TEXAS A&M UNIVERSITY
Engineering

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis

Mu-Ruei Tseng, Harshavardhana Srinivasan

Background

TEXAS A&M UNIVERSITY Engineering

Diffusion Model

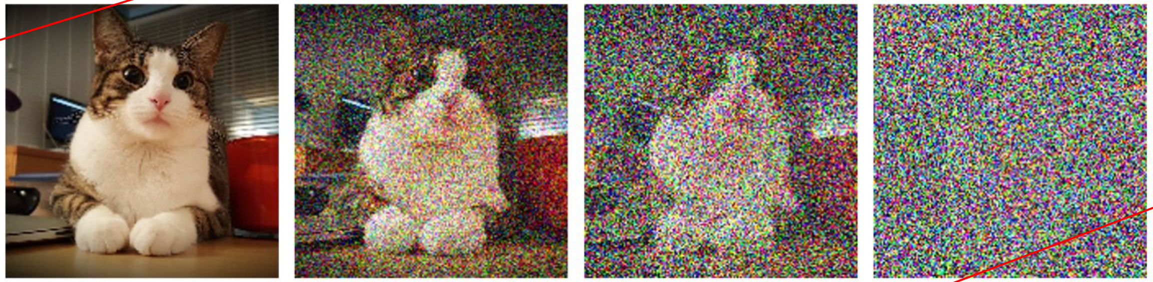
- A generative model that create data from noise

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \boxed{\epsilon \sim \mathcal{N}(0, I)}$$

what the model $\epsilon_\theta(x_t, t)$ tries to predict

x_0 forward process x_t

Cumulative noise schedule at timestep t



The model's estimate of the original image

reverse process $\hat{x}_0 = \dfrac{x_t - \sqrt{1-\bar{\alpha}_t} \times \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}$

Background

Ex: a linear noise schedule $\alpha_t = 1 - \frac{t}{T}$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

$\bar{\alpha}_t$ is nonlinear with t

x_0 $\qquad x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$ $\qquad x_t$

$t = 0$	$t = 250$	...	$t = 750$	$t = 1000$
$\alpha_0 = 1$	$\alpha_{250} = 0.75$		$\alpha_{750} = 0.25$	$\alpha_{1000} = 0$

Background

TEXAS A&M UNIVERSITY Engineering

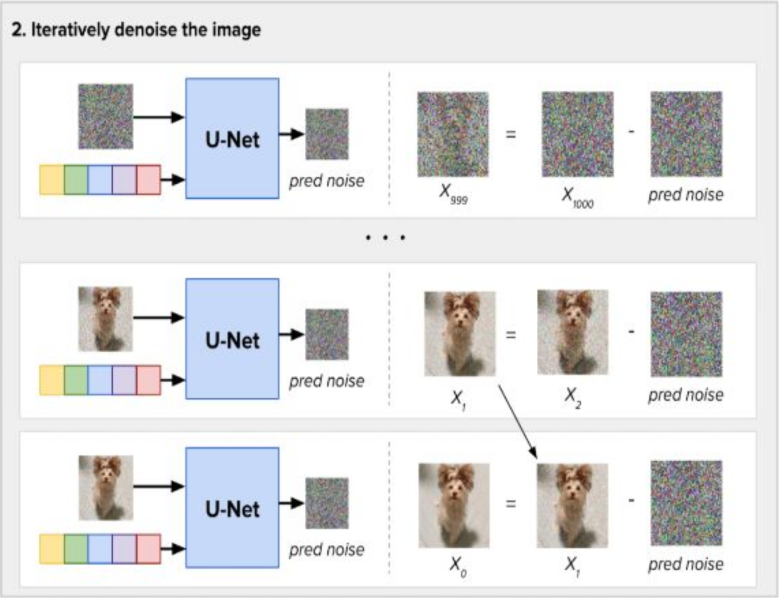
Reverse Process (inferencing)

