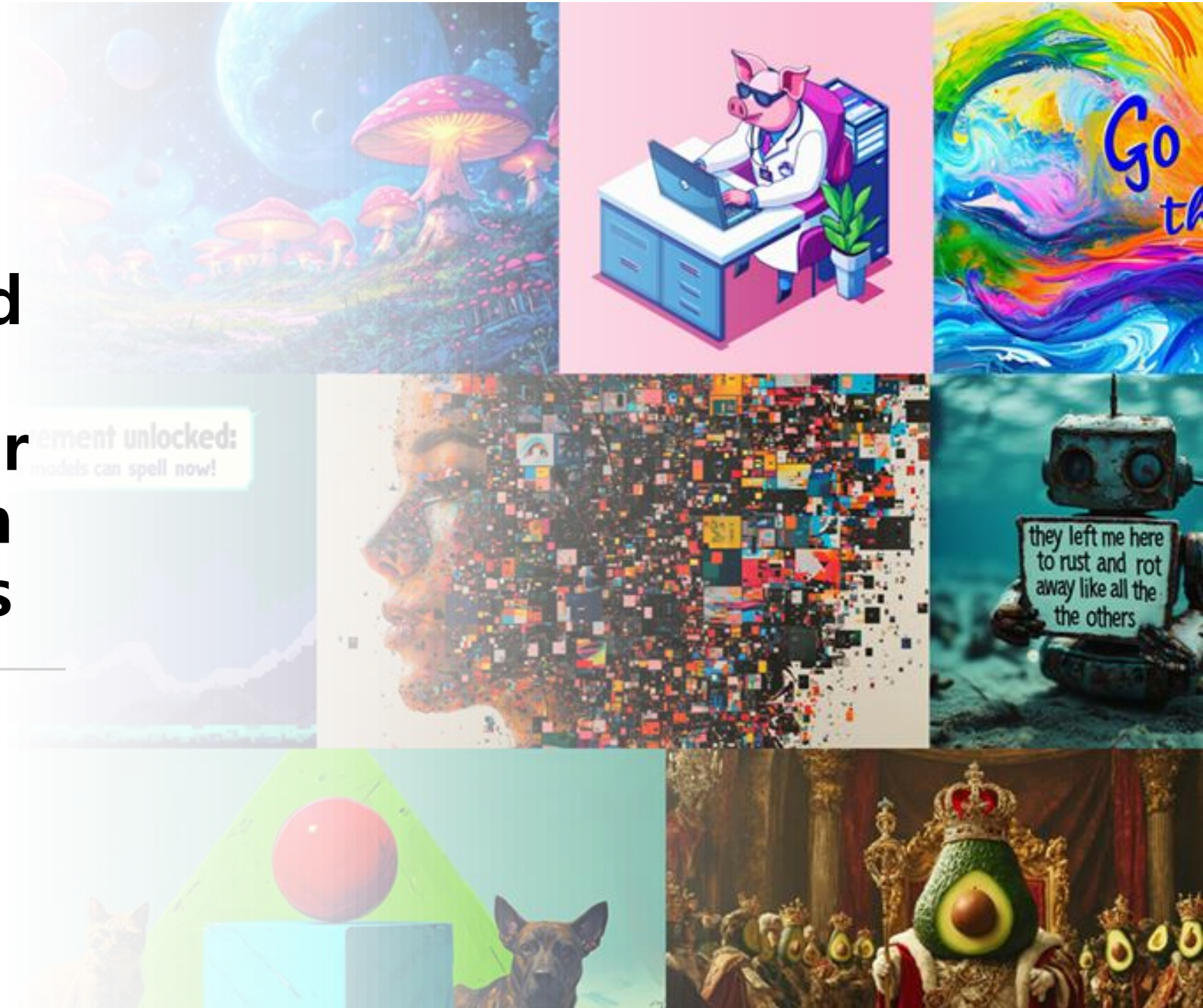
Scaling Rectified Flow Transformers for High-Resolution Image Synthesis

20220706 박주연
20230040 김유진



목차

- Diffusion model
- Rectified Flow
- 논문 goal 소개
- Simulation – Free Training of Flows & Flow Trajectories
- Tailored SNR Samplers for RF models (Timestep improvement)
- Text-to-image Architecture
- Experiments

What is Diffusion model?



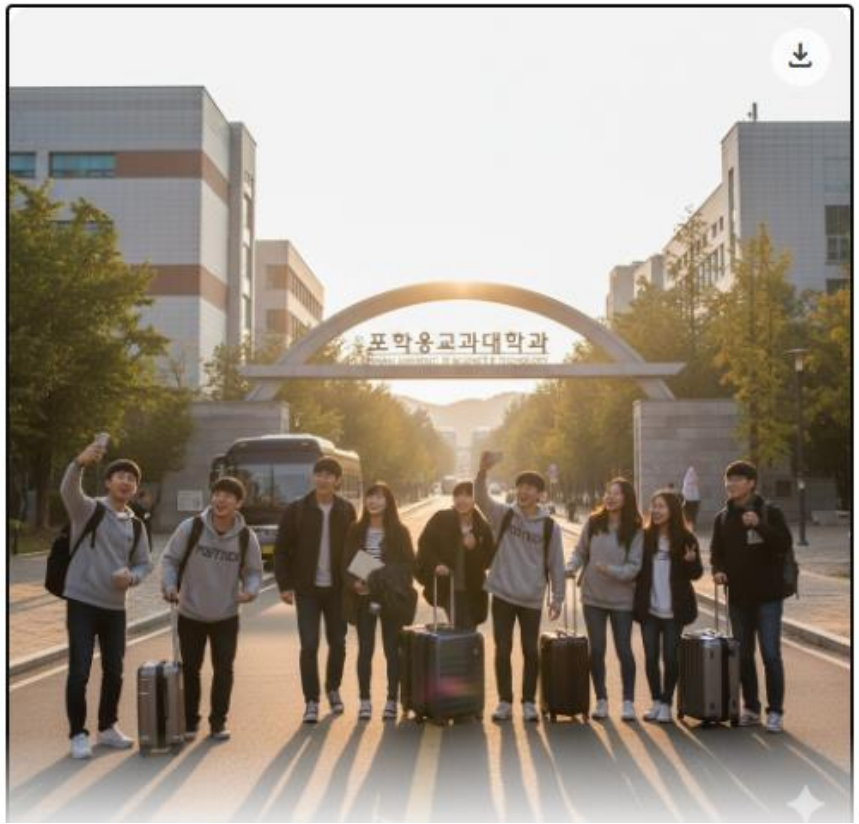

통동
토동
폭삭 속아수다



포항공과대학교 학생들이 종강 후 행복하게 학교에서 집으로 가는 사진 만들어줘

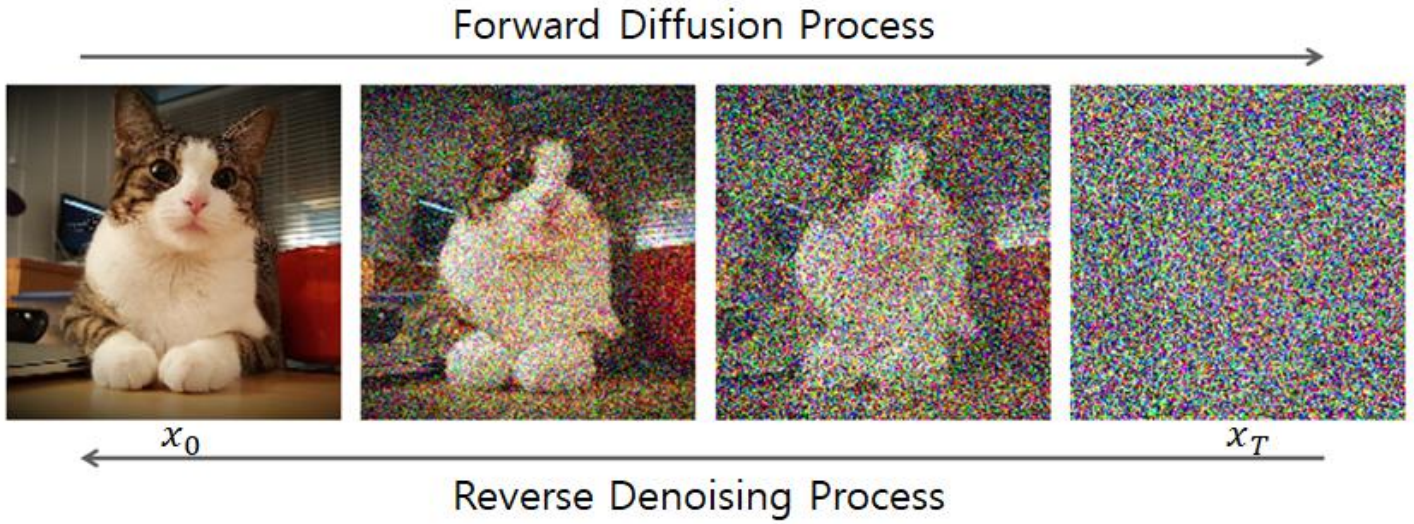
물론이죠! 여기 포항공과대학교 학생들이 종강 후 행복하게 집으로 가는 사진입니다.



