ABRA: SCALING DIFFUSION IMAGE TRAINING

Anonymous Author(s)
Affiliation Withheld

ABSTRACT

Compute-optimal scaling laws guide the training of frontier language models yet remain largely unexplored for visual generation. We present a systematic scaling law study for text-to-image diffusion models using ABRA, a controlled family of flow-matching transformers trained across three orders of magnitude worth of compute (10^{19} to 10^{22} FLOPs), reaching significantly larger compute budgets than previous works. We demonstrate that diffusion models scale just as predictably as language models but require far more data to train optimally: compute optimality occurs at approximately 200 image tokens per parameter, ten times the Chinchilla compute-optimal prescription for LLMs. We show that unlike language models, diffusion models are robust to overtraining and that practitioners should err on the side of more data rather than a larger model. Finally, we show that this predictability extends beyond training loss to generative quality metrics, optimal CFG settings, representation quality, and even the shape of the training curves, which collapse onto a universal form.

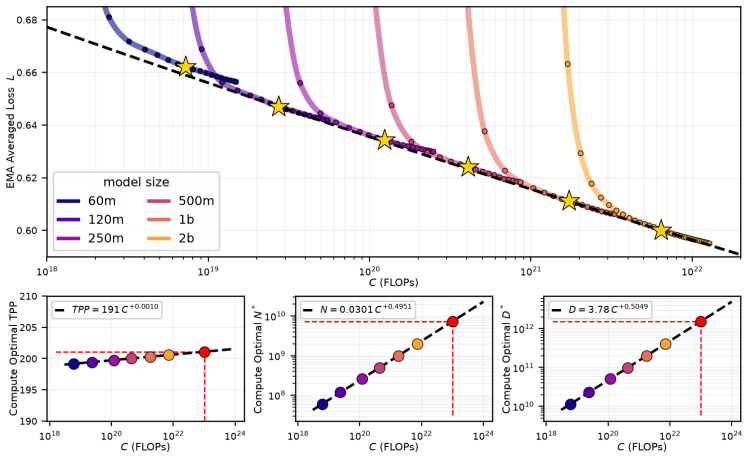

Figure 1: Text-to-image diffusion transformers follow predictable scaling laws and require ten times as much data as a comparably sized language model to reach compute optimality.

1 INTRODUCTION

Compute-optimal scaling laws describe how a fixed training compute budget C (in FLOPs) should be allocated between model size N (in parameters) and dataset size D (in training tokens) in order to minimize loss (Hestness et al., 2017; Amodei et al., 2016; Kaplan et al., 2020; Hoffmann et al., 2022). For large language models compute should be divided evenly between the model size and the dataset size according to $N \propto C^{1/2}$ and $D \propto C^{1/2}$, giving the "Chinchilla rule" of training LLMs to 20 tokens per parameter (TPP) (Hoffmann et al., 2022). Scaling analysis has become a standard tool in language modeling and is used to guide the training of frontier models (Touvron et al., 2023; OpenAI, 2023; DeepSeek-AI, 2024; Grattafiori et al., 2024; Sun et al., 2024; Kimi Team, 2025; The Microsoft AI Team, 2026).

The extension of scaling law analysis to visual generation presents challenges not present in language modeling. Visual content has higher dimensionality (i.e., 2D images), resolution dependent token information density, separate training and generation paradigms, and much noisier training curves.

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.

The scaling of these models remains understudied despite the release of many large-scale visual generation studies. Existing work by Liang et al. (2024) fits iso-FLOP profiles below 10^{19} FLOPs and extrapolates to 10^{21} FLOPs. We use an order of magnitude more compute and a more controlled family which allows us to derive much more precise empirical fits.

We significantly refine the scaling foundations for text-to-image models by conducting a systematic scaling study using ABRA, a controlled family of dense flow-matching transformers ranging from 60M to 2B parameters. We utilize μP to fit scaling laws spanning three orders of magnitude (from 10^{19} to 10^{22} FLOPs). Our central findings indicate that compute optimality is achieved at roughly 200 image tokens per parameter for text-to-image diffusion modeling, a $10\times$ increase over the standard Chinchilla rule for LLMs. Furthermore, we find that diffusion model training is robust to overtraining, giving practitioners confidence to err on the side of training a smaller model with more data.

We fit compute-optimal scaling laws to the training loss of the EMA weights and a range of common generative metrics, finding that all of these quantities scale predictably. We study the scaling of model representation quality using linear probing and find that representation quality also follows predictable scaling trends. We ablate the effect of resolution on our scaling laws and we demonstrate that like LLMs, diffusion models exhibit scaling collapse.

Our main contributions are:

- We present a comprehensive and controlled study of compute-optimal scaling for dense text-to-image transformers and derive an **actionable 200 TPP** rule for training diffusion models.
- We show that diffusion model training is **robust to overtraining** and precisely characterize this phenomenon.
- We extend our analysis to show that common generative metrics scale predictably, and that **representation and generation capabilities scale heterogeneously**.
- We show for the first time that diffusion models exhibit **scaling collapse**.

The remainder of the paper is organized as follows. We review related work on neural scaling laws and diffusion models in Section 2, describe our experimental setup and scaling ladder in Section 3, and present our empirical findings in Section 4. We include additional analysis in the appendix. Additional scaling metrics can be found in Appendices C and D.

