Stretching Each Dollar: Diffusion Training from Scratch on a Micro-Budget

Supplementary Material

A. Related Work

The landscape of diffusion models [32, 71–73] has rapidly evolved in the last few years, with modern models trained on web-scale datasets [4, 17, 60, 61]. In contrast to training in pixel space [48, 61], the majority of large-scale diffusion models are trained in a compressed latent space, thus referred to as latent diffusion models [60]. Similar to autoregressive models, transformer architectures [83] have also been recently adopted for diffusion-based image synthesis. While earlier models commonly adopted a fully convolutional or hybrid UNet network architecture [14, 48, 60, 61], recent works have demonstrated that diffusion transformers [52] achieve better performance [4, 8, 17]. Thus, we also use diffusion transformers for modeling latent diffusion models.

Since the image captions in web-scale datasets are often noisy and of poor quality [6, 64], recent works have started to recaption them using vision-language models [42, 44, 84]. Using synthetic captions leads to significant improvements in the diffusion models' image generation capabilities [4, 8, 17, 24]. While text-to-image generation models are the most common application of diffusion models, they can also support a wide range of other conditioning mechanisms, such as segmentation maps, sketches, or even audio descriptions [86, 89]. Sampling from diffusion models is also an active area of research, with multiple novel solvers for ODE/SDE sampling formulations to reduce the number of iterations required in sampling without degrading performance [35, 37, 45, 72]. Furthermore, the latest approaches enable single-step sampling from diffusion models using distillation-based training strategies [63, 75]. The sampling process in diffusion models also employs an additional guidance signal to improve prompt alignment, either based on an external classifier [14, 48, 65] or self-guidance [31, 38]. The latter classifier-free guidance approach is widely adopted in large-scale diffusion models and has been further extended to large-language models [62, 90].

Since the training cost of early large-scale diffusion models was noticeably high [47, 58, 60], multiple previous works focused on bringing down this cost. Gokaslan et al. [24] showed that using common tricks from efficient deep learning can bring the cost of stable-diffusion-2 models under \$50K. Chen et al. [8] also reduced this cost by training a diffusion transformer model on a mixture of openly accessible and proprietary image datasets. Cascaded training of diffusion models is also used by some previous works [25, 53, 61], where multiple diffusion models are employed to sequentially upsample the low-resolution generations from the base diffusion model. A key limitation of cascaded training is the strong influence of the low-resolution base model on overall image fidelity and prompt alignment. Most recently, Pernias et al. [53] adopted the cascaded training approach (Würstchen) while training the base model in a $42\times$ compressed latent space. Though Würstchen achieves low training cost due to extreme image compression, it also achieves significantly lower image generation performance on the FID evaluation metric. Alternatively, patch masking has been recently adopted as a means to reduce the computational cost of training diffusion transformers [22, 91], taking inspiration from the success of patch masking in contrastive models [26]. Patch masking is straightforward to implement in transformers, and the diffusion transformer successfully generalizes to unmasked images in inference, even when patches for each image were masked randomly during training.

B. Additional Details on Experimental Setup

We use `DiT-Tiny/2` and `DiT-Xl/2` diffusion transformer architectures [52] for small and large scale training setups, respectively. We use four and six transformer blocks in the patch-mixer for the `DiT-Tiny/2` and `DiT-Xl/2` architectures, respectively. The patch-mixer comprises approximately 13% of the parameters in the backbone diffusion transformer. We use half-precision layernorm normalization and SwiGLU activation in the feedforward layers for all transformers. Initially, half-precision layernorm led to training instabilities after 100K steps of training. Thus, we further apply layer normalization to query and key embeddings in the attention layers [11], which stabilizes the training. We reduce the learning rate for expert layers by half, as each expert now processes a fraction of all patches. We provide an exhaustive list of our training hyperparameters in Table 3. We use the default configuration for EDM [37] diffusion framework ($\sigma_{\text{max}} = 80, \sigma_{\text{min}} = 0.002, S_{\text{max}} = \infty, S_{\text{noise}} = 1, S_{\text{min}} = 0, S_{\text{churn}} = 0$), except that we increase σ_{data} to match the standard deviation of our image datasets in latent space. We generate images using deterministic sampling from Heun's 2^{nd} order method [20, 37] with 30 sampling steps. Unless specified, we use classifier-free guidance of 3 and sample 30K images in quantitative evaluation on the cifar-captions dataset. We reduce the guidance scale to 1.5 for large-scale models. We find that these guidance values achieve the best FID score under the trade-off of FID and Clip-score. In qualitative generations, we recommend using a guidance scale of 5 for better

photorealism and prompt adherence. By default, we use 512×512 pixel resolution when generating synthetic images from large-scale models and 256×256 pixel resolution with small-scale models trained on the cifar-captions dataset.

We consider the computation of text embeddings for captions and latent compression of images as a one-time cost that amortizes over multiple training runs of a diffusion model. Thus, we compute them offline and do not account for these costs in our estimation of training costs. We use a four-channel variational autoencoder (VAE) from the Stable-Diffusion-XL [54] model to extract image latents. We also consider the latest 16-channel VAEs to test their performance in large-scale micro-budget training [17]. We use the EDM framework from Karras et al. [37] as a unified training setup for all diffusion models.

