How far can we go with ImageNet for Text-to-Image generation?

Anonymous authors
Paper under double-blind review

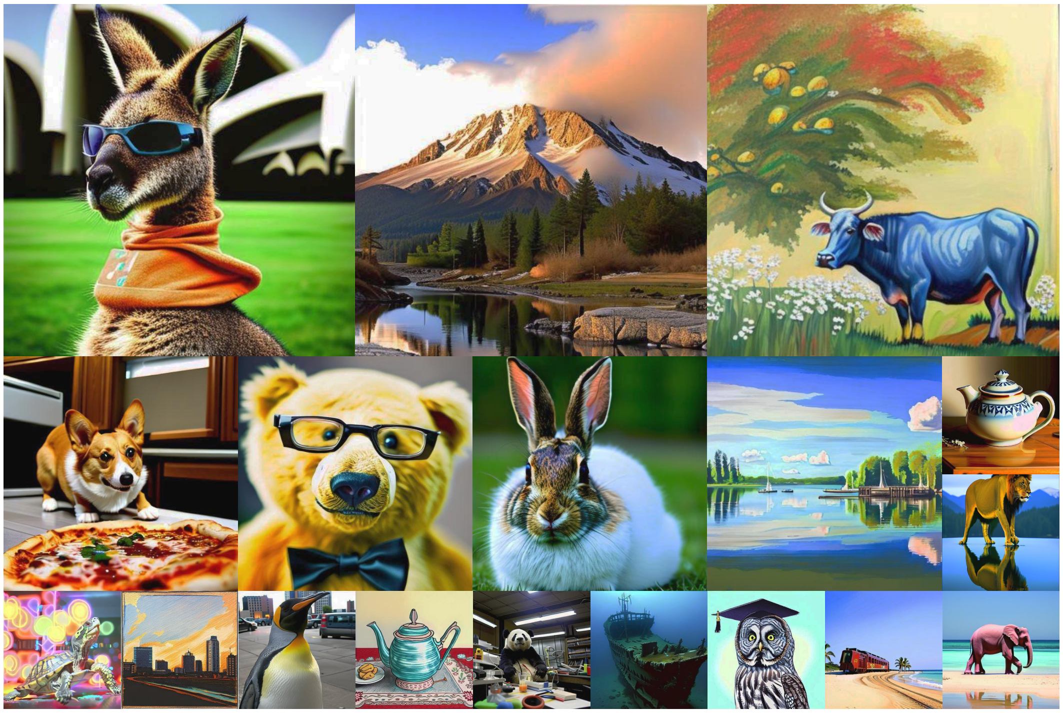


Figure 1: **Images** generated by our 400M parameters text-to-image model trained solely on ImageNet. Text prompts are taken from `PartiPrompts` Yu et al. (2022).

Abstract

Recent text-to-image (T2I) generation models have achieved remarkable sucess by training on billion-scale datasets, following a 'bigger is better' paradigm that prioritizes data quantity over availability (closed vs open source) and reproducibility (data decay vs established collections). We challenge this established paradigm by demonstrating that one can achieve capabilities of models trained on massive web-scraped collections, using only ImageNet enhanced with well-designed text and image augmentations. With this much simpler setup, we achieve a +6% overall score over SD-XL on GenEval and +5% on DPGBench while using just *1/10th the parameters and 1/1000th the training images*. We also show that ImageNet pre-trained models can be fine-tuned on task specific datasets (like for high resolution aesthetic applications) with good results, indicating that ImageNet is sufficient for acquiring general capabilities. This opens the way for more reproducible research as ImageNet is widely available and the proposed standardized training setup only requires 500 hours of H100 to train a text-to-image model.

1 Introduction

The prevailing wisdom in text-to-image (T2I) generation holds that larger training datasets inevitably lead to better performance. This "bigger is better" paradigm has driven the field to billion-scale image-text paired datasets like LAION-5B (Schuhmann et al., 2022), DataComp-12.8B (Gadre et al.,

2023) or ALIGN-6.6B (Pham et al., 2023). While this massive scale is often justified as necessary to capture the full text-image distribution, in this work, we challenge this assumption and argue that data quantity overlooks fundamental questions of data efficiency and quality in model training.

Our critique of the data-scaling paradigm comes from a critical observation: current training sets are either closed-source or rapidly decaying which makes results impossible to fully reproduce, let alone compare fairly. As such, the community of T2I generation is in dire need of a standardized training setup to foster open and reproducible research. Luckily, Computer Vision already has such dataset in ImageNet (Russakovsky et al., 2015) that has been the gold standard in many tasks for many years. It is widely available and its strength and limitations are well known. Furthermore, it is heavily used in class-conditional image generation (Peebles & Xie, 2023; Jabri et al., 2023), which makes its evaluation metrics more familiar. This begs the question of *how far can we go with ImageNet for text-to-image generation?*

Our findings are that we can indeed get a surprisingly competitive model by training solely on ImageNet. As shown on Figure 1, we can achieve excellent visual quality. Additionally we also achieve very competitive scores on common benchmarks such as GenEval (Ghosh et al., 2024) and DPGBench (Hu et al., 2024), matching or even surpassing popular models that are trained on much more data and at a far greater cost, such as SDXL (Podell et al., 2023), Pixart-α (Chen et al., 2023) and Pixart-Σ (Chen et al., 2024) (see Figure 2). However, this does not come without any hurdles. In this paper we analyze the challenges of training using ImageNet only and propose successful strategies to overcome them. Our strategies allow us to train models of smaller size (about 300M-400M parameters) on a reasonable compute budget (about 500 H100 hours) making it accessible to more research teams. We further show these models can be successfully fine-tuned for specific tasks, namely high-resolution aesthetic image generation.

Our contributions are thus the following:

- We analyze the shortcomings of training T2I diffusion models on ImageNet and propose mitigation strategies.
- Then, we propose a standardized training setup using only images from ImageNet, providing accessible and reproducible research for T2I generation.
- We provide several models in the 300M-400M parameters range generating high quality images and outperforming competing models that are 10 times the size and trained on 1000 times more data.
- We show that models trained with ImageNet act as strong pretraining checkpoints for task-specific fine-tuning (like high-resolution aesthetic image generation)

To commit to open and reproducible science, all our training data are hosted at `https://huggingface.co/datasets/anonymized_for_review` and all our code and models are hosted at `https://github.com/anonymized_for_review`.

In the next section, we outline the challenges in using ImageNet for T2I generation and evaluate mitigation strategies for each of them. We then combine them in a complete training recipe and use it to train several models that we compare against the state of the art. We further show that such a model can be further fine-tuned on a task-specific dataset for higher aesthetic quality. Finally, we discuss the related work before we conclude.

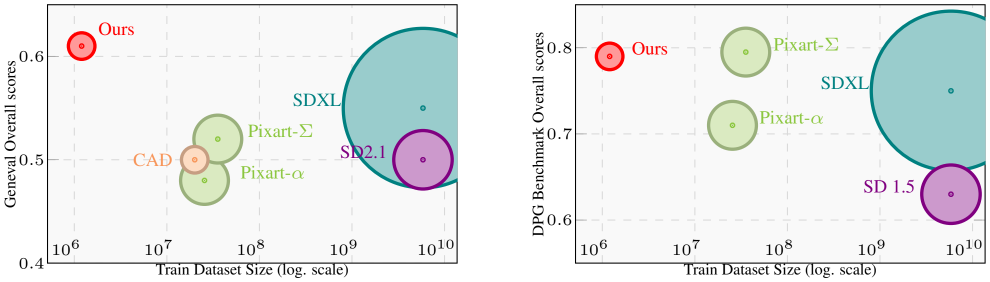


Figure 2: **Quantitative results** on GenEval (left) and DPGBench (right). The size of the bubble represents the number of parameters. In both cases, we outperform models of 10× the parameters and trained on 1000× the number of images.

Model	TA	FID Inc.↓		Prec.↑	Rec.↑	Den.↑	Cov.↑	Jina CLIP↑		GenEval↑
		IN-Val	COCO					IN-Val	COCO	
DiT-I	✗	20.14	71.00	0.67	0.29	0.71	0.39	31.21	22.42	0.11
	✓	6.29	45.71	0.77	0.76	0.82	0.72	38.45	38.39	0.55
CAD-I	✗	84.77	46.35	0.75	0.05	1.40	0.10	20.55	14.06	0.17
	✓	6.16	49.89	0.80	0.72	0.89	0.76	38.01	37.85	0.55

Table 1: **Image quality and compostionality** of AIO models (✗) and TA models (✓). FID reported is FID `Inception v3`. Precision, Recall, Density and Coverage are computed using `DINOv2` features on ImageNet Val . Additional values on COCO test set are reported in Table 9.

2 Adopting ImageNet for Text-to-Image generation

We focus on training text-to-image models using ImageNet, a small, open-source and widely accepted data collection. We first discuss the evaluation criterions and then gradually pinpoint the major limitations in setting up a T2I diffusion model using ImageNet. To overcome these limitations, we show that well-crafted augmentations can bring forth a compositionally accurate T2I model while training under hard data constraints. For our analysis, we leverage two architectures: (1) **DiT-I** (our adaptation of DiT (Peebles & Xie, 2023) to handle text) and (2) **CAD-I** (Dufour et al., 2024a). The suffix "I" is added to indicate the model is trained only on ImageNet.

Evaluation: *Image-quality*. We specifically assess the generation quality of both in-distribution (w.r.t Imagenet-50k validation set) and zero-shot (MSCOCO-30k captions validation set (Lin et al., 2014)). Specifically, we adopt: **(1) FID** (Heusel et al., 2017) using both standard `Inception-v3` and `Dino-v2` backbones, (2) **Precision** (Kynkäänniemi et al., 2019), (3) **Recall** (Kynkäänniemi et al., 2019), (4) **Density** (Naeem et al., 2020), and (5) **Coverage** (Naeem et al., 2020). These are all calculated with `Dino-v2` features.