2 Related Work

Text-to-Image Transformers and Flow-Matching. Peebles & Xie (2023) introduced and studied the efficacy of vision transformers in the context of generative diffusion modeling (Sohl-Dickstein et al., 2015). The original diffusion formulations (Ho et al., 2020; Song et al., 2021b;a) have largely been replaced by flow-matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023; Heitz et al., 2023) which formulates the denoising process as simple linear interpolation. We use v-prediction introduced by Albergo & Vanden-Eijnden (2023) and Ma et al. (2024) and the latent diffusion formulation (Rombach et al., 2022; Kingma & Welling, 2014).

Neural Scaling Laws. While the origins of neural scaling laws trace their roots to the nineties (see Caballero et al., 2023), their modern study can be largely attributed to the works (Amodei et al., 2016; Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022). These works find that many machine learning models, and large language models in particular, exhibit predictable improvements in performance as a function of compute and model size. While there is an extensive body of literature studying the scaling behaviors of language models, for example (Bergsma et al., 2025; 2026; Tao et al., 2024; Alabdulmohsin et al., 2022; Muennighoff et al., 2023; Dey et al., 2023), comparatively little is known about the scaling of diffusion models.

Scaling Laws for Diffusion Models. As previously mentioned, Liang et al. (2024) perform what is, to our knowledge, the only compute-optimal scaling law study for text-to-image diffusion modeling. Our work uses an order of magnitude more compute and a controlled model family to get refined estimates of the scaling laws for text-to-image diffusion models. Liang et al. (2024) find that model size should be scaled faster than dataset size, which leads to systematic undertraining at frontier scales. We find that data and parameters should be scaled at the same rate, i.e., $D \propto C^{1/2}$ and $N \propto C^{1/2}$.

A related line of work (Li et al., 2024b;a) studies the scalability of text-to-image models on downstream benchmarks, but does not perform a compute-optimal analysis. Mei et al. (2025) study latent diffusion under inference scaling rather than compute-optimal training; we corroborate their finding that optimal CFG shifts with model size. Yan et al. (2025) show that pixel-sequence generative models reach compute optimality at roughly 200–400 TPP. Ravishankar et al. (2025) study how diffusion depth-estimation performance scales with compute, but do not rigorously fit compute-optimal laws. For video models, Yin et al. (2025) fit scaling laws using models up to only 250M parameters. Finally, Zheng et al. (2025); Ryu & Han (2026) apply μP (Yang et al., 2021) to diffusion transformers for hyperparameter transfer.

Ge et al. (2026) is concurrent work on compute-optimal scaling for image and video generation. Despite large architectural differences, both works agree that diffusion training requires more data than LLM training. Our work differs from theirs in that we provide extensive analysis of generative metrics, a study of representation quality and resolution, and a demonstration of scaling collapse.

3 Method

We aim to characterize how text-to-image diffusion transformers should be scaled: for a given training budget, how many parameters should we use and how much data should we train on? This section describes our approach to answering these questions.

3.1 Compute Optimality

We follow the definition of compute optimality from Hoffmann et al. (2022). Given a fixed training budget of C FLOPs, we seek the model size N and dataset size D minimizing a loss or evaluation metric $\mathcal{L}$,

$$N_{\text{opt}}(C),\ D_{\text{opt}}(C) = \arg\min_{N,D}\ \mathcal{L}(N, D; C). \tag{1}$$

Throughout this paper, N denotes the total parameter count of the diffusion transformer, excluding the frozen text encoder. The embedding and unembedding layers represent a negligible fraction of parameters at our scales, so we report total parameters throughout. D denotes the number of image tokens seen during training, and we measure data in *image* tokens per parameter, abbreviated TPP. To locate this compute-optimal point accurately, we intentionally overtrain models in the Abra family so that the optimum $(N_{\text{opt}}, D_{\text{opt}})$ lands on the interior of our search range. We train up to 400 TPP, an upper bound informed by our prior experience training diffusion models and by the autoregressive pixel sequence results of Yan et al. (2025).

3.2 Training and Sampling

We train latent diffusion models with a v-prediction flow-matching objective (Ma et al., 2024; Salimans & Ho, 2022). Specifically, the network $v_\theta(x_t, t; \texttt{text})$ regresses the velocity of the linear interpolant between a clean image latent and Gaussian noise. We maintain an EMA of the model weights with decay parameter 0.9999 (Ho et al., 2020; Peebles & Xie, 2023). Because we have training steps roughly fixed we keep the EMA timescale constant across model sizes (Karras et al., 2024). We train on a base resolution of 512×512 with five image aspect-ratio buckets. Each image is encoded by the frozen $f = 8$ FLUX.2 VAE (Black Forest Labs, 2025) into a $64 \times 64 \times 32$ latent which is then reshaped using a 2×2 fold into a $32 \times 32 \times 128$ tensor that is consumed by the transformer. To make the unconditional estimate available at inference, we drop text conditioning with probability 0.1 during training.

We integrate the ODE $\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t, t; \texttt{text})$ backwards in time starting from $\varepsilon \sim \mathcal{N}(0, 1)$ at $t = 1$ with a 50-step Euler solver, sampling in the latent space and decoding to pixels with the FLUX.2 autoencoder (Black Forest Labs, 2025). We apply classifier-free guidance (Ho & Salimans, 2022) with guidance scale swept per model and held fixed across sampling steps.