We conduct all experiments on an 8×H100 GPU node. We train and evaluate all models using `bfloat16` mixed-precision mode. We did not observe a significant speedup when using FP8 precision with the Transformer Engine library[2]. We use PyTorch 2.3.0 [51] with PyTorch native flash-attention-2 implementation. We use the DeepSpeed [59] flops profiler to estimate total FLOPs in training. We also use just-in-time compilation (`torch.compile`) to achieve a 10-15% speedup in training time. We use StreamingDataset [79] to enable fast data loading.

Table 2. Comparing the computational cost and storage overhead of CLIP [18, 55] and T5 [56, 77] text encoders. We use the state-of-the-art CLIP model from Fang et al. [18]. We report the compute and storage cost for one million image captions. Even though T5 embeddings achieve better generation quality, especially for text generation, computing them is an order of magnitude slower than CLIP embeddings and even precomputing them for our dataset (37M images) poses high storage overheads (36.4 TB). In this table, we use one H100 GPU to estimate the time to process 1M image captions. We save the embeddings in `float16` precision.

Text encoder	Sequence length	Embedding size	Time (min:sec)	Storage (GB)
CLIP (`ViT-H-14`)	77	1024	3:20	157
T5 (`T5-xxl`)	120	4096	33:04	983

Choice of text encoder. To convert discrete token sequences in captions to high-dimensional feature embeddings, two common choices of text encoders are CLIP [33, 55] and T5 [56], with T5-xxl[3] embeddings narrowly outperforming equivalent CLIP model embeddings [61]. However, T5-xxl embeddings pose both compute and storage challenges: 1) Computing T5-xxl embeddings costs an order of magnitude more time than a CLIP ViT-H/14 model while also requiring 6× more disk space for pre-computed embeddings. Overall, using T5-xxl embeddings is more demanding (Table 2). Following the observation that large text encoders trained on higher quality data tend to perform better [61], we use state-of-the-art CLIP models as text encoders [18][4].

Training process. We use a `DiT-Xl/2` transformer with eight experts in alternate transformer blocks for our large-scale training setup. We provide details on training hyperparameters in Table 3. We conduct the training in following two phases. We refer to any diffusion transformer trained using our micro-budget training as `MicroDiT`.

- *Phase-1:* In this phase, we pretrain the model on 256×256 resolution images. We train for 250K optimization steps with 75% patch masking and finetune for another 30K steps without any patch masking.
- *Phase-2:* In this phase, we finetune the Phase-1 model on 512×512 resolution images. We first finetune the model for 50K steps with 75% patch masking followed by another 5K optimization steps with no patch masking.

Cost analysis. We translate the wall-clock time of training to financial cost using a $30/hour budget for an 8×H100 GPU cluster. Since the cost of H100 fluctuates significantly across vendors, we base our estimate on the commonly used cost estimates for A100 GPUs [8], in particular $1.5/A100/hour[5]. We observe a 2.5x reduction in wall-clock time on H100 GPUs, thus assume a $3.75/H100/hour cost. We also report the wall-clock time to benchmark the training cost independent of the fluctuating GPU costs in the AI economy.

Datasets. We use the following five datasets, comprising a total of 37 million images, to train our large-scale models.

- *Conceptual Captions (real).* This dataset was released by Google and includes 12M pairs of image URLs and captions [7]. Our downloaded version of the dataset includes 10.9M image-caption pairs.
- *Segment Anything-1B (real).* Segment Anything comprises 11.1M high-resolution images originally released by Meta for segmentation tasks [40]. Since the images do not have corresponding real captions, we use synthetic captions generated by the LLaVA model [8, 44].
- *TextCaps (real).* This dataset comprises 28K images with natural text in the images [68]. Each image has five associated descriptions. We combine them into single captions using the Phi-3 language model [1].
- *JourneyDB (synthetic).* JourneyDB is a synthetic image dataset comprising 4.4M high-resolution Midjourney image-caption pairs [76]. We use the train subset (4.2M samples) of this dataset.
- *DiffusionDB (synthetic).* DiffusionDB is a collection of 14M synthetic images generated primarily by Stable Diffusion models [85]. We filter out poor quality images

[2] https://github.com/NVIDIA/TransformerEngine
[3] https://huggingface.co/DeepFloyd/t5-v1_1-xxl
[4] We use the DFN-5B text encoder: https://huggingface.co/apple/DFN5B-CLIP-ViT-H-14-378
[5] https://cloud-gpus.com/

Table 3. **Training hyperparameters.** Hyperparameters across both phases of our large-scale training setup.

