DIFFUSIONBENCH: ON HOLISTIC EVALUATION OF DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion transformer (DiT) research on image generation has converged to a single evaluation setup: class-conditional generation on ImageNet. While methods improve the FID and related metrics, it is increasingly unclear whether they reflect real progress in generative modeling. The natural alternative, i.e., text-to-image (T2I) generation, is perceived as too costly or inconvenient to train and evaluate and is often skipped. We argue that this perception no longer holds. We introduce NANOGEN, a unified DiT training and evaluation framework. NANOGEN matches state-of-the-art DiT baselines on ImageNet and, with 12 lines of configuration change, also trains competitive text-to-image models. It currently supports RAE, VAE, pixel-space, and MeanFlow diffusion methods under both ImageNet and T2I setups. Under NANOGEN, training T2I requires comparable compute to ImageNet. After training 21 latent diffusion models with NANOGEN, we observe that method ranking shows no strong correlation between ImageNet and T2I generation: Pearson correlation is between -0.377 and -0.580 across three metrics. This suggests that a method which improves class-conditional ImageNet FID may show no corresponding improvement on T2I, clearly indicating the necessity of evaluating DiTs on both tasks. To this end, we summarize ImageNet and text-to-image results, which yields DIFFUSIONBENCH, a holistic benchmark for DiT research. We recommend reporting DIFFUSIONBENCH in place of ImageNet alone: improvements on DIFFUSIONBENCH are more likely to reflect broader progress.

1 INTRODUCTION

Diffusion transformers (DiTs) have become the dominant method for image generation, with rapid progress over the last few years across architecture design (Peebles & Xie, 2023; Ma et al., 2024; Wang et al., 2025b; Yao et al., 2025), training objectives (Liu et al., 2023; Li & He, 2025; Geng et al., 2025a), and representation learning (Yu et al., 2024; Jiang et al., 2025; Wu et al., 2025b; Singh et al., 2025; Leng et al., 2025). Over the same period, the community has converged on a very narrow set of datasets for measuring this progress, *e.g.*, most prominently, class-conditional ImageNet (Deng et al., 2009) generation at 256 and 512 resolutions. The use of ImageNet-FID (Heusel et al., 2017) as a reporting standard has value: it makes method-method comparisons cheap and has accelerated progress on important modelling questions in image generation.

However, with better modelling methods, it becomes increasingly hard to tell whether a reported gain on ImageNet indicates broadly better modelling or some kind of overfitting to the benchmark itself. The natural correction would be extending the evaluation to text-to-image (T2I) generation, a critical application of diffusion models, but in practice this does not happen very often. This is likely because training T2I models is perceived as high-cost and high-friction: they require different data pipelines, different evaluation procedures, and often an entirely different codebase.

We challenge the premise that evaluating beyond ImageNet requires a separate and high-cost research programme. We introduce NANOGEN, a DiT training framework whose ImageNet configuration matches state-of-the-art methods, and whose text-to-image configuration is reachable from the ImageNet configuration with roughly 12 lines of config changes, covering the dataset and conditioning module. Recent works such as i1 (Zeng et al., 2026) and MiniT2I (Wang et al., 2026) investigate simple recipes for training a strong T2I model, while our goal is to build a unified framework and compare the effectiveness of recent methods across tasks. With the shared backbone, optimiser,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

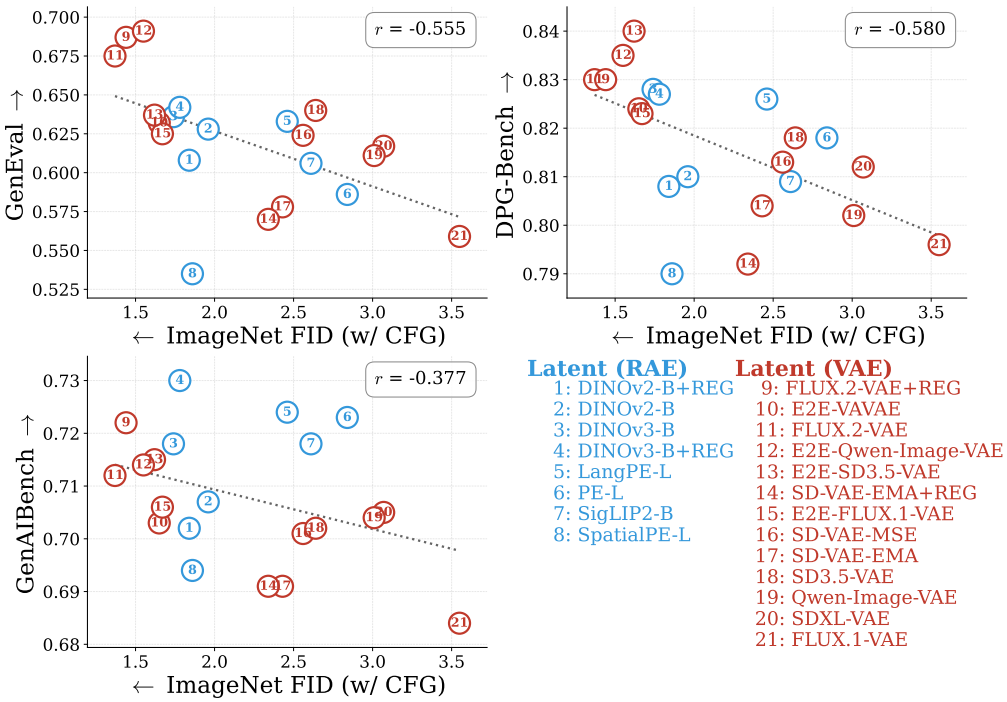


Figure 1: **Class-conditional ImageNet FID is not strongly correlated with T2I metrics for VAE and RAE methods.** Pearson correlations between ImageNet FID and a set of text-to-image evaluation metrics: GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and GenAIBench (Li et al., 2024) across both RAE and VAE latent spaces. Results on ImageNet are evaluated under the best CFG scale of each method. We find no evidence of strong correlation across the three T2I metrics, indicating that ImageNet FID does not reliably predict T2I quality. We focus on latent-space methods in this comparison, as including pixel-space methods would inflate the overall correlation. The corresponding without CFG version and the variants that additionally include three pixel-space methods are shown in Fig. 5 and Fig. 4 in the appendix.

training loop, and evaluation suite, this paper asks a scientific question: if a method improves class-conditional ImageNet FID, does this imply a corresponding improvement on T2I generation?

Among frontier DiT models, better ImageNet-FID does not consistently translate to better T2I performance. Using NanoGen, we train 21 latent diffusion models for ImageNet and T2I generation with a shared training recipe. As shown in Fig. 1, ImageNet-FID performance does not reliably predict T2I performance measured by metrics such as GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and GenAIBench (Li et al., 2024). As further shown in Fig. 5 in the Appendix, this problem is consistent regardless if classifier-free guidance is applied. Therefore, a technique that improves over existing methods on ImageNet generation may not exhibit such improvement on T2I generation. Without more holistic evaluation, progress in image generation may not generalize.

Using NanoGen, we are able to obtain and combine ImageNet and T2I generation results using a range of metrics into a single benchmark, DiffusionBench. Since ImageNet rankings do not reliably predict T2I performance (Fig. 1), our core recommendation is that future DiT work report DiffusionBench rather than ImageNet alone. Methods that improve DiffusionBench are then more likely to reflect broadly useful progress.