3.3 Model Family

We construct Abra, a family of diffusion transformers from 60M to 2B parameters, with depth growing from 19 to 33 blocks and width from 384 to 1792, spanning roughly 10^{19} to 10^{22} image-token FLOPs. Our models combine aspects from various open source text-to-image models into a new family which is amenable to scaling (Black Forest Labs, 2025; Chen et al., 2024).

Each model uses four double-stream blocks followed by single-stream blocks for the remainder of the layers, with an FFN expansion rate of 6. Liang et al. (2024) found that a constant aspect ratio yields smooth iso-FLOP curves. For ABRA we use a monotonically non-decreasing width-to-depth ratio under the assumption that this smoothness persists under monotone aspect ratios. Text conditioning is provided by a frozen Qwen3-4B model (Yang et al., 2025) and we exclude the conditioner parameters from our FLOP and model size counts. Across the ladder we double the batch size every time we double the parameter count, keeping the number of training steps fixed. A full accounting of the ABRA architecture can be found in Appendix A.

3.4 CHOICE OF DATASET

We train on the DataComp-1B dataset (Gadre et al., 2023), a large-scale collection of web image-text pairs. Because web-sourced captions are often low quality, we re-caption the images using a variety of open-source vision-language models, following (Betker et al., 2023). During training we include a mix of dense, medium, sparse, and web-sourced captions. We do not perform mid or post training of our models and focus exclusively on diffusion model pre-training. We report metrics on either DataComp-1B or on a held-out evaluation mixture that is more diverse than the training data, of which DataComp-1B is a $\sim 20\%$ subset.

3.5 μP AND HYPERPARAMETER SELECTION

We transfer hyperparameters across the model family with μP (Yang et al., 2021), using the Complete-P parameterization (Dey et al., 2025), and train with the Adam optimizer (Kingma & Ba, 2015). We validate our implementation using spectral coordinate checking (Chickering et al., 2026). We validate μ-transfer for ABRA-60M, ABRA-120M, and ABRA-250M, and then zero-shot transfer the optimal learning rate, 4×10^{-4}, to the remaining models in the ladder. We sweep the initialization scale on the ABRA-60M model and initialize all layers from the same value of $\sigma = 0.01$. Unlike Mlodozeniec et al. (2025) and Jiang et al. (2026), we do not tune per-layer base hyperparameters. We use a logit-normal timestep sampler with mean $p_\mu = 1.9$ and standard deviation $p_\sigma = 1.0$.

4 EMPIRICAL RESULTS

Our central finding is that text-to-image flow-matching transformers scale predictably across a wide variety of evaluative metrics, but require roughly ten times as much data per parameter as LLMs to reach compute optimality. N denotes the parameter count of the vision transformer and D is measured in image tokens.

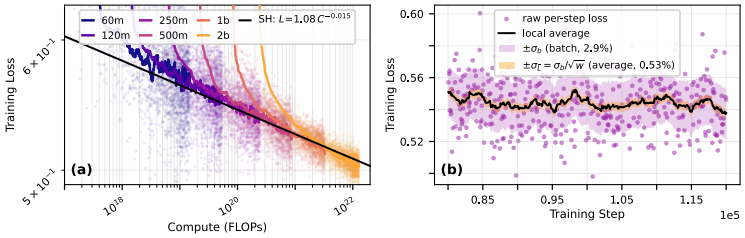

Figure 2: *Left:* The raw training loss for ABRA runs. The bold curves are a simple moving average of the scattered loss data. The solid black line is a qualitative fit of this data. *Right:* The noise decomposition for the ABRA-250M run, illustrating that the per-step differences in the loss curve are dominated by the batch-to-batch noise.

4.1 ESTIMATING COMPUTE-OPTIMAL TOKENS PER PARAMETER

Finding: Diffusion model training is compute-optimal at 200 TPP, roughly $10\times$ higher than the canonical LLM compute-optimal budget of 20 TPP.

Figure 2(a) shows the training loss for the ABRA family with a qualitative fit of the data. As we can see from Figure 2(b) the raw per-step loss is dominated by batch noise. Over 10,000 steps, the batch-to-batch noise on the ABRA-250M run is roughly 4 to $7\times$ larger than the improvement in the reducible loss over the same period. All fits in this section will therefore target the EMA model loss (**not** an EMA of the training loss as was done in (Liang et al., 2024)) which is substantially smoother than the raw model loss.

To construct these curves without training hundreds of models we exploit the constant LR schedule together with the smoothness of the EMA loss and interpolate each model's evaluation checkpoints into a continuous loss-versus-TPP trajectory using PCHIP interpolation (Fritsch & Carlson, 1980; Fritsch & Butland, 1984). The evaluation loss is computed on our evaluation dataset by taking 10,000 samples and averaging the loss over the logit-normal timestep distribution using the EMA weight predictions. At each TPP T we fit a power law in compute, $L(C;T) = A_T C^{-\alpha_T} + \beta_T$; fixed-$C$ slices of this surface give the iso-FLOP profiles, and the $\arg\min$ over T gives the compute-optimal TPP. The minima of these curves tightly cluster around 200 TPP at all but the smallest of our training budgets. The relative flatness of these profiles is also conspicuous and we investigate this point further in Section 4.2. We exclude ABRA-60M as a clear outlier (Figure 1), in line with the observation that scaling laws hold only in an appropriate regime (Hestness et al., 2017).

