Scaling Rectified Flow Transformers for High-Resolution Image Synthesis



Figure 1: generated samples

Patrick Esser et al.

ICML 2024

Presented by: Qi Shi

Motivation: Efficient Scaling

Scaling is all you need

Source: Esser et al., ICML 2024 oral presentation

Scaling is not free



Source: Esser et al., ICML 2024 oral presentation

Key Contributions

EFFICIENCY

Large-scale comparison of training formulations
For the best-performing setting with few steps

SCALABILITY

01

Reweighted rectified flow with tailored timestep sampling

02

MM-DiT with joint attention for text and image tokens

03

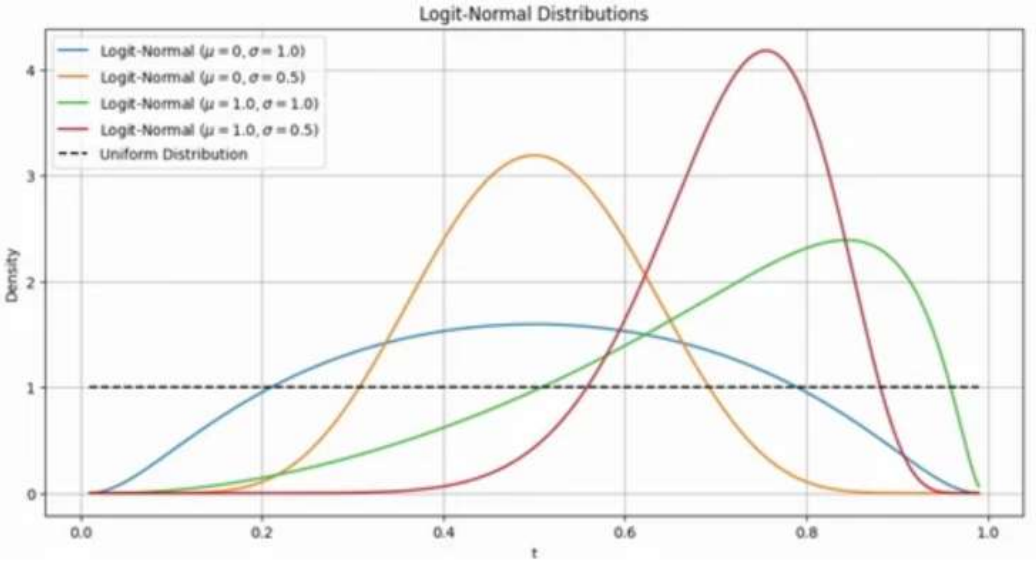
Timestep shifting for synthesis at higher image resolutions

04

Predictable gains from model size and training compute

Method 1: Reweighting rectified flow

$$L_{\mathrm{CFM}}(z_t) = \mathbb{E}_{t\sim p(t),\varepsilon\sim\mathcal{N}(0,1)} \left\| v_\Theta(z_t, t) - (\varepsilon - x_0) \right\|_2^2$$



Logit-normal sampling

$u \sim N(\mu, \sigma^2)$

$t = \mathrm{sigmoid}(u)$

Intuition: when t = 0 or 1, optimal prediction is obvious.