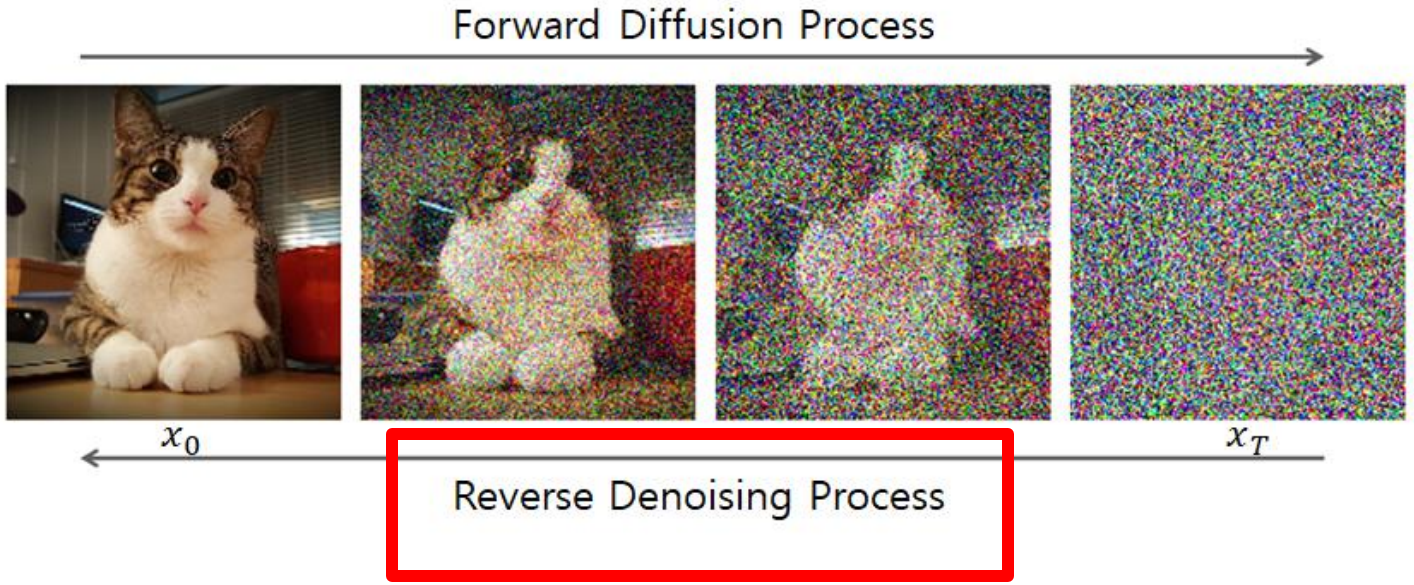




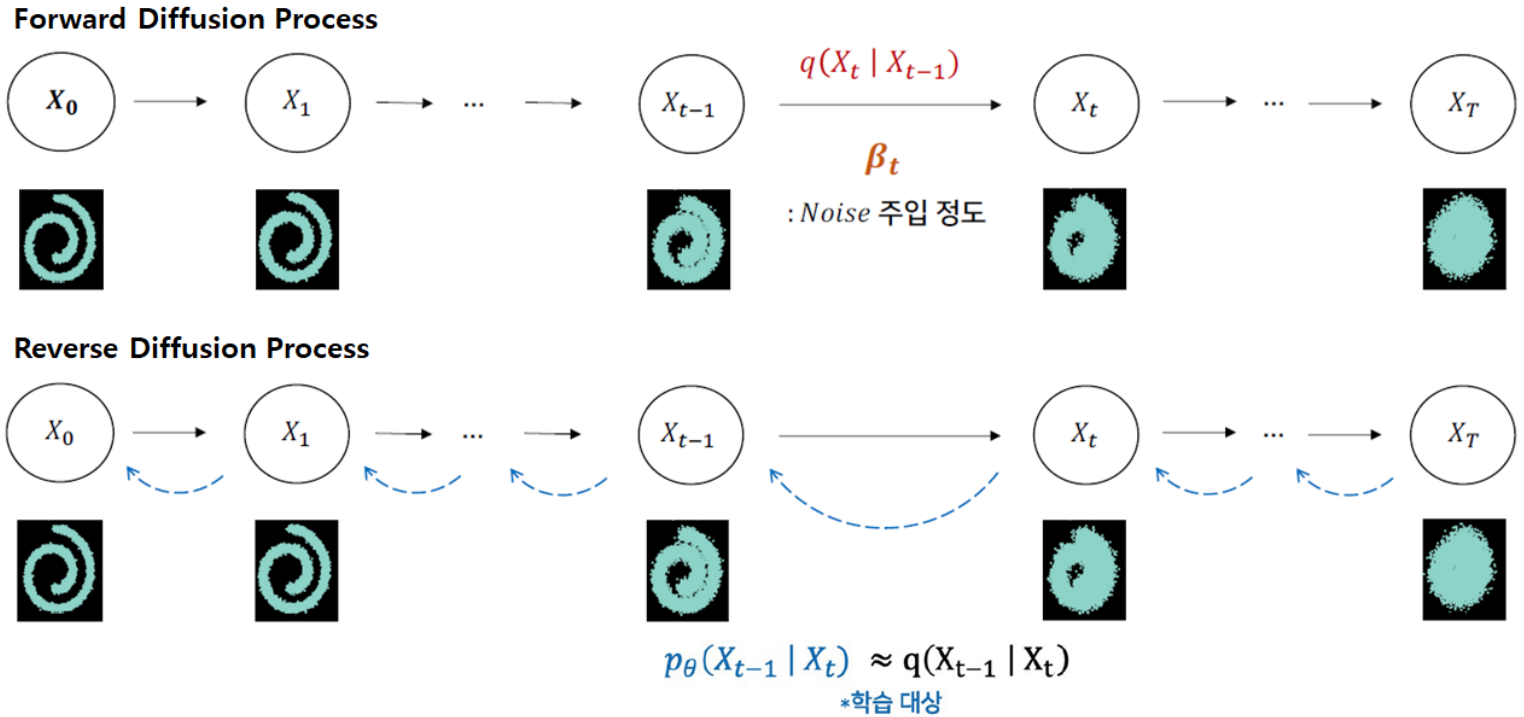
What is Diffusion model?



What is Diffusion model?

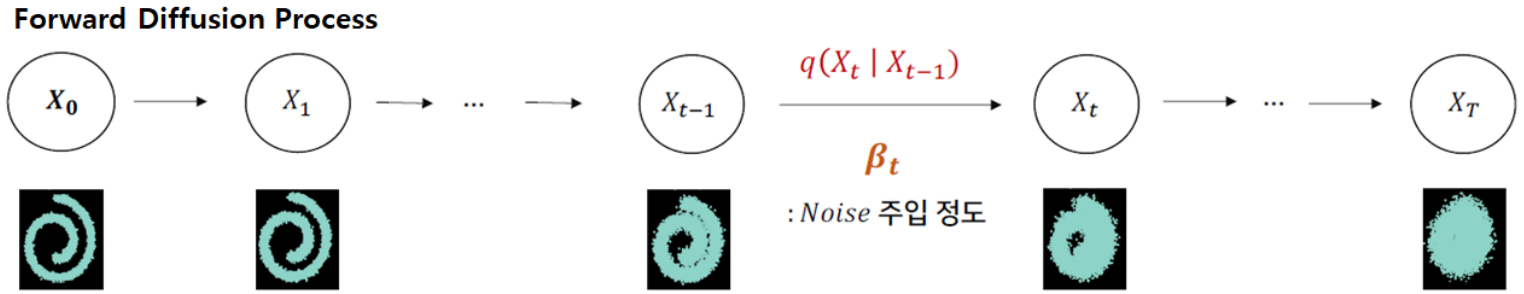


What is Diffusion model?



$$q(X_t \mid X_{t-1}) := N(X_t ; \mu_{X_{t-1}}, \Sigma_{X_{t-1}}) := N\left(X_t ; \sqrt{1-\beta_t}\, X_{t-1},\ \beta_t \cdot I\right)$$

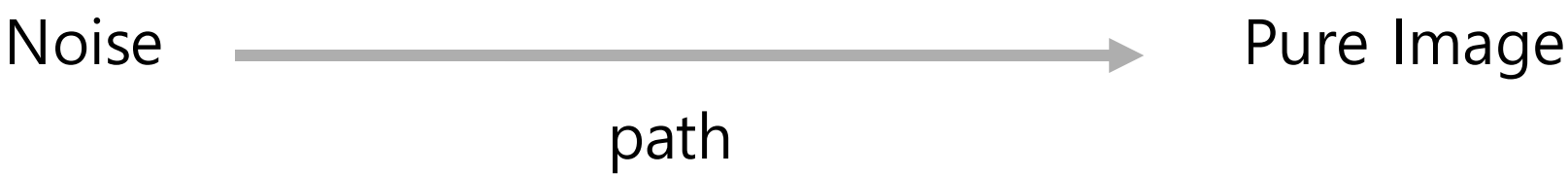
What is Diffusion model?



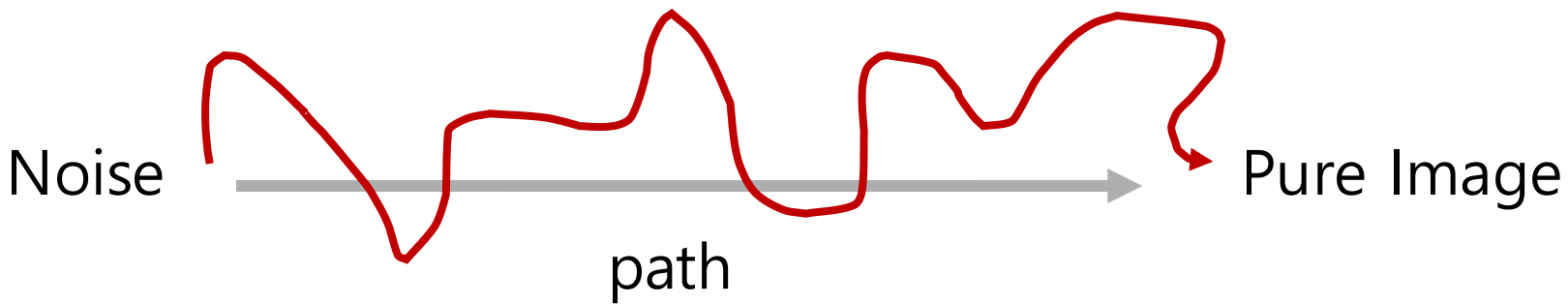
$$q(x_{t-1}|x_t) = \frac{q(x_t|x_{t-1}) \cdot q(x_{t-1})}{q(x_t)}$$

$$q(X_t \mid X_{t-1}) := N(X_t ; \mu_{X_{t-1}}, \Sigma_{X_{t-1}}) := N\left(X_t ; \sqrt{1-\beta_t}\, X_{t-1},\ \beta_t \cdot I\right)$$

What is Diffusion model?