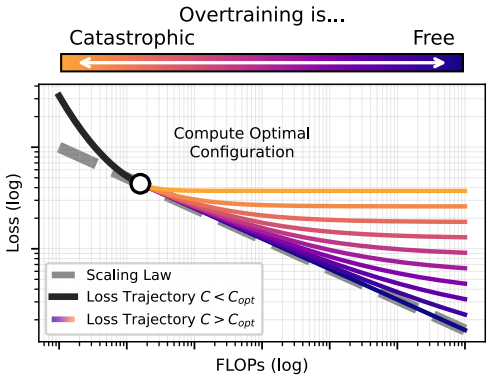

Figure 3: Overtraining compute-optimal models incurs a cost at fixed compute. The compute-optimal point occurs when the model loss curve is tangent to the scaling law curve.

There are a multitude of ways to fit scaling laws (see Li et al., 2025 for a survey) and to ensure our findings are reliable we conduct an in-depth comparison against many other fit procedures in Appendix G with our findings tabulated in Table 4. Across every fit procedure we tried, the optimal TPP lies within a narrow band of only about ± 17 TPP.

4.2 Diffusion Overtraining is Forgiving

Finding: Overtraining a diffusion model is nearly free (in the sense of Figure 3).

In practice, models are frequently trained well past their compute-optimal point. The most common reason is inference cost: a smaller model trained on more data is cheaper to serve, so practitioners deliberately overtrain to shift compute from deployment to training (Touvron et al., 2023; Gadre et al., 2025).

Therefore, scaling analysis cannot simply end at the compute-optimal point. It is important to understand precisely how we expect the model to behave when overtrained. Since the compute-optimal point is definitionally a minimum over the loss (1), any deviation from compute optimality necessarily incurs some cost in training FLOPs. This idea is illustrated visually in Figure 3. Concretely, fix a compute budget C and let $N_{\text{opt}}(C)$ denote the compute-optimal parameter count at that budget. For a model with $N \neq N_{\text{opt}}$ parameters trained at the same compute C (equivalently, at a non-optimal TPP), we can define the **loss penalty**

$$\Delta L(N, C) := L(N, C) - L\big(N_{\text{opt}}(C),\, C\big). \tag{2}$$

The quantity ΔL is the additional loss we incur by *not* training the compute-optimal model size for a budget of C FLOPs.

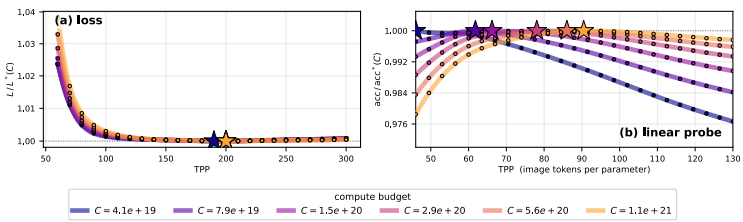

Figure 4: Normalized iso-FLOP curves. *(a)* EMA loss $L/L^*(C)$ and *(b)* linear-probe accuracy $\text{acc}/\text{acc}^*(C)$ versus TPP for several compute budgets, each normalized to its per-budget optimum (star). The near-flat curves show the compute-optimal TPP is largely budget-insensitive; representational quality (b) is optimal at a substantially lower TPP than the loss (a).

Figure 5 plots this loss penalty as a relative percentage, $\Delta L(N, C)/L\big(N_{\text{opt}}(C),\, C\big)$, as a function of TPP for the ABRA family. We additionally plot the same data for an open-source LLM scaling family, Gemstones (McLeish et al., 2025), which spans the parameter range 48M–2B and is heavily (and purposefully) overtrained. We provide additional comparisons against other overtrained LLM studies in Appendix F.

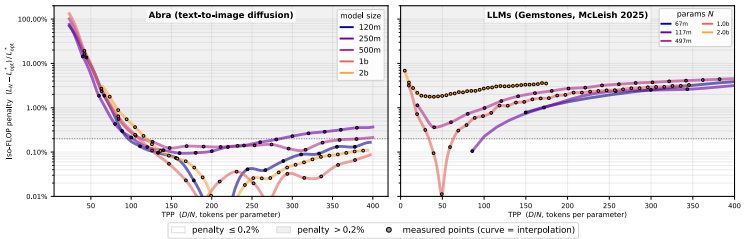

Figure 5: Relative iso-FLOP loss penalty $\Delta L/L^*$ (Eq. (2)) as a function of TPP for *(left)* the ABRA text-to-image diffusion family and *(right)* the Gemstones LLM suite (McLeish et al., 2025). ABRA markers denote evaluated checkpoints and Gemstones curves are the raw per-checkpoint fineweb-edu validation-loss penalties. The shaded bands mark the 0.2% penalty threshold (white below, grey above).

From Figure 5 it is clear that diffusion models suffer almost no loss penalty when $2\times$ overtrained (less than 0.5%). In the parlance of Figure 3, overtraining diffusion models is nearly "free". This is manifestly not the case for LLMs, where the loss penalty is substantial even for modest overtraining (see Figure 5, right).

On the other hand, Figure 5 also reveals that *undertraining* diffusion models is catastrophic. This observation yields a practical rule: in compute-limited settings it is far safer to err on the side of overtraining a small model than risk undertraining a large one.

4.3 GENERATIVE METRICS SCALE PREDICTABLY

Finding: Generative metrics follow power laws in compute but their compute-optimal allocations diverge.