Slide Source: Esser et al., ICML 2024 oral presentation

Experiments: Comparing training formulations

61 formulations, two training datasets, 24 evaluation conditions

	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: average rank

	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 at 25 steps

Experiments: Testing performance at few steps

Euler sampling

5 to 50 steps

COCO validation

FID on CLIP features

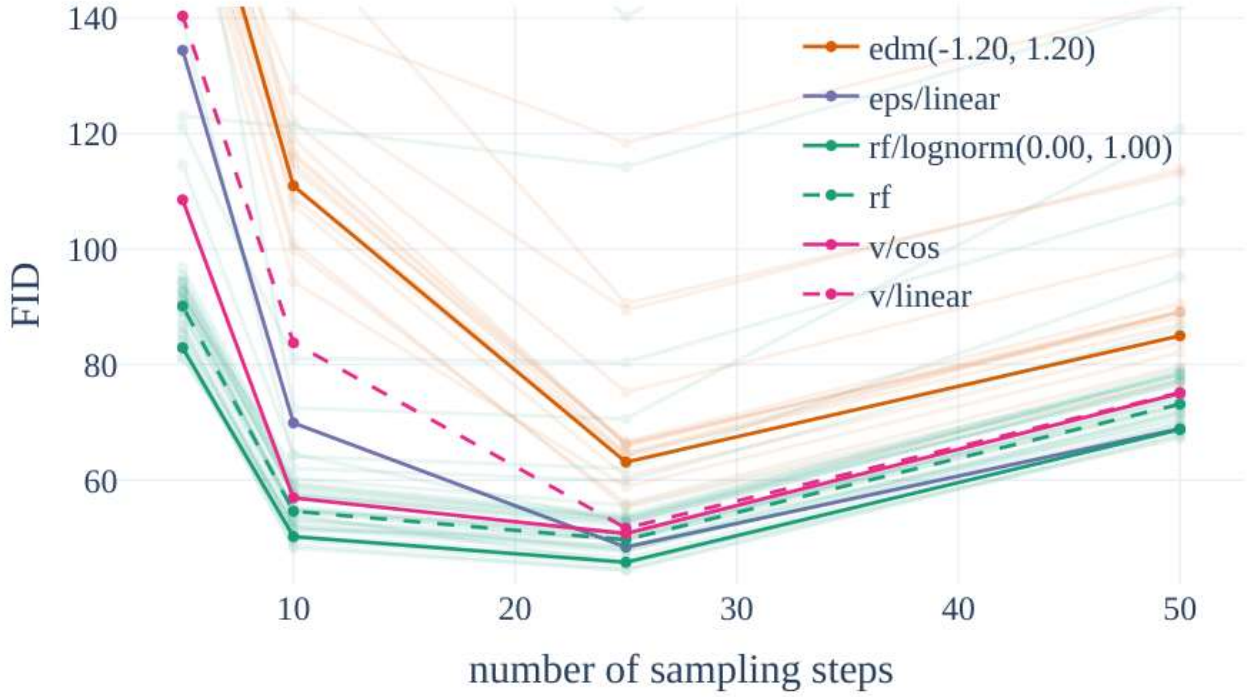
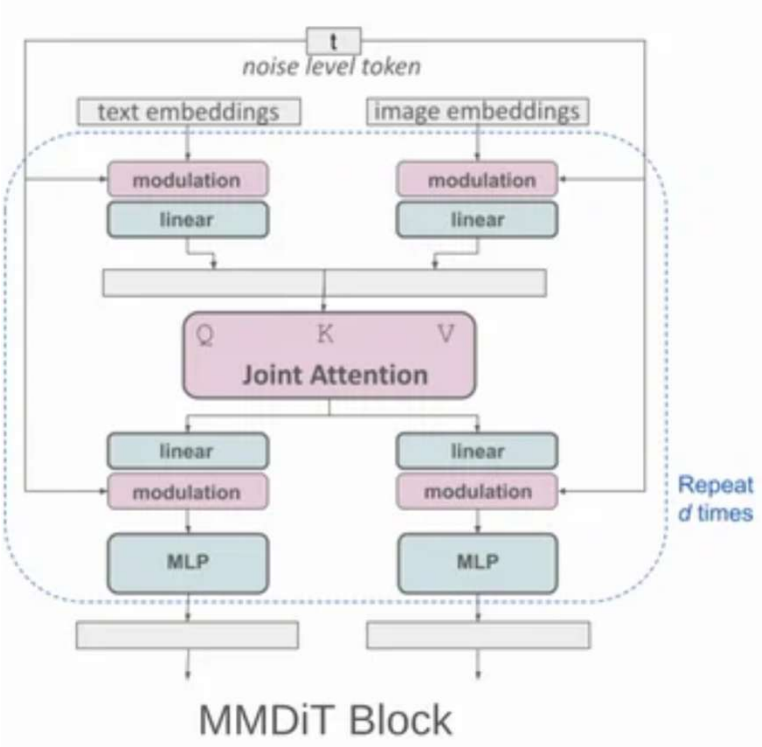
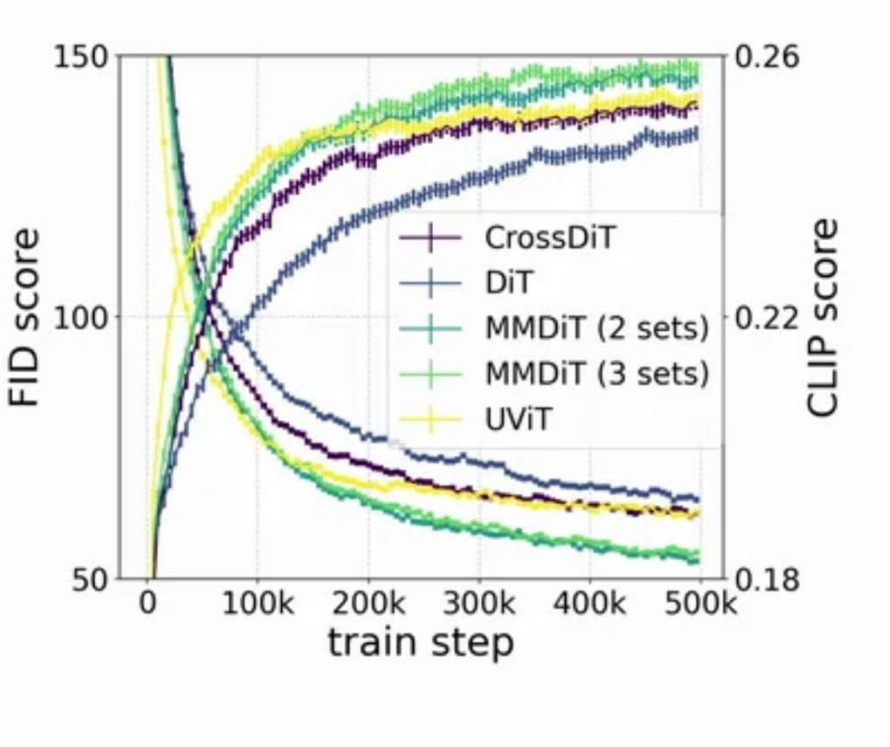