Start with pure noise x_T

For each timestep $t = T, T-1, \dots 1$

Calculate $\hat{x}_0 = \dfrac{x_t - \sqrt{1-\bar{\alpha}_t} \times \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}$

compute $x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1-\alpha_{t-1}} \times \epsilon$



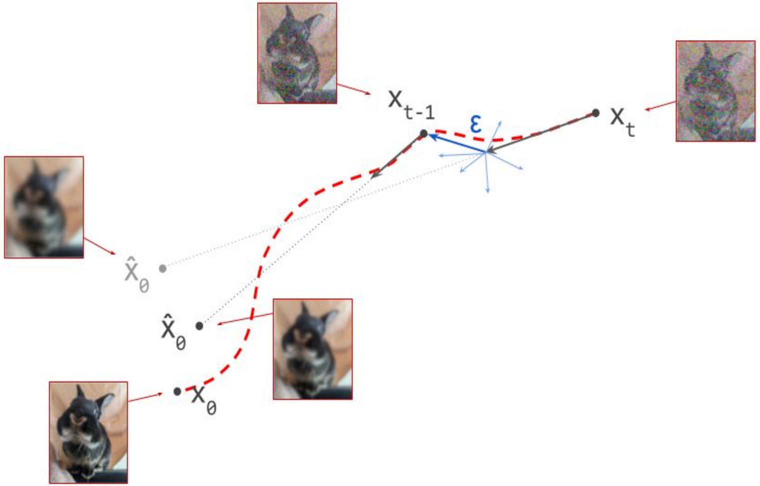
Denoising process for a single image

Background

The reverse process aims to estimate and follow a complex nonlinear curved path

At each stage

- Predict the tangent direction
- Take a small step along the direction
- Can lead to suboptimal result due to the **accumulated error**

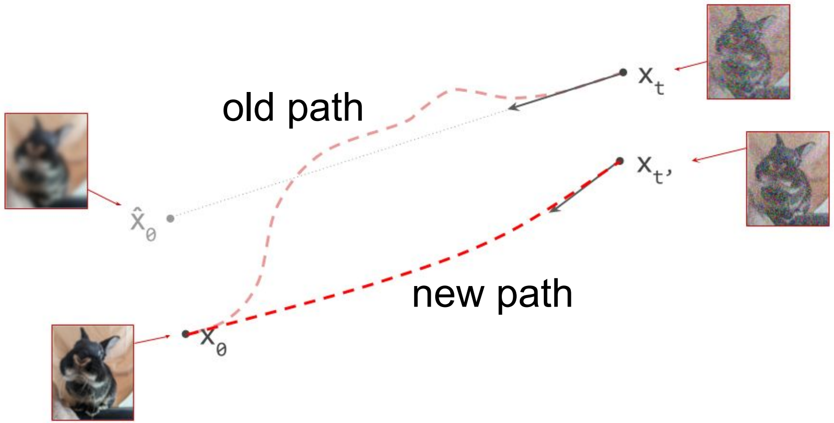


Contribution

TEXAS A&M UNIVERSITY Engineering

Introduce Rectified Flow for Noise Addition

- Reformulate the forward process as straight-line paths between the data and noise distributions



Contribution

Design Tailored SNR Samplers

- Emphasize the importance of focusing on intermediate timesteps in the diffusion process

Novel Transformer-Based Architecture

- A bidirectional transformer architecture with separate streams for image and text tokens
- Outperform SOTA models like DALL-E 3 and SDXL

Comprehensive Experiments on Noise Methods

- Conduct large-scale experiments comparing different noise-adding methods and sampling techniques

Methodology - Flow Based Method

TEXAS A&M UNIVERSITY Engineering

- Proposed Flow Based Generative model, which is used to map samples from a noise distribution p_1 to data distribution p_0 through an ODE equation

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

- The vector field u, maps the data to noise through intermediate steps as defined in the forward process as follows:

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

- A typical solution to the above equation involves solving the marginal vector field $u_t(z|\epsilon)$ that averages across all noise samples, which becomes costly in computation.