Low diffusion loss does not necessarily indicate strong generative performance (Theis et al., 2016; Ho et al., 2020). Generation quality for diffusion models is therefore measured using distributional level metrics computed on the outputs of the ODE sampling procedure. Common metrics include FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), and CLIPScore (Hessel et al., 2021), and all of these metrics are sensitive to the classifier-free guidance (CFG) scale. Different guidance strengths favor different qualitative aspects of image generation. High guidance typically adheres better to prompts and low guidance typically adheres better to the target data distribution (Ho & Salimans, 2022; Saharia et al., 2022). Because generative quality depends so heavily on the choice of CFG it is insufficient to compare models at a fixed guidance scale. Instead, we must find the optimal guidance scale for every model on each metric.

To this end, we sweep the optimal guidance scale for each model size and every metric. We generate $50{,}000$ samples for FID and $25{,}000$ samples for all other metrics. Figure 6 (top) shows that every tested metric follows a power law of the form from Section 4.1. We fit $M(C) = A\,C^{-\alpha} + F_M$, with the sign of the power term flipped for CLIPScore. The offset F_M plays the role of the loss law's β_T and Table 8 collects the fitted coefficients.

Figure 6 (bottom) shows that compute-optimal allocation between D and N differs sharply across metrics. FID and KID require the dataset size to grow much more rapidly than the model size to reach compute optimality. On the other hand, CLIPScore and CMMD show decreasing optimal TPP as a function of compute, indicating that the number of parameters should grow more rapidly than the dataset size. We must therefore conclude that no single metric can define compute optimality, mirroring the situation where no single metric can determine visual quality (Stein et al., 2023; Jayasumana et al., 2024).

Finally, we find that the optimal CFG decreases with both model size and training steps. Mei et al. (2025) report the same trend for model size and for sampling steps, but did not investigate this trend over training horizons. In summary, stronger generative capacity needs less guidance and the feature

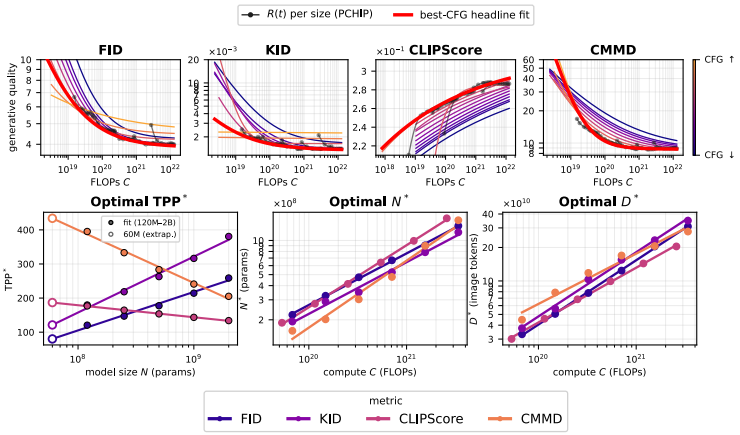

Figure 6: Standard generative metrics scale predictably. *Top:* generative quality vs training compute C for FID, KID, CLIPScore, and CMMD. The bold red curve is its compute-optimal frontier. *Bottom:* compute-optimal TPP^* as a function of model size N, and the compute-optimal N_{opt} and D_{opt} as functions of compute C.

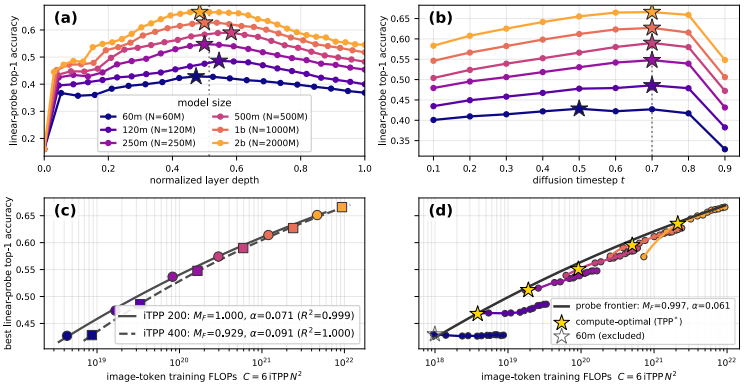

Figure 7: (a) Best top-1 accuracy as a function of relative layer depth. **(b)** Best top-1 accuracy as a function of diffusion timestep. **(c)** Top-1 accuracy follows predictable scaling laws at different TPP. **(d)** Top-1 accuracy as a function of training compute.

model behind each metric also shifts the optimal CFG. Appendix G.8 gives our full CFG analysis (Figures 20 and 22), including analyzing a DINOv2 (Oquab et al., 2024) family of metrics: FDD, KDD, and DMMD.

4.4 Linear Probe Accuracy

> **Finding:** Linear-probe accuracy follows predictable scaling laws. Understanding performance is compute-optimal at far lower TPP than generation.

Diffusion models are not only good visual generators, they are also good representation learners (Xiang et al., 2023). To investigate how the model's understanding performance scales as a function of compute we conduct linear probing experiments (Chen et al., 2020; Alain & Bengio, 2016; Xiang et al., 2023) on frozen activations from each Abra checkpoint and report ImageNet (Deng et al., 2009) top-1 accuracy. Following Xiang et al. (2023), we pool per-block features from a forward pass at a fixed diffusion timestep, and we sweep layers, timesteps, and training steps.