We highlight the main contributions of this paper below:

- We release NanoGen, a unified DiT training framework (Sec. 2) that matches state-of-the-art methods on ImageNet (Tab. 1) and extends to T2I with roughly 12 lines of config changes.
- We show empirically that ImageNet rankings do not reliably predict text-to-image performance (Fig. 1, Tab. 2 and Tab. 3), with magnitudes large enough to flip conclusions from ImageNet.
- We incorporate both tasks into DiffusionBench (Sec. 3; Tab. 2 and Tab. 3), present results of many existing methods, and recommend its adoption as a default DiT benchmark.

2 THE NANOGEN TRAINING AND EVALUATION FRAMEWORK

NANOGEN is a diffusion model training and evaluation framework that supports both class-conditional ImageNet generation and text-to-image generation under a single codebase. Its goal is to make the additional cost of evaluating a method on the T2I task as low as possible, so that "Did this idea also help on T2I?" is a question any author can answer without re-inventing the wheel.

2.1 OVERALL ARCHITECTURE

Design principles. NANOGEN uses one DiT backbone, one optimiser, one training loop, one evaluation harness, and one config format for both ImageNet and T2I tasks. Switching tasks requires two changes: (i) the data pipeline switches datasets: class-labelled ImageNet for class-conditional generation, or captioned images for T2I; (ii) the conditioning module is swapped accordingly: a class embedder for ImageNet, or a frozen text encoder for T2I. Everything else remains the same. As a result, moving NANOGEN to a new task is equivalent to switching a dataset and a conditioner, rather than rewriting the stack. NANOGEN supports many recent diffusion methods, including RAE (Zheng et al., 2025), VAE (Kingma & Welling, 2013), pixel-space, REG (Wu et al., 2025b), and MeanFlow (Geng et al., 2025a) methods. It also supports RAEv2 (Singh et al., 2026) encoders and tokenizers.

Backbone architecture. We use a standard diffusion transformer with three deliberate modifications relative to the common DiT recipe.

- Decoupled Diffusion Transformer (DDT) (Wang et al., 2025b). We use the DDT backbone as RAE (Zheng et al., 2025) does, splitting the model into an encoder and a decoder. The encoder takes the noisy input together with the conditioning tokens and produces a semantic representation. The decoder is a shallow but wide transformer, which takes that representation and the noisy entity as input and predicts the diffusion target. This split increases effective width without the quadratic FLOPs cost of a uniformly wide DiT.
- No AdaLN in the encoder. Similar to iMeanFlow (Geng et al., 2025b), i1 (Zeng et al., 2026), and MM-JiT (Wang et al., 2026), we remove the AdaLN (Peebles & Xie, 2023) modules from encoder blocks and retain AdaLN in the decoder. The modulation in the decoder is computed from the semantic output of the encoder rather than directly from the timestep.
- In-context conditioning. The encoder receives timestep and other conditioning tokens prepended to the visual tokens, without task-specific modulation.

Because all conditioning is in-context, adding or removing conditioning for a new task only requires changing the conditioning tokens. The rest of the architecture is identical across tasks.

Task-specific conditioning tokens. Only token counts and meanings differ between tasks:

- ImageNet (class-conditional). 4 timestep tokens plus 8 class-conditioning tokens.
- Text-to-image. 4 timestep tokens plus 256 text-conditioning tokens.

Everything else, such as backbone, loss formulation, optimiser, and EMA, is shared. This allows for moving a method to T2I generation with a small change rather than extensive engineering effort.

Training recipe. Wherever possible we keep optimisation-level choices fixed across tasks. We use the AdamW (Loshchilov & Hutter, 2017) optimiser with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, and a learning rate that linearly warms up to $2{\times}10^{-4}$ and then linearly decays to $2{\times}10^{-5}$. We apply gradient clipping at 1.0 and maintain an exponential moving average (EMA) of model weights with decay 0.9995. For the diffusion schedule, training timesteps are sampled from a logit-normal distribution with mean 0 and standard deviation 1. Following SD3 (Esser et al., 2024) and RAE (Zheng et al., 2025), we additionally apply dimension-dependent timestep shifting $t_m = \frac{\alpha t_n}{(1+(\alpha-1)t_n)}$, where $\alpha = \sqrt{\frac{m}{n}}$, $n = 4,096$, and m is the effective input dimension. We use v-prediction by default. For methods whose original recipe specifies x-prediction, *e.g.*, JiT (Li & He, 2025), PixelGen (Ma et al., 2026), we preserve x-prediction for faithful reproduction. For sampling, we default to an Euler sampler with 50 function evaluations (NFEs). On ImageNet we report results both with and without

Table 1: **ImageNet-256 reproducibility across latent-space and pixel-space methods.** For each method, we use AdamW and follow the original architecture. We build a model of similar size by adjusting the transformer width and depth, and train it for the same number of epochs as in the original paper. Evaluation follows the original setup of each method. Using a unified codebase, we achieve FID and IS scores comparable to or better than the published results.

Method	Epochs	#Params	Pred.	NFE	with Guidance	FID↓	IS↑
Latent-space							
RAE (DINOv2-B)	80	839M	v	50	×	2.16	214.8
Ours	80	847M	v	50	×	**2.07**	**213.5**
E2E-VAVAE	80	675M	v	250	×	5.26	-
Ours	80	680M	v	250	×	**3.64**	**152.5**
E2E-VAVAE + REPA	80	675M	v	250	×	3.46	159.8
Ours	80	681M	v	250	×	**2.88**	**165.4**
Pixel-space							
PixNerd	160	458M	v	100	✓	2.64	297.0
Ours	160	446M	v	100	✓	**2.58**	**299.3**
JiT	200	131M	x	50	✓	8.62	-
Ours	200	88M	x	50	✓	**5.49**	**231.6**
PixelGen	40	459M	x	50	×	7.53	131.7
Ours	40	458M	x	50	×	**7.52**	**123.5**

classifier-free guidance (CFG) (Ho & Salimans, 2022); on T2I we report only the with-CFG results. Data-dependent hyperparameters such as batch size, total training budget, and warmup duration also differ by task and are specified per section. For MeanFlow (Geng et al., 2025a) models, we use 75% flow-matching loss mixed with 25% MeanFlow loss. We set $\kappa = 0$ for the without-CFG experiments and $\kappa = 0.5$ for the with-CFG experiments.

Evaluation protocols. NANOGEN supports unified online evaluation during training. The evaluation harness depends only on the HuggingFace Transformers library and requires no other packages. On ImageNet, we support FID (Heusel et al., 2017), IS (Salimans et al., 2016), FDr (Yang et al., 2026), and MIND (Berthet et al., 2026). On T2I, GenAIBench (VQAScore) (Lin et al., 2024), DPG-Bench (Hu et al., 2024), and GenEval (Ghosh et al., 2023) are supported. Users can track training progress across all metrics in real time.

2.2 IMAGENET GENERATION IN NANOGEN

Before evaluating on the text-to-image task, we first confirm that NANOGEN establishes trustworthy ImageNet baselines, where the implemented methods match reported numbers in their papers.