$$\psi_t(\cdot|\epsilon) : x_0 \mapsto a_t x_0 + b_t \epsilon$$

$$u_t(z|\epsilon) := \psi_t'(\psi_t^{-1}(z|\epsilon)|\epsilon)$$

Methodology - Objective Function

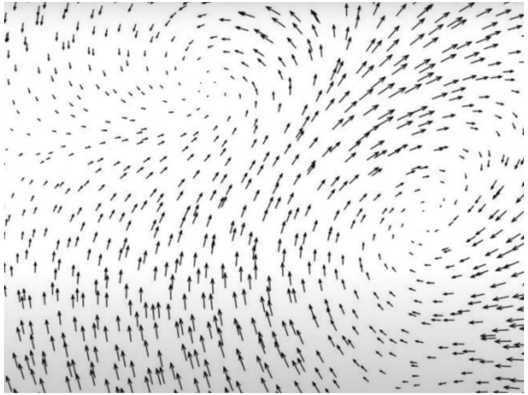
TEXAS A&M UNIVERSITY Engineering

Objective Function

- The objective function trains the model to predict the velocity field $v(z, t)$, which describes how samples move along the trajectory from data (p_0) to noise (p_1) over time t

- The model minimizes the difference between the predicted velocity $v(z, t)$ (output of the neural network) and the target velocity $u(z, t)$. For example: $||v_\theta(z, t) - u_t(z)||_2^2$. The original loss function computes this difference, averaged over all timesteps and noise samples.

velocity field



Methodology - Objective Function

TEXAS A&M UNIVERSITY Engineering

Instead of learning $||v_\theta(z,t) - u_t(z)||_2^2$, reform the loss to minimize the noise added at each timestamp:

- Simpler and more direct to predict
- Aligns more naturally with the reverse process (as the forward process explicitly adds noise at each step)

The final objective incorporates weights derived from the signal-to-noise ratio (SNR) and timestep density

- Ensure that the model prioritizes important timesteps for improved performance.

Weighted Loss Function:

$$\mathcal{L}_w(x_0) = -\frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}(t),\epsilon\sim\mathcal{N}(0,I)}\left[w_t\lambda_t'\|\epsilon_\Theta(z_t,t) - \epsilon\|^2\right]$$

where $w_t = -\frac{1}{2}\lambda_t' b_t^2$ corresponds to $\mathcal{L}_{CFM}$.

Methodology - Flow Trajectories

TEXAS A&M UNIVERSITY Engineering

Rectified Flow: Defines a forward process as straight paths between the data distribution and a standard normal distribution as $z_t = (1 - t)x_0 + t\epsilon$

EDM: Uses the forward process of the form $z_t = x_0 + b_t\epsilon$ where $b_t = \exp F_{\mathcal{N}}^{-1}(t|P_m, P_s^2)$

Cosine: Has the forward process of the form $z_t = \cos(\frac{\pi}{2}t)x_0 + \sin(\frac{\pi}{2}t)\epsilon$

LDM (Linear): Uses a modification of the DDPM schedule. Both are variance preserving schedules given by $b_t = \sqrt{1 - a_t^2}$

The diffusion coefficients of at and bt: $a_t = (\prod_{s=0}^{t}(1 - \beta_s))^{\frac{1}{2}}$ and $\left(\sqrt{\beta_0} + \frac{t}{T-1}(\sqrt{\beta_{T-1}} - \sqrt{\beta_0})\right)^2$

Methodology - Tailored SNR Samplers

TEXAS A&M UNIVERSITY Engineering

Intuitively, the error in the intermediate timestamp is harder to learn

- Involves a more complex balance between signal and noise
- The model should focus more in the intermediate timesteps

Tailored SNR Sampler (logit-normal sampling, mode sampling)

- Sample intermediate timesteps with higher frequency (opposed to uniform distribution)
- Outperform uniform sampling and diffusion baselines like EDM and LDM-Linear.