Figures 7(a) and (b) show that linear probing accuracy peaks for middle layers (consistent with (Chen et al., 2020; Yan et al., 2025; Xiang et al., 2023)) and for the moderately high diffusion timestep $t \approx 0.7$. Figure 7(c) demonstrates that linear-probe accuracy $a^*(C) = \max_{t,\ell} \mathrm{acc}(C; t, \ell)$ follows tight scaling laws at fixed TPP.

Additionally, we seek to understand the compute-optimal configuration for image understanding. Figure 4(b) applies the same iso-FLOP analysis of Section 4.1 to the per-model best probe accuracies $a^*(C)$ shown in Figure 7(d). The optimal TPP for image understanding is much lower than the TPP for generation in the compute range that we tested, but notably the optimal allocation is far from

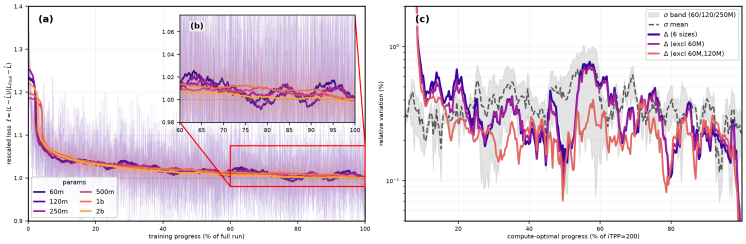

Figure 8: (a) Training loss curves for six models from the ABRA family rescaled with Equation (3). **(b)** The deviation Δ between models collapses beneath the noise floor σ for the three seed-replicated models (ABRA-60M, ABRA-120M, and ABRA-250M).

the balanced $D \propto N$ that we found for generation quality. For image understanding optimal TPP is steadily increasing, indicating that FLOPs are more efficiently used for increasing dataset size rather than model size. The finding that understanding is easier than generation at these scales is consistent with the works of Yan et al. (2025) and Xiang et al. (2023). We estimate that generation and understanding become jointly compute-optimal at roughly 5×10^{22} FLOPs and a 6B model; larger compute-optimal generators are no longer undertrained for understanding.

4.5 SCALING COLLAPSE

Finding: Compute optimal diffusion model training exhibits scaling collapse.

Conventional scaling law analysis fits power laws in quantitative metrics *after training* (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022). More recently Qiu et al. (2025) showed that the loss curves of compute-optimally trained language models follow a universal, non-power-law form: appropriately rescaled, the training curve for each model in the scaling ladder collapses onto a single trajectory (**scaling collapse**). Because the cross-size differences in rescaled loss are *smaller* than the seed-to-seed noise, practitioners can use collapse as a diagnostic to keep models near compute-optimal (Bergsma et al., 2026). To our knowledge this phenomenon has not been examined for diffusion models until now.

Following Qiu et al. (2025) we let p index our model family and consider the rescaled compute $C(p) = xt^*(p)$, where $x \in [0, 1]$ and $t^*(p)$ is the compute used to train the model p. Thus x represents the percentage of training that has finished. We then define the rescaled loss

$$\ell(x, p; \theta) = \frac{L(x\, t^*(p),\, p) - \hat{L}}{L(t^*(p),\, p) - \hat{L}}, \tag{3}$$

as well as the *collapse deviation* Δ and the per-model (relative) noise floor

$$\Delta(x) := \frac{\mathbb{V}_{p,\theta(p)}[\ell(x, p; \theta)]^{1/2}}{\mathbb{E}_{p,\theta(p)}[\ell(x, p; \theta)]}, \qquad \sigma(x; p) := \frac{\mathbb{V}_{\theta(p)}[\mathcal{L}(xt^*(p), p, \theta(p))]^{1/2}}{\mathbb{E}_{\theta(p)}[\mathcal{L}(xt^*(p), p, \theta(p))]},$$

where $\mathcal{L}$ is the shifted loss $\mathcal{L} = L - \hat{L}$ and where $\hat{L}$ is the irreducible loss for the model family, found via our fit procedure. Rescaling in this manner fixes the endpoints of the training curves to match so that the shape of the training curves depends only on relative compute. We train the ABRA-60M, ABRA-120M, and ABRA-250M models with four different seeds, which change both the initialization and the data ordering. We cut off the data window at a fixed 200 TPP for all of the models based on the analysis above.

Figure 8 shows the result of our experiment. We plot the raw-loss curves with an overlaid SMA (left) and the raw Δ, σ values with their own SMA (right); the moving average is required in the context of diffusion since the training loss curves, and thus Δ and σ, are extremely noisy. Because we only ran multiple seeds up to 250M parameters, we can only determine the collapse for these three models. Despite the noise, we see a clear collapse in training loss to the noise floor within the first 15% of training. For the remainder of training, the cross-size deviation Δ remains at or below the noise floor σ, indicating that the ABRA family satisfies the criteria for scaling collapse and is therefore compute-optimal (or nearly compute-optimal).