Evaluation: *Compositionality*. To understand the text-image alignment capabilities and image composition prowess, we adopt (1) **CLIPScore** (Hessel et al., 2021) and (2) **Jina-CLIP Score** (Koukounas et al., 2024) on both MSCOCO-30k and ImageNet validation set, (3) **GenEval** (Ghosh et al., 2024) and (4) **DPGBench** (Hu et al., 2024).
Based on these evaluation strategies that assess both image quality and compositional accuracy, we can systematically identify the key limitations of training T2I diffusion models on ImageNet. Our analysis reveals several ImageNet-specific challenges that must be addressed to achieve high-performance generation with limited data.

2.1 Text challenges

Challenge: *Absence of captions*. Class-conditional models trained with ImageNet have shown exceptional generation capabilities (Peebles & Xie, 2023; Ma et al., 2024) However, extending this use to T2I generation is difficult since ImageNet, being a classification dataset, lacks any sort of caption corresponding to its images. To adopt ImageNet for T2I generation, similar to prior works (Radford et al., 2021), one could build captions by a very simple strategy of `'An image of <class name>'` (denoted AIO). However, this results in very poor generation capabilities as shown in Table 1. This can be mainly attributed to the two major shortcomings of AIO captions for ImageNet: First, AIO captions lack vocabulary. They contain only roughly a thousand words corresponding to the concepts of the classes and thus lack attributes, spatial relations, etc. This constraint on the diversity in the text-condition space leads to a clear bottleneck in text understanding. Second, there is often more content in the image than just the class. For example, a caption "an image of *golden retriever*" mentions the class name but leaves out details and concepts that could be in the background. This lack of details leads to spurious correlation where the model can learn to associate unrelated visual pattern (e.g., grass texture) to the class name (e.g., *golden retriever*) because the text for this concept is never mentioned in the text space. Finally, despite the presence of humans in the images, ImageNet does not contain a '*person*' class, resulting in humans not being represented in the AIO text space. This issue extends to many categories (*road*, *water*, etc), as ImageNet is an object-centric dataset.

Solution: *Long informative captions.* To overcome this challenge, we employ a synthetic captioner (Liu et al., 2024b) (TA for *Text Augmentation*) to generate comprehensive captions that capture: (i) *Scene composition* and *spatial relationships*; (ii) *Background elements* and *environmental context*; (iii) *Secondary objects* and *participants*; (iv) *Visual attributes* (color, size, texture); and (v) *Actions* and *interactions* between elements.

We compare the gains attributed to long captions both quantitatively (Table 1) as well as qualitatively (Figure 4: row 1-2). For ImageNet-Val set, we observe that models trained with long captions significantly improves performance, resulting in lower FIDs of **6.29** for DiT-I and **6.16** for CAD-I in contrast to **20.14** for DiT-I and **85** for CAD-I on AIO captions. As a point of reference, we remind the reader that models of this size (below 0.5B parameters) typically have an FID of 9 using the class-conditional setup (Peebles & Xie, 2023). Additional evidence of image quality improvement is found with the Precision (P), Recall (R), Density (D), and Coverage (C) metrics. DiT-I achieves better performance with text-augmentation over AIO on all four of the P,R,D,C metrics, whereas, CAD-I shows improvement in three out of four of these metrics, indicating that TA is quite vital for transforming ImageNet into a T2I specific dataset. For COCO test set – which is a zero-shot task for our training, this trend is all the more dramatic. The TA models are the only ones able to correctly follow the prompt as attested by the much improved Jina CLIP score (DiT-I from **22.42** to **38.45**; CAD-I from **14.06** to **37.85**), while keeping similar image quality.

Regarding text-image alignment and compositionality, models trained with longer captions benefit from the added information, evidenced by the improvement in GenEval overall score from (DiT-I from **0.11** to **0.55**; CAD-I from **0.17** to **0.55**) (see Table 1, last column).

2.2 IMAGE CHALLENGES

Using long, informative captions (TA) significantly enhances both generation quality and compositional alignment. But training text-to-image diffusion models on ImageNet only still faces two critical limitations: early overfitting and poor compositional generalization.

Limitation: *Early overfitting.* Models trained on ImageNet with long captions (TA) demonstrate promising initial performance. However, due to the relatively small scale of ImageNet (only 1.2 million images) they begin overfitting at approximately 200k training steps (see Figure 3).

Limitation: *Restricted Complex Compositionality Abilities.* ImageNet's object-centric nature presents a challenge for learning complex compositions. Even with enhanced textual descriptions via TA captioning, models still struggle with spatial relationships, attribute binding, and multi-object compositions. This limitation manifests in lower GenEval scores for compositional prompts involving multiple objects, color attribution, and positional relationships, as shown in Figure 8 (Column 1 and 2) and Table 2.

Solution: *Image Augmentation (IA).* To reduce overfitting and improve compositional reasoning, we investigate the use of image augmentations during training. We experiment with two augmentation strategies. Details about implementation and training pipeline are given in Appendices E and F.

- **CutMix** (Yun et al., 2019): For each image in the dataset, we randomly select an image from a different class and overlay a smaller version of it onto the original image. A caption is

Model	IA	Overall↑	One obj.↑	Two obj.↑	Count.↑	Col.↑	Pos.↑	Col. attr.↑
DiT-I	✖	0.55	0.95	0.61	0.36	0.80	0.28	0.33
	Crop	0.54	0.96	0.56	0.38	0.79	0.22	0.33
	CutMix	0.58	0.95	0.67	0.43	0.80	0.30	0.35
CAD-I	✖	0.55	0.97	0.60	0.42	0.74	0.26	0.35
	Crop	0.54	0.96	0.61	0.40	0.71	0.23	0.33
	CutMix	0.57	0.94	0.68	0.40	0.70	0.35	0.36

Table 2: **GenEval scores** of TA and TA + IA models. All models are trained with long captions. A Prompt Extender was used before generating images. Models are evaluated at 256^2 resolution.

Model	IA	FID Inc.↓		Prec.↑	Rec.↑	Den.↑	Cov.↑	Jina CLIP↑	
		IN-Val	COCO					IN-Val	COCO
DiT-I	✖	6.29	45.71	0.77	0.76	0.82	0.72	38.45	38.39
	Crop	6.20	44.04	0.77	0.75	0.83	0.74	38.45	38.39
	CutMix	7.30	49.12	0.79	0.74	0.88	0.75	38.77	36.80
CAD-I	✖	6.16	49.89	0.80	0.72	0.89	0.76	38.01	37.85
	CutMix	6.62	49.31	0.80	0.70	0.90	0.76	38.17	37.71

Table 3: **Image quality** of TA models and TA + IA models. All models are trained with long captions. FID reported is FID `Inception v3`. Precision, Recall, Density and Coverage are computed using `DINOv2` features on ImageNet Val. Values on COCO test set are reported in Table 9.

generated using the CutMix image as input. This technique introduces additional variability in the training data.

- **Crop** : During training, we randomly mask a portion of the image tokens such that the model is exposed only to a local crop of the original image. We add crop coordinates tokens to the captions of the image. This augmentation encourages the model to decouple object features from their background context, and to learn correspondences between partial visual elements and specific text tokens.

In Figure 3, we plot the evolution of FID and GenEval scores over training steps for the CAD-I architecture. Training with TA alone leads to early overfitting: we observe a sharp rise in FID after 200k steps. In contrast, models trained with TA+CutMix or TA+Crop maintain significantly lower FID curves for longer, with a delayed onset of overfitting.

Table 2 assesses the impact of image augmentation on GenEval metrics. Image augmentation (both CutMix and Crop) leads to a notable improvement in the GenEval overall score, with an increase of 2 points. Notably, the Two Objects sub-task sees a +6 point increase for both architectures, CAD-I sees a +9 point gain in Position, and DiT-I gains +2 points in Color Attribution.

These improvements in compositional metrics are achieved while maintaining or improving image quality, as measured by FID scores and Prec., Rec., Den., and Cov. in Table 3. The qualitative examples in Figure 4 further demonstrate the enhanced compositional capabilities, with models trained using augmentation techniques producing more accurate representations of complex prompts. This improvement is particularly evident in the teddy bear scene: while the TA model generates a teddy bear awkwardly positioned on a motor bike, the TA + IA model creates a more natural composition with the teddy bear appropriately driving the motor bike. Similarly, the "goat on top of a mountain" example shows more refined details and aesthetically pleasing composition with TA + IA, whereas the TA model struggles with the scene's layout. Additional examples are given in Figure 8.

2.3 Scaling to higher resolution

All experiments discussed thus far were conducted at a resolution of 256^2. We now explore whether higher resolution generation such as 512^2 is feasible under the same data constraints without requiring additional supervision or sacrificing performance.

Starting from a DiT-I checkpoint at 256^2 resolution, trained with TA + IA for $250k$ steps, we further train it at 512^2 resolution for an additional $50k$ steps on the same data. Both the pretraining and the fine-tuning use the CutMix image augmentation strategy for simplicity. This fine-tuning procedure requires no changes to the text encoder or transformer backbone, aside from adjusting the image tokenization to handle the larger input size. Images generated at 512^2 are shown in Figure 6.

Table 4 summarizes the results on the GenEval benchmark. Further training at higher resolutions preserves the model's compositional capabilities while improving performance, with the overall score increasing from 0.58 at 256^2 to 0.61 at 512^2.