What is Diffusion model?



curved paths require many integration steps to simulate the process

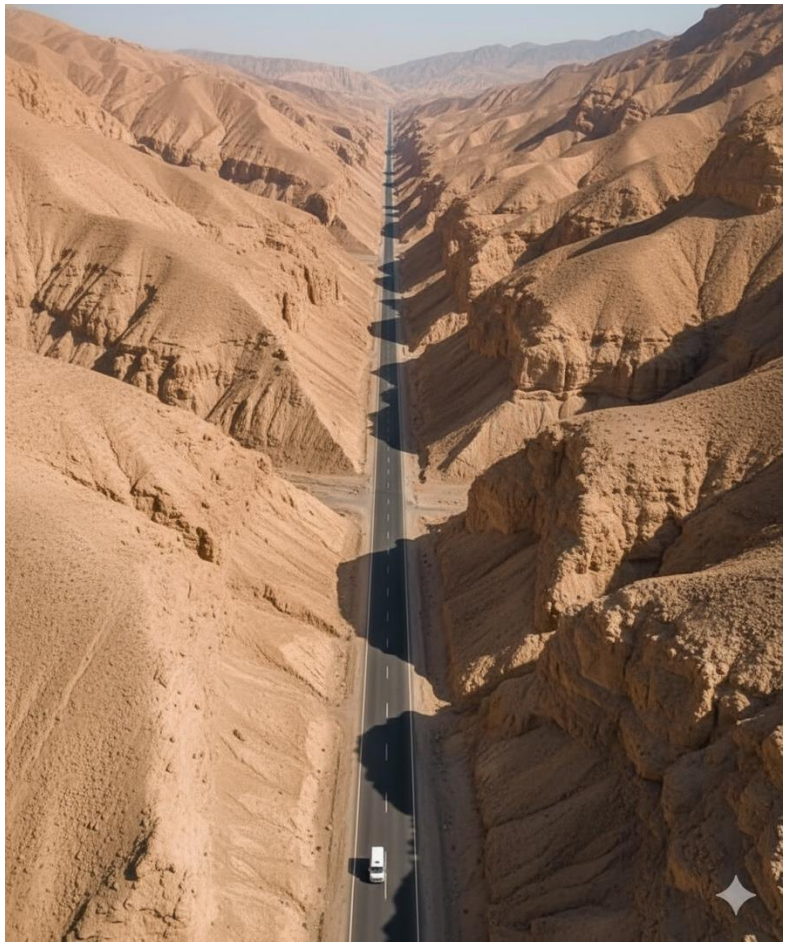
➔ 이미지 생성 시 많은 시간 소요

What is Diffusion model? & Rectified Flow



What is Diffusion model? & Rectified Flow





What is Rectified Flow?

consider generative models that define a mapping between samples x1 from a noise distribution p1 to samples x0 from a data distribution p0 in terms of an ordinary differential equation (ODE)

$$dy_t = v_\Theta(y_t, t)\, dt$$

What is Rectified Flow?

consider generative models that define a mapping between samples x1 from a noise distribution p1 to samples x0 from a data distribution p0 in terms of an ordinary differential equation (ODE)

$$dy_t = v_\Theta(y_t, t)\, dt$$

Efficient alternative : Directly regress u_t

What is Rectified Flow?

consider generative models that define a mapping between samples x1 from a noise distribution p1 to samples x0 from a data distribution p0 in terms of an ordinary differential equation (ODE)

$$dy_t = v_\Theta(y_t, t)\, dt$$

우리가 만든 이상적인 경로가 t시점에 가져야 할 속도

Efficient alternative : Directly regress u_t

What is Rectified Flow?

consider generative models that define a mapping between samples x1 from a noise distribution p1 to samples x0 from a data distribution p0 in terms of an ordinary differential equation (ODE)

$$dy_t = v_\Theta(y_t, t)\, dt$$

우리가 만든 이상적인 경로가 t시점에 가져야 할 속도

Efficient alternative : Directly regress u_t

Forward Process $z_t = a_t x_0 + b_t \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, I)$

What is Rectified Flow?

$$z_t = a_t x_0 + b_t \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, I)$$

$$a_t := 1 - t \text{ and } b_t := t$$

What is Rectified Flow?

$z_t = a_t x_0 + b_t \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, I)$

$a_t := 1 - t \text{ and } b_t := t$



$$z_t = (1 - t)x_0 + t\epsilon$$

define the **forward process** as **straight paths** between the data distribution and a standard normal distribution

What is Rectified Flow?

Mathematical Part of this paper , Section 2

: Regress velocity → Regress Noise

z_t can be written as solution to the ODE $z'_t = u_t(z_t|\epsilon)$

$$\mathcal{L}_{FM} = \mathbb{E}_{t,p_t(z)}||v_\Theta(z,t) - u_t(z)||_2^2. \qquad (7)$$

intractable

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)}||v_\Theta(z,t) - u_t(z|\epsilon)||_2^2 \,, \qquad (8)$$

What is Rectified Flow?

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)} ||v_\Theta(z,t) - \frac{a'_t}{a_t} z + \frac{b_t}{2}\lambda'_t \epsilon||_2^2 \quad (11)$$

$$= \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)} \left(-\frac{b_t}{2}\lambda'_t\right)^2 ||\epsilon_\Theta(z,t) - \epsilon||_2^2 \quad (12)$$

where we defined $\epsilon_\Theta := \frac{-2}{\lambda'_t b_t}(v_\Theta - \frac{a'_t}{a_t} z)$.

$$u_t(z_t|\epsilon) = \frac{a'_t}{a_t} z_t - \frac{b_t}{2}\lambda'_t \epsilon \quad (10)$$

What is Rectified Flow?

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)} ||v_\Theta(z,t) - \frac{a'_t}{a_t} z + \frac{b_t}{2}\lambda'_t \epsilon||_2^2 \quad (11)$$

$$= \mathbb{E}_{t,p_t(z|\epsilon),p(\epsilon)} \left(-\frac{b_t}{2}\lambda'_t\right)^2 ||\epsilon_\Theta(z,t) - \epsilon||_2^2 \quad (12)$$

where we defined $\epsilon_\Theta := \frac{-2}{\lambda'_t b_t}(v_\Theta - \frac{a'_t}{a_t} z)$.

$$u_t(z_t|\epsilon) = \frac{a'_t}{a_t} z_t - \frac{b_t}{2}\lambda'_t \epsilon \quad (10)$$

실제 노이즈와 모델이 예측한 노이즈 차이가 줄도록 손실함수 설계

What is Rectified Flow?

- 이론적으로 단순
- 소수의 스텝으로 샘플링 가능(i.e. 이미지 생성 가능)
- 작은 혹은 중간 사이즈의 데이터에서는 이점이 증명되었지만 class-conditional model에 국한

Goal of this paper

기존의 Rectified Flow 방법을

1. Timestep sampling
2. Text-to-Image Architecture
3. Scaling



Goal of this paper

기존의 Rectified Flow 방법을

1. Timestep sampling
2. Text-to-Image Architecture
3. Scaling

→

고해상도 Text-to-Image:
기존 모델 대비 성능 우위 확보

Timestep improvement

RF loss trains the velocity **uniformly** on all timesteps in [0,1]

Timestep improvement

RF loss trains the velocity **uniformly** on all timesteps in [0,1]

Hypothesis: 중간 시간대(t)의 속도 예측이 양 끝단보다 더 어렵다

1. Logit-Normal Sampling
2. Mode Sampling with Heavy Tails
3. CosMap

Timestep improvement-Logit-Normal Sampling

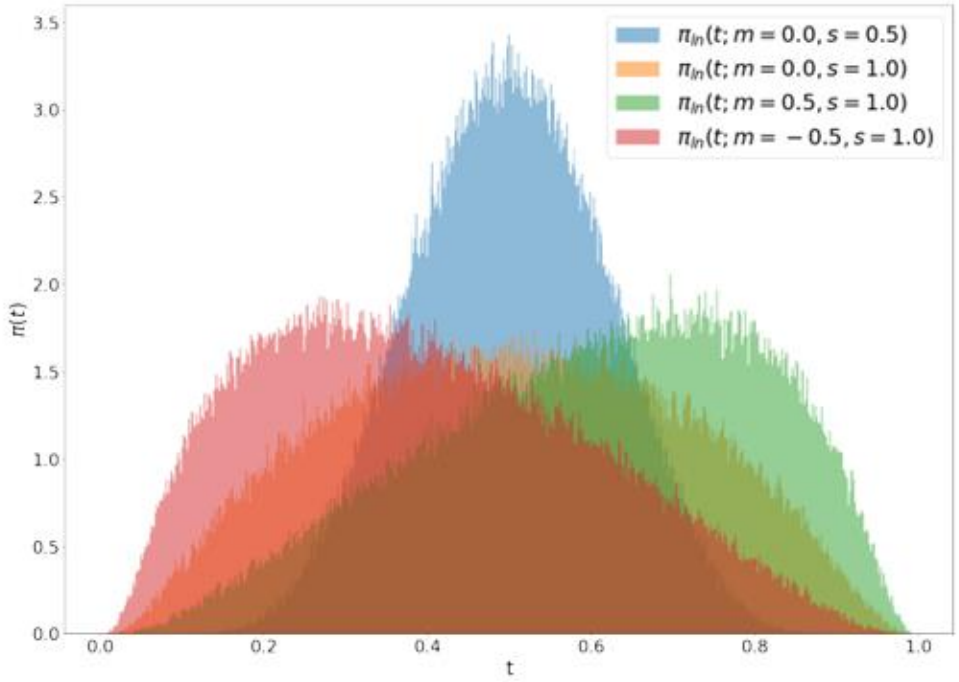
$$\pi_{\ln}(t; m, s) = \frac{1}{s\sqrt{2\pi}} \frac{1}{t(1-t)} \exp\left(-\frac{(\text{logit}(t) - m)^2}{2s^2}\right), \qquad \text{where } \text{logit}(t) = \log \frac{t}{1-t} \tag{19}$$

Parameter = m and s

Timestep improvement-Logit-Normal Sampling

$$\pi_{\ln}(t; m, s) = \frac{1}{s\sqrt{2\pi}} \frac{1}{t(1-t)} \exp\left(-\frac{(\text{logit}(t) - m)^2}{2s^2}\right), \qquad \text{where } \text{logit}(t) = \log \frac{t}{1-t} \tag{19}$$

Parameter = m and s



m= 좌우 편향
s= 높이

Timestep improvement- Mode Sampling with Heavy Tails

$$f_{\text{mode}}(u; s) = 1 - u - s \cdot \left(\cos^2\left(\frac{\pi}{2}u\right) - 1 + u\right). \quad (20)$$