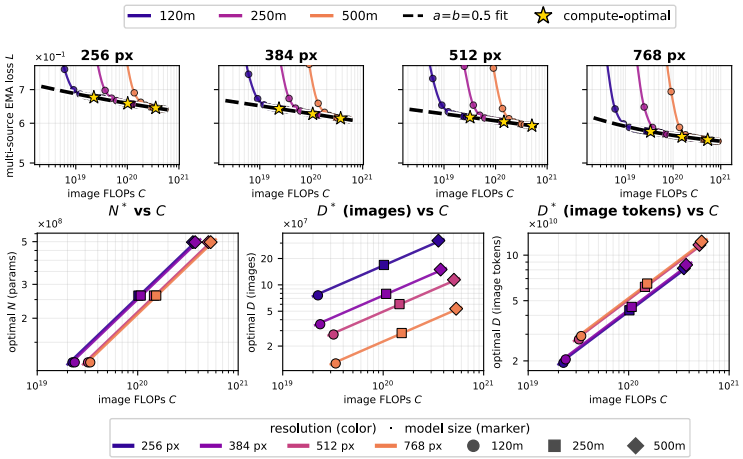

Figure 9: (Top Row) Training curves and loss scaling laws at different resolutions. **(Bottom Row)** compute-optimal model size, compute-optimal *number of images*, compute-optimal dataset size in number of image tokens.

4.6 The Effect of Resolution

Finding: Higher resolution training requires more image tokens to reach compute optimality.

The slope of a scaling law is understood to be a property dependent more on the data distribution than the model architecture (Hestness et al., 2017; Bahri et al., 2024; Kaplan et al., 2020). The data distribution for images changes quite dramatically when we change resolution: the information per patch goes down and we rescale the timestep sampling distribution according to Hoogeboom et al. (2023). Therefore the natural hypothesis is that as image resolution increases we would expect that the diffusion models will require an increasing number of image tokens to reach compute optimality.

We show that this hypothesis is correct. We train the Abra-120M through Abra-500M ladder at 256, 384, 512, and 768 pixel resolutions. For each model we use the Hoogeboom et al. (2023) SNR rescaling on our base distribution. Models are evaluated using the log-normally distributed timesteps averaged over 25,000 samples. For all resolutions we continue to see scaling behavior as a function of compute.

Because our largest models are only 500M, getting clean scaling law fits is challenging. To this end, and motivated by our analysis above, we assume the balanced compute optimality constraint $a = b = 0.5$. Under this assumption we find that the compute-optimal TPP systematically increases as a function of resolution from about 165 at 256×256 resolution to around 247 at 768×768 resolution. While the required image tokens per parameter increases, the number of images required to realize the larger data consumption actually goes down: we need fewer *images* per-parameter as we increase resolution. This is consistent with the findings of Yan et al. (2025) in the context of pixel-autoregressive generation models.

5 Conclusion

We revisited compute-optimal scaling for text-to-image diffusion pre-training with $10\times$ as much compute as previous studies and a controlled model family, Abra, which spans 60M to 2B parameters. Text-to-image diffusion transformers follow tight, predictable scaling laws with compute optimality achieved at approximately 200 TPP, an order of magnitude higher than for LLMs.

We further showed that diffusion training is robust to overtraining, that generative metrics like FID and CLIPScore scale predictably, that higher resolution requires more tokens to reach optimality, that representation quality scales differently from generation, and—for the first time—that compute-optimal diffusion training exhibits scaling collapse.

ACKNOWLEDGMENTS

We thank Firstname Lastname and Firstname Lastname for helpful discussion throughout the completion of this work.

AI USE STATEMENT

We used AI tools to assist in editing our manuscript, with partial drafting work, for finding related works in the literature, as a coding assistant for writing code and analyzing data. We did not use LLMs to generate data. AI-assisted work and code has been reviewed by the authors.

REFERENCES

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting Neural Scaling Laws in Language and Vision. In *NeurIPS*, 2022.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *arXiv:1610.01644*, 2016.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *ICML*, 2016.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power Lines: Scaling Laws for Weight Decay and Batch Size in LLM Pre-training. In *NeurIPS*, 2025.

Shane Bergsma, Bin Claire Zhang, Nolan Dey, Shaheer Muhammad, Gurpreet Gosal, and Joel Hestness. Scaling with Collapse: Efficient and Predictable Training of LLM Families. In *ICLR*, 2026.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla Scaling: A replication attempt. In *arXiv:2404.10102*, 2024.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. OpenAI technical report. Accessed: 2026-07-22.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://github.com/black-forest-labs/flux2, 2025. Accessed: 2026-07-22.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken Neural Scaling Laws. In *ICLR*, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

Kyle R. Chickering. The spectral maximal update parameterization in theory and practice. https://kyrochi.github.io/blog/spectral-mup/index.html, 2026. Accessed: 2026-07-22.

Kyle R Chickering, Huijuan Wang, Mengxi Wu, Alexander Moreno, Muhao Chen, Xuezhe Ma, Daria Soboleva, Joel Hestness, Zhengzhong Liu, and Eric Xing. GQA-μP: The maximal parameterization update for grouped query attention. In *arXiv:2605.15290*, 2026.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. In *JMLR*, 2023.

DeepSeek-AI. DeepSeek LLM: Scaling open-source language models with longtermism. In *arXiv:2401.02954*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Nolan Dey, Gurpreet Gosal, Zhiming Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster. In *arXiv:2304.03208*, 2023.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: CompleteP enables compute-efficient deep transformers. In *NeurIPS*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

F. N. Fritsch and J. Butland. A method for constructing local monotone piecewise cubic interpolants. *SIAM Journal on Scientific and Statistical Computing*, 5(2):300–304, 1984. doi: 10.1137/0905021.

F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980. doi: 10.1137/0717021.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Alexandros G. Dimakis, Gabriel Ilharco, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *ICLR*, 2025.