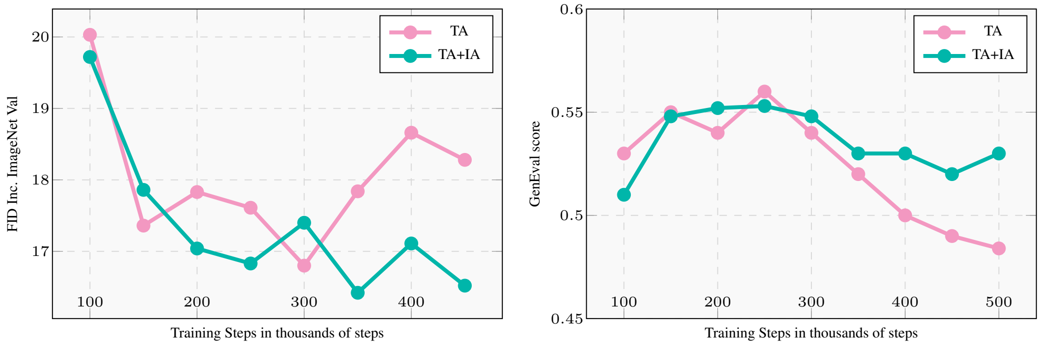
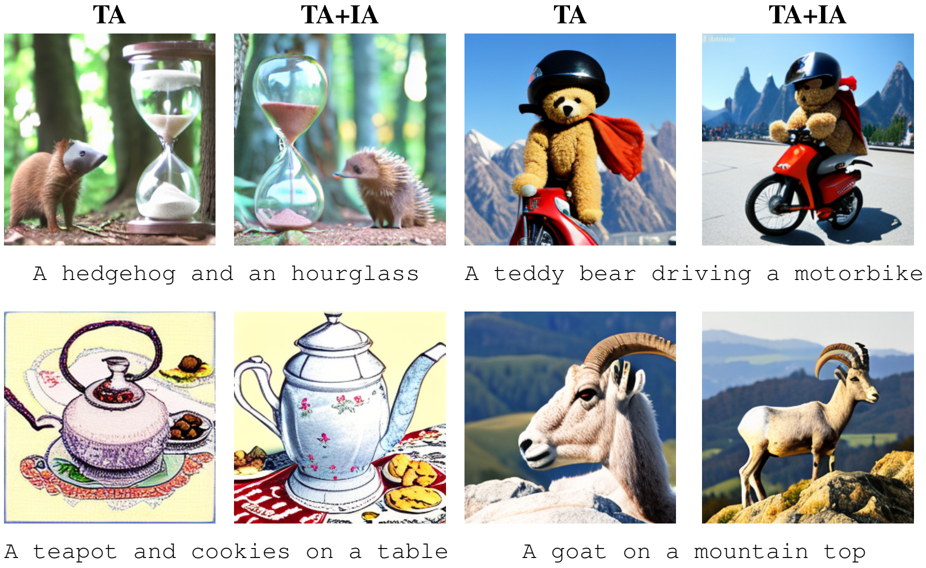


Figure 3: **Training dynamics showing FID and GenEval scores vs training steps**. **TA + IA** maintains better scores throughout training compared to **TA** only, demonstrating improved resistance to overfitting. Lower FID scores indicate better image quality. Better GenEval scores indicate better compositionality abilities.



A hedgehog and an hourglass

A teddy bear driving a motorbike

A teapot and cookies on a table

A goat on a mountain top

Figure 4: **Qualitative comparison: Text-Augmentation (TA, first, third columns) vs Text+Image Augmentation (TA+IA second, last columns)) for four prompts (left and right blocks per row).** Image augmentation improves text comprehension, compositionality and overall image quality.

3 State-of-the-art comparison

We now investigate how good a model trained solely on ImageNet fares against the state of the art.

Quantitative results: Comparison to the state of the art on GenEval and DPG benchmarks. We test the composition ability of our 512^2 model trained with TA + IA on the GenEval and DPGBench benchmarks and compare our performances to the ones of popular state-of-the-art models.

Resolution	Overall↑	One obj.↑	Two obj.↑	Count.↑	Col.↑	Pos.↑	Col. attr.↑
DiT-I 256^2	0.58	0.95	0.67	0.43	0.80	0.30	0.35
DiT-I 512^2	0.61	0.98	0.73	0.43	0.76	0.34	0.40

Table 4: **GenEval scores** of models with different resolution. The 512^2 is finetuned from the 256^2.

Model	**#params**	**#train data**	**Overall↑**	**One obj.↑**	**Two obj.↑**	**Count.↑**	**Col.↑**	**Pos.↑**	**Col. attr.↑**
SD v1.5	0.9B	5B+	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-α	0.6B	25M	0.48	0.98	0.50	0.44	0.80	0.08	0.07
PixArt-Σ (512)	0.6B	35M+	0.52	0.98	0.59	0.50	0.80	0.10	0.15
SD v2.1	0.9B	5B+	0.50	0.98	0.51	0.44	0.85	0.07	0.17
SDXL	3.5B	5B+	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SD3 M (512)	2B	1B+	0.62	0.98	0.74	0.63	0.67	**0.34**	0.36
SANA-0.6	0.6B	⊘	0.64	**0.99**	0.71	0.63	**0.91**	0.16	0.42
FLUX-dev	12B	⊘	**0.67**	**0.99**	**0.81**	**0.79**	0.74	0.20	**0.47**
Ours (512^2)	0.4B	1.2M	0.61	0.98	0.73	0.43	0.76	**0.34**	0.40

Table 5: **Results on GenEval**. Results are reported from their papers. **Bold** indicates best, underline second best.

Model	**#params**	**#train data**	**Overall↑**	**Entity↑**	**Attribute↑**	**Relation↑**	**Other↑**	**Global↑**
SDv1.5	0.9B	5B+	63.2	74.2	75.4	73.5	67.8	74.6
Pixart-α	0.6B	25M	71.1	79.3	78.6	82.6	77.0	75.0
CAD	0.4B	20M	77.6	85.3	84.7	**91.5**	74.8	84.5
Pixart-Σ (512)	0.6B	35M	79.5	87.1	86.5	84.0	86.1	87.5
Pixart-Σ (1024)	0.6B	35M	80.5	82.9	**88.9**	86.6	87.7	86.9
Sana-0.6	0.6B	⊘	**84.3**	90.0	88.6	90.1	**91.9**	82.6
SDXL	3.5B	5B+	74.7	82.4	80.9	86.8	80.4	83.3
SD3-Medium	2B	1B+	84.1	**91.0**	88.8	80.7	88.7	**87.9**
Janus	1.3B	1B+	79.7	87.4	87.7	85.5	86.4	82.3
FLUX-dev	12B	⊘	84.0	89.5	88.7	91.1	89.4	82.1
Ours (512^2)	0.4B	1.2M	78.6	86.1	84.9	**91.5**	76.8	78.4

Table 6: **Results on DPG-Bench**. We compare our models to the results reported in Wu et al. (2024). **Bold** indicates best, underline second best.

GenEval. Table 5 reports the results on GenEval benchmark. We observe that our 512^2 model (**0.61**) performs better on average than SD1.5 (**0.43**), Pixsart-α (**0.48**), SD2.1 (**0.50**), PixArt-Σ-512^2 (**0.52**) and SDXL (**0.55**) in their native resolution. The striking improvements of our model are in the position attribute and color attribution where our model achieves more than **+10** w.r.t SDXL and even matches SD3 M (512^2).

DPGBench. Table 6 reports the results on DPGBench, a recent benchmark similar to Geneval but with a more complex prompt set. We observe similar trends as for GenEval: compared to the current leaderboard, we achieve an overall accuracy of **78.6%** with our 512^2 model, which improves over SDXL by **+3.9%**. and PixArt-α by **+7.5%**. Impressively, our models reach accuracies comparable to that of Janus Wu et al. (2024), a 1.3B parameters VLM with generation capabilities. Notably, our model is particularly good at *Relation*, achieving state-of-the-art of **91.5%**.

4 Task specific finetuning: Aesthetics

Here, we show that our models trained on ImageNet possess general capabilities that can be further exploited by task specific fine-tuning, such as high aesthetic image generation. Note that ImageNet does not contain many high aesthetic images, making this task challenging. Starting from the 512^2 model, we upscale the model and fine-tune it on LAION-POP, a dataset curated for high aesthetics images, for 100k steps.

Quantitative results of this fine-tuning are shown in Table 7 using PickScore (Kirstain et al., 2023), Aesthetics Score (Schuhmann et al., 2022), HPSv2 (Wu et al., 2023) and ImageReward (Xu et al.,

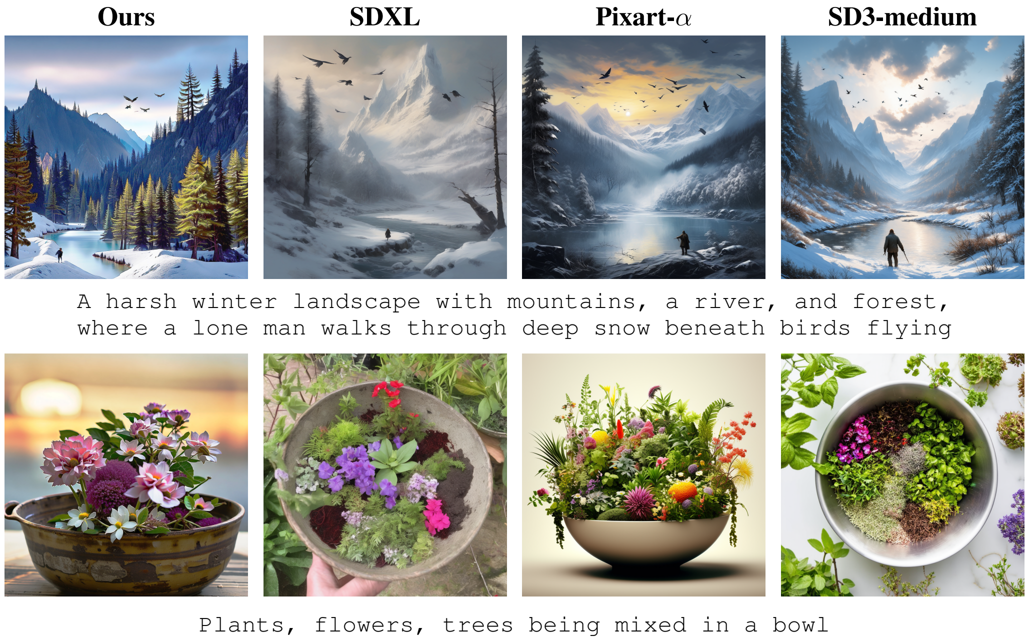

Ours
SDXL
Pixart-α
SD3-medium
A harsh winter landscape with mountains, a river, and forest, where a lone man walks through deep snow beneath birds flying
Plants, flowers, trees being mixed in a bowl

Figure 5: **Comparison with SOTA models at 1024^2 resolution**. Each row shows the same prompt rendered by four different models: Ours, SDXL, Pixart-α, and SD3-Medium. The prompt is taken from `ImageRewards` Xu et al. (2023). Additional comparisons are shown in Figure 7

Resolution	Finetuning	TTS	PickScore↑	Aes.Score↑	HPSv2.1↑	ImageReward↑
512^2	✗	✗	20.94	5.46	0.24	0.20
1024^2	Laion-POP	✗	21.04	5.67	0.25	0.24
1024^2	Laion-POP	✓	21.57	6.28	0.29	0.64

Table 7: **Aesthetic metrics** of TA models and TA+IA models. All models are trained with long captions. Text prompts are taken from `PartiPrompts` Yu et al. (2022). TTS denotes Test-Time Scaling using HPSv2 as selection criterion.