For $-1 \leq s \leq \frac{2}{\pi - 2}$

$$\pi_{\text{mode}}(t; s) = \left|\frac{d}{dt} f^{-1}_{\text{mode}}(t)\right|$$

Timestep improvement- Mode Sampling with Heavy Tails

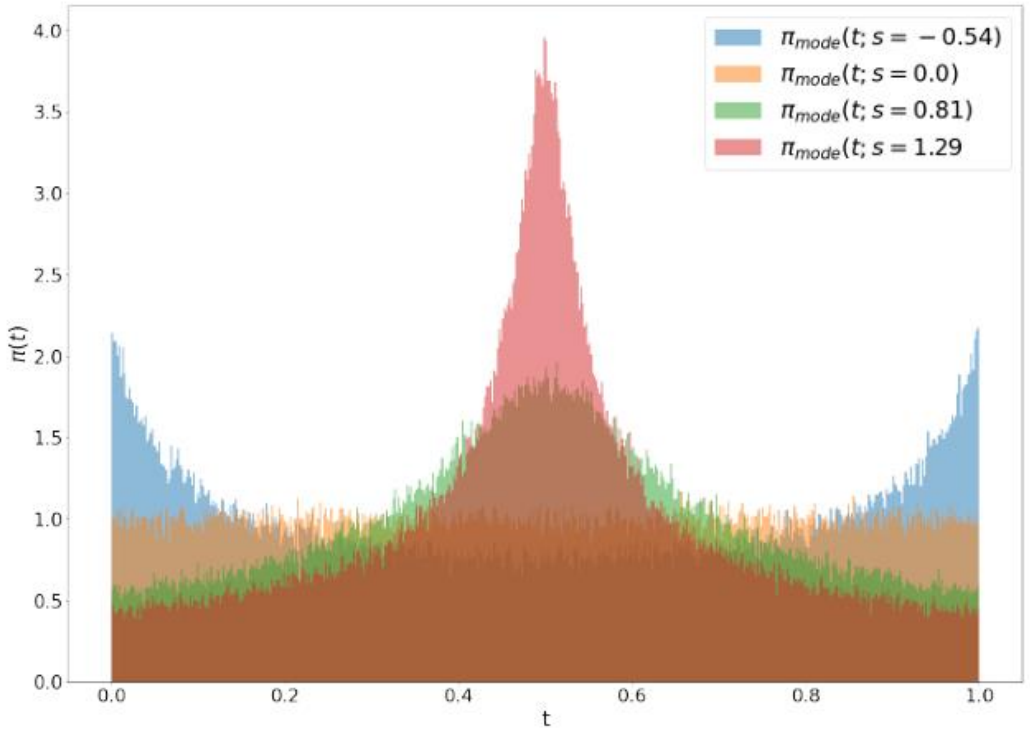
$$f_{\text{mode}}(u; s) = 1 - u - s \cdot \left(\cos^2\left(\frac{\pi}{2}u\right) - 1 + u\right). \quad (20)$$

For $-1 \leq s \leq \frac{2}{\pi - 2}$

$$\pi_{\text{mode}}(t; s) = \left|\frac{d}{dt} f^{-1}_{\text{mode}}(t)\right|$$

Parameter = s

S(+) : 가운데를 선호
S(-) : 양 끝을 선호



Timestep improvement- Mode Sampling with Heavy Tails

$$f_{\text{mode}}(u; s) = 1 - u - s \cdot \left(\cos^2\left(\frac{\pi}{2}u\right) - 1 + u\right). \quad (20)$$

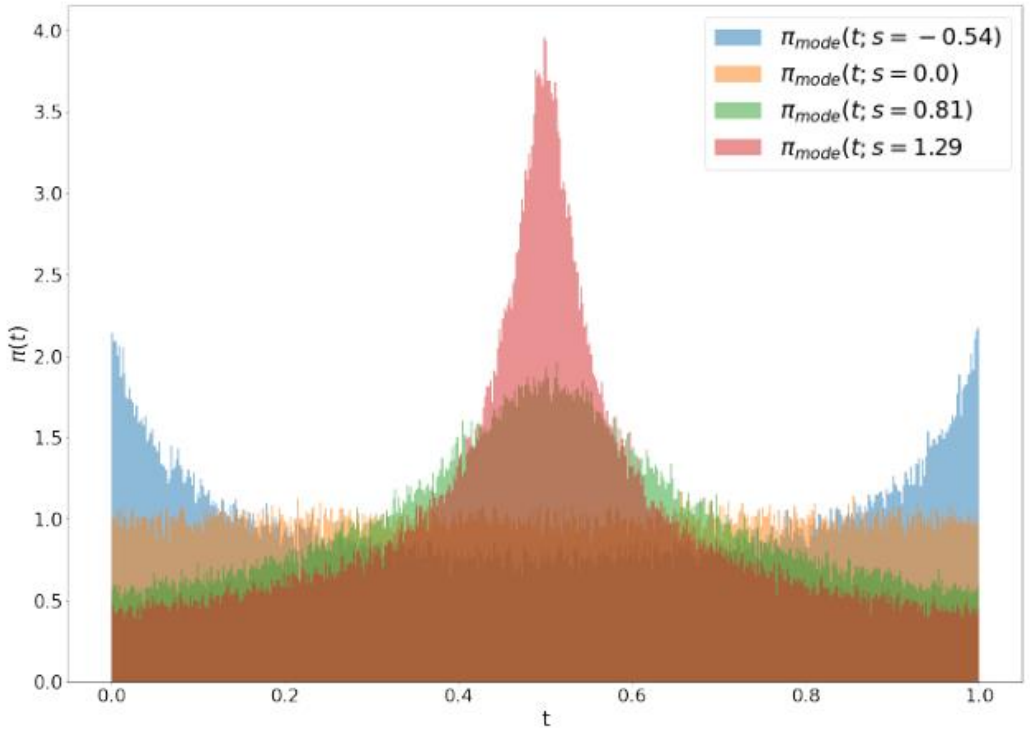
For $-1 \leq s \leq \frac{2}{\pi - 2}$

$$\pi_{\text{mode}}(t; s) = \left|\frac{d}{dt} f^{-1}_{\text{mode}}(t)\right|$$

Parameter = s

S(+) : 가운데를 선호
S(-) : 양 끝을 선호

Remark: for s = 0,
the probability density function is $\mathcal{U}(t)$



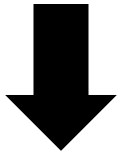
Timestep improvement- CosMap

$$t = f(u) = 1 - \frac{1}{\tan(\frac{\pi}{2}u) + 1}, \qquad (21) \qquad u \sim \mathcal{U}(u)$$

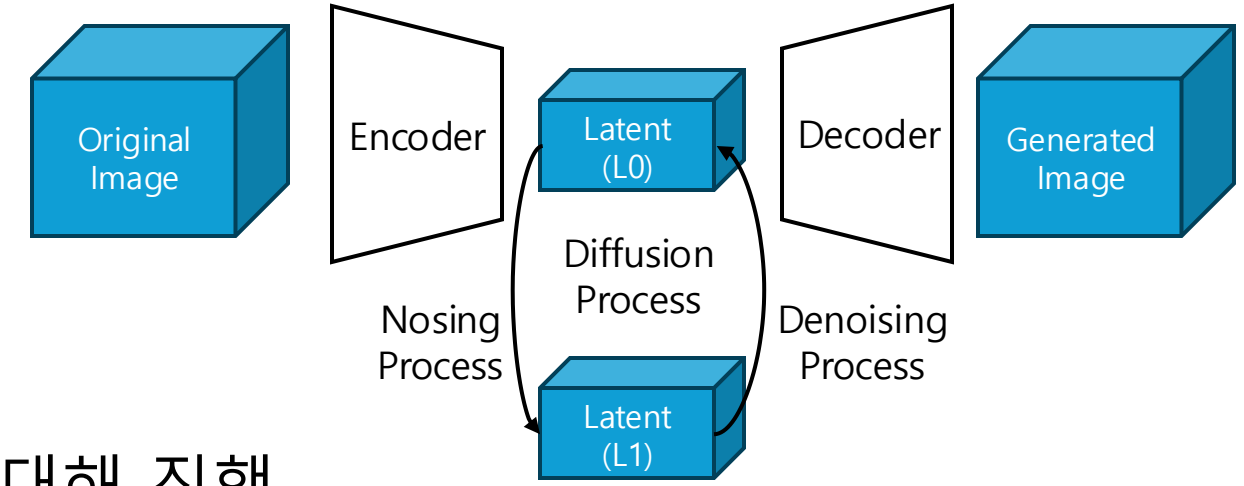
$$\pi_{\text{CosMap}}(t) = \left| \frac{d}{dt} f^{-1}(t) \right| = \frac{2}{\pi - 2\pi t + 2\pi t^2}. \qquad (22)$$