Resolution	256×256 (Phase-I)		512×512 (Phase-II)	
Masking ratio	0.75	0	0.75	0
Training steps	250000	30000	50000	5000
Batch size	2048	2048	2048	2048
Learning rate	2.4×10^{-4}	8×10^{-5}	8×10^{-5}	8×10^{-5}
Weight decay	0.1	0.1	0.1	0.1
Optimizer	AdamW	AdamW	AdamW	AdamW
Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$
Optimizer epsilon	1×10^{-8}	1×10^{-8}	1×10^{-8}	1×10^{-8}
Lr scheduler	Cosine	Constant	Constant	Constant
Warmup steps	2500	0	500	0
Gradient clip norm	0.25	0.25	0.5	0.5
EMA	no ema	no ema	0.99975	0.9975
Hflip	False	False	False	False
Precision	bf16	bf16	bf16	bf16
Layernorm precision	bf16	bf16	bf16	bf16
QK-normalization	True	True	True	True
$(P_{\text{mean}}, P_{\text{std}})$	$(-0.6, 1.2)$	$(-0.6, 1.2)$	$(0, 0.6)$	$(0, 0.6)$

from this dataset, resulting in 10.7M samples, and use this dataset only in the first phase of training.

CIFAR-Captions. We construct a text-to-image dataset, named cifar-captions, that closely resembles the existing web-curated datasets and serves as a drop-in replacement of existing datasets in our setup. Cifar-captions is closed-domain and only includes images of ten classes (airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks), imitating the widely used CIFAR-10 classification dataset [41]. In contrast to other small-scale text-to-image datasets that are open-world, such as subsets of conceptual captions [7], we observe fast convergence of diffusion models on this dataset. We create this dataset by first downloading 120M images from the coyo-700M [6] dataset. We observe a success rate of approximately 60% at the time of downloading the dataset. Next, we use a ViT-H-14 [18] to measure CLIP score by averaging it over the eighteen prompt templates (such as 'a photo of a {}.') used in the original CLIP model [55]. We select images with CLIP scores higher than 0.25 (1.25% acceptance rate) that results in a total of 1.3M images. As the real captions are highly noisy and uninformative, we replace them with synthetic captions generated with an LLaVA-1.5 model [44].

Billion-image synthetic dataset. To capture the asymptotic performance of training solely on synthetic images, we also consider a dataset comprising 1B images generated from diffusion models[6]. It mainly includes generations from various versions of stable diffusion models and their adaptations.

Evaluation metrics. We use the following evaluation metrics to assess the quality of synthetic images generated by the text-to-image models.

- *FID*. Fréchet Inception Distance (FID) measures the 2-Wasserstein distance between real and synthetic data distributions in the feature space of a pretrained vision model. We use the clean-fid[7] [50] implementation for a robust evaluation of the FID score.
- *Clip-FID*. Unlike FID that uses an Inception-v3 [78] model trained on ImageNet [12], Clip-FID uses a CLIP [55] model, trained on a much larger dataset than ImageNet, as an image feature extractor. We use the default ViT-B/32 CLIP model from the clean-fid library to measure Clip-FID.
- *Clip-score*. It measures the similarity between a caption and the generated image corresponding to the caption. In particular, it measures the cosine similarity between normalized caption and image embeddings calculated using a CLIP text and image encoder, respectively.

Human preference based evaluation. We ask each individual to choose one of the two given images based on the following criteria: 1) choose based on image quality in the absence of the prompt, and 2) choose based on the alignment between the prompt and the generated images. We use the identical prompt to sample from both models. We report the win rate based on the percentage of comparisons in which the image from the winning model is preferred over the other model. We further follow the evaluation pipeline of Betker et al. [4] and use DrawBench extended prompts to generate samples. We slightly compress these prompts to accommodate the 77-token context window of CLIP text encoders. We list these prompts at the end of the paper.

C. Background on Layer-wise Scaling and Mixture-of-experts

Layer-wise scaling. In contrast to using identical transformer blocks throughout the network, layer-wise scaling

[6] https://huggingface.co/datasets/bigdata-pw/Diffusion1B

[7] https://github.com/GaParmar/clean-fid

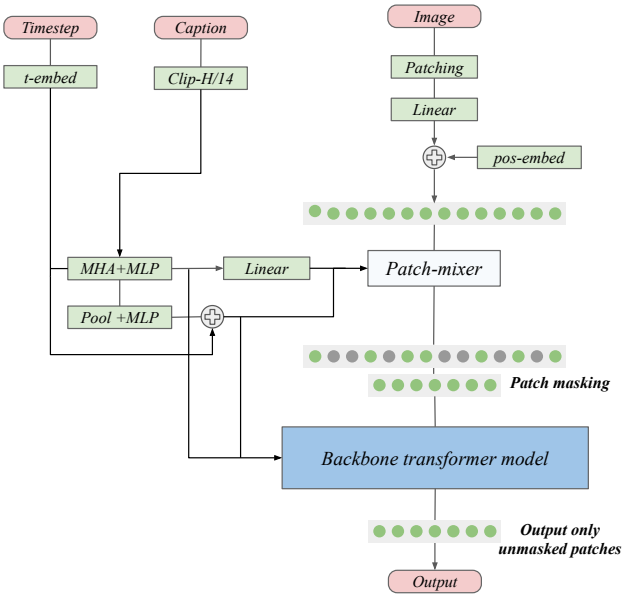


Figure 6. **Overall architecture of our diffusion transformer.** We prepend the backbone transformer model with a lightweight patch-mixer that operates on all patches in the input image before they are masked. Following contemporary works [4, 17], we process the caption embeddings using an attention layer before using them for conditioning. We use sinusoidal embeddings for timesteps. Our model only denoises unmasked patches, thus the diffusion loss (Eq. 3) is calculated only on these patches. We modify the backbone transformer using layer-wise scaling on individual layers and use mixture-of-expert layers in alternate transformer blocks.