Methodology - Tailored SNR Samplers

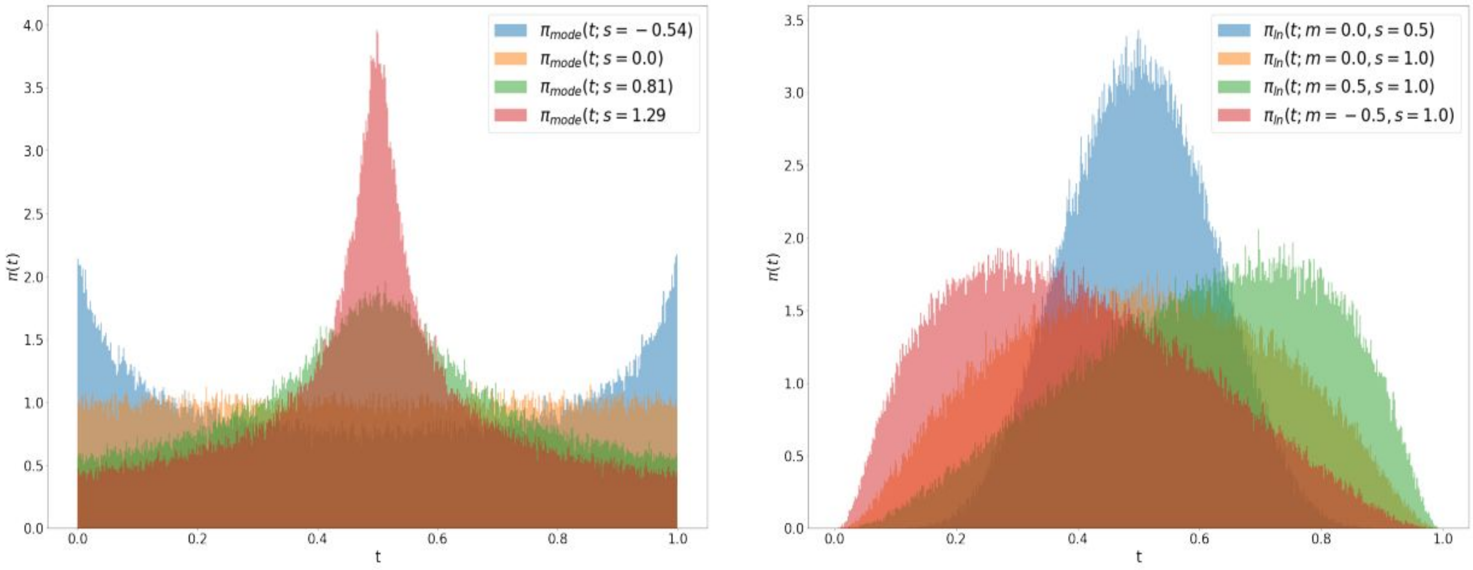
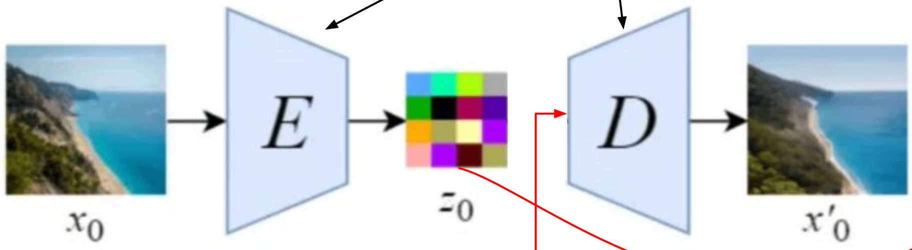