Figure 3: performance at fewer steps

Method 2: MM-DiT: enable complex text-to-image generation





Adapting existing multimodal learning methods

Slide Source: Esser et al., ICML 2024 oral presentation

Experiments: Evaluating prompt comprehension

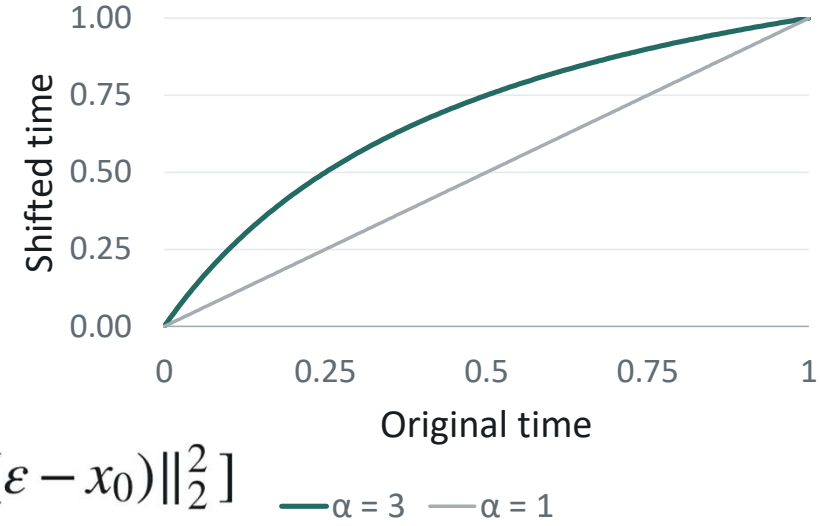
All reported models and GenEval submetrics

Model	Overall	Objects		Counting	Colors	Position	Color Attribution
		Single	Two				
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