Text-to-Image Architecture- LDM

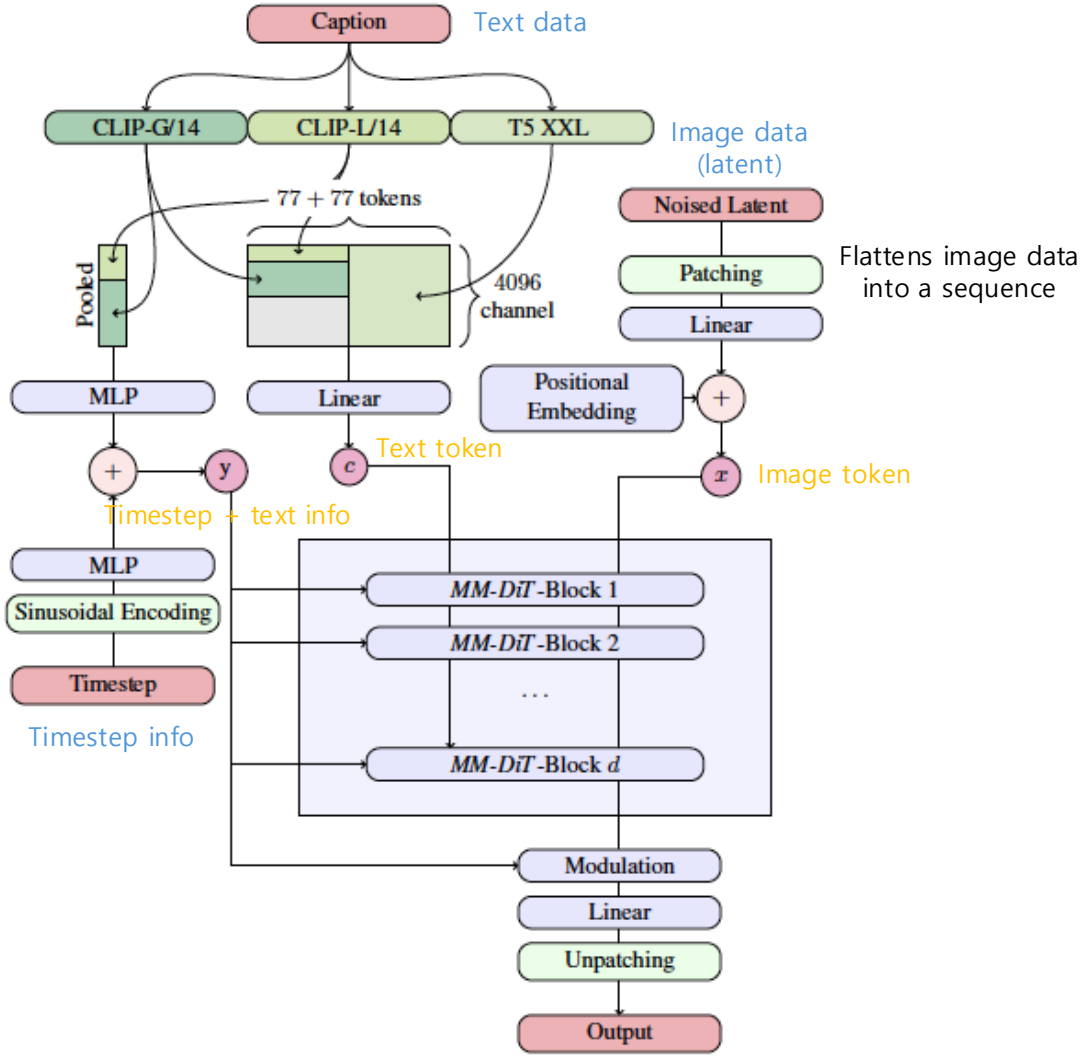
Diffusion process를 **각 픽셀**에 대해 실행
→ Slow
→ Requires large computing power



Latent Diffusion Model (LDM)
Diffusion process를 **압축된 이미지**에 대해 진행
→ Autoencoder (VAE) 활용



Text-to-Image Architecture- MM-DiT



(a) Overview of all components.

Input to the model

- Text data (Caption)

  : 답지 – 생성된 이미지가 나타내야 하는 이미지의 정보
- Timestep info (how much noise is in the image)

  : 얼마나 많은 denoising step이 필요한지
- Image data (Image encoded by LDM method)

  : noise가 있는 압축된 이미지

Input to MM-DiT

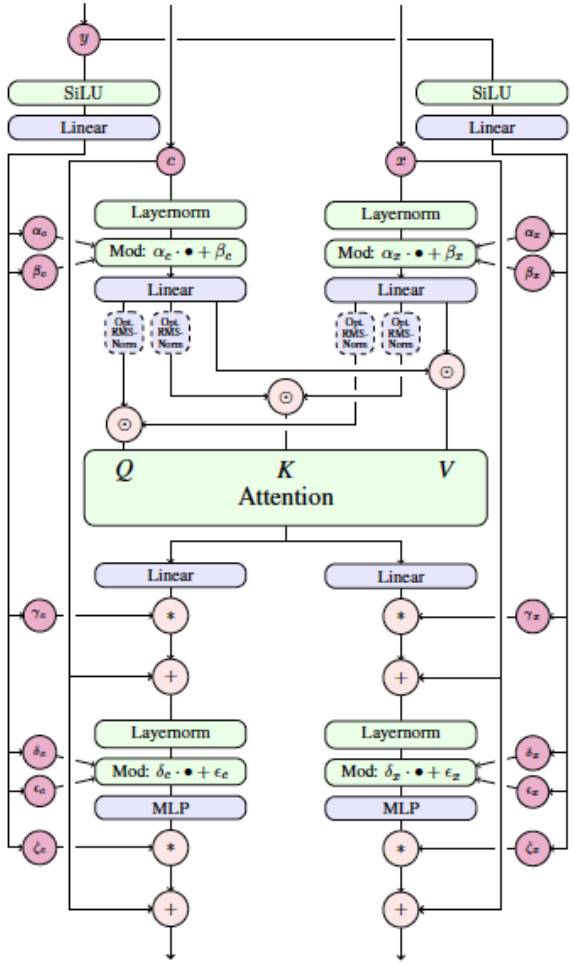
- Text token (c)

  : 3가지 text encoder를 통해 얻은 token들을 합한 벡터
- Timestep + text info (y)

  : Timestep info와 축소된 Text token을 더한 벡터
- Image token (x)

  : Patching을 통해 sequence가 된 noised latent

Text-to-Image Architecture- MM-DiT



(b) One *MM-DiT* block

Multi-Modal Diffusion Transformer (MM-DiT)

- **기존 모델 (Stable Diffusion 2.x)**

  : Denoising에 **U-Net (CNN 모델)** 사용

- **DiT (Diffusion Transformer)**

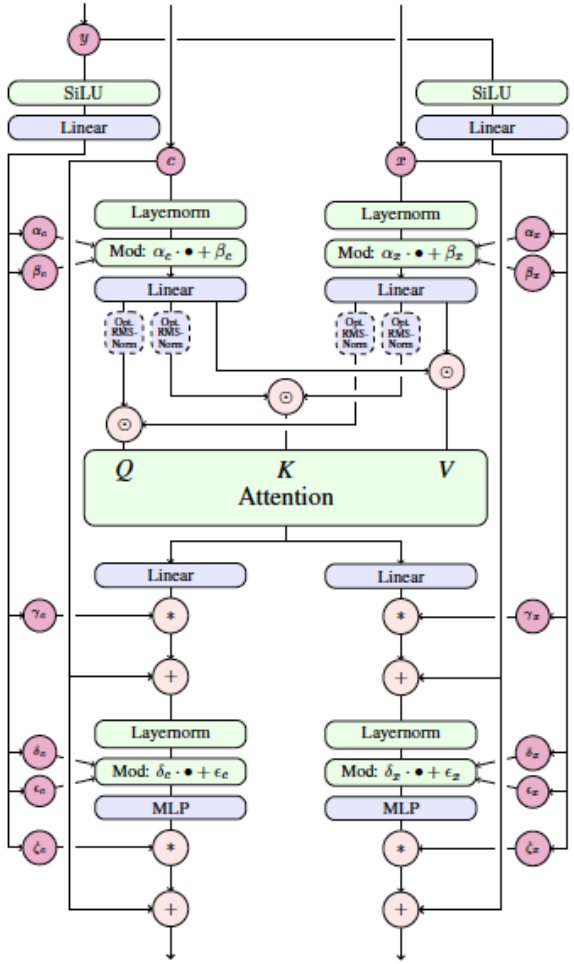
  : U-Net 대신 **Transformer** 사용

  : 하지만 image token들 사이에만 Self-Attention을 사용한 뒤 text token–image token 사이에 Cross-Attention 사용

  -> text token과 image token에 모두 Attention이 실행되는 것이 아닌, 특정 과정마다 text token이 '참고'됨

  : 복잡한 프롬프트(text token)을 처리하지 못하고, 글자가 들어가는 이미지를 정확하게 생성하지 못함 (오타 등)

Text-to-Image Architecture- MM-DiT



(b) One *MM-DiT* block

Multi-Modal Diffusion Transformer (MM-DiT)

- **MM-DiT**

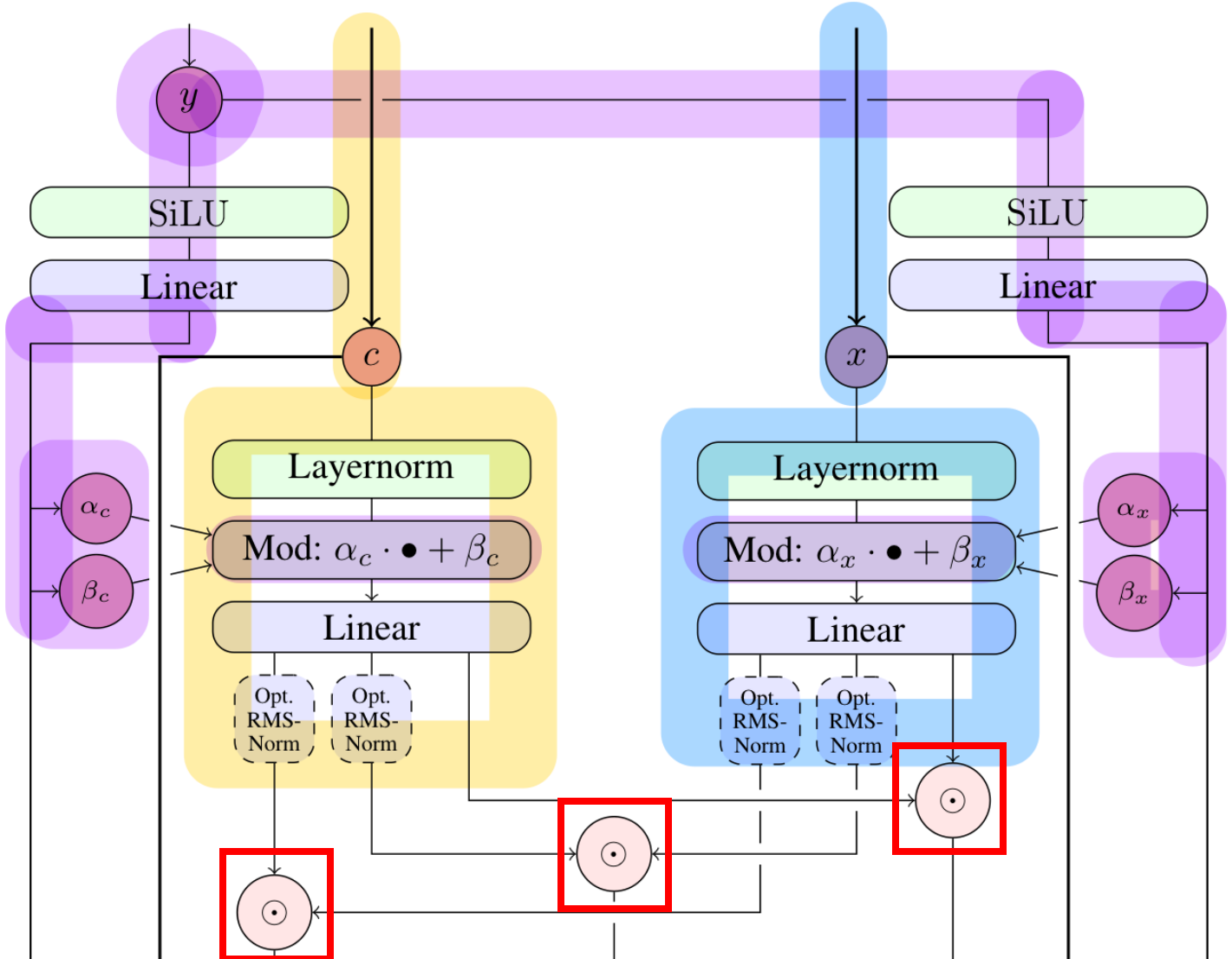
  : **Joint Self-Attention**을 사용해 image token과 text token을 **하나의 sequence로 이어붙여** transformer에 입력됨