2023). All metrics are improved, which shows that fine-tuning the model on data with a task-oriented curation is possible. Table 11 compares our finetuned models to the state-of-the-art on Aesthetic metrics. We further experiment with test-time scaling, using the Random Search protocol from Ma et al. (2025) with HPSv2 as a criterion, and obtain much higher results, suggesting that the model has hidden aesthetics capabilities.

We show qualitative examples in Figure 5, comparing to SDXL, Pixart-α and SD3-M. Our model shows very competitive results at a fraction of the training cost. More example are shown in Figure 7.

5 Related Work

Diffusion Models. Song et al. (2020); Ho et al. (2020); Sohl-Dickstein et al. (2015) have demonstrated remarkable success across various domains Huang et al. (2023); Courant et al. (2025); Dufour et al. (2024b). While image generation remains their most prominent application Dhariwal & Nichol (2021); Song et al. (2020); Karras et al. (2022), text-to-image (T2I) synthesis Rombach et al. (2022); Saharia et al. (2022); Ramesh et al. (2022) has emerged as a particularly impactful use case. These models operate by learning to reverse a gradual Gaussian noise corruption process. At extreme noise levels, the model effectively samples from a standard normal distribution to produce realistic images. The core optimization objective is:

$$\min_{\theta} \mathbb{E}_{(x_0,c)\sim p_{\text{data}},\epsilon\sim\mathcal{N}(0,1)} \left[\left\| \epsilon - \epsilon_\theta(x_t, c, t) \right\|^2 \right] \quad (1)$$

where $x_t{=}\sqrt{\gamma(t)}x_0{+}\sqrt{1-\gamma(t)}\epsilon$ denotes the noised image at timestep t, x_0 the original image, c the corresponding condition (such as text), ϵ is standard normal noise, ϵ_θ the learned noise predictor, and $\gamma(t)$ the variance schedule.

Computational Efficiency. Traditional diffusion models require substantial computational resources, with leading implementations consuming hundreds of thousands of GPU hours Rombach et al. (2022). Recent advances have significantly improved training efficiency. Wei et al. (2023); Yu et al. (2024) identified limitations in the diffusion loss's representation learning capabilities, demonstrating that supplementary representation losses accelerate convergence. Chen et al. (2023) achieved dramatic compute reduction by repurposing class-conditional models for text-to-image generation. Dufour et al. (2024a) introduced architectural improvements and coherence-aware mechanisms, matching Stable Diffusion's performance Rombach et al. (2022) with 100x fewer GPU hours.

Data Efficiency. T2I models relied on billion-scale web-scraped datasets (Rombach et al., 2022), creating accessibility barriers due to storage requirements and reproducibility challenges from copyright restrictions. Chen et al. (2023) pioneered data curation using 20M high-quality images from recaptioned SAM data (Kirillov et al., 2023), though portions remain proprietary. Subsequent work explored CC12M (Changpinyo et al., 2021; Gu et al., 2023; Dufour et al., 2024a) and YFCC100M's public subset (Thomee et al., 2016; Gokaslan et al., 2024), revealing overfitting below 10M samples. Our approach diverges by leveraging ImageNet (Russakovsky et al., 2015) – a reproducible, well-established benchmark with standardized metrics (Heusel et al., 2017). We transform this classification dataset into T2I training data through synthetic captions and image augmentations.

Synthetic captions. Synthetic image captioning has benefited several tasks. For instance, visual question answering Sharifzadeh et al. (2024) and visual representation learning Tian et al. (2023) achieve state-of-the-art performances by enhancing the captioning output of Vision-Language Models Lai et al. (2024); Sharifzadeh et al. (2024). Similarly, training with synthetic captions for text-to-image generation is becoming the defacto protocol for large diffusion models, such as DALL-E Betker et al. (2023), Pixart-α Chen et al. (2023) and Stable Diffusion-3 Esser et al. (2024). More recently, some approaches Liu et al. (2024a); Li et al. (2024) extend this approach by training text-to-image (T2I) models on multi-level captions. Inspired by these, we deploy the popular LLaVA captioner Liu et al. (2024b) to augment existing textual captions and use them to train T2I generators.

6 Conclusion

In this work, we challenged the prevailing wisdom that billion-scale datasets are necessary to unlock text-to-image generation and suggest this is merely a sufficient condition. These large-scale datasets are usually either closed sourced or rapidly decaying, which threatens openness and reproducibility in text-to-image generation research. Instead, we show that it is possible to train smaller models to high quality using ImageNet only and get general text-to-image capabilities.

We propose a standardized text-to-image training setup on ImageNet that leads to models capable of generating high quality images while being excellent at prompt following. This is attested by results on common benchmarks outclassing models widely recognized as good text-to-image generators such as SDXL (61% or +6% on GenEval and 78.6% or +3.9% on DPGBench).

The implications of our work extend beyond just computational efficiency, open science and reproducibility. By showing that smaller datasets can achieve state-of-the-art results, we open new possibilities for specialized domain adaptation where large-scale data collection is impractical. Our work also suggests a path toward more controllable and ethical development of text-to-image models, as smaller datasets enable more thorough content verification and bias mitigation.

Looking forward, we believe our results will encourage the community to reconsider the "bigger is better" paradigm. Future work could explore additional augmentation strategies, investigate the theoretical foundations of data efficiency, and develop even more compact architectures optimized for smaller datasets. Ultimately, we hope this work starts a shift toward more sustainable and responsible development of text-to-image generation models.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI*, 2023.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *European Conference on Computer Vision*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

Nicolas Dufour, Victor Besnier, Vicky Kalogeiton, and David Picard. Don't drop your samples! coherence-aware training benefits conditional diffusion. In *CVPR*, 2024a.

Nicolas Dufour, David Picard, Vicky Kalogeiton, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. *arXiv*, 2024b.

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 *Proc. ICML*, 2024.

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. *NeurIPS*, 2023.

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, 2024.

Aaron Gokaslan, A Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Commoncanvas: Open diffusion models trained on creative-commons images. In *CVPR*, 2024.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua M Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *ICLR*, 2023.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv*, 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. *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 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*, 2024.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv*, 2023.

Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. *Proc. ICML*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

Diederik P Kingma. Auto-encoding variational bayes. *ICLR*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

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. *arxiv*, 2023.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, et al. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 2019.

Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, Meng Cao, and Yinfei Yang. Revisit large-scale image-caption data in pre-training multimodal foundation models. *arxiv*, 2024.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with llama-3? *arxiv*, 2024.

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 *ECCV*, 2014.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arxiv*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024b.

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.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proc. ICML*, 2020.

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

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *arXiv*, 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. *arXiv*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. Synth^2: Boosting visual-language models with synthetic captions and image embeddings. *arxiv*, 2024.

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

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

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *CVPR*, 2023.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017.

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *ICCV*, 2023.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv*, 2024.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

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. *arXiv*, 2022.

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*, 2024.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