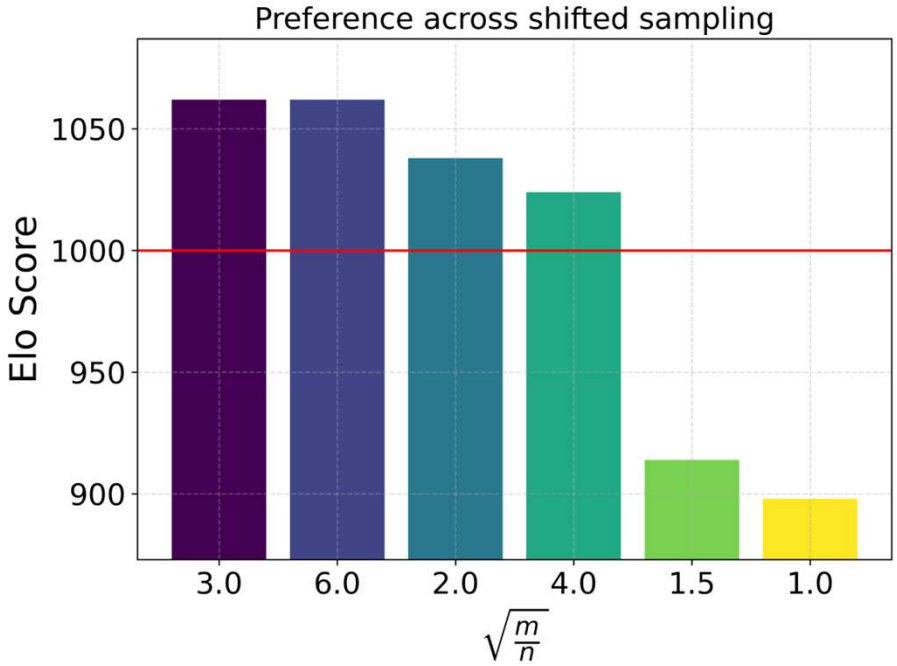
Method 3: Resolution-dependent shifting: enable scaling in resolution

$$t_m = \frac{\sqrt{\frac{m}{n}}\, t_n}{1 + (\sqrt{\frac{m}{n}} - 1) t_n}$$



$$\mathcal{L}_{\text{high-res}} = \mathbb{E}_{t_n \sim p(t),\, \varepsilon} \left[\| v_\theta((1 - t_m) x_0 + t_m \varepsilon,\ t_m) - (\varepsilon - x_0) \|_2^2 \right]$$