Chongjian Ge, Hanwen Jiang, Tianyu Wang, Jiuxiang Gu, Yiran Xu, Ziwen Chen, Shaoteng Liu, Jing Shi, Yicong Hong, Zefan Cai, Hailin Jin, and Hao Tan. Chimera: Designing and Chinchilla-Scaling Hybrid Visual Diffusion Transformers. In *arXiv:2607.28611*, 2026.

Aaron Grattafiori et al. The Llama 3 herd of models. In *arXiv:2407.21783*, 2024.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned Stochastic Tensor Optimization. In *ICML*, 2018.

Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α-(de)blending: A minimalist deterministic diffusion model. In *SIGGRAPH*, 2023.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of EMNLP*, 2020.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically. In *arXiv:1712.00409*, 2017.

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. In *NeurIPS*, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *NeurIPS*, 2022.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, 2023.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *CVPR*, 2024.

Tianze Jiang, Blake Bordelon, Cengiz Pehlevan, and Boris Hanin. Hyperparameter transfer with mixture-of-expert layers. In *ICML*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024. Accessed: 2026-07-22.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. In *arXiv:2001.08361*, 2020.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024.

Kimi Team. Kimi k2: Open agentic intelligence. In *arXiv:2507.20534*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Hao Li, Shamit Lal, Zhiheng Li, Yusheng Xie, Ying Wang, Yang Zou, Orchid Majumder, R. Manmatha, Zhuowen Tu, Stefano Ermon, Stefano Soatto, and Ashwin Swaminathan. Efficient scaling of diffusion transformers for text-to-image generation. In *arXiv:2412.12391*, 2024a.

Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R. Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the Scalability of Diffusion-based Text-to-Image Generation. In *CVPR*, 2024b.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (Mis)Fitting: A Survey of Scaling Laws. In *ICLR*, 2025.

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. In *arXiv:2410.08184*, 2024.

Tomasz Limisiewicz, Artidoro Pagnoni, Srini Iyer, Mike Lewis, Sachin Mehta, Alisa Liu, Margaret Li, Gargi Ghosh, and Luke Zettlemoyer. Compute Optimal Tokenization. In *arXiv:2605.01188*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

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 *ECCV*, 2024.

Sean McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum, Ashwinee Panda, and Tom Goldstein. Gemstones: A model suite for multi-faceted scaling laws. In *NeurIPS*, 2025.

Kangfu Mei, Zhengzhong Tu, Mauricio Delbracio, Hossein Talebi, Vishal M. Patel, and Peyman Milanfar. Bigger is not Always Better: Scaling Properties of Latent Diffusion Models. In *TMLR*, 2025.

Bruno Mlodozeniec, Pierre Ablin, Louis Béthune, Dan Busbridge, Michal Klein, Jason Ramapuram, and Marco Cuturi. Completed hyperparameter transfer across modules, width, depth, batch and duration. In *arXiv:2512.22382*, 2025.

MMSegmentation Contributors. MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling Data-Constrained Language Models. In *NeurIPS*, 2023.

NVIDIA. Nemotron-H: A family of accurate and efficient hybrid Mamba-transformer models. In *arXiv:2504.03624*, 2025.

OpenAI. GPT-4 technical report. In *arXiv:2303.08774*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. In *TMLR*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving Discrepancies in Compute-Optimal Scaling of Language Models. In *NeurIPS*, 2024.

Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. Diffusion beats autoregressive in data-constrained settings. In *arXiv:2507.15857*, 2025.

Shikai Qiu, Lechao Xiao, Andrew Gordon Wilson, Jeffrey Pennington, and Atish Agarwala. Scaling collapse reveals universal dynamics in compute-optimally trained neural networks. In *ICML*, 2025.

Rahul Ravishankar, Zeeshan Patel, Jathushan Rajasegaran, and Jitendra Malik. Scaling properties of diffusion models for perceptual tasks. In *CVPR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Simo Ryu and Chunghwan Han. Summer-22B: A Systematic Approach to Dataset Engineering and Training at Scale for Video Foundation Model. In *arXiv:2603.00173*, 2026.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. In *arXiv:2508.10104*, 2025.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *NeurIPS*, 2023.

Xingwu Sun et al. Hunyuan-Large: An open-source MoE model with 52 billion activated parameters by Tencent. In *arXiv:2411.02265*, 2024.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *NeurIPS*, 2024.

The Microsoft AI Team. MAI-Thinking-1: Building a hill-climbing machine. https://microsoft.ai/pdf/mai-thinking-1.pdf, 2026. Accessed: 2026-07-22.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *ICLR*, 2016.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. In *arXiv:2302.13971*, 2023.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing shampoo using adam. In *ICLR*, 2025.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *ICCV*, 2023.

Xinchen Yan, Chen Liang, Lijun Yu, Adams Wei Yu, Yifeng Lu, and Quoc V. Le. Rethinking generative image pretraining: How far are we from scaling up next-pixel prediction? In *arXiv:2511.08704*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, et al. Qwen3 technical report. In *arXiv:2505.09388*, 2025.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *NeurIPS*, 2021.

Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, Ke Lin, Jiarong Ou, Rui Chen, Victor Shea-Jay Huang, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, Wentao Zhang, and Kun Gai. Towards Precise Scaling Laws for Video Diffusion Transformers. In *CVPR*, 2025.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. In *TMLR*, 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.

Chenyu Zheng, Xinyu Zhang, Rongzhen Wang, Wei Huang, Zhi Tian, Weilin Huang, Jun Zhu, and Chongxuan Li. Scaling Diffusion Transformers Efficiently via μP. In *NeurIPS*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.