A Appendix

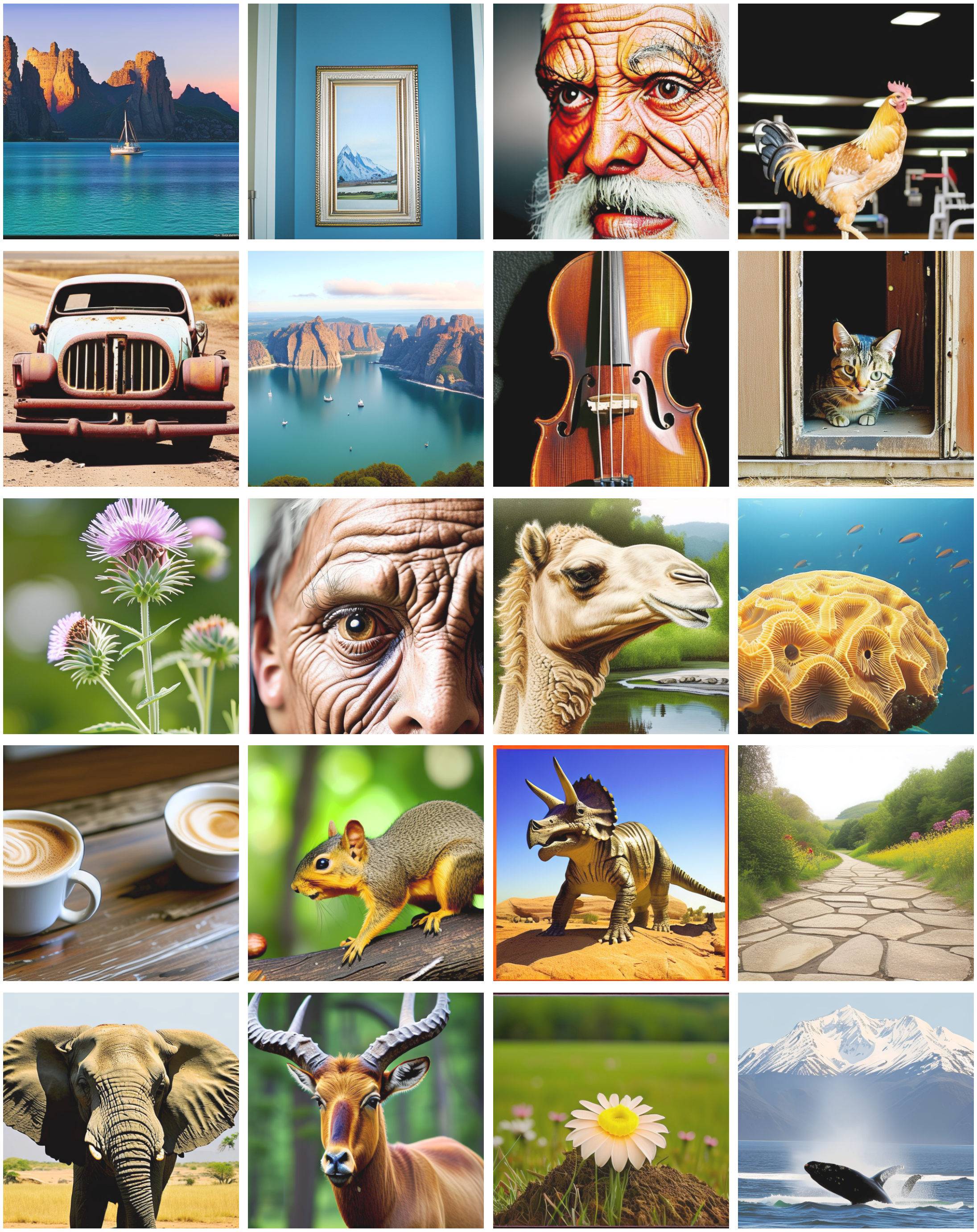
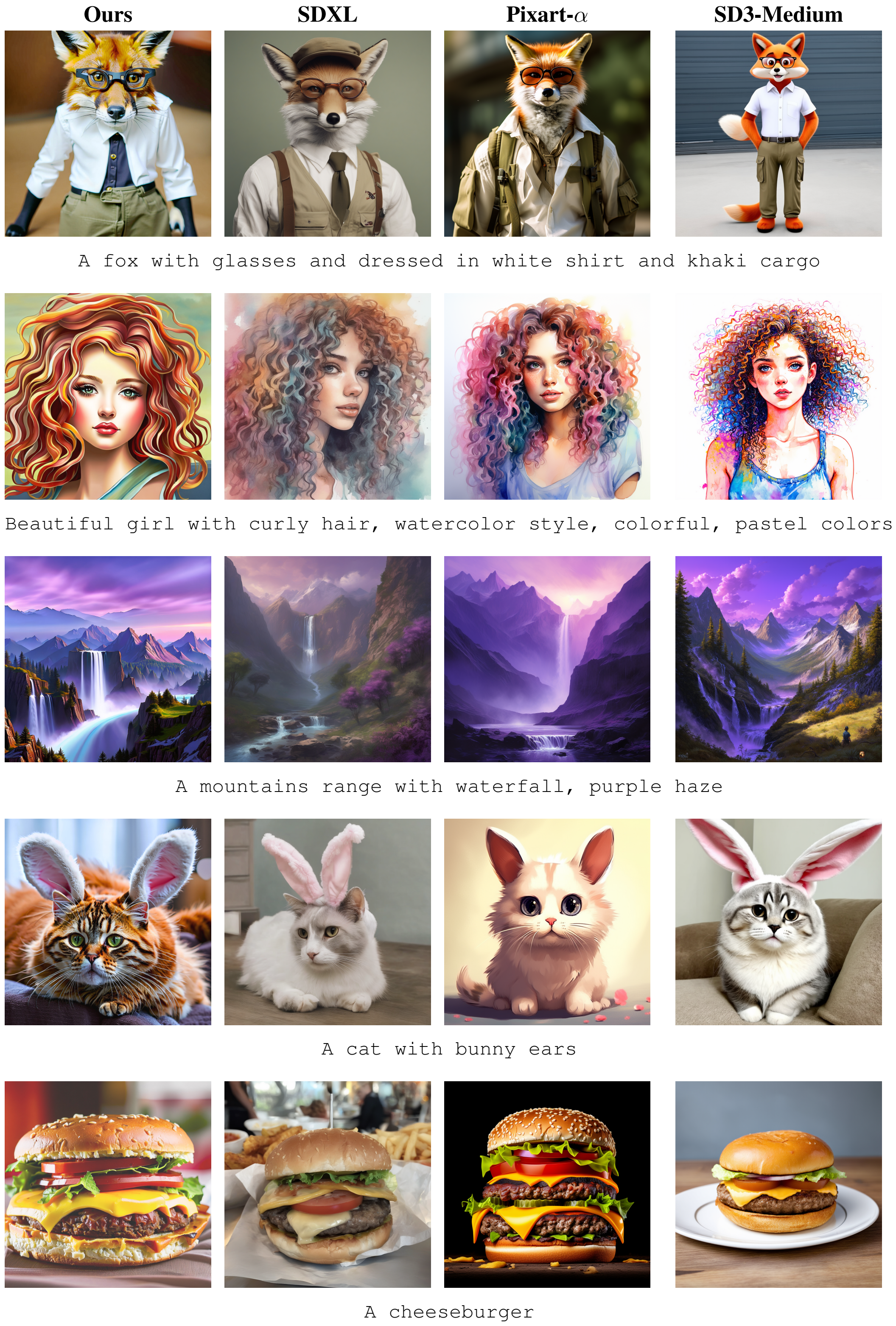


Figure 6: Additional Qualitative Results at 512^2 resolution



A fox with glasses and dressed in white shirt and khaki cargo

Beautiful girl with curly hair, watercolor style, colorful, pastel colors

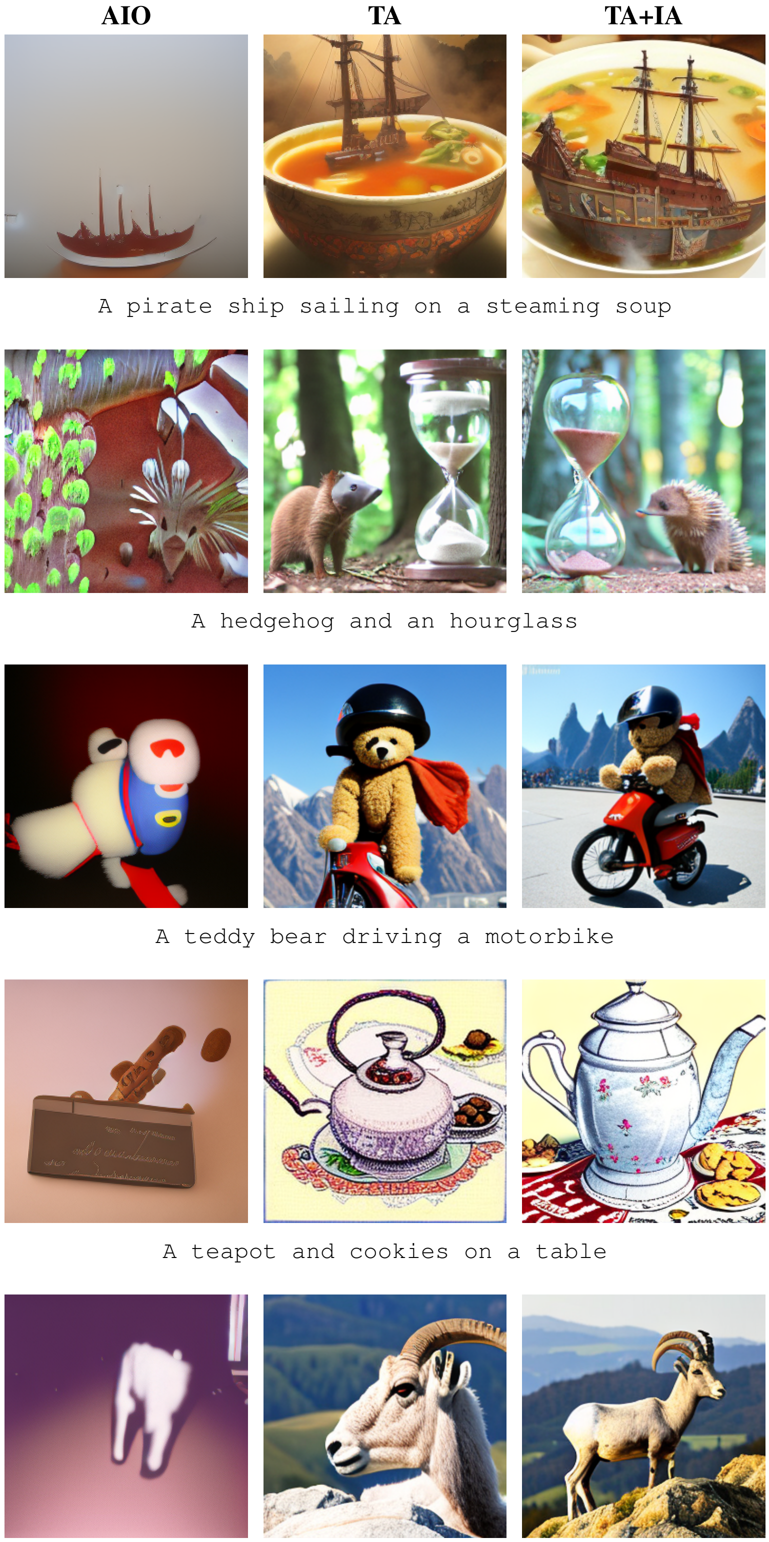
A mountains range with waterfall, purple haze

A cat with bunny ears

A cheeseburger

Figure 7: **Comparison with SOTA models at** 1024^2 **resolution**. Each row shows the same prompt rendered by four different models: Ours, SDXL, Pixart-α, and SD3-Medium. The prompts are taken from `ImageRewards`

.