dynamically increases the transformer block size with the depth of the network. In a text-to-image transformer, each block consists of a self-attention layer, a cross-attention layer, and a feedforward layer. Layer-wise scaling increases the hidden dimension of feedforward layers by a factor m_f, thus linearly increasing the number of parameters and FLOPs of the feedforward block. Similarly, layer-wise scaling dynamically increases the number of heads in the attention layers by scaling the embedding size by a factor of m_a. By default, $m_a = 1.0$ and $m_f = 4.0$ lead to a canonical transformer network. We use $m_a \in [0.5, 1.0]$ and $m_f \in [0.5, 4.0]$ to reduce the size of transformer blocks. Note that we do not apply any scaling to cross-attention layers ($m_a = 1.0$, $m_f = 4.0$) to avoid degrading the controllability of captions on image generation.

Mixture-of-experts. Mixture-of-experts enable the construction of much larger transformers, referred to as sparse transformers, with minimal impact on the training and inference cost, thus making them highly applicable for micro-budget training. A sparse model modifies the transformer block to include replicas of the feedforward layer, referred to as experts. The input patch embeddings to the feedforward layers are first fed into a router that determines the configuration of patches processed by each expert. We use the expert-choice (EC) routing mechanism [92] where each expert selects the top-k patches using the importance score determined by the routing network for each expert (Figure 7). We favor EC routing over conventional routing due to its simplicity, as it does not require any auxiliary loss to balance the load across experts [67, 93]. We use a linear layer as a router and it is trained jointly with the rest of the network.

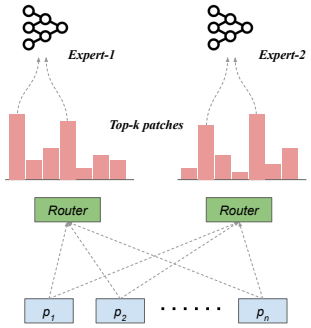


Figure 7. **Mixture-of-experts (MoE).** Expert-choice routing based mixture-of-experts [92]. Each patch is passed to a patch router that determines the top-k patches routed to each expert.

D. Additional Results on Effectiveness of Deferred Masking

In this section, we supplement the results and discussion from the main paper on evaluating the effectiveness of deferred masking. We use the `DiT-Tiny/2` model and the cifar-captions dataset (256×256 image resolution) for all experiments in this section. We train each model for 60K optimization steps using the AdamW optimizer and exponential moving average with a 0.995 smoothing coefficient enabled for the last 10K steps.

D.1. Out-of-the-box performance: Making high masking ratios feasible with deferred masking

In this section, we provide further details on the out-of-the-box performance of deferred masking. We evaluate its out-of-the-box performance with common training parameters for up to 87.5% masking ratios. As a baseline, we train a network with no patch-mixer, i.e., naive masking (Figure 2c) for each masking ratio.

We use commonly used default hyperparameters to simulate the out-of-the-box performance for both our approach and the baseline. We train both models with the AdamW optimizer with an identical learning rate of 1.6×10^{-4}, 0.01 weight decay, and a cosine learning rate schedule. We set $(P_{\text{mean}}, P_{\text{std}})$ to $(-1.2, 1.2)$ following the original work [37]. We provide the results in Figure 8.

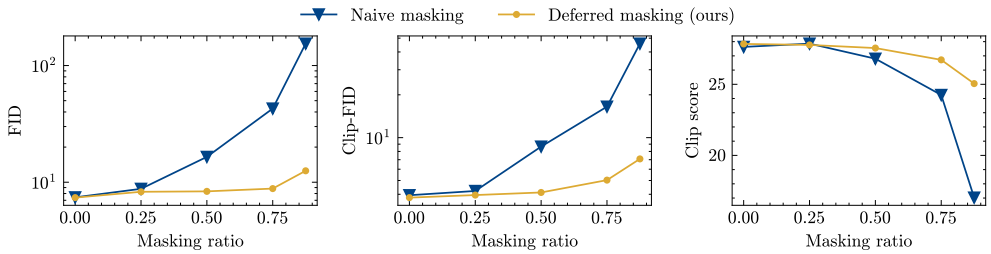


Figure 8. **Out-of-the-box performance of deferred masking.** Without any hyperparameter optimization, we compare the performance of our deferred masking with a naive masking strategy. We find that deferred masking, i.e., using a patch-mixer before naive masking, tremendously improves image generation performance, particularly at high masking ratios.

(a) *Ablating the choice of β_2 in AdamW optimizer.* Unlike the LLM training where β_2 if often set to 0.95 [5, 81], we find that image generation quality consistently degrades as we reduce β_2.