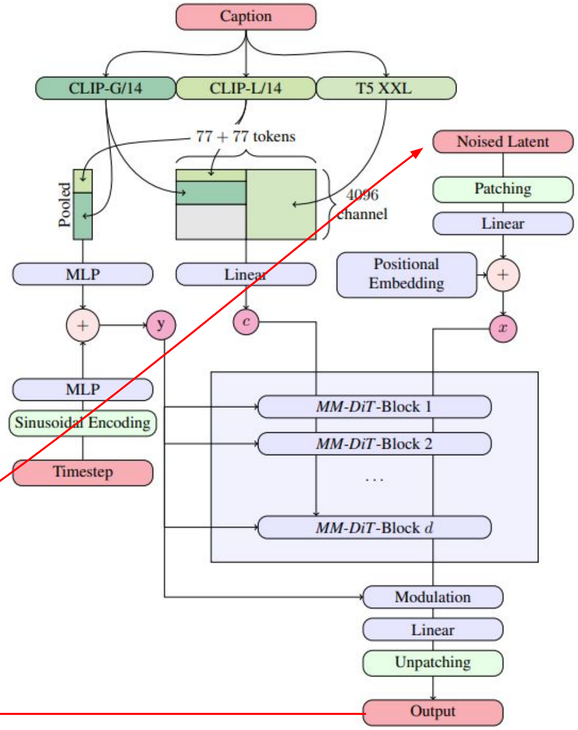
Figure 11. The mode (left) and logit-normal (right) distributions that we explore for biasing the sampling of training timesteps.

Similar to the Latent Diffusion Model

- Training the diffusion model in the latent space

Pretrained



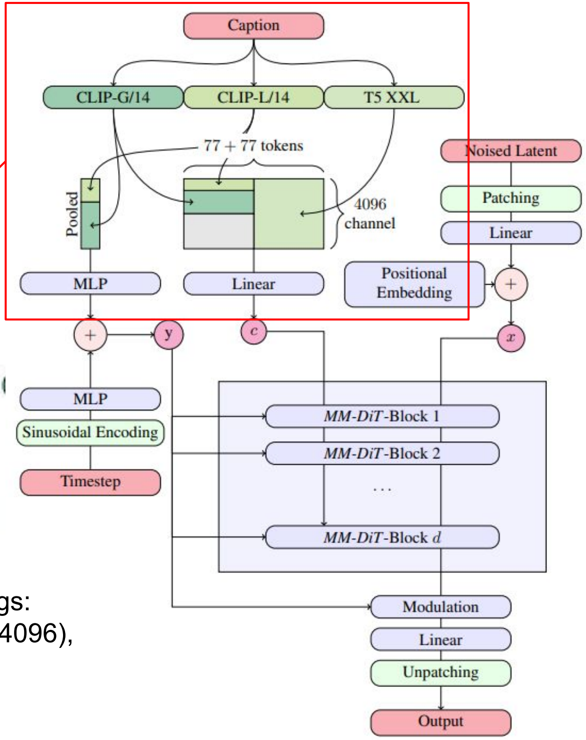


(a) Overview of all components.

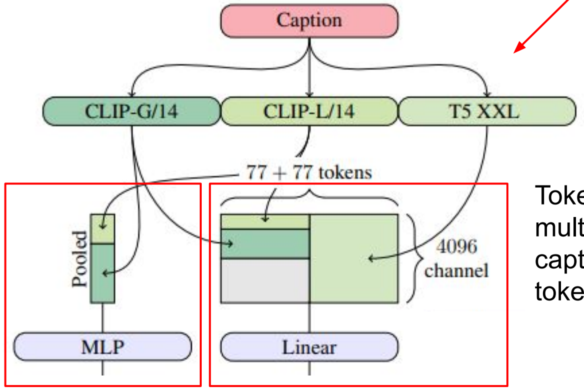
Methodology - Text-to-Image Model

Text embedding

- Uses pretrained text encoders to obtain high-quality text embeddings
  - Uses multiple text encoders and create a unified text representation (like ensemble)
  - A 46.3% dropout rate is applied to the text embeddings during training (more flexible)



(a) Overview of all components.