  -> image token과 text token이 하나의 Self-Attention 과정을 거치며 서로 참고됨

  : image token과 text token이 각각 **동일하게 Attention 과정을 거침**

  -> 복잡한 프롬프트(text token)을 잘 처리하고, 글자가 들어가는 이미지를 정확하게 생성할 수 있음

Text-to-Image Architecture- MM-DiT

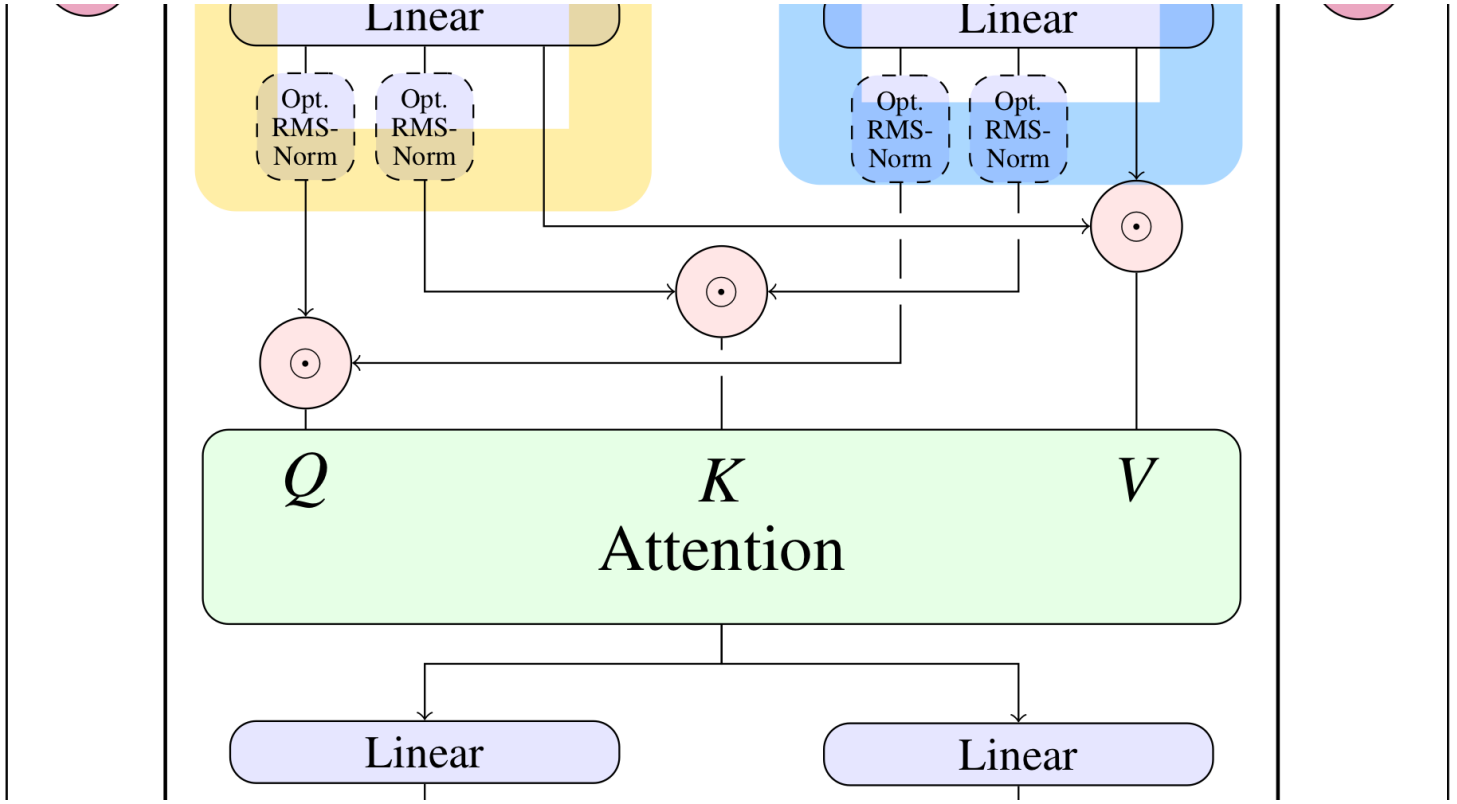


timestep + text token (y)
: 모델에게 **timestep과 text token**의 전반적인 내용을 계속해서 전달

Mod (Modulation)
: token의 분포를 y token의 정보에 기반해 이동시킴

Concatenation
: image token과 text token을 이어붙여 하나의 input sequence로 만듦

Text-to-Image Architecture- MM-DiT



Concatenation
: text token과 image token을 이어붙여 하나의 input sequence로 만듦
-> 이론적으로 text와 image 외에도 video, audio 등의 data도 추가 가능
=> Multi-Modal

각각의 concatenated sequence들은 각각 Query (Q), Key (K), Value (V)가 되어 Attention 모델에 입력됨

Output은 다시 text token과 image token으로 나뉘어 과정을 거쳐감

Improving Rectified Flows

What is improved?

- Training process for Rectified Flow models
- What is the most efficient and effective method for training (flow trajectory)?

What method is the best?

- RF loss with Logit-Normal Sampling (rf/lognorm(0.00, 1.00))

***FID Score** (낮을수록 좋음)
: Image가 얼마나 현실적인가?
– 얼마나 현실의 이미지와 비슷한지, 얼마나 image의 quality가 좋은지
***CLIP Score** (높을수록 좋음)
: Image가 얼마나 text와 일치하는가?
- 생성된 image가 얼마나 prompt를 잘 표현하는지, 잘 반영하는지

	rank averaged over		
variant	all	5 steps	50 steps
rf/lognorm(0.00, 1.00)	1.54	1.25	1.50
rf/lognorm(1.00, 0.60)	2.08	3.50	2.00
rf/lognorm(0.50, 0.60)	2.71	8.50	1.00
rf/mode(1.29)	2.75	3.25	3.00
rf/lognorm(0.50, 1.00)	2.83	1.50	2.50
eps/linear	2.88	4.25	2.75
rf/mode(1.75)	3.33	2.75	2.75
rf/cosmap	4.13	3.75	4.00
edm(0.00, 0.60)	5.63	13.25	3.25
rf	5.67	6.50	5.75
v/linear	6.83	5.75	7.75
edm(0.60, 1.20)	9.00	13.00	9.00
v/cos	9.17	12.25	8.75
edm/cos	11.04	14.25	11.25
edm/rf	13.04	15.25	13.25
edm(-1.20, 1.20)	15.58	20.25	15.00

Table 1. **Global ranking of variants.** For this ranking, we apply non-dominated sorting averaged over EMA and non-EMA weights, two datasets and different sampling settings.

	ImageNet		CC12M	
variant	CLIP	FID	CLIP	FID
rf	0.247	49.70	0.217	94.90
edm(-1.20, 1.20)	0.236	63.12	0.200	116.60
eps/linear	0.245	48.42	0.222	*90.34*
v/cos	0.244	50.74	0.209	97.87
v/linear	0.246	51.68	0.217	100.76
rf/lognorm(0.50, 0.60)	**0.256**	80.41	0.233	120.84
rf/mode(1.75)	*0.253*	**44.39**	0.218	94.06
rf/lognorm(1.00, 0.60)	0.254	114.26	**0.234**	147.69
rf/lognorm(-0.50, 1.00)	0.248	45.64	0.219	**89.70**
rf/lognorm(0.00, 1.00)	0.250	*45.78*	*0.224*	89.91

Table 2. **Metrics for different variants.** FID and CLIP scores of different variants with 25 sampling steps. We highlight the **best**, second best, and *third best* entries.

Improving Autoencoders

What is improved?

- VAE improved using different number of latent channels

What method is the best?

- Increasing the number of latent channels to 16 channels

Metric	4 chn	8 chn	16 chn
FID (↓)	2.41	1.56	**1.06**
Perceptual Similarity (↓)	0.85	0.68	**0.45**
SSIM (↑)	0.75	0.79	**0.86**
PSNR (↑)	25.12	26.40	**28.62**

Table 3. **Improved Autoencoders.** Reconstruction performance metrics for different channel configurations. The downsampling factor for all models is $f = 8$.