Experiments: Choosing the timestep shift: validating method 3



512^2 examples, shift 1.0



512^2 examples, shift 3.0



Figure 6: resolution-dependent shifting

Experiments: Measuring human preference

Stable Diffusion 3 beats all baselines in human preference evaluation

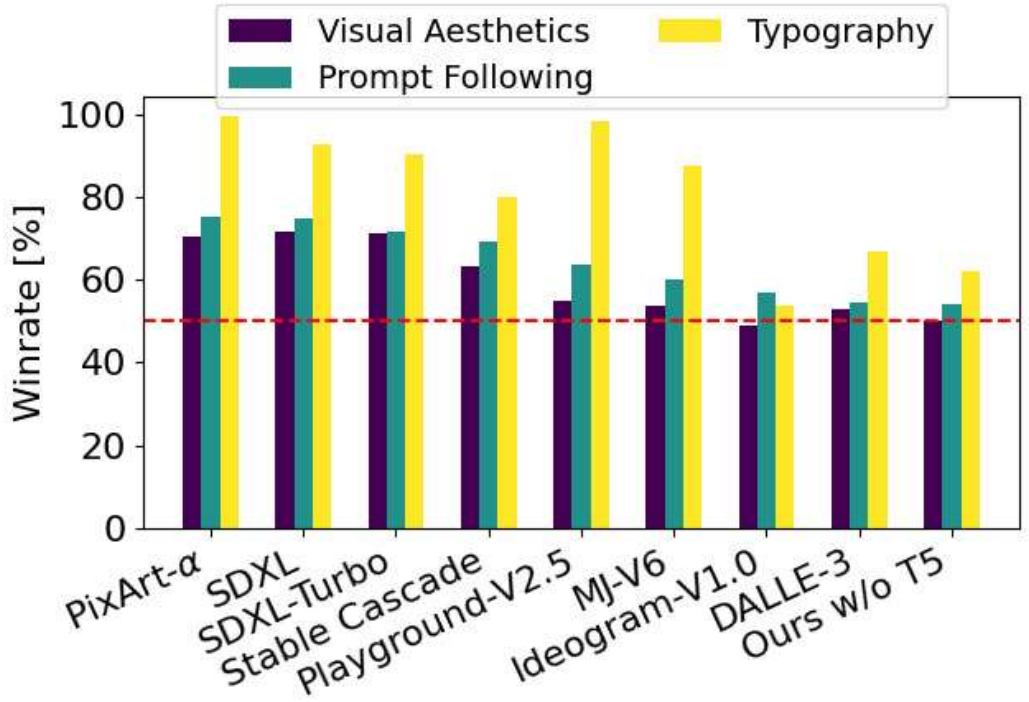


Figure 7: human preference evaluation

Scalable study

256^2 images, 500k updates, batch 4096, up to 8B parameters

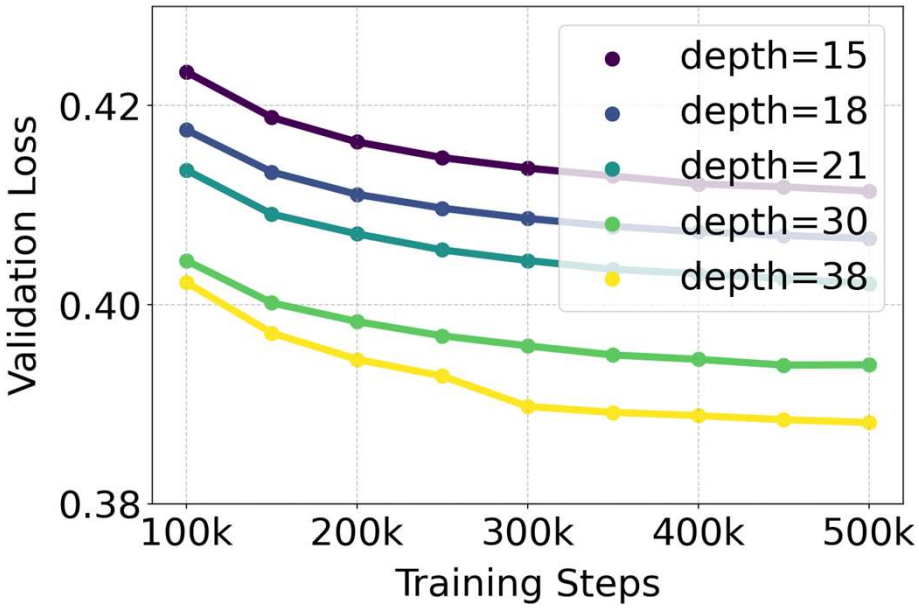


Figure 8: loss versus training steps

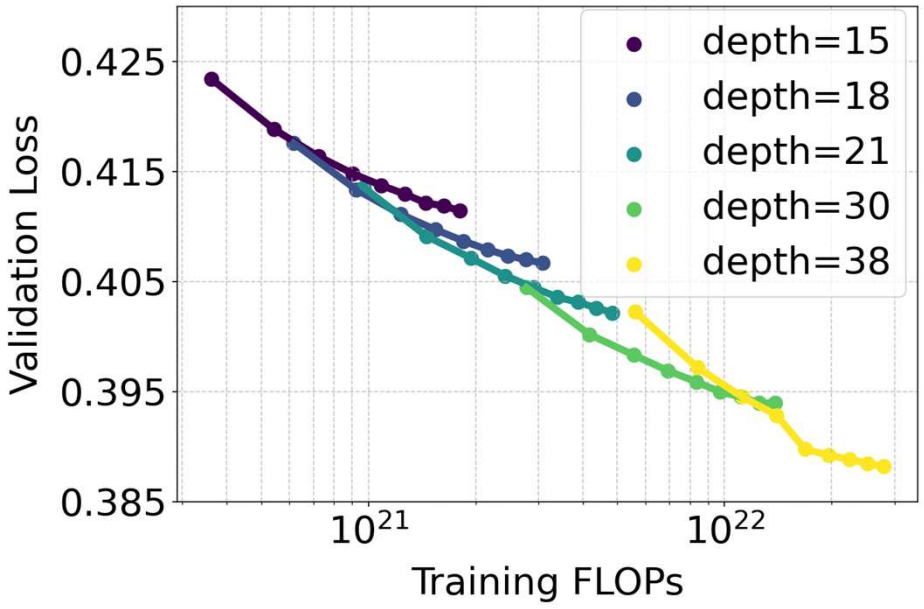


Figure 8: loss versus training FLOPs

Scalable study

Image-model panels, validation loss and evaluation scores

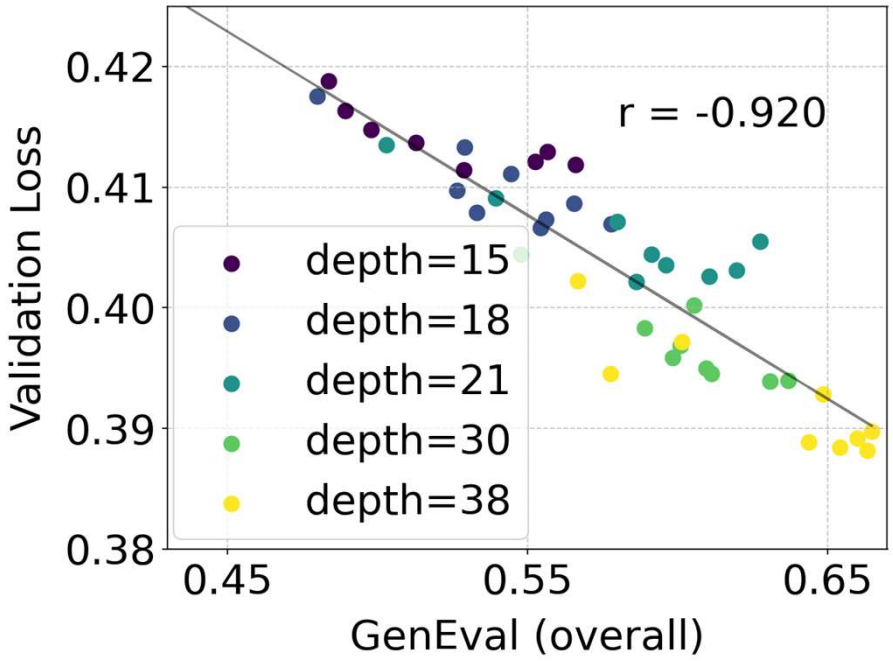


Figure 8: loss and GenEval

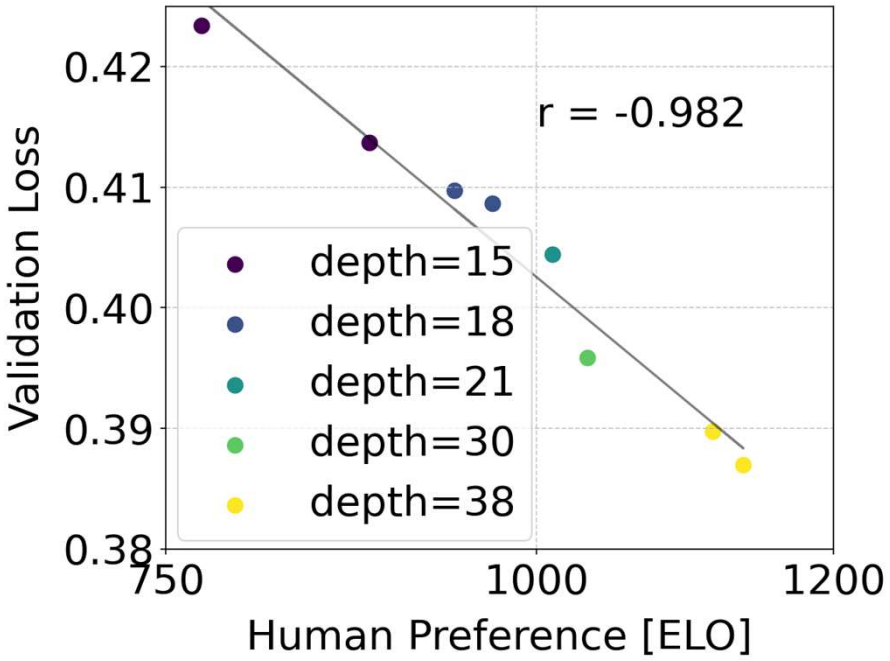


Figure 8: loss and human preference

Impact

Aug 2024



FLUX.1

12B

Dec 2024



HunyuanVideo

13B

Aug 2025



Qwen-Image

20B

[1] Black Forest Labs. FLUX (2024). github.com/black-forest-labs/flux

[2] Weijie Kong et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv:2412.03603, 2024.

[3] Chenfei Wu et al. Qwen-Image Technical Report. arXiv:2508.02324, 2025.

Discussion (or weakness)

- Why is the velocity target $\boldsymbol{\varepsilon - x_0}$ harder to predict at intermediate t, and why does emphasizing those timesteps improve training?
- Why do **FID and CLIP** rank samplers so differently, and which metric should we trust?
- Does the scaling trend continue **beyond 8B parameters**, or does it eventually saturate?