β_2	FID (↓)	Clip-FID (↓)	Clip-score (↑)
0.999	**8.53**	**4.85**	**26.88**
0.99	8.63	4.94	26.75
0.95	8.71	5.02	26.71
0.9	8.81	5.13	26.61

(b) *Ablating the choice of weight-decay in `AdamW` optimizer.* In resonance with transformer training in LLMs, we observe improvement in performance with increase in weight decay regularization.

wd	FID (↓)	Clip-FID (↓)	Clip-score (↑)
0.00	8.77	5.03	26.82
0.01	8.73	5.03	26.82
0.03	8.53	**4.85**	26.88
0.10	**8.38**	4.90	**27.00**

(c) *Ablating the parameters of noise distribution.* We used patch-mixer with 0.999 β_2 and 0.1 weight decay. We observe a tradeoff between FID and Clip-score in first two choices and set (m, s) to $(-0.6, 1.2)$ in all followup experiments.

(m, s)	FID (↓)	Clip-FID (↓)	Clip-score (↑)
$(-1.2, 1.2)$	**8.38**	**4.90**	27.00
$(-0.6, 1.2)$	8.49	4.93	**27.47**
$(-0.6, 0.6)$	9.05	6.72	26.95
$(-0.25, 0.6)$	10.44	7.51	27.46
$(0.0, 0.6)$	12.76	9.00	27.40

(d) *Ablating the block-size in block sampling.* We observe consistent performance degradation with block masking (at 75% masking ratio). At block size of 8 (latent image with $16 \times 16 = 256$ patches), block sampling sampling collapse to sampling a quadrant, thus we sample a single continuous square patch for it.

$Block - size$	FID (↓)	Clip-FID (↓)	Clip-score (↑)
1	**8.49**	**4.93**	**27.47**
2	8.89	4.64	27.42
4	9.80	5.09	26.91
square	12.71	7.11	26.14

(e) *Ablating the size of patch-mixer.* We increase the width of path embeddings, while also varying the multipliers for attention layers (m_a) and feedforward layers (m_f). We also report the total wall-clock time (in hours:minutes) for training.

(w, m_a, m_b)	Time	FID (↓)	Clip-FID (↓)	Clip-score (↑)
$(384, 0.5, 0.5)$	3:19	8.49	4.93	27.47
$(384, 1.0, 4.0)$	3:20	7.72	4.80	27.72
$(512, 1.0, 4.0)$	3:24	7.40	4.57	**27.79**
$(768, 1.0, 4.0)$	3:39	**7.09**	**4.39**	27.76

(f) *Ablating the choice of feedforward layers: GELU vs SwiGLU activation.* We find that replacing GELU activation with a SwiGLU based activation in feedforward layer improves all three performance metrics. SwiGLU activation is also commonly used in transformers for large language models [81].

$Block - size$	FID (↓)	Clip-FID (↓)	Clip-score (↑)
GELU	7.82	4.61	27.33
SwiGLU	**7.40**	**4.57**	**27.79**

(g) *Testing the cyclic learning rate scheduler.* We consider the cosine learning rate with warm restarts, while varying the duration of the base cycle (t) and the multiplier (m_t), where the duration of each subsequent cycle increases by the multiplier factor.

lr schedule	FID (↓)	Clip-FID (↓)	Clip-score (↑)
1-cycle ($t = 60{,}000, m_t = 1$)	**7.40**	4.57	27.79
3-cycles ($t = 20{,}000, m_t = 1$)	7.64	4.52	27.75
4-cycles ($t = 4{,}000, m_t = 2$)	7.49	**4.43**	**27.81**

(h) *Using higher learning rate for each batch for better performance.* We observe training instabilities after 2K steps for learning rates higher than 3.2×10^{-4} in mixed-precision training.

lr	FID (↓)	Clip-FID (↓)	Clip-score (↑)
1.6×10^{-4}	7.40	4.57	27.79
3.2×10^{-4}	**7.09**	**4.10**	**28.24**

Figure 9. **Ablating individual components in our training pipeline.** In each subsequent ablation, we use the overall best performing model from previous ablation. In each ablation, we train a `DiT-Tiny/2` model for 60K training steps using 75% masking ratio.

D.2. Ablation study of our training pipeline

In this section, we further discuss the improvements over out-of-the-box performance (Figure 9). When ablating the learning rate, we observe better performance with higher learning rates. However, we run into training instabilities after approximately 2K steps, despite using qk-normalization [11], with mixed precision training. Overall, we recommend ablating the choice of learning rate for each network and using the maximum attainable value without causing training instability. We also consider a fast training strategy using cyclic learning schedules [69]. Cyclic learning has been previously observed to improve the convergence rate for image classifiers [69, 70]. We consider a cyclic cosine schedule with base cycle time t and each subsequent cycle duration multiplied by a multiplier (m_t). We only observe a marginal benefit of cyclic schedules (Table 9g). In favor of simplicity, we continue to use a single cycle schedule.

Comparison with training hyperparameters of LLMs. Since the diffusion transformer architectures are very similar to transformers used in large language models (LLMs), we compare the hyperparameter choices across the two tasks. Similar to common LLM training setups [9, 34, 81], we find that SwiGLU activation [66] in the feedforward layer outperforms the GELU [27] activation function. Similarly, higher weight decay leads to better image generation performance. However, we observe better performance when using a higher running average coefficient for the AdamW second moment (β_2), in contrast to large-scale LLMs where $\beta_2 \approx 0.95$ is preferred. As we use a small number of training steps, we find that increasing the learning rate to the maximum possible value until training instabilities also significantly improves image generation performance.