Implementation details. Using NANOGEN, we train DiT models on ImageNet (Deng et al., 2009) at 256×256 resolution for both latent-space and pixel-space generation. For data pre-processing, we follow ADM (Dhariwal & Nichol, 2021) and center-crop images to 256×256. For the model backbone, we use a DDT with a 28-layer encoder of width 1,152 and a 2-layer decoder of width 2,048, yielding ∼615M parameters. We use a budget of 80 epochs at batch size 1,024 and 40 epochs of warmup for the learning-rate schedule. For evaluation, we follow the standard ImageNet protocol and generate 50,000 images, with 50 samples per class. We report FID (Heusel et al., 2017), IS (Salimans et al., 2016), FDr (Yang et al., 2026), and MIND (Berthet et al., 2026) metrics.

ImageNet reproducibility validation. We implement and re-train six existing methods using NANOGEN, including three latent-space methods: RAE (Zheng et al., 2025) and two E2E-VAEs (Leng et al., 2025), and three pixel-space methods: PixNerd (Wang et al., 2025a), JiT (Li & He, 2025), and PixelGen (Ma et al., 2026). We summarize their reported numbers and the NANOGEN results in Tab. 1, where we try our best to keep major hyperparameters similar, *e.g.*, model size, inference NFEs, and CFG settings. We observe that the NANOGEN results are competitive with published numbers and sometimes slightly superior. This validates the effectiveness of NANOGEN. We treat this validation as a prerequisite for cross-task analysis, not a contribution.

2.3 TEXT-TO-IMAGE GENERATION IN NANOGEN

Text-to-image generation is the second axis of DiT training and evaluation in NANOGEN. Because ranking of methods measured on ImageNet does not appear consistent on T2I, as shown in Fig. 1,

it is important to evaluate method effectiveness not only on ImageNet, but also on T2I. This section first shows that moving from ImageNet training to T2I training is a small change in NANOGEN. We then use existing T2I metrics, such as GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024), to compare DiT methods in Sec. 3.

Implementation details. The T2I configuration of NANOGEN is reached from the ImageNet configuration by replacing the class-embedding conditioner with a text-encoder conditioner and switching the dataset loader to a captioned image corpus. We use Qwen3-0.6B (Team, 2025a) as the text encoder and take its final hidden states as the text-conditioning tokens. Pre-training uses the JourneyDB (Sun et al., 2023), Long-Caption, and Short-Caption splits of BLIP-3o (Chen et al., 2025). We use a batch size of 1024 and 10% conditioning dropout for CFG and train for 100K iterations. All other choices, including AdamW optimiser, learning-rate schedule, EMA, gradient clipping, v-prediction objective, and sampler, are inherited from the recipe in Sec. 2. For evaluation, we apply a classifier-free guidance scale of 6.0 across the entire timestep interval. To avoid metric hacking on specific T2I benchmarks, we report results from the pre-training stage only and skip supervised fine-tuning on datasets like BLIP-3o-60K.

3 DIFFUSIONBENCH: A HOLISTIC BENCHMARK

DIFFUSIONBENCH evaluates DiT methods on both class-conditional ImageNet and text-to-image generation. We train each method using NANOGEN under the task-specific protocols detailed in Sec. 2 and report multiple benchmark metrics.

3.1 BENCHMARKING IMAGENET GENERATION METHODS

Latent-space generation. In the latent-space regime we train NANOGEN on both RAE (Zheng et al., 2025) and VAE latents. For RAE, we replace the latent encoder with six frozen pretrained vision encoders spanning two backbone scales and several pretraining objectives: ViT-B encoders DINOv2-B (Oquab et al., 2024), DINOv3-B (Siméoni et al., 2025), and SigLIP2-B (Tschannen et al., 2025); and ViT-L encoders PE-L, LangPE-L, and SpatialPE-L from the Perception Encoder family (Bolya et al., 2025). For VAE, we train a range of widely used VAEs, including SD-VAE (Rombach et al., 2022), SDXL-VAE (Podell et al., 2024), SD3.5-VAE (Esser et al., 2024), FLUX.1-VAE (Labs, 2024), FLUX.2-VAE (Labs, 2025), Qwen-Image-VAE (Wu et al., 2025a), and VA-VAE (Yao et al., 2025), together with their REPA-E (Leng et al., 2025) end-to-end variants. Results are summarized in Tab. 2, evaluated with the best per-method CFG scale over the timestep interval $[0.0, 0.9]$. Our goal is not a new ImageNet state of the art. We make three observations.

First, the best FID=1.37 is achieved by FLUX.2-VAE (Labs, 2025), followed by DiTs trained with the REPA-E VAE family (Leng et al., 2025), with FID around 1.5 and 1.6. It is unclear how the FLUX.2-VAE is trained, but its architecture shares the same batch normalization layer as in REPA-E (Leng et al., 2025) design, so perhaps they share similar mechanisms of end-to-end VAE and DiT tuning. **Second**, the RAE family has slightly higher FID, with the better ones around 1.7-1.9. DINOv3-B has the best FID=1.74 among RAEs, while DINOv2-B has an FID = 1.96. Comparing results in Table 2 with Table 4, it remains unclear how RAE benefits further from CFG, an open direction we leave to future work. **Third**, traditional VAEs such as SD-VAE (Rombach et al., 2022) and SD3.5-VAE (Esser et al., 2024) lag behind. We note that the performance gap at 80 epochs is largely due to the faster convergence from well-structured latent spaces, such as RAE and REPA-E latents. We expect the gap with standard VAEs to narrow with longer training.

Pixel-space generation and MeanFlow. For pixel-space generation, NANOGEN operates directly on pixels without any latent tokenizer. In Tab. 2, we train PixNerd (Wang et al., 2025a), JiT (Li & He, 2025), and PixelGen (Ma et al., 2026) using NANOGEN. We observe that pixel-space FID is typically higher than latent-space FID at 80 training epochs. Similar to traditional VAE methods, the evaluated pixel-space methods do not show accelerated convergence at 80 epochs.

NANOGEN also supports one-/few-step generation. We train MeanFlow (Geng et al., 2025a) on the SD-VAE-MSE (AI, n.d.) latent, and evaluate the model for one or two inference steps. MeanFlow reaches FID 6.60 and 5.40 with one or two steps, respectively. Even so, MeanFlow still lags behind the multi-step methods on either latent-space or pixel-space generation.

Table 2: **Systematic comparison on ImageNet-256 with CFG.** A single NANOGEN backbone and training recipe applied across diverse latent-space tokenizers and pixel-space architectures. All models are trained for 80 epochs with ∼615M parameters and reported with classifier-free guidance applied at the best CFG scale of each method, selected per method via a sweep over the guidance interval fixed at $[0.0, 0.9]$. FDr and MIND are computed against five vision encoders: Inception, ConvNeXt, DINOv2, MAE, and SigLIP. Except for the REPA-E family where the E2E VAEs are frozen after end-to-end training, REPA is not used.