Pooling:
creates a single, compact embedding that summarizes the entire text
ex: mean pooling (avg the tokens)

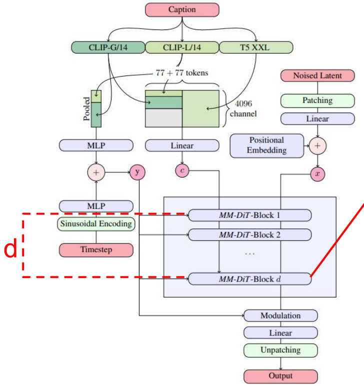
Token-Level Embeddings:
multiple vectors (154 x 4096), captures fine-grained token-specific details

Methodology - Text-to-Image Model

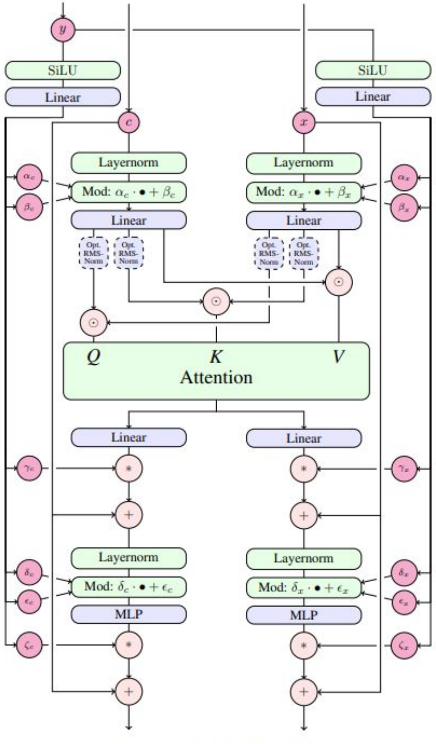
TEXAS A&M UNIVERSITY
Engineering

MMDiT Block

- Unlike DiT, MMDiT uses separate weights for two modalities
  - Like having two transformers but share the attention operation
  - Intuition: Both representation can work in their own space while taking the other into account (during the attention part)
- Model Scaling:
  - Ensures a balanced architecture as the model grows deeper.
    - Depth: d
    - Hidden size: 64 x d
    - MLP size: 4 x 64 x d
    - Attention heads: d



(a) Overview of all components.



(b) One *MM-DiT* block

Evaluation Metric

CLIP score

- Measures how well a generated image aligns with a given text prompt
- Calculate the similarity between text and image embeddings

$$\text{CLIP score} = cos(z_{\text{text}}, z_{\text{image}}) = \frac{z_{\text{text}} \cdot z_{\text{image}}}{||z_{\text{text}}|| \times ||z_{\text{image}}||}$$

FID (Fréchet Inception Distance)

- Evaluates the quality and diversity of generated images by comparing their distribution to a reference distribution
- In this paper, they use CLIP features instead of features generated from Inception-v3

Evaluation Metric

Public Benchmarks (T2I-CompBench, GenEval)

Human Ratings

- Prompt following:
  - *Which image looks more representative to the provided text?*
- Visual Aesthetic:
  - *Given the prompt, which image is of higher-quality and aesthetically more pleasing?*
- Typography:
  - *Which image more accurately shows the text specified in the description?*

Results

TEXAS A&M UNIVERSITY Engineering

The paper shares extensive results covering two areas primarily

- Different samplers and trajectories
  - Employs global ranking and selective ranking from the best samplers

- Model architecture scaling
  - Comparison across different architecture and the performances

Results - Global Ranking of Variants

TEXAS A&M UNIVERSITY Engineering

	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

- Ranks 61 variants based on their overall performance across datasets (ImageNet, CC12M), sampling steps (5 and 50), and sampler settings.
- Uses Pareto- based global ranking derived from averaged ranks over CLIP and FID scores across various configurations.
- It identifies the top-performing variants globally, considering a wide range of scenarios (e.g., **rf/lognorm(0, 1)**, **rf/lognorm(1, 0.6)**, and others).