A pirate ship sailing on a steaming soup

A hedgehog and an hourglass

A teddy bear driving a motorbike

A teapot and cookies on a table

A goat on a mountain top

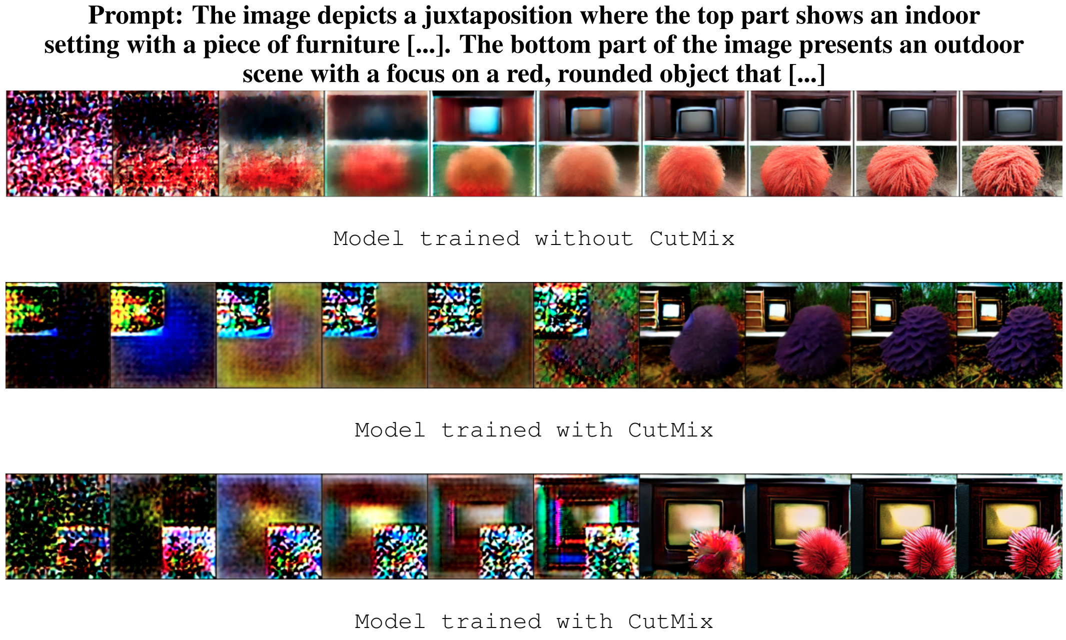
Figure 8: **Qualitative comparison across models: AIO, TA, and TA+IA**. Image and text augmentations improve text comprehension and overall image quality.

A FAILURES CASES WITH CUTMIX

Figure 9 exposes a failure case of models trained with CutMix augmentation when confronted with prompts juxtaposing semantically disjoint concepts. Rather than synthesizing a coherent scene, these models sometimes denoise the input as two independent images, resulting in visible artifacts: jagged seams, color bleeding, or distorted transitions along the composition boundary (row 2).

Intriguingly, CutMix-trained models also exhibit a learned adaptability to disjoint concepts—as seen in row 3, where the model attempts to "bend" the scene to accommodate both prompts (e.g., blending indoor lighting with the outdoor object).

Our finetuned 512^2 model substantially reduces these discontinuities, generating more globally coherent images. This improvement suggests that higher-resolution training helps the model reconcile disjoint regions by leveraging finer spatial details.



Prompt: The image depicts a juxtaposition where the top part shows an indoor setting with a piece of furniture [...]. The bottom part of the image presents an outdoor scene with a focus on a red, rounded object that [...]

Model trained without CutMix

Model trained with CutMix

Model trained with CutMix

Figure 9: **Qualitative examples of generated images with descriptions.** Each image is followed by a textual description to highlight key features.

B Additional quantitative results

Model	TA	IA	FID Inc.↓	FID DINOv2↓	Prec.↑	Rec.↑	Den.↑	Cov.↑	CS↑	Jina-CS↑
DiT-I	✗	✗	71.00	1107.22	0.46	0.07	0.37	0.08	18.03	22.42
	✓	✗	45.71	631.91	0.64	0.44	0.52	0.29	25.52	36.63
	✓	Crop	44.02	627.78	0.63	0.41	0.51	0.29	25.46	38.39
	✓	CutMix	49.12	631.01	0.65	0.45	0.54	0.30	25.68	36.80
CAD-I	✗	✗	46.35	858.43	0.52	0.18	0.45	0.15	12.89	14.06
	✓	✗	46.93	655.37	**0.66**	**0.42**	**0.61**	0.28	26.37	35.72
	✓	CutMix	49.41	**646.51**	**0.66**	0.41	0.57	**0.29**	**26.60**	**36.51**

Table 8: **Ablation study** on COCO dataset. Precision, Recall, Density and Coverage are computed using `DINOv2` features. **Bold** indicates best, underline second best.

Model	TA	IA	FID Inc.↓	FID DINOv2↓	Prec.↑	Rec.↑	Den.↑	Cov.↑	CS↑	Jina-CS↑
DiT-I	✗	✗	71.00	1107.22	0.46	0.07	0.37	0.08	18.03	22.42
	✓	✗	45.71	631.91	0.64	0.44	0.52	0.29	25.52	36.63
	✓	Crop	44.02	627.78	0.63	0.41	0.51	0.29	25.46	38.39
	✓	CutMix	49.12	631.01	0.65	0.45	0.54	0.30	25.68	36.80
CAD-I	✗	✗	46.35	858.43	0.52	0.18	0.45	0.15	12.89	14.06
	✓	✗	46.93	655.37	**0.66**	**0.42**	**0.61**	0.28	26.37	35.72
	✓	CutMix	49.41	**646.51**	**0.66**	0.41	0.57	**0.29**	**26.60**	**36.51**

Table 9: **Ablation study** on COCO dataset. Precision, Recall, Density and Coverage are computed using `DINOv2` features. **Bold** indicates best, underline second best.

Model	IA	Res.	Dataset	PickScore↑	Aes.Score↑	HPSv2.1↑	ImageReward↑
DiT-I	✗	256	ImageNet	20.74	5.29	0.25	0.18
	Crop	256	ImageNet	20.63	5.17	0.24	0.12
	CutMix	256	ImageNet	20.81	5.34	0.25	0.26
	Crop	256	Laion-Pop	19.50	4.17	0.19	-0.98
	Crop	256	ImNet+COCO	20.58	5.17	0.24	0.04
DiT-I finetuned	Crop	512	ImageNet	20.94	5.46	0.24	0.20
	Crop	1024	ImageNet	20.36	4.96	0.22	-0.42
	Crop	1024	Laion-Pop	21.04	5.67	0.25	0.24
CAD-I	✗	256	ImageNet	20.03	5.17	0.24	0.22
	CutMix	256	ImageNet	20.03	5.16	0.24	0.30
CAD-I Flow	✗	256	ImageNet	20.61	4.96	0.24	0.10
	CutMix	256	ImageNet	20.57	5.00	0.24	0.09

Table 10: **Aesthetic metrics** of TA models and TA+IA models. All models are trained with long captions. Text prompts are taken from `PartiPrompts` Yu et al. (2022).

Model	**#params**	**#train data**	**Aes. Score$\uparrow$**	**PickScore$\uparrow$**	**HPSv2.1$\uparrow$**	**ImageReward$\uparrow$**
SD v1.5	0.9B	5B+	5.68	21.3	0.25	0.24
SD v2.1	0.9B	5B+	5.81	21.5	0.26	0.38
PixArt-α	0.6B	25M	6.47	22.6	0.29	0.97
PixArt-Σ	0.6B	35M+	6.44	22.5	0.29	1.02
CAD	0.35B	-	5.56	21.4	0.26	0.69
Sana-0.6B	0.6B	-	6.31	22.8	**0.30**	**1.23**
Sana-1.6B	1.6B	-	6.36	22.8	**0.30**	**1.23**
SDXL	2.6B	5B+	5.94	22.0	0.25	0.46
SD3-Medium	2B	1B+	6.18	22.5	0.30	1.15
FLUX-dev	12B	-	**6.56**	**22.9**	**0.30**	1.19
Ours (ft. Laion-POP 1024^2)	0.4B	1.5M	6.28	21.6	0.29	0.64

Table 11: **Results on Reward Metrics.** Results are computed using the `PartiPrompts` Yu et al. (2022). SOTA scores are computing using HuggingFace checkpoints at their native resolution. **Bold** indicates best, underline second best.

C Implementation details

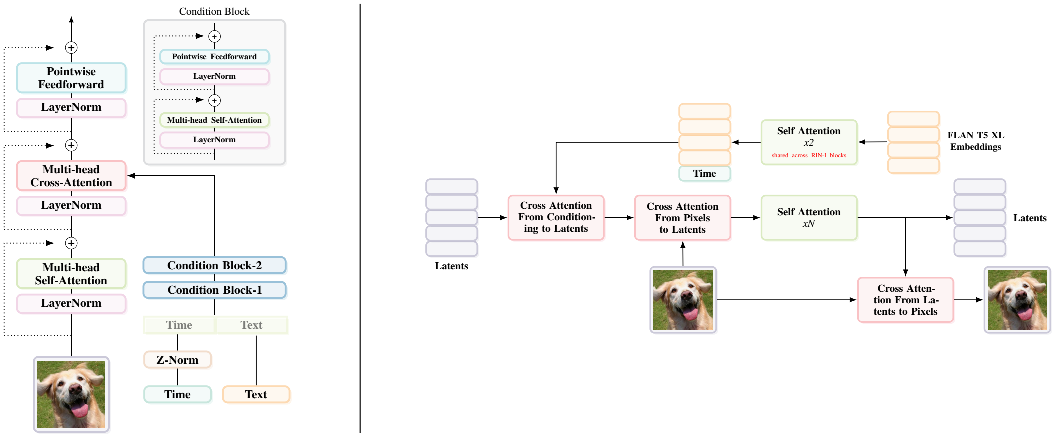


Figure 10: **Fundamental architecture blocks used in our experiments**. *Left*: DiT-I block and *Right*: CAD-I block.

In this work, we use both DiT Peebles & Xie (2023) and RIN Jabri et al. (2023) architectures. To adapt DiT for text-conditional setting, we replace `AdaLN-Zero` conditioning with *cross-attention* to input the text condition into the model, as in Chen et al. (2023). Before feeding the text condition to the model, we refine it using *two* self-attention layers. Similar to Esser et al. (2024), we add QK-Normalization Henry et al. (2020) in each of the *self-attention* and *cross-attention* blocks to mitigate sudden growths of attention entropy and reduce training loss instability. We also add LayerScale Touvron et al. (2021) to each of the residual blocks of DiT for further stability. Figure 10 details our DiT-I architecture, Table 11 (left) gives the hyperparameters of the DiT-I model.
To adapt the RIN Jabri et al. (2023) for the text-conditional setting, we used the off-the-shelf architecture CAD from Dufour et al. (2024a), an adaptation of the RIN architecture detailed in the Appendix of Dufour et al. (2024a). Figure 10 and Table Table 11 (right) details our CAD-I architecture.
We use the framework of latent diffusion Rombach et al. (2022). For encoding the images into the latent space, we use the pre-trained variational autoencoder Kingma (2014); Van Den Oord et al. (2017) provided by the authors of Stable Diffusion Rombach et al. (2022). The checkpoint used is available on HuggingFace: `https://huggingface.co/stabilityai/sd-vae-ft-ema`. For text conditions, we encode the captions using the T5 text encoder. The checkpoint is available on HuggingFace: `https://huggingface.co/google/flan-t5-xl`.