Method	FDr↓					MIND↓					FID↓	IS↑
	Incep.	ConvNeXt	DINOv2	MAE	SigLIP	Incep.	ConvNeXt	DINOv2	MAE	SigLIP		
Latent-space (RAE) (Zheng et al., 2025)												
DINOv2-B	1.22	2.20	3.26	6.19	7.76	2.03	66.49	27.74	0.43	6.52	1.96	224.1
DINOv2-B + REG	1.15	2.15	3.21	6.43	7.71	1.90	64.30	27.40	0.44	6.43	1.84	236.2
DINOv3-B	1.09	2.10	3.30	6.53	7.66	1.80	65.83	29.35	0.43	6.03	1.74	244.2
DINOv3-B + REG	1.11	2.12	3.41	6.54	7.88	1.88	65.81	29.97	0.43	6.17	1.78	248.1
SigLIP2-B	1.59	3.01	7.33	10.61	11.38	1.86	97.91	76.01	0.87	9.89	2.61	222.9
PE-L	1.72	2.80	5.91	10.90	9.88	2.86	90.07	56.80	0.88	8.19	2.84	221.5
LangPE-L	1.48	2.99	6.27	8.87	10.24	2.12	108.83	61.86	0.68	9.03	2.46	196.7
SpatialPE-L	1.16	1.82	4.67	6.62	8.57	1.35	55.46	45.79	0.51	7.23	1.86	247.1
Latent-space (VAE)												
SD-VAE-EMA	1.38	1.92	7.71	6.94	19.15	2.92	48.52	87.44	0.62	20.39	2.43	259.6
SD-VAE-EMA + REG	1.32	1.74	7.24	7.55	18.47	2.04	55.02	84.72	0.69	20.14	2.34	271.6
SD-VAE-MSE	1.45	2.15	7.61	7.82	21.82	3.10	55.66	86.97	0.70	25.01	2.56	259.7
SDXL-VAE	1.69	2.74	9.26	8.64	21.55	3.38	70.85	109.22	0.75	22.66	3.07	256.0
SD3.5-VAE	1.51	2.14	7.76	6.19	15.16	3.04	75.53	89.20	0.55	14.04	2.64	262.9
FLUX.1-VAE	2.04	3.89	9.25	8.19	19.54	3.37	107.70	105.18	0.82	18.62	3.55	245.7
FLUX.2-VAE	0.89	1.07	4.32	3.90	10.75	0.90	24.98	43.59	0.31	9.84	1.37	272.7
FLUX.2-VAE + REG	0.92	0.95	4.27	3.91	10.17	1.06	37.56	42.47	0.31	9.42	1.44	294.1
Qwen-Image-VAE	1.85	4.56	9.77	9.19	25.46	2.75	159.26	118.42	0.90	25.97	3.01	238.9
E2E-VAVAE	1.08	1.99	4.51	4.71	9.83	2.13	62.17	50.75	0.36	9.17	1.65	275.4
E2E-FLUX.1-VAE	1.07	1.83	5.12	4.68	11.76	1.16	50.73	53.43	0.36	10.75	1.67	266.3
E2E-SD3.5-VAE	1.16	1.30	4.57	5.49	11.12	1.10	42.14	48.18	0.42	10.25	1.62	265.4
E2E-Qwen-Image-VAE	1.06	2.26	4.57	4.63	11.33	1.54	61.19	48.81	0.37	10.24	1.55	261.4
Pixel-space												
JiT	2.38	4.59	9.57	13.70	22.51	3.97	146.37	113.63	1.23	21.21	4.08	231.2
PixNerd	2.45	4.01	8.33	11.67	20.65	4.24	104.21	86.10	0.96	18.94	4.17	213.8
PixelGen	2.26	4.34	7.80	13.14	17.97	3.96	138.33	86.50	1.12	15.73	3.97	247.4
One-/Few-step												
MeanFlow (SD-VAE-MSE, NFE=1)	3.71	4.71	17.35	15.54	43.69	6.19	116.13	203.45	1.35	49.72	6.60	206.7
MeanFlow (SD-VAE-MSE, NFE=2)	3.19	3.25	13.19	9.84	28.42	7.93	69.05	148.75	0.83	27.45	5.40	226.5

3.2 BENCHMARKING T2I GENERATION METHODS

We take the same diffusion methods evaluated on ImageNet (Sec. 2.2) and re-train each into a T2I model under a common protocol. For each method we report results using T2I metrics including GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and GenAIBench (Li et al., 2024). All variants follow the training recipe of Sec. 2.3. Tab. 3 summarises the systematic comparison of all methods under the latent-space and pixel-space setups. We have five major observations.

First, in the state-of-the-art frontier, ImageNet ranking does not robustly predict the T2I ranking. For example, RAE with SpatialPE-L has very good ImageNet FID, but its T2I performance is among the worst across various metrics. **Second**, different metrics to some extent disagree with each other. For example, E2E-Qwen-Image-VAE is one of the strongest if we look at GenEval and DPG-Bench metrics but it falls into the second tier under the GenAIBench metric. **Third**, the class-conditional ImageNet trend is consistent with the T2I-metric trend if we look at broader method category ranking. That is, improved latent-space methods (RAE, FLUX.2-VAE, and REPA-E) > traditional latent-space methods > pixel-space methods > MeanFlow (Tab. 2, Tab. 3, and Fig. 4 [a,b]). From this perspective, ImageNet signals are useful. End-to-end VAE tuning improves both ImageNet FID and T2I metrics, including FLUX.1-VAE and Qwen-Image-VAE. But a better ImageNet FID does not predict a better T2I score across different methods, which remains the case without CFG (Fig. 5). Most state-of-the-art methods report FID between 1 and 2, which fall into the most uncorrelated regions (Fig. 4b). **Fourth**, our pre-trained T2I models generally score lower than the public T2I models in Table 3.

Table 3: **DiT comparison on text-to-image generation.** We use NANOGEN for training DiTs across the latent space, pixel space, and MeanFlow under a unified backbone and training recipe. We report GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and GenAIBench (Li et al., 2024). For reference, we additionally report the benchmark results of several public T2I models, such as SD3.5-Large (Esser et al., 2024), FLUX-1 (Labs, 2024), FLUX-2 (Labs, 2025), Qwen-Image (Wu et al., 2025a), and Z-Image-Turbo (Team, 2025b).

Method	Iters	#Params	GenEval↑	DPG-Bench↑	GenAIBench↑
Public models					
SD-3.5-Large	-	8B	0.691	0.842	0.767
FLUX-1	-	12B	0.654	0.838	0.748
FLUX-2	-	32B	0.854	0.870	0.841
Qwen-Image	-	20B	0.848	0.888	0.803
Z-Image-Turbo	-	6B	0.736	0.847	0.759
Latent-space (RAE) (Zheng et al., 2025)					
DINOv2-B	100K	615M	0.628	0.810	0.707
DINOv2-B + REG	100K	619M	0.608	0.808	0.702
DINOv3-B	100K	615M	0.636	0.828	0.718
DINOv3-B + REG	100K	619M	0.642	0.827	0.730
SigLIP2-B	100K	615M	0.606	0.809	0.718
PE-L	100K	617M	0.586	0.818	0.723
SpatialPE-L	100K	617M	0.535	0.790	0.694
LangPE-L	100K	617M	0.633	0.826	0.724
LangPE-L	200K	617M	0.635	0.824	0.715
Latent-space (VAE)					
SD-VAE-EMA	100K	611M	0.578	0.804	0.691
SD-VAE-EMA + REG	100K	615M	0.570	0.792	0.691
SD-VAE-MSE	100K	611M	0.624	0.813	0.701
SDXL-VAE	100K	611M	0.617	0.812	0.705
SD3.5-VAE	100K	612M	0.640	0.818	0.702
Qwen-Image-VAE	100K	612M	0.611	0.802	0.704
E2E-FLUX.1-VAE	100K	612M	0.625	0.823	0.706
E2E-SD3.5-VAE	100K	612M	0.637	0.840	0.715
E2E-Qwen-Image-VAE	100K	612M	0.691	0.835	0.714
FLUX.1-VAE	100K	612M	0.559	0.796	0.684
FLUX.1-VAE	200K	612M	0.544	0.816	0.687
FLUX.2-VAE	100K	612M	0.675	0.830	0.712
FLUX.2-VAE	200K	612M	0.625	0.841	0.713
FLUX.2-VAE + REG	100K	616M	0.687	0.830	0.722
E2E-VAVAE	100K	611M	0.632	0.824	0.703
E2E-VAVAE	200K	611M	0.679	0.836	0.716
Pixel-space					
JiT	100K	615M	0.516	0.782	0.674
PixNerd	100K	615M	0.484	0.777	0.643
PixelGen	100K	615M	0.554	0.798	0.678
One-/Few-step					
MeanFlow (SD-VAE-MSE, NFE=1)	100K	613M	0.287	0.688	0.582
MeanFlow (SD-VAE-MSE, NFE=2)	100K	613M	0.341	0.721	0.602

Last, when we train T2I for 200K steps, the performance generally remains similar or improves slightly under the three metrics. Yet images at 200K steps look better than at 100K (Fig. 3). We suspect better metrics are needed.