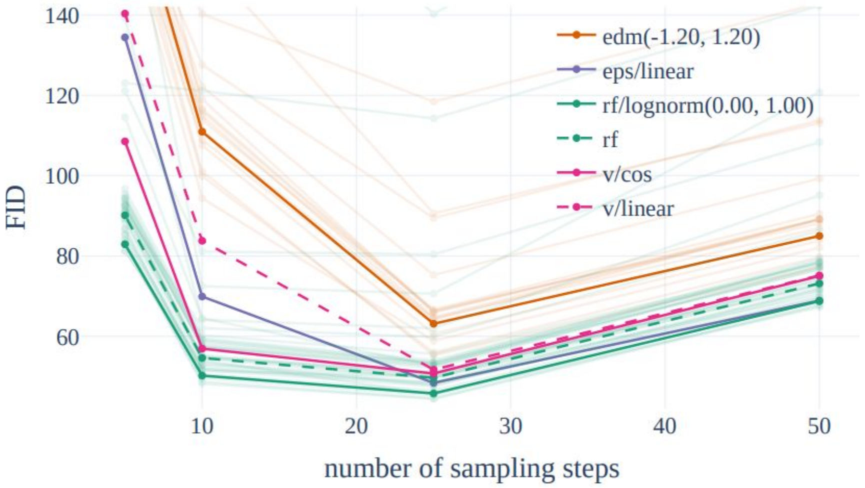
Results - Effective Samplers

TEXAS A&M UNIVERSITY Engineering

variant	ImageNet CLIP	ImageNet FID	CC12M CLIP	CC12M 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

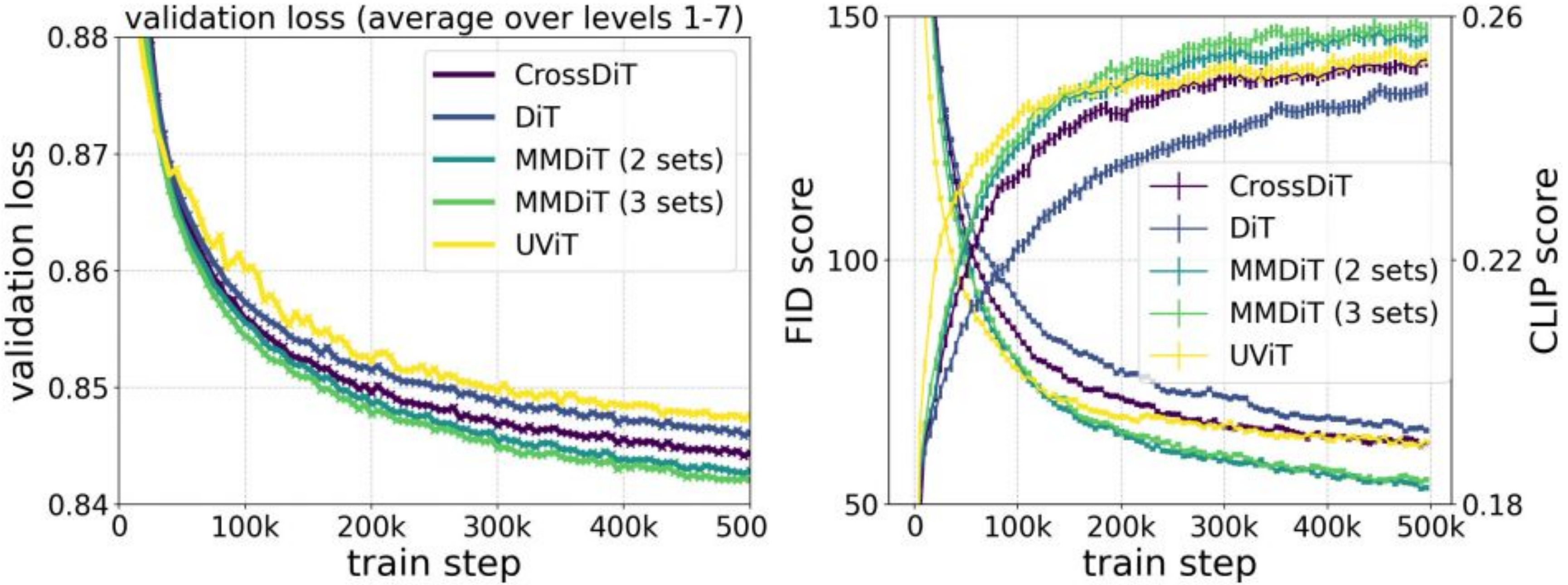
- Shows how individual variants, like **rf/lognorm(0.50, 0.60)** or **rf/lognorm(0, 1)**, perform on ImageNet and CC12M in terms of direct evaluation metrics.