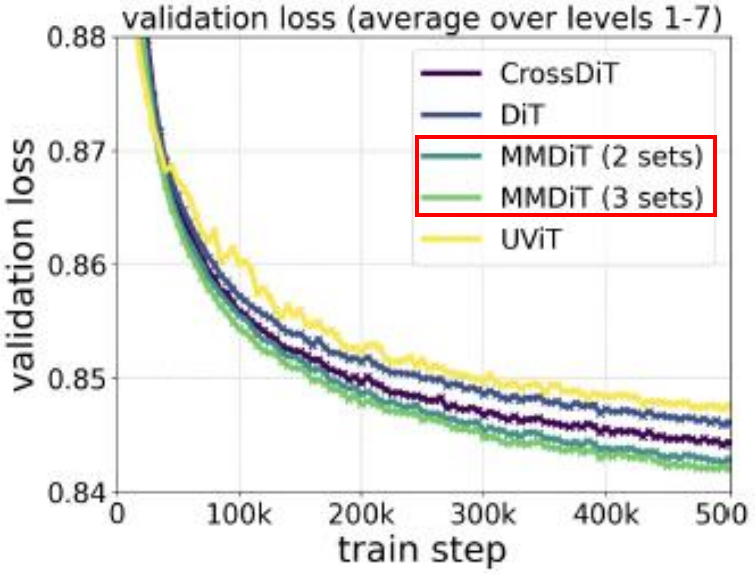
Improving the Text-to-Image Backbones

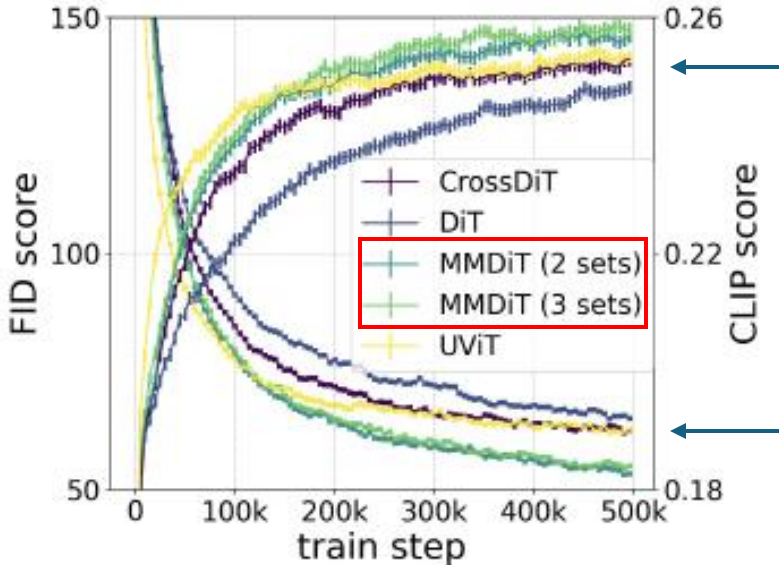
What is improved?

- Neural network that does the denoising
- Which is better: **existing models vs. this paper's model (MM-DiT)?**

What method is the best?

- MM-DiT (novel model introduced by this paper)





CLIP Score

FID Score

Improving Scaling for High Resolutions

What is improved?

- Added more noise for higher resolution images

: we need more noise to destroy high resolution images because they have more pixels.

What method is the best?

- Resolution-dependent timestep shift with shift value = 3.0

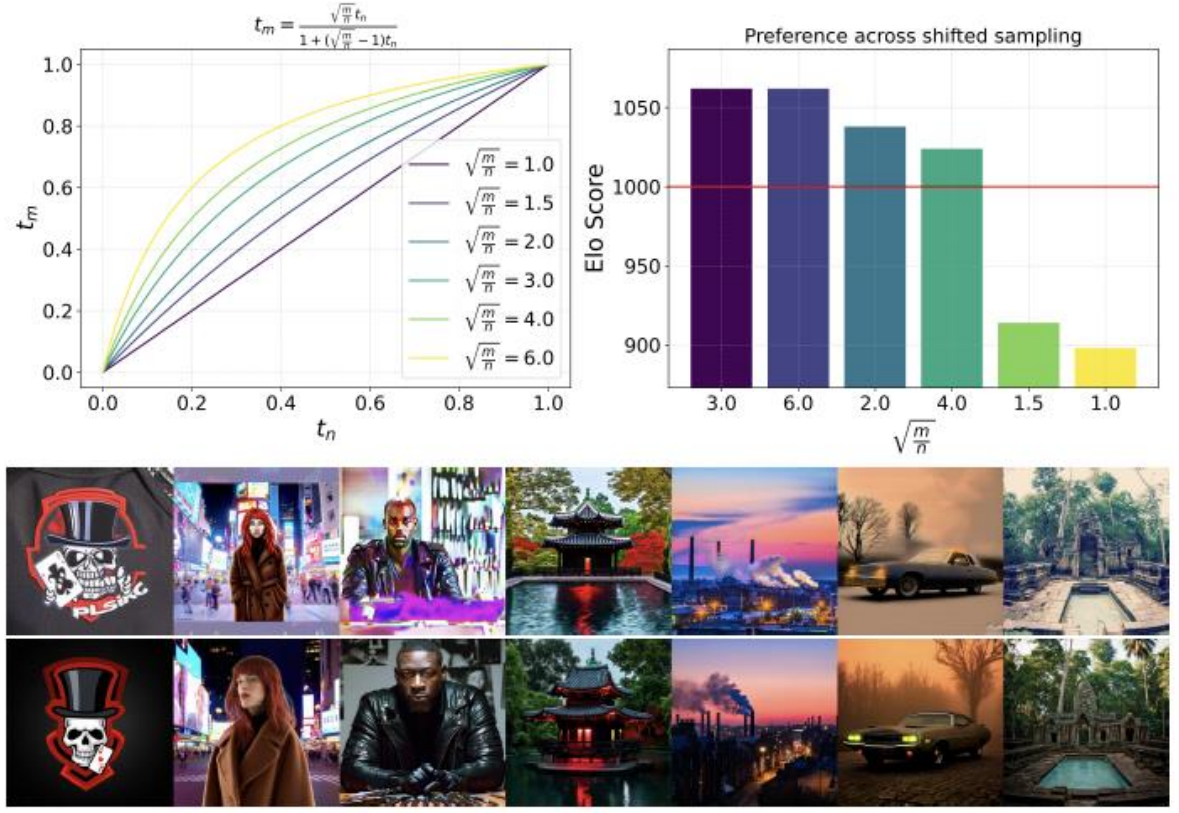


Figure 6. **Timestep shifting at higher resolutions.** *Top right:* Human quality preference rating when applying the shifting based on Equation (23). *Bottom row:* A 512^2 model trained and sampled with $\sqrt{m/n} = 1.0$ (*top*) and $\sqrt{m/n} = 3.0$ (*bottom*). See Section 5.3.2.

Results: Comparison with Existing Models

Model	Overall	Objects					Color Attribution
		Single	Two	Counting	Colors	Position	
minDALL-E	0.23	0.73	0.11	0.12	0.37	0.02	0.01
SD v1.5	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-alpha	0.48	0.98	0.50	0.44	0.80	0.08	0.07
SD v2.1	0.50	0.98	0.51	0.44	0.85	0.07	0.17
DALL-E 2	0.52	0.94	0.66	0.49	0.77	0.10	0.19
SDXL	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SDXL Turbo	0.55	**1.00**	0.72	0.49	0.80	0.10	0.18
IF-XL	0.61	0.97	0.74	*0.66*	0.81	0.13	0.35
DALL-E 3	0.67	0.96	0.87	0.47	*0.83*	**0.43**	*0.45*
Ours (depth=18), 512^2	0.58	0.97	0.72	0.52	0.78	0.16	0.34
Ours (depth=24), 512^2	0.62	0.98	0.74	0.63	0.67	*0.34*	0.36
Ours (depth=30), 512^2	0.64	0.96	0.80	0.65	0.73	0.33	0.37
Ours (depth=38), 512^2	*0.68*	0.98	0.84	*0.66*	0.74	0.40	0.43
Ours (depth=38), 512^2 w/DPO	0.71	0.98	0.89	**0.73**	*0.83*	*0.34*	0.47
Ours (depth=38), 1024^2 w/DPO	**0.74**	0.99	**0.94**	0.72	**0.89**	0.33	**0.60**

Table 5. **GenEval comparisons**. Our largest model (depth=38) outperforms all current open models and DALLE-3 (Betker et al., 2023) on GenEval (Ghosh et al., 2023). We highlight the **best**, second best, and *third best* entries. For DPO, see Appendix C.

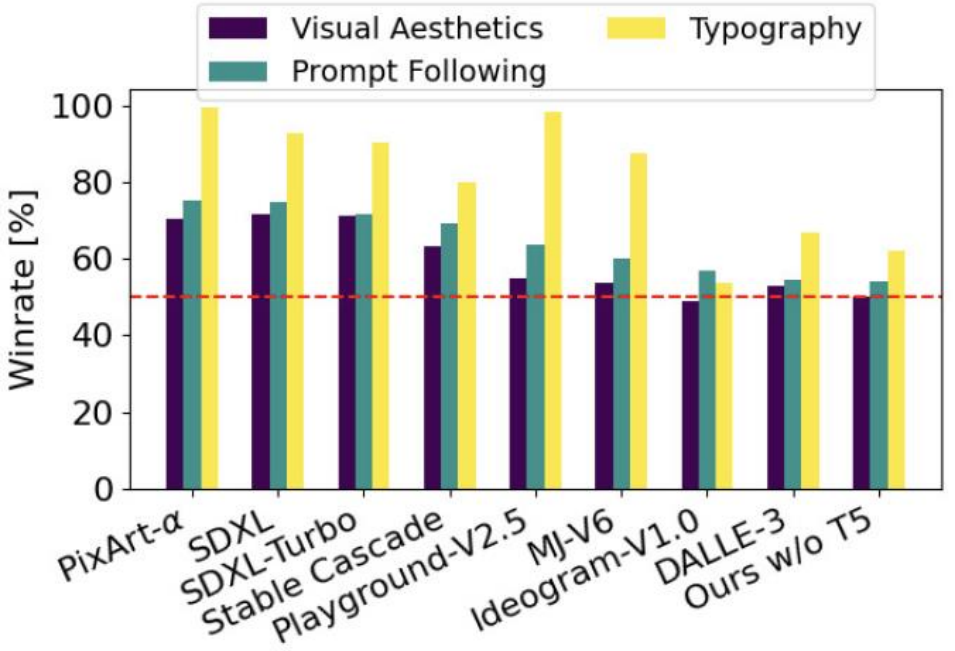


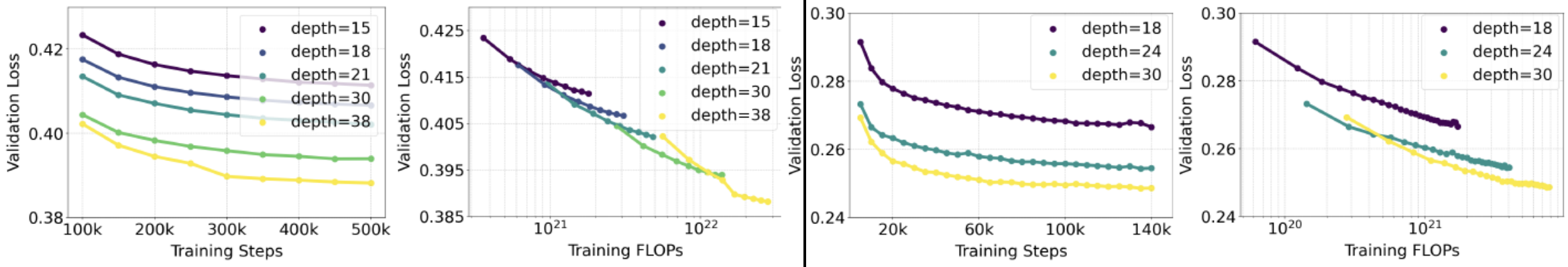
Figure 7. **Human Preference Evaluation against currrent closed and open SOTA generative image models.** Our 8B model compares favorable against current state-of-the-art text-to-image models when evaluated on the parti-prompts (Yu et al., 2022) across the categories *visual quality*, *prompt following* and *typography generation*.