Are the T2I results competitive? As shown in Fig. 2, we use a small compute budget around 10 hours of wall-clock time with 32 H200 GPUs, though all setups are runnable on 8 H200 GPUs. RAEv2 (Singh et al., 2026) reports a GenEval score of 0.624 using a SigLIP2-B encoder and an 875M diffusion model, pre-trained on the same dataset for 150K iterations; in comparison, we report a GenEval score of 0.691 using E2E-Qwen-Image-VAE and 0.633 using RAE based on LangPE-L. We did not find many other public numbers of pre-trained-only T2I models under similar compute. On the other hand, many papers report T2I models after supervised fine-tuning on the BLIP-3o-60K (Chen et al., 2025) dataset, where their GenEval scores are around 0.85-0.90. We did similar fine-tuning on top of our FLUX.1-VAE T2I checkpoint and can achieve 0.90 GenEval score. While GenEval scores can be very high, this might be due to metric hacking, and the resulting models may not be universally better. We call for more hack-resistant evaluation for T2I models.

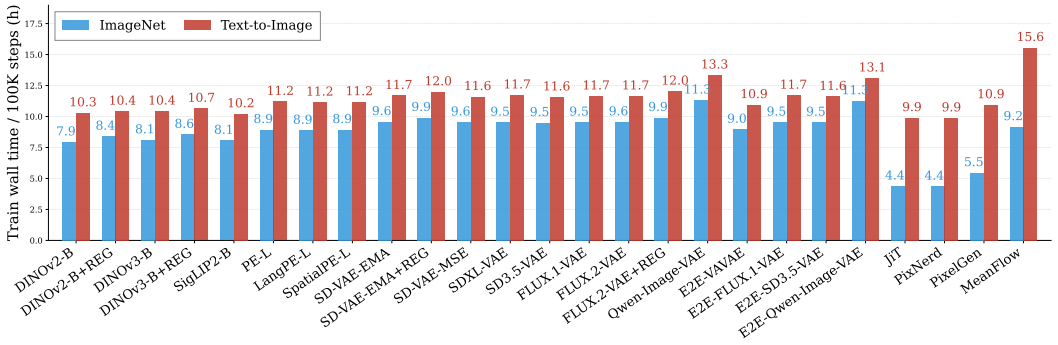


Figure 2: **Wall-clock training time comparison of ImageNet and T2I setups.** We time 100K training steps for 25 DiT methods on 32 H200 GPUs using a unified NanoGen recipe. Training T2I remains efficient across all methods. Moreover, training cost is comparable across latent-space methods, while pixel-space methods such as JiT (Li & He, 2025), PixNerd (Wang et al., 2025a), and PixelGen (Ma et al., 2026), are much cheaper to train on ImageNet because they do not compute latents from VAEs. RAE (Zheng et al., 2025) methods are marginally faster to train than VAE methods, because RAE relies on transformer-based vision encoders whereas VAEs mainly use a convolution-based U-Net structure. MeanFlow (Geng et al., 2025a) is much slower than other T2I methods, as it computes the MeanFlow objective with `torch.jvp`, which adds substantial computational overhead. Unless stated otherwise, ImageNet and T2I models are trained for 100K steps.

Recommended usage. Our recommendation is that future DiT papers report DiffusionBench, which includes both ImageNet and T2I generation, rather than any single axis. Methods that improve DiffusionBench are more likely to reflect broadly useful progress; methods that improve one axis but regress another may still be valuable, but should be labelled as task-specific improvements.

4 Background and Related Work

We discuss the most relevant work here; Appendix A.5 provides further background.

Diffusion transformers (DiTs) (Peebles & Xie, 2023) replace the U-Net backbone of earlier diffusion models with a transformer over image patches. Later works have refined it, including SiT (Ma et al., 2024), long skip connections (Bao et al., 2023), joint text–image attention in MMDiT (Esser et al., 2024), and encoder–decoder split in DDT (Wang et al., 2025b). NanoGen currently is built on the DDT backbone (Sec. 2). Some other works remove the latent tokenizer and train diffusion transformers directly in pixel space, including PixNerd (Wang et al., 2025a), JiT (Li & He, 2025), and PixelGen (Ma et al., 2026). Recent works also train text-to-image models with simple and fully open recipes, including i1 (Zeng et al., 2026) and MiniT2I (Wang et al., 2026). These works focus on ablating training recipes for a single strong T2I model. In contrast, we fix a unified recipe and compare a wide range of recent diffusion methods in a fair setup, with minimal engineering effort and compute cost. We do not aim to achieve state-of-the-art performance on T2I tasks.

ImageNet metrics and their limitation. Generation quality on ImageNet is typically measured by Frechet Inception Distance (FID) (Heusel et al., 2017), Inception Score (Salimans et al., 2016), and precision/recall (Kynkäänniemi et al., 2019). FID has limitations: it is sensitive to image resizing and preprocessing, depends on a frozen Inception-v3 (Szegedy et al., 2016) classifier trained on a related domain, and has now saturated. Alternatives to FID include sFID (Nash et al., 2021), FDr (Yang et al., 2026), and MIND (Berthet et al., 2026).

Holistic benchmarking. Multi-task evaluation has gained ground in language modeling through HELM (Liang et al., 2022) and BIG-Bench (Srivastava et al., 2023). In image generation, HEIM (Lee et al., 2023) evaluates *already trained* T2I models. Existing infrastructure does not support *training and evaluating* the same DiT method on *both ImageNet and T2I*.