Results - Samplers vs Training Steps

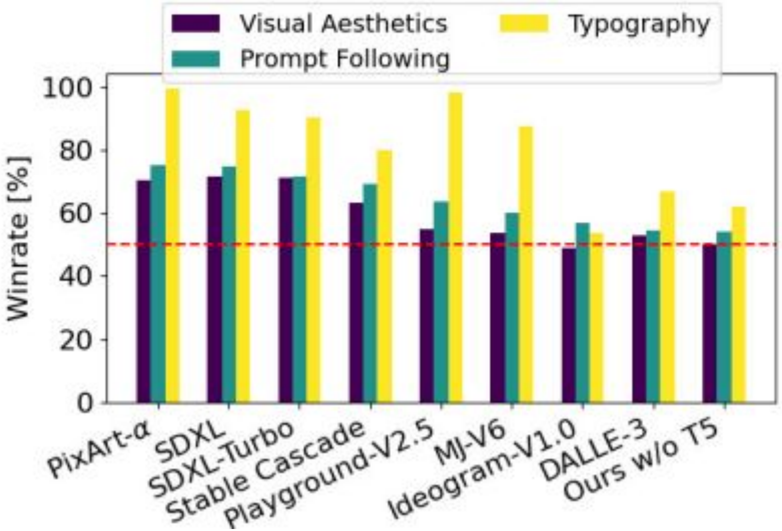


- **Rectified Flows (RF)** are highly sample-efficient, performing better than other formulations when the number of sampling steps is fewer than 25.
- For 25 or more steps, **rf/lognorm(0.00, 1.00)** continues to perform competitively, matching or surpassing **eps/linear**, showcasing its robustness across varying step counts.

Results - Samplers vs Training Steps



Results - Human Preference vs Closed And Open Model Evaluation



The 8B model performs favorably against current SOTA text-to-image models in human evaluations on the Parti-prompts, excelling in **visual quality**, **prompt adherence**, and **typography generation**.

Results - Scalability and Performance of MM-DiT Across Image and Video Generation

TEXAS A&M UNIVERSITY Engineering

Model	Overall	Objects Single	Objects Two	Counting	Colors	Position	Color Attribution
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**

- MM-DiT models improve validation loss with larger sizes and more training steps, aligning with evaluation metrics and human preferences.
- Larger models outperform SOTA models in prompt comprehension and overall quality

Discussion

Key Contributions

- Rectified Flow
  - Resolves curved forward paths, improving reverse process efficiency and accuracy
- Tailored SNR Sampler (during training)
  - Focuses on intermediate timesteps as their errors are more difficult to model
- MM-DiT architecture
  - Multimodal transformer with separate weights for text and image modalities improves integration

Strengths

- Outperforms SOTA models (e.g., SDXL, DALL-E 3) on CLIP, FID, and human preferences
- Comprehensive evaluation

Discussion

TEXAS A&M UNIVERSITY Engineering

Limitation

- Performance Degradation with limited steps
  - Effectiveness of the model decreases significantly when the number of training steps is reduced
- Multi-Modal beyond text-image
  - The study focuses on high-resolution text-to-image tasks, with limited exploration of audio-text and video-text applications

Future Directions

- Exploring other multi-modal capabilities beyond text-to-image, such as text-to-video or audio-to-text
- Investigating efficient scaling methods to significantly reduce computational overhead while maintaining performance

References

- The paradox of diffusion distillation
- An Introduction to Diffusion Models and Stable Diffusion
- Stable Diffusion 3