Results: Sampling Efficiency

	relative CLIP score decrease [%]			
	5/50 steps	10/50 steps	20/50 steps	path length
depth=15	4.30	0.86	0.21	191.13
depth=30	3.59	0.70	0.24	187.96
depth=38	2.71	0.14	0.08	185.96

Table 6. **Impact of model size on sampling efficiency.** The table shows the relative performance decrease relative to CLIP scores evaluated using 50 sampling steps at a fixed seed. Larger models can be sampled using fewer steps, which we attribute to increased robustness and better fitting the straight-path objective of rectified flow models, resulting in shorter path lengths. Path length is calculated by summing up $\|v_\theta \cdot dt\|$ over 50 steps.

Results: Impact of Model Size on Performance



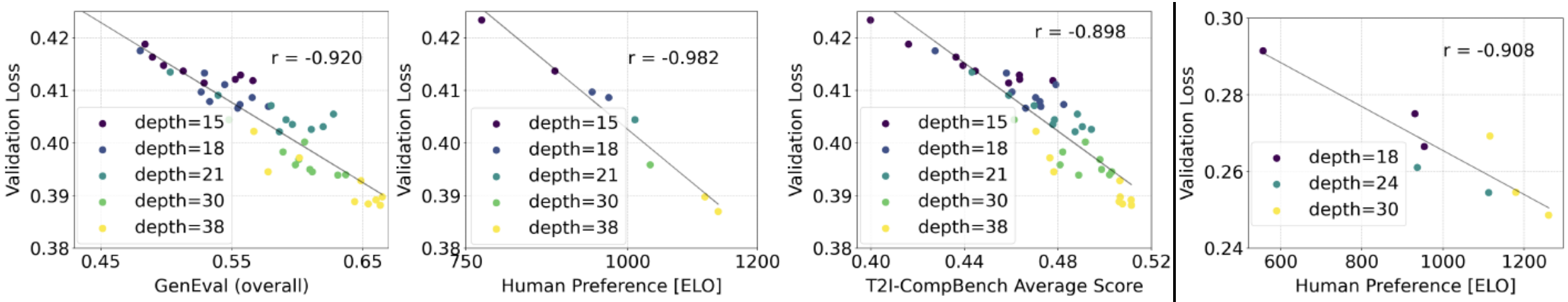
Image

Video

Validation loss가 Training Steps와 Model Size가 증가할수록 감소

*Training FLOPs: total number of operations / total computational cost of a model

Results: Use of Validation Loss

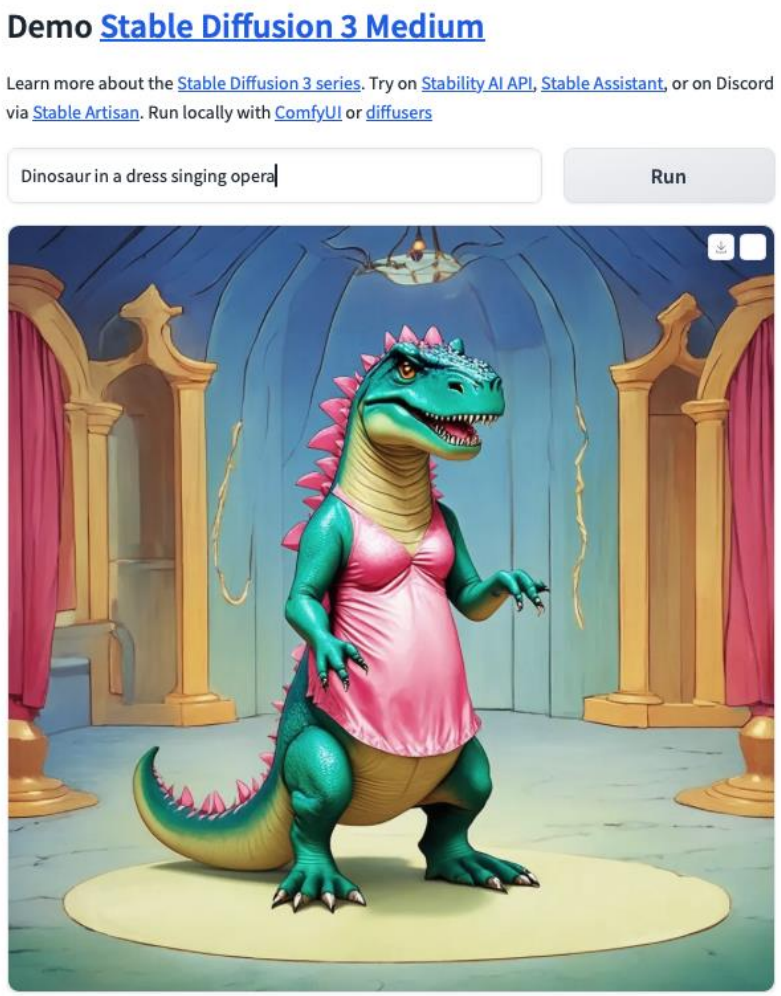


Image

Video

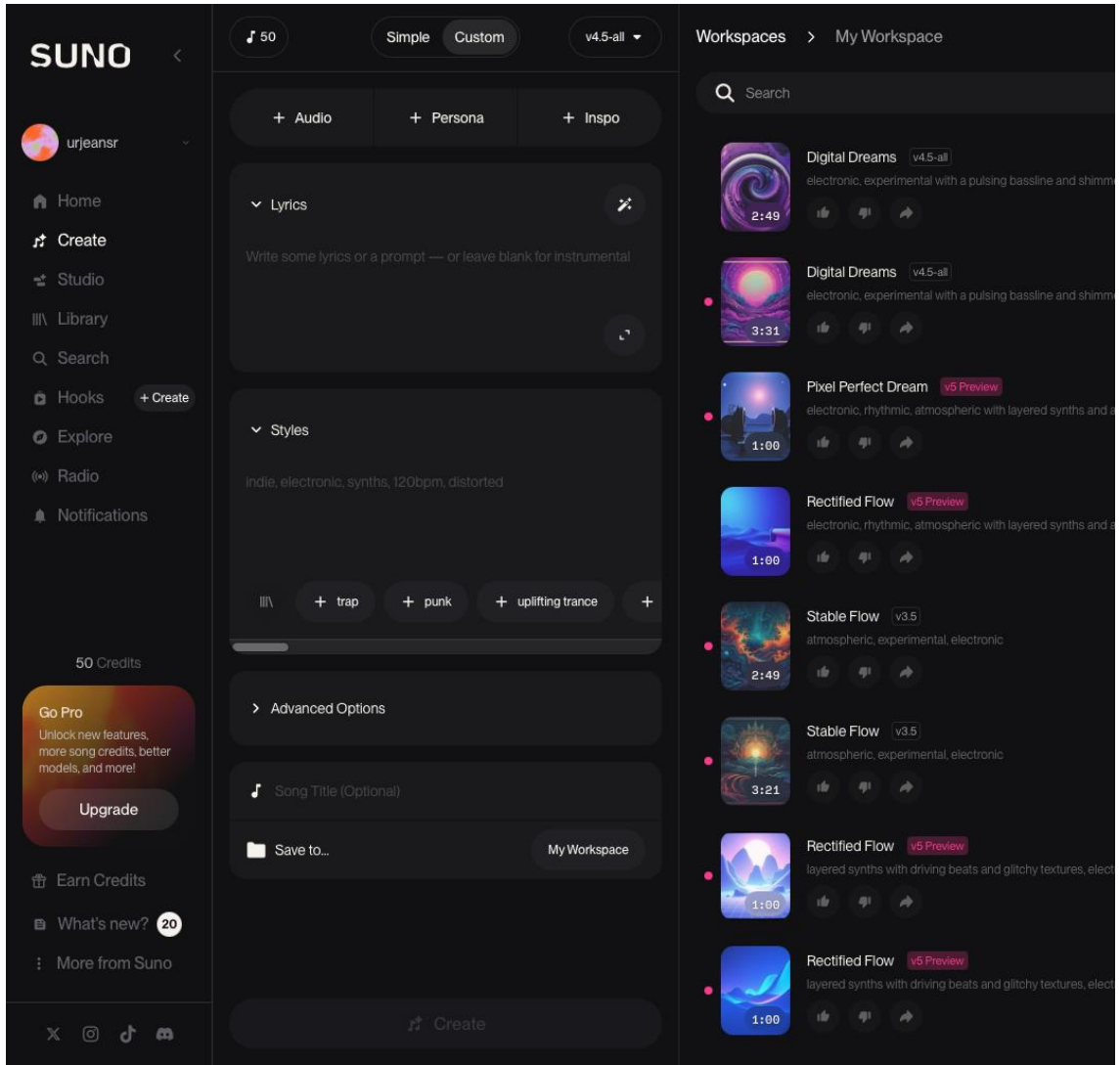
Validation loss는 Model Performance를 효과적으로 나타낼 수 있는 중요한 지표임

Demo: Stable Diffusion 3



https://huggingface.co/spaces/stabilityai/stable-diffusion-3-medium

Demo: SUNO



https://suno.com/create

Conclusion

This paper

1. **Proposed a novel timestep sampling method (RF loss Logit-Normal Sampling) for RF training**
2. **Proposed MM-DiT architecture**
3. **Performed a scaling study with various variants to fine-tune the model**
4. **Showed that validation loss reflects both model performance and human preference evaluation**