Text-to-image evaluation. Early text-to-image evaluation uses FID on MS-COCO (Lin et al., 2014) for image fidelity and diversity. They also use CLIPScore (Hessel et al., 2021) to evaluate prompt

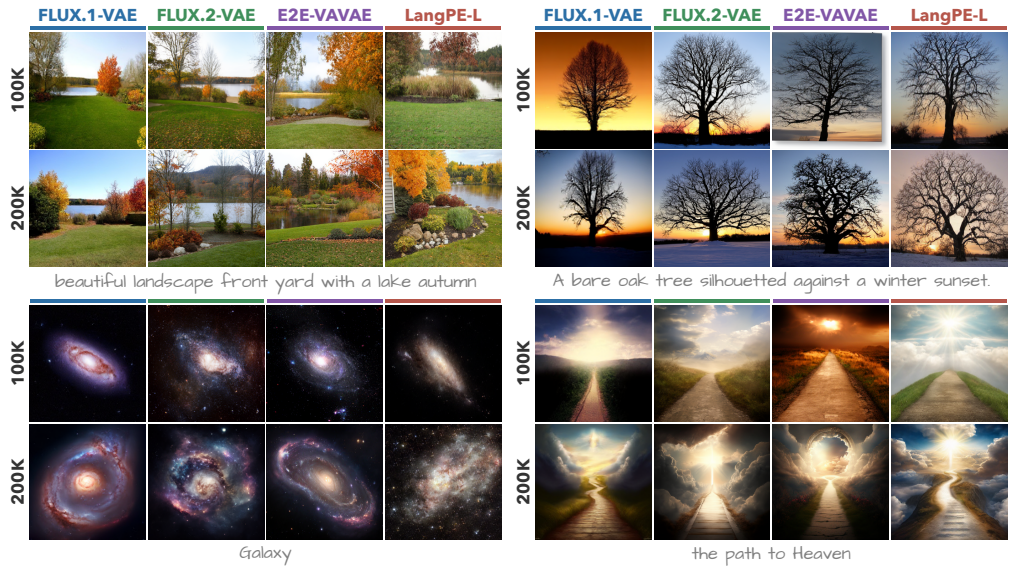


Figure 3: **Text-to-image qualitative samples at** 256×256**.** Curated qualitative samples from NANOGEN latent-space methods trained for 100K and 200K iterations at batch size 1024, evaluated on a shared set of text prompts. Quantitative scores for the same methods are reported in Tab. 3.

alignment. These metrics are coarse and do not accurately reflect model quality. Metrics with more accurate measurement of prompt alignment include reward models such as ImageReward (Xu et al., 2023), HPSv2 (Wu et al., 2023), and PickScore (Kirstain et al., 2023), compositional benchmarks such as GenEval (Ghosh et al., 2023) and DPGBench (Hu et al., 2024), and VLM-based scorers such as VQAScore (Lin et al., 2024), UnifiedReward (Wang et al., 2025c), and Qwen-Image-Bench (Li et al., 2026). These metrics aim to fully reflect human preference. The latter is currently best captured by large-scale human-labeled arenas such as the Artificial Analysis Arena (Artificial Analysis, 2026). It aggregates human pairwise votes into ELO ratings.

5 CONCLUSION

Diffusion transformer research has matured to the point where single-benchmark evaluation is no longer enough. In this paper we introduce NANOGEN, a training and evaluation framework that removes the engineering barrier to training and evaluating DiT methods on the T2I task, and use it to show that ImageNet rankings do not reliably predict text-to-image performance. Finally, we package the two evaluation axes: ImageNet and T2I generation, into DIFFUSIONBENCH and argue for its adoption as the default DiT benchmark. We hope that making holistic evaluation cheap in engineering and compute will encourage broadly useful progress.

What we are not claiming. We do not claim that ImageNet and FID are no longer useful; they are still a good platform for generative modeling. Besides, we do not claim that DIFFUSIONBENCH is a permanent benchmark. It should be refreshed as methods begin to saturate it.

Limitations. First, the ImageNet-T2I correlation (Fig. 1) is measured at the scale and compute we could afford and may look different at other scales. Second, the benchmarking results (Tab. 2 and Tab. 3) are obtained after 100K iterations with a batch size of 1,024, which we can afford. Longer training will improve the generated image quality (see Fig. 3).

Future work. There are several promising directions. First, DIFFUSIONBENCH can be broadened to other generative modalities such as world models, videos, and 3D, so that cross-task evaluation captures the full scope of diffusion-based generation. Second, current T2I metrics could be hacked through fine-tuning on a curated dataset, so better hack-resistant mechanisms are needed for T2I metrics. Finally, we envisage DIFFUSIONBENCH as a living and community-maintained leaderboard that is periodically refreshed to keep pace with advances in methodology.

AI USE STATEMENT

We used Claude Code (Anthropic, 2025) to draft the paper outline, prepare experiment configurations, and launch experiments. For all figures, we provided plotting templates and used Claude Code to adapt them to our experiment logs and checkpoints. We also used Cursor to assist with drafting parts of the paper. During initial proof-of-concept development, we used a coding agent to scaffold the codebase. The authors implemented the final versions of the core methods, with AI assistance limited to Cursor's code autocompletion. The authors manually checked the experiment configurations, figures, and AI-assisted text and take full responsibility for the final manuscript.

REFERENCES

Stability AI. Improved autoencoders ... https://huggingface.co/stabilityai/sd-vae-ft-mse, n.d. Accessed: April 11, 2025.

Anthropic. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet, 2025.

Artificial Analysis. Artificial analysis: Independent analysis of AI models and API providers. https://artificialanalysis.ai/, 2026. Accessed June 2026.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Quentin Berthet, Yu-Han Wu, Clement Crepy, Romuald Elie, Klaus Greff, and Michael Eli Sander. Mind: Monge inception distance for generative models evaluation. *arXiv preprint arXiv:2605.06797*, 2026.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models–architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025a.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-image-bench: From generation to creation in text-to-image evaluation, 2026. URL https://arxiv.org/abs/2605.28091.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Pixel diffusion beats latent diffusion with perceptual loss. *arXiv preprint arXiv:2602.02493*, 2026.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *International Conference on Machine Learning*, pp. 7958–7968. PMLR, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In *Advances in Neural Information Processing Systems*, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Qwen Team. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer, 2025b.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation. https://peppaking8.github.io/#/post/minit2i, 2026.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier

than you think. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 7714–7743. Curran Associates, Inc., 2025b. doi: 10.52202/085713-0264. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0b99315234cc95e6ef281f9155b68832-Paper-Conference.pdf.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Boya Zeng, Tianze Luo, Shu Pu, Jucheng Shen, Taiming Lu, Gabriel Sarch, and Zhuang Liu. i1: A simple and fully open recipe for strong text-to-image models. *arXiv preprint arXiv:2606.11289*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690, 2025. URL https://arxiv.org/abs/2510.11690.

A ADDITIONAL RESULTS

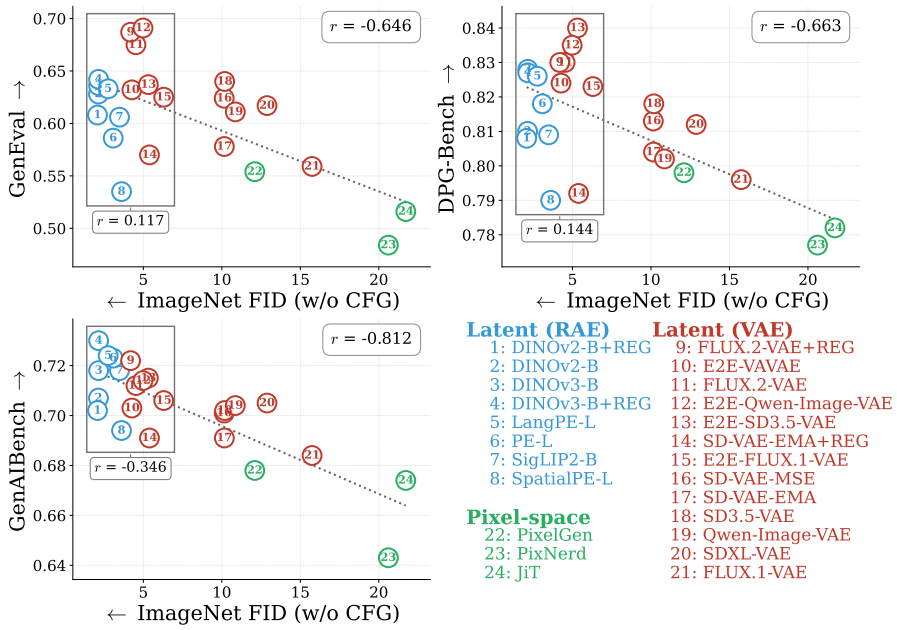
A.1 IMAGENET SYSTEMATIC COMPARISON WITHOUT CFG

Table 4: **Systematic comparison on ImageNet-256 without CFG.** All training setup is kept the same as Tab. 2, but evaluated without classifier-free guidance.