Parameter	Value
Num. blocks	24
Embedding dimension	1024
Num. heads	16
Patch size	2
Input size (256)	(4,32,32)
Num. text registers	16
Num. condition self-attn blocks	2
Condition self-attn heads	16
Condition FFN expansion	4
DiT FFN expansion	4
Unconditional drop prob.	0.1

DiT-I parameters

Parameter	Value
Input size (256)	(4,32,32)
Num. latents	128
Latent dimension	512
Num. processing layers	2
Num. blocks	3
Patch size	2
Read/Write heads	8
Compute heads	16
Data positional embedding	learned
Num. text registers	16
Unconditional drop prob.	0.1

CAD-I parameters

Figure 11: **Key architecture parameters** of DiT-I (left) and CAD-I (right).

D Captioning details

Captioning efficiently with LLaVA To caption images, we use the checkpoint `llama3-llava-next-8b-hf` (available on HuggingFace: `https://huggingface.co/llava-hf/llama3-llava-next-8b-hf`) with the prompt *"Describe this image"*. LLaVA encodes images using a dynamic resolution scheme. It processes both the entire image and four distinct patches as unique images and concatenates them. For 256x256 images, LLaVA uses around 2500 image tokens. To make the captioning process more efficient, we prune the image tokens, retaining only the tokens of the entire image and discarding patch-specific tokens. This optimization increased inference speed by a factor of 2.7, without compromising performances. Examples of long captions generated by LLaVA are given in Figure 12.

Captioning CutMix images We caption CutMix images from $\mathrm{CM}^{1/2}$ with similar settings used for captioning the original ImageNet images. However, to ensure that LLaVA does not describe both the base and the CutMix images independently, we use a different prompt: *"Describe this image. Consider all the objects in the picture. Describe them, describe their position and their relation. Do not consider the image as a composite of images. The image is a single scene image"*.
For settings $\mathrm{CM}^{1/4}$, $\mathrm{CM}^{1/9}$ and $\mathrm{CM}^{1/16}$, LLaVA tends to either ignore the smaller CutMix image or describe the image as a composite of two images. To avoid this behaviour, we encode the image by using the entire image patch and add tokens from the patch to which the CutMix image belongs. We use the following prompt: *"Describe this image. Consider all the objects in the picture. Describe them, describe their position and their relation. Do not consider the image as a composite of images. The image is a single scene image"*. Examples of long captions generated by LLaVA for CutMix images are given in Figure 12.

E CUTMIX DETAILS

The CutMix framework systematically combines concepts while preserving object centrality. Our framework defines four precise augmentation patterns, each designed to maintain visual coherence while introducing novel concept combinations. These are briefly described below:

1. **CM**$^{1/2}$ (Half-Mix):
   Scale: Both images maintain their original resolution.
   Position: Deterministic split along height or width.
   Coverage: Each concept occupies 50% of final image.
   Preservation: Both concepts maintain full resolution.
2. **CM**$^{1/4}$ (Quarter-Mix):
   Scale: CutMix image resized to 50% side length.
   Position: Fixed placement at one of four corners.
   Coverage: 2nd concept occupies 25% of final image.
   Preservation: Base image center region remains intact.
3. **CM**$^{1/9}$ (Ninth-Mix):
   Scale: CutMix image resized to 33.3% side length.
   Position: Fixed placement along image borders.
   Coverage: 2nd concept occupies 11.1% of final image.
   Preservation: Base image center, corners remain intact.
4. **CM**$^{1/16}$ (Sixteenth-Mix):
   Scale: CutMix image resized to 25% side length.
   Position: Random placement not central 10% region.
   Coverage: 2nd concept occupies 6.25% of final image.
   Preservation: Base image center region remains intact.

Each augmentation strategy generates 1,281,167 samples, matching ImageNet's training set size. Figure 12 shows examples of the different structured augmentations.
We also define **CM**^{all}, which uniformly samples from all four patterns. The CM^{all} variant combines equal proportions (25%) from each pattern to maintain the same total sample count. Post-augmentation, we apply LLaVA captioning to all generated images, ensuring semantic alignment between visual and textual representations. This produces detailed descriptions that accurately reflect the augmented content while maintaining natural language fluency.

E.1 TRAINING WITH CUTMIX IMAGES

Because the CutMix image augmentations have strong artefacts corresponding to the boundaries of the mixing, we have to prevent the model from learning those salient features and reproducing them. To that end, we propose to train on image augmentation only at timesteps t where the noisy image x_t is sufficiently noisy that the artifacts no longer matter. In practice, this corresponds to sampling either from the original image training set $\mathcal{A}$ or from the augmented image training set $\mathcal{A}_{\text{IA}}$ conditionally to t, compared to an additional hyperparameter τ deciding whether t is sufficiently large for image augmentation. This extra condition leads to replacing the original diffusion loss in Equation 1 with

$$\min_\theta \mathbb{E}_{\substack{t\sim\mathcal{U}(0,T),\\ \rho\sim\mathcal{B}_(\tau,p)(t),\\ (x,c)\sim\mathcal{A}(\rho),\\ \epsilon\sim\mathcal{N}(0,1)}} \left[\|\epsilon - \epsilon_\theta(x_t, c, t)\|^2\right]. \tag{2}$$

In this novel loss, the timestep $t \sim \mathcal{U}(0, T)$ is still sampled uniformly. We introduce a new random variable ρ that is sampled conditionally to t, where $\mathcal{B}_{\tau,p}(t)$ denotes a specific distribution that corresponds to:

$$\mathcal{B}_{\tau,p}(t) = \begin{cases} 0, \text{ if } t \leq \tau, \\ \mathcal{B}_p, \text{ else.} \end{cases} \tag{3}$$

Here, $\mathcal{B}_p$ a Bernoulli distribution of parameter p. The text-image pair (x_0, c) is then sampled conditionally to ρ, where $\mathcal{A}(\rho)$ is a distribution that uniformly samples from the original or the augmented datasets depending on ρ:

$$\mathcal{A}(\rho) = \begin{cases} \mathcal{A}, \text{ if } \rho = 0, \\ \mathcal{A}_{\text{IA}}, \text{ else.} \end{cases} \tag{4}$$

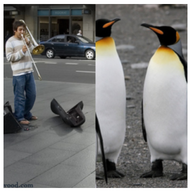
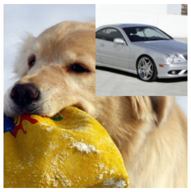
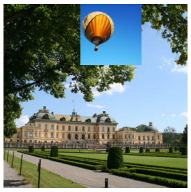
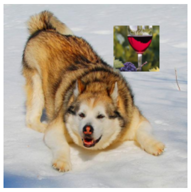
Setting	Long Caption
LLaVA w/o CM 	In the heart of the verdant background, a delicate white butterfly with a hint of yellow around its edges is captured in mid-flight. Its wings, a blend of two shades of cream and white, add a soft contrast to the vibrant colors of the flower it’s hovering over. The flower, a stunning shade of purple, is adorned with a yellow center that speaks of the sun’s rays it might have been basking in. Surrounding it is a lush green foliage, adding depth to this tranquil scene. The butterfly, positioned slightly to the left of the flower’s center, seems to be in the process of alighting or perhaps taking off, adding a dynamic element to this otherwise serene tableau.
CM$^{1/2}$ 	On the left side, there is a person playing the trumpet on a street. The individual is standing on the ground, wearing a long-sleeved shirt and jeans. On the ground next to the player, there is a trash can. The street is made of asphalt, with buildings in the background and a clear sky overhead. On the right side of the image, there are two penguins standing on a gravel surface, possibly a road or a rocky beach. The penguins are frozen in place, displaying a contrast to the dynamic scene on the left side. The penguins’ yellow beaks, black bodies, and white feathers are clearly visible. The image captures an everyday moment of music played on the street, juxtaposed with the surreal and unexpected presence of the penguins. The image does not provide information about the time of day, specific location, or any actions being performed by the penguins or the person. The image is static, presenting two moments captured in time rather than a sequence of events.
CM$^{1/4}$ 	The image shows a large, muscular dog in the foreground and a silver sports car in the background. The dog, appearing to be a Golden Retriever, is on the left side of the frame, occupying a significant portion of the space. It has a yellow hue with darker markings, and its ears are floppy. The dog is heavily coated with snowflakes, suggesting that the photo was taken in a snowy environment. The sports car, positioned on the right, is a two-door coupe with sleek curves and a notable design, featuring the Mercedes-Benz logo on its front grille. The car has a silver finish, and the photo captures it from a perspective that shows the front and side profile. The car is parked on an asphalt surface, possibly a parking lot or a driveway. The dog is facing the camera with direct gaze, while the car is positioned slightly towards the side, away from the viewer’s perspective.
CM$^{1/9}$ 	The image depicts a picturesque outdoor scene featuring an ornate building, which appears to be a palace or manor house, with classical architectural elements including symmetrical windows, a central cupola, and multiple chimneys. In front of the building is a well-maintained garden with pathways and neatly trimmed hedges or borders. Above the garden, there is a clear blue sky with a few scattered clouds. In the sky, there is a single hot air balloon with a bright orange and yellow pattern. The balloon is floating at a considerable height above the garden and the building, suggesting it might be part of a leisure activity or a special event. The image is a photograph with natural lighting, indicative of a sunny day.
CM$^{1/16}$ 	The image is a photograph featuring a husky dog resting in the snow. The dog has a light coat with darker markings around its face and ears, and it is lying on its side with its head up, looking directly at the camera. Its eyes are open and its mouth is slightly open, showing teeth and a pink tongue, which suggests the dog might be panting or in a relaxed state. Next to the dog’s side, there is a wine glass with red wine and a few purple flowers, which could be lilacs, positioned on the left side of the glass stem. The wine glass and flowers are set against a blurred background that gives the impression of greenery.

Figure 12: **Long captions generated by our synthetic LLaVA captioner**. The captions generated are highly diverse and add in much more intricate details of *compositionality*, *colors* as well as *concepts* which are not present in the original ImageNet dataset. The captions generated for our augmented images are also highly coherent and explain the scene in a much more realistic way.