Design choices in masking and patch-mixer. We observe a consistent improvement in performance with a larger patch mixer. However, despite the better performance of larger patch-mixers, we choose to use a small patch mixer to lower the computational budget spent by the patch mixer in processing the unmasked input. We also update the noise distribution $(P_{\text{mean}}, P_{\text{std}})$ to $(-0.6, 1.2)$ as it improves the alignment between captions and generated images. In ablating the choice of masking, we shift from masking patches randomly to retaining a continuous region of image patches with block masking 3b. We perform this ablation on 256×256 image resolution with a corresponding latent resolution of 32×32 and 256 patches with a patch size of two in the `DiT-Tiny/2` model. Note that at a block size of 8 and a 75% masking ratio, block sampling collapses to masking quadrants of image patches. Thus, for this configuration, we resort to sampling a single continuous square patch of non-masked patches. Overall, we find that any amount of block masking degrades performance compared to random masking of each patch. It is likely because despite latent compression and patching, there exists some redundancy in visual information between neighboring patches, and the diffusion transformer model receives more global semantic information about the image with random masking.

Table 4. **Using layer-wise scaling.** Layer-wise scaling of transformer architecture is a better fit for masked training in diffusion transformers. We validate its effectiveness in the canonical naive masking with 75% masking ratio.

Arch	FID (↓)	Clip-FID (↓)	Clip-score (↑)
Constant width	19.6	9.9	26.7
Layer-wise scaling	**15.9**	**7.4**	**27.1**

D.3. Validating improvements in diffusion transformer architecture

Layer-wise scaling. We investigate the impact of this design choice in a standalone experiment where we train two variants of a `DiT-Tiny` architecture, one with a constant width transformer and the other with layer-wise scaling of each transformer block. We use naive masking for both methods. We select the width of the constant-width transformer such that its computational footprint is identical to layer-wise scaling. We train both models for an identical number of training steps and wall-clock time. We find that the layer-wise scaling approach outperforms the baseline constant width approach across all three performance metrics (Table 4), demonstrating a better fit of the layer-wise scaling approach in masked training of diffusion transformers.

Mixture-of-experts layers. We train a `DiT-Tiny/2` transformer with expert-choice routing based mixture-of-experts (MoE) layers in each alternate transformer block. On the small-scale setup, it achieves similar performance to baseline model trained without MoE blocks. While it slightly improves Clip-score from 28.11 to 28.66, it degrades FID score from 6.92 to 6.98. We hypothesize that the lack of improvement is due to the small number of training steps (60K), as using multiple experts effectively reduces the number of samples observed by each router. In top-2 routing for 8-experts, each expert is trained effectively for one fourth number of epochs over the dataset.

D.4. Deferred masking as pretraining + unmasked finetuning.

We find that deferred masking also acts as a strong pretraining objective, and using it with an unmasked finetuning schedule achieves better performance than training an unmasked network under an identical computational budget. We first train a network without any masking and another identical network with 75% deferred masking. We finetune the latter with no masking and measure performance as we increase the number of finetuning steps. We mark the

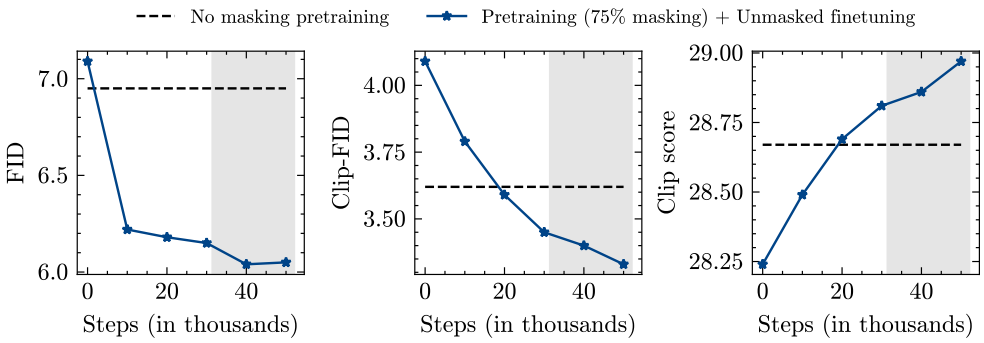


Figure 10. **Deferred masking in pretraining**. We test the advantage of using deferred masking with finetuning over training a model with no masking for a given computational budget. After pretraining with 75% deferred masking, we increase the number of unmasked finetuning steps and compare its performance with another model trained completely without masking. The shaded region represents steps after the isoflops threshold, where deferred masking pretraining and finetuning have higher computation cost. At the isoflops threshold, the finetuned model achieves better performance than training the diffusion model without any masking.

Masking ratio	Masking ratio	FID	Clip-FID	Clip-score
Downscaled network	-	**6.60**	3.85	28.49
Deferred masking	0.5	6.74	**3.45**	**28.86**
Downscaled network	-	7.09	4.56	27.68
Deferred masking	0.75	**6.96**	**4.17**	**28.16**
Downscaled network	-	**7.52**	**5.03**	26.98
Deferred masking	0.875	8.35	5.26	**27.00**

Table 5. **IsoFLOPs training.** Both patch masking and model downscaling, i.e., reducing the size of the model, reduce the computational cost in training and can be complementary to each other. However, it is natural to ask how the two paradigms compare to each other in training diffusion transformers. Under identical training setup and wall-clock time, we compare the effectiveness of model downscaling and our deferred masking approaches. We find that except at extremely high masking ratios, deferred masking achieves better performance across at least two performance metrics. Based on this finding, we do not use a masking ratio higher than 75% in our models.

isoflops threshold when the combined cost of masked pretraining and unmasked finetuning is identical to the model trained with no masking. We find that at the isoflops threshold, the finetuned model achieves superior performance across all three performance metrics. The performance of the model also continues to improve with unmasked finetuning steps beyond the isoflops threshold.

E. Additional Results on Micro-budget Training of Large-scale Models

Training MicroDiT in higher dimensional latent space. We replace the default four-channel autoencoder with one that has sixteen channels, resulting in a $4\times$ higher dimensional latent space. Recent large-scale models have adopted high dimensional latent space as it provides significant improvements in image generation abilities [10, 17]. Note that the autoencoder with higher channels itself has superior image reconstruction capabilities, which further contributes to overall success. Intriguingly, we find that using a higher dimensional latent space in micro-budget training hurts performance. For two `MicroDiT` models trained with identical computational budgets and training hyperparameters, we find that the model trained in four-channel latent space achieves better FID, Clip-score, and GenEval scores (Table 7). We hypothesize that even though an increase in latent dimensionality allows better modeling of data distribution, it also simultaneously increases the training budget required to train higher-quality models.

E.1. Challenge with canonical evaluation metrics in determining effect of synthetic data.

To determine the effect of synthetic data, we train two models under identical training setup and cost: 1) model trained only on real images (total 22M images) 2) model trained on the combined real and synthetic images (total 37M images).

Under canonical performance metrics, both models apparently achieve similar performance. For example, the model trained on real-only data achieved an FID score of 12.72 and a CLIP score of 26.67, while the model trained on both real and synthetic data achieved an FID score of 12.66 and a CLIP score of 28.14. Even on GenEval [23], a benchmark that evaluates the ability to generate multiple objects and modelling object dynamics in images, both models achieved an identical score of 0.46. These results seemingly suggest that incorporating a large amount of synthetic samples didn't yield any meaningful improvement in image generation capabilities.

However, we argue that this observation is heavily influenced by the limitations of our existing evaluation metrics. In a qualitative evaluation, we found that the model trained

Table 6. **Breakdown of computational cost.** Computational cost of the two-stage training of our large-scale model. Our total computational cost is 3.45×10^{20} FLOPs, amounting to a total cost of $1,890 and 2.6 training days on an 8×H100 GPU machine.

Resolution	Masking ratio	Training steps	Total FLOPs	8×A100 days	8×H100 days	Cost ($)
256×256	0.75	250000	1.47×10^{20}	2.77	1.11	800
	0.00	30000	4.53×10^{19}	0.94	0.38	271
512×512	0.75	50000	1.18×10^{20}	2.18	0.88	630
	0.00	5000	3.48×10^{19}	0.65	0.26	189

(a) Measuring fidelity and prompt alignment of generated images on COCO dataset.

Channels	FID-30K (↓)	Clip-FID-30K (↓)	Clip-score (↑)
4	**12.65**	**5.96**	**28.14**
16	13.04	6.84	25.63

(b) Measuring performance on the GenEval benchmark.

		Objects					
Channels	Overall	Single	Two	Counting	Colors	Position	Color attribution
4	**0.46**	**0.97**	**0.47**	**0.33**	**0.78**	**0.09**	**0.20**
16	0.40	0.96	0.36	0.27	0.72	0.07	0.09

Table 7. **Why prefer 4-channel image encoders over 16-channel image encoders in micro-budget training?** We ablate the dimension of latent space by training our `MicroDiT` models in four and sixteen channel latent space, respectively. Even though training in higher dimensional latent space is being adopted across large-scale models [10, 17] we find that it underperforms when training on a micro-budget.

Table 8. **Zero-shot FID on COCO2014 validation split**. We report total training time in terms of number of days required to train the model on a machine with eight A100 GPUs. We observe a 2.5× reduction in training time when using H100 GPUs. Our micro-budget training takes 14.2× less training time than state-of-the-art low-cost training approach while simultaneously achieving competitive FID compared to some open-source models.

Model	Params (↓)	Sampling steps (↓)	Open-source	Training images(↓)	8×A100 GPU days (↓)	FID-30K (↓)
CogView2 [15]	6.00B	–	–	–	–	24.0
Dall-E [57]	12.0B	256	–	–	–	17.89
Glide [48]	3.50B	250	–	–	–	12.24
Parti-750M [87]	0.75B	1024	–	3690M	–	10.71
Dall-E.2 [58]	6.50B	–	–	650M	5208.3	10.39
Make-a-Scene [21]	4.00B	1024	–	–	–	11.84
GigaGAN [36]	1.01B	1	–	980M	597.8	9.09
ImageN [61]	3.00B	–	–	860M	891.5	7.27
Parti-20B [87]	20.0B	1024	–	3690M	–	7.23
eDiff-I [2]	9.10B	25	–	11470M	–	6.95
Stable-Diffusion-2.1[a][60]	0.86B	50	✓	3900M	1041.6	9.12
Stable-Diffusion-1.5 [60]	0.86B	50	✓	4800M	781.2	11.18
Würstchen [53]	0.99B	60	✓	1420M	128.1	23.60
PixArt-α [8]	0.61B	20	✓	25M[b]	94.1[c]	7.32
MicroDiT (our work)	1.16B	30	✓	37M	**6.6**	12.66

[a] As the FID scores for the stable diffusion models are not officially reported [60], we calculate them using the official release of each model. We achieve slightly better FID scores compared to the scores reported in Würstchen [53]. We use our FID scores to represent the best performance of these models.

[b] Includes 10M proprietary high-quality images.

[c] PixArt-α training takes 85 days on an 8×A100 machine when only training till 512×512 resolution.

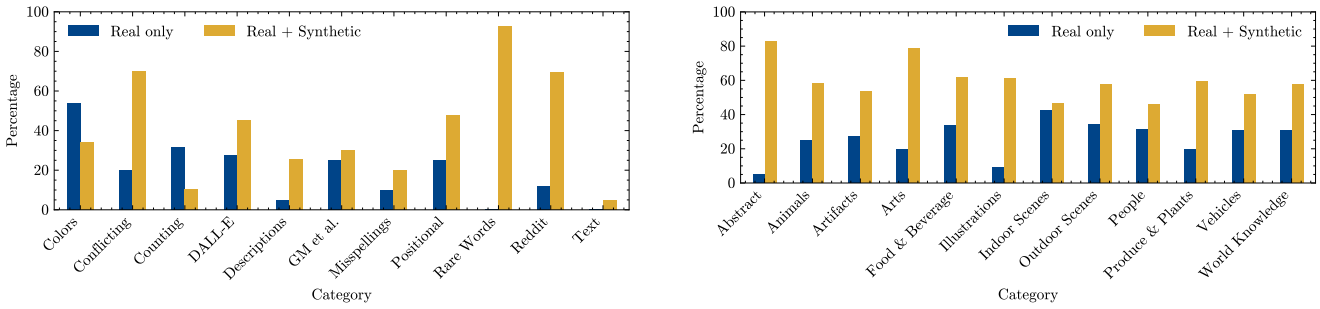
on the combined dataset achieved much better image quality (Figure 13). The real data model often fails to adhere to the prompt, frequently hallucinating key details and often failing to synthesize the correct object. Metrics, such as FID, fail to capture this difference because they predominantly measure distribution similarity [53]. Thus, we focus on using human visual preference as an evaluation metric. To automate the process, we use GPT-4o [49], a state-of-the-art multimodal model, as a proxy for human preference. We supply the following prompt to the model: *Given the prompt '{prompt}', which image do you prefer, Image A or Image B, considering factors like image details, quality, realism,*

Table 9. Comparing performance on compositional image generation using GenEval [23] benchmark. Table adopted from Esser et al. [17]. Higher performance is better.

Model	Open-source	Overall	Objects		Counting	Colors	Position	Color attribution
			Single	Two				
DaLL-E.2 [58]	–	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DaLL-E.3 [4]	–	0.67	0.96	0.87	0.47	0.83	0.43	0.45
minDALL-E [39]	✓	0.23	0.73	0.11	0.12	0.37	0.02	0.01
Stable-Diffusion-1.5 [60]	✓	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-α [8]	✓	0.48	0.98	0.50	0.44	0.80	0.08	0.07
Stable-Diffusion-2.1 [60]	✓	0.50	0.98	0.51	0.44	0.85	0.07	0.17
Stable-Diffusion-XL [54]	✓	0.55	0.98	0.74	0.39	0.85	0.15	0.23
Stable-Diffusion-XL-Turbo [63]	✓	0.55	1.00	0.72	0.49	0.80	0.10	0.18
IF-XL	✓	0.61	0.97	0.74	0.66	0.81	0.13	0.35
Stable-Diffusion-3 [17]	✓	0.68	0.98	0.84	0.66	0.74	0.40	0.43
MicroDiT (our work)	✓	0.46	0.97	0.47	0.33	0.78	0.09	0.20

Table 10. **Scaling synthetic dataset to half billion images**. Comparing performance on compositional image generation using GenEval [23] benchmark for `MicroDiT` models trained with different configuration of training dataset.

Model (training dataset)	Overall	Objects		Counting	Colors	Position	Color attribution
		Single	Two				
MicroDiT (37M - combined real and synthetic images)	0.46	0.97	0.47	0.33	0.78	0.09	0.20
MicroDiT (490M - only synthetic images)	0.52	0.98	0.61	0.31	0.89	0.11	0.17



(a) DrawBench (b) PartiPrompts

Figure 12. **Assessing image quality with GPT-4o.** We supply the following prompt to the GPT