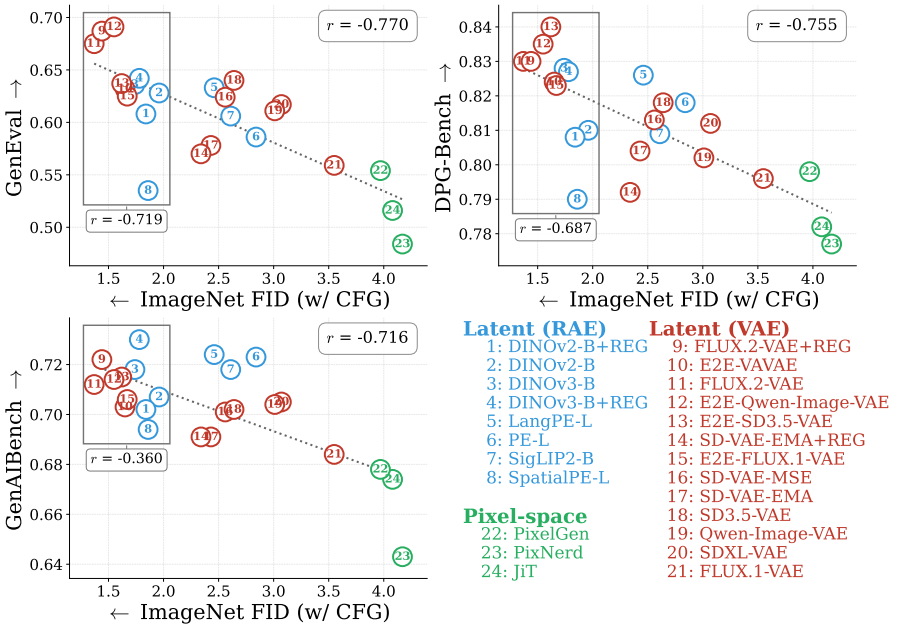
Method	FDr↓					MIND↓					FID↓	IS↑
	Incep.	ConvNeXt	DINOv2	MAE	SigLIP	Incep.	ConvNeXt	DINOv2	MAE	SigLIP		
Latent-space (RAE) (Zheng et al., 2025)												
DINOv2-B	1.32	2.33	3.14	6.33	7.56	2.54	70.53	26.52	0.44	6.30	2.14	211.9
DINOv2-B + REG	1.29	2.45	3.26	6.42	7.82	2.23	76.20	27.98	0.45	6.47	2.08	207.7
DINOv3-B	1.33	2.82	3.89	6.80	8.17	2.16	97.15	33.62	0.45	6.45	2.15	200.8
DINOv3-B + REG	1.32	2.71	3.79	6.71	8.11	2.35	90.39	32.33	0.45	6.38	2.15	204.4
SigLIP2-B	2.11	4.20	8.55	10.97	11.76	2.45	162.71	85.81	0.90	10.15	3.48	179.4
PE-L	1.86	3.15	6.17	11.01	9.87	3.05	97.69	58.40	0.89	8.08	3.08	206.6
LangPE-L	1.65	3.48	6.66	8.97	10.28	2.53	138.39	64.01	0.69	9.07	2.76	182.2
SpatialPE-L	2.18	4.17	7.30	7.01	10.74	2.35	206.52	69.92	0.54	9.16	3.61	160.4
Latent-space (VAE)												
SD-VAE-EMA	5.97	10.14	18.17	11.11	35.43	9.56	688.82	245.95	0.99	38.87	10.16	113.4
SD-VAE-EMA + REG	3.14	5.14	11.30	9.58	25.63	3.45	246.05	145.13	0.87	28.41	5.39	162.7
SD-VAE-MSE	5.97	10.36	17.84	12.03	38.19	9.34	676.48	243.23	1.07	43.99	10.15	112.4
SDXL-VAE	7.50	12.58	21.50	13.57	40.39	12.96	858.54	297.59	1.21	44.20	12.88	104.7
SD3.5-VAE	5.96	10.21	17.97	10.39	30.85	8.50	634.85	240.52	0.96	30.70	10.18	111.7
FLUX.1-VAE	9.28	16.66	23.27	14.02	41.46	14.81	1072.69	316.55	1.41	42.58	15.75	86.6
FLUX.2-VAE	2.76	4.65	8.56	5.15	17.41	2.53	234.79	96.75	0.41	16.46	4.53	146.9
FLUX.2-VAE + REG	2.56	4.06	7.91	5.17	16.74	2.11	183.74	88.65	0.41	15.82	4.19	155.8
Qwen-Image-VAE	6.52	13.25	19.78	13.45	41.51	9.34	804.53	270.50	1.31	44.01	10.86	108.9
E2E-VAVAE	2.64	5.90	8.59	6.33	16.30	2.20	277.26	103.04	0.51	15.54	4.27	147.6
E2E-FLUX.1-VAE	3.83	6.89	10.92	6.67	21.28	4.23	373.50	129.48	0.53	20.39	6.30	134.3
E2E-SD3.5-VAE	3.37	5.10	9.21	7.07	18.70	3.19	266.87	108.44	0.55	17.69	5.32	140.8
E2E-Qwen-Image-VAE	3.11	6.80	9.57	6.46	19.85	2.84	337.95	113.68	0.53	18.88	4.98	138.4
Pixel-space												
JiT	12.82	22.87	28.16	23.59	53.76	24.19	1632.56	412.33	2.15	55.40	21.72	65.0
PixNerd	12.18	21.42	25.08	19.67	48.24	22.77	1581.55	345.19	1.71	49.33	20.61	63.9
PixelGen	7.01	13.93	17.89	18.98	34.49	9.07	769.26	222.02	1.68	32.35	12.10	104.0
One-/Few-step												
MeanFlow (SD-VAE-MSE, NFE=1)	14.56	24.09	37.39	21.15	70.61	31.56	1993.71	571.19	1.88	81.85	24.83	61.4
MeanFlow (SD-VAE-MSE, NFE=2)	12.24	21.97	34.62	16.23	58.31	24.85	1860.67	524.10	1.42	62.69	20.58	63.0

A.2 IMAGENET-FID AND T2I METRICS CORRELATION INCLUDING PIXEL-SPACE METHODS

Our main analysis excludes pixel-space methods. We include three pixel-space methods in Figure 4: JiT (Li & He, 2025), PixNerd (Wang et al., 2025a), and PixelGen (Ma et al., 2026), with both without and with CFG settings. Pixel-space methods are much worse than latent-space methods on both ImageNet FID and the T2I metrics, which would artificially raise the overall correlation. We therefore focus the main analysis on the latent-space frontier, where the correlation is not driven by these outliers. We also exclude MeanFlow methods, which would raise the correlation further.