The noise ϵ is sampled from the Normal distribution, as in the usual diffusion equation. Similarly, the noisy image x_t is obtained by $x_t{=}\sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon$.
This novel loss function is more involved than the regular diffusion training; yet, in practice, it is very easy to implement and can be done entirely during the mini-batch construction as described in Algorithm 1.
Figure 13 illustrates our complete CutMix pipeline.

Algorithm 1 Batch with CutMix image augmentation

```
Input: dataset $\mathcal{A}$, $\mathcal{A}_{\text{IA}}$, augmentation time $\tau$, augmentation probability $p$, batch size $m$
$B \leftarrow \{\}$
for $i = 1$ to $m$ do
    $t \sim \mathcal{U}(0, T)$
    $(x_0, c) \sim \mathcal{A}$
    if $t > \tau$ then
        $\rho \sim \mathcal{B}_p$
        if $\rho$ then
            $(x_0, c) \sim \mathcal{A}_{\text{IA}}$
        end if
    end if
    $\epsilon \sim \mathcal{N}(0, 1)$
    $x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1 - \gamma(t)}\epsilon$
    $B \leftarrow B \cup \{(x_t, c, t)\}$
end for
Return: $B$
```

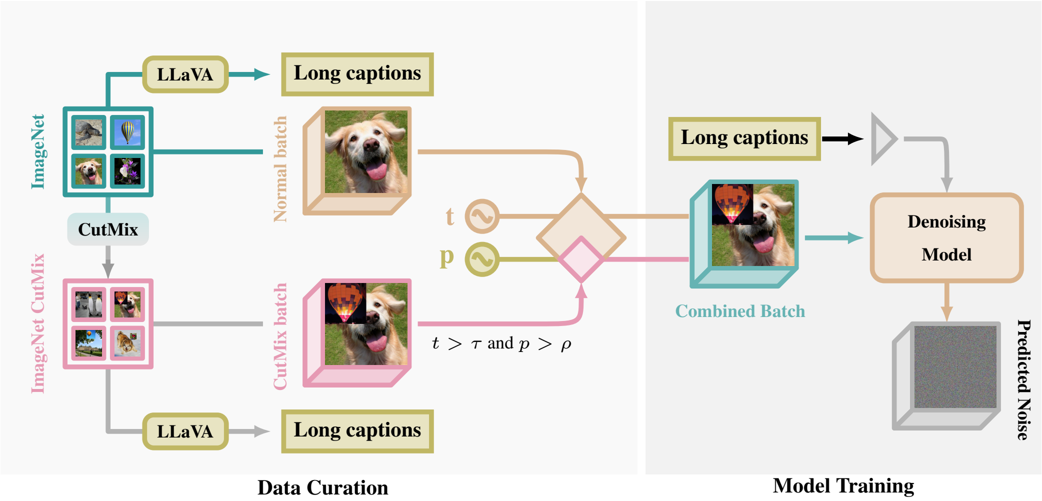


Figure 13: **Pipeline of our Cutmix Data Curation and Training process.** Starting from ImageNet, we a) use LLaVa VLM to caption the images into long detailed caption (top branch left) and b) use several CutMix strategies to create new images combining several ImageNet concepts and caption them using LLaVa into long and detailed captions (bottom branch left). During training, we sample batches of normal and CutMix images and we select from each batch depending on the timestep t at which the CutMix strategy is valid and a probability p of sampling CutMix images.

E.2 CutMix Augmentation Ablations

E.2.1 Ablation on CutMix settings

First, we analyse the performances of the pixel augmentations for $\{$**CM**$^{1/2}$, **CM**$^{1/4}$, **CM**$^{1/9}$, **CM**$^{1/16}$, **CM**^{all} $\}$ settings. We fix the probability of using a pixel-augmented image in the batch when $t > \tau$ to $p = 0.5$ and we measure both image quality and composition ability. Results are reported in Table 12.

For image quality, all settings seem to perform similarly, with **CM**$^{1/2}$ being the best at 6.13 FID and **CM**^{all} being the worst at 6.81 FID. This indicates that all settings are able to avoid producing uncanny images that would disturb the training too much.

For composition ability, **CM**$^{1/16}$ can improve over the baseline on extended prompts, whereas **CM**^{all} can improve over the baseline on original prompts. Overall, only **CM**^{all} manages to keep closer performances between the original prompts and the extended ones. Since **CM**^{all} is a mixture of all other settings, it also has the most diverse training set and is thus harder to overfit. As such, we consider **CM**^{all} for the best models.

Model	**CutMix Settings**	**FID↓**	**GenEval↑**	
			◇	⋆
CAD-I	CM$^{1/2}$	6.13	0.46	0.55
	CM$^{1/4}$	6.41	0.49	0.53
	CM$^{1/9}$	6.63	0.51	0.51
	CM$^{1/16}$	6.42	0.47	0.56
	CM^{all}	6.81	0.53	0.55

Table 12: **Ablation study** on CutMix settings. The probability of sampling CutMix images used here is $\rho = 0.5$. Models are trained for 250k steps. FID is computed on the ImageNet val set with long prompts, using the `Inception-v3` backbone. ◇ means original GenEval prompts. ⋆ means extended GenEval prompts.

Model	ρ	**FID↓**	**GenEval↑**	
			◇	⋆
CAD-I	0	6.16	0.51	0.55
	0.25	5.99	0.55	0.58
	0.5	6.41	0.49	0.53
	0.75	6.71	0.45	0.53
	1	6.07	0.48	0.49

Table 13: **Ablation study** on probability ρ of sampling a CutMix image during training. The CutMix setting is CM$^{1/4}$. Models are trained for 250k steps. FID is computed on ImageNet val set with long prompts, using the `Inception-v3` backbone. ◇ means original GenEval prompts. ⋆ means extended GenEval prompts.

Model	τ	**FID↓**	**GenEval↑**	
			◇	⋆
CAD-I	300	6.99	0.51	0.53
	400	6.62	0.55	0.57
	500	6.16	0.48	0.55
	600	5.90	0.50	0.55

Table 14: **Ablation study** on timestep threshold τ. The CutMix setting is **CM**^{all} . Models are trained for 250k steps. FID is computed on ImageNet val set with long prompts, using the `Inception-v3` backbone. ◇ means original GenEval prompts. ⋆ means extended GenEval prompts.

E.2.2 Ablation on CutMix probability

Next, we analyse the influence of the probability p of using a pixel augmented image in the batch, when the condition on t is met. Results for $p \in \{0.25, 0.5, 0.75, 1.0\}$ are shown in Table 13, using **CM**$^{1/4}$ pixel augmentations.

As we can see in terms of image quality, the FID is slightly degraded by having too frequent pixel augmentation ($p > 0.5$). This can be explained by the fact that pixel-augmented images are only seen when $t > \tau$. As such, a high value for p creates a distribution gap between the images seen for $t > \tau$ and the images seen for $t \leq \tau$.

Composition ability shows a similar behaviour with the GenEval overall score decreasing when p increases for both the original and the extended prompts. As such, we consider $p \leq 0.5$ for the best models.

E.2.3 Ablation on threshold τ

Finally, we analyse the influence of the threshold τ, which enables CutMix images to be sampled in training batches. Table 14 shows the FID Inception on ImageNet Val and the GenEval scores of models trained with different τ values.

We find that $\tau = 400$ results in the highest GenEval score of 0.55 on original prompts and 0.57 on extended prompts, while $\tau = 600$ yields the lowest FID on ImageNet Val. As such, we use $\tau = 400$ for the best models.

F Cropping details

Our cropping training methodology (see Figure 14) removes spurious concept correlations due to its masking scheme. We maintain the original captions and force the model to independently identify relevant textual elements. This creates a more challenging learning task for the model that enhances text-image alignment. During training, we only consider tokens corresponding to small portion of the image and mask out the rest from both the loss function and cross-attention layers. Given we do this online, this is highly efficient and also allows an infinite training set based on ImageNet to train on. For making the model understand the full dynamics of the training data, in our training scheme with crops, we only feed cropped images to the model with a probability, $p = 0.5$. In rest of the cases, we use entire image. To keep the training scheme as simple as possible, for cropped versions, we only use crop resolution of >50% of the normal resolution.

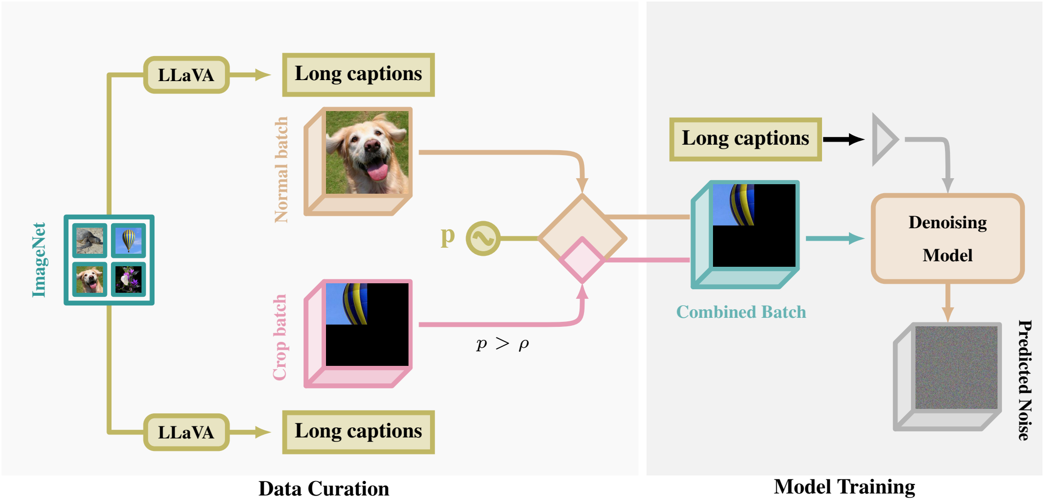


Figure 14: **Pipeline of our Cropping Data Curation and Training process.** Starting from ImageNet, we a) use LLaVa VLM to caption the images into long detailed caption and b) use cropping strategies to create new images from ImageNet by cropping. We keep the same captions as if we were using the original image. During training, we do cropping online with a probability p of sampling cropped images. The crop images can have any resolution >50% of the original resolution.