(a) without CFG



(b) with CFG

Figure 4: **Correlation between ImageNet FID and T2I metrics including pixel-space methods.** Same setup as Fig. 1, but with three pixel-space methods, JiT (Li & He, 2025), PixNerd (Wang et al., 2025a), and PixelGen (Ma et al., 2026), added. (a) Without classifier-free guidance and (b) under the best CFG scale of each method over a timestep interval of $[0.0, 0.9]$.

A.3 IMAGENET-FID AND T2I METRICS CORRELATION WITHOUT CFG

Our main correlation uses ImageNet FID with CFG, which matches the T2I protocol. We repeat the analysis in Figure 5 with ImageNet FID evaluated without CFG. The Pearson r values change across metrics, but there is still no evidence of a strong correlation between ImageNet FID and the T2I metrics. This finding does not depend on the CFG protocol used on ImageNet.

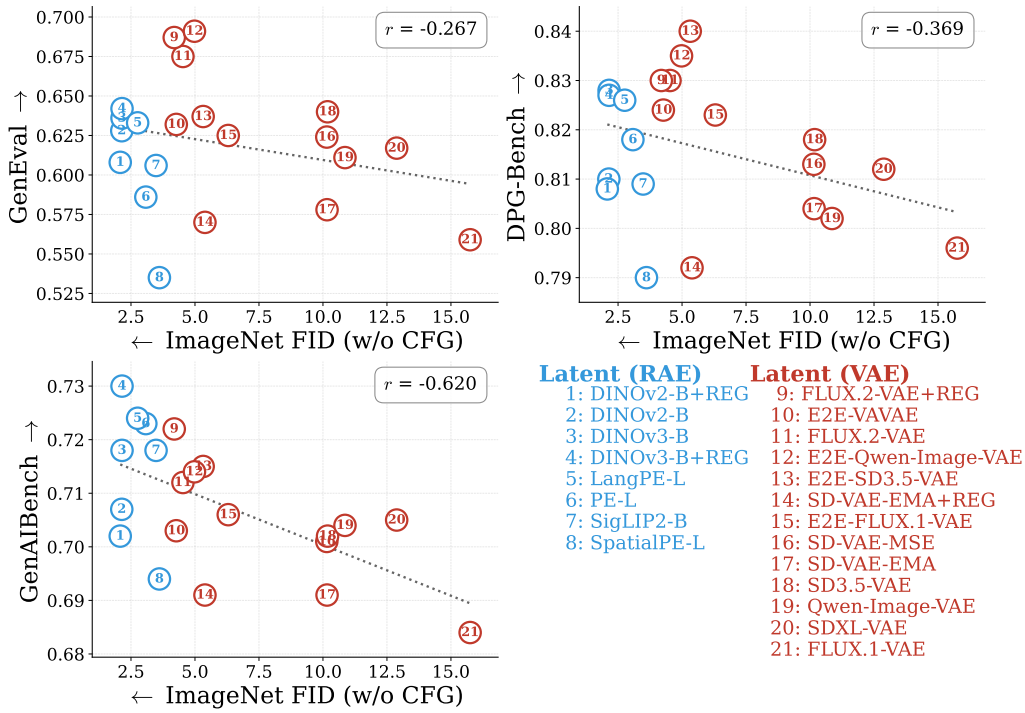


Figure 5: **Correlation between ImageNet FID and T2I metrics without CFG.** Same setup as Fig. 1, but with ImageNet evaluated without classifier-free guidance.

A.4 Additional Text-to-image qualitative samples

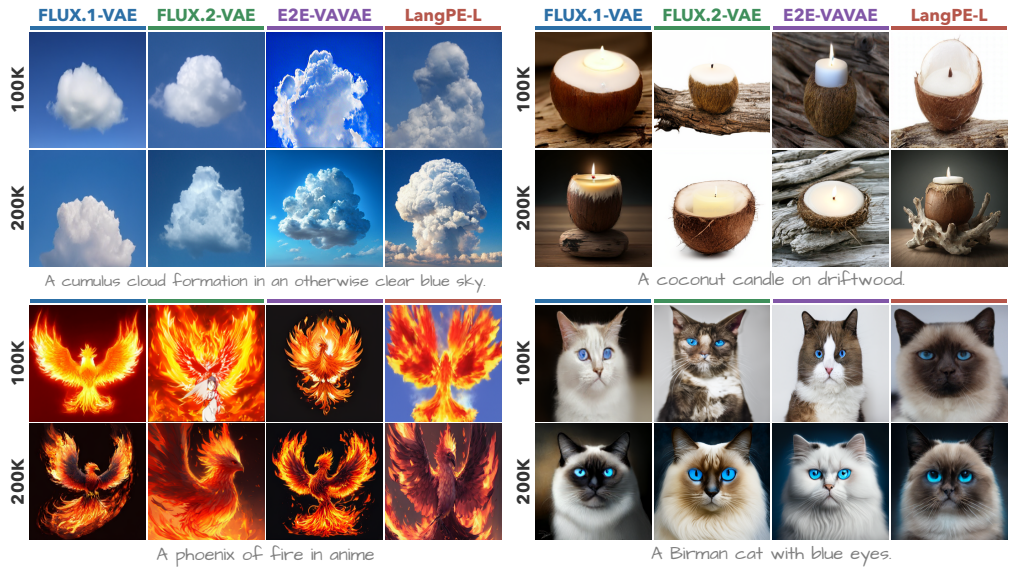


Figure 6: **Additional Text-to-image qualitative samples.** We keep the same setup as used in Fig. 3.

A.5 Additional related work

Diffusion and flow matching. Diffusion models (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2020) have become the dominant framework for visual generation. They corrupt data with Gaussian noise along a forward path $x_t = \alpha_t x_0 + \sigma_t \epsilon$ and train a network predicting the noise ϵ to reverse it. Flow matching (Lipman et al., 2023; Liu et al., 2023) instead regresses the velocity field along the same path. The prediction target (the noise ϵ, the clean signal x_0, or the velocity v) is interchangeable up to a time-dependent reweighting (Karras et al., 2022; Salimans & Ho, 2022; Lu & Song, 2025). We primarily adopt v-prediction (Sec. 2). At inference, samples are generated by integrating the ordinary differential equation (ODE) of the learned velocity field from noise to data.

Tokenizers and representation alignment. Latent diffusion models (Rombach et al., 2022) diffuse in the lower-dimensional latent space of a pretrained VAE (Kingma & Welling, 2013), which makes high-resolution training computationally tractable. Pretrained vision-encoder representations have been shown to substantially accelerate and improve diffusion training: representation alignment (REPA) (Yu et al., 2024) aims to align the internal features of DiTs with those of a frozen vision encoder; REPA-E (Leng et al., 2025) uses this alignment signal to additionally tune the VAE end to end. Representation autoencoders (RAE) (Zheng et al., 2025; Singh et al., 2026) merge their ideas and use a frozen vision encoder directly as tokenizer, such as DINOv2 (Oquab et al., 2024), DINOv3 (Siméoni et al., 2025), SigLIP2 (Tschannen et al., 2025), or the Perception Encoder (Bolya et al., 2025), so a single representation provides both compression and semantic structure. Recent analysis studies which properties of these encoders benefit generation most (Singh et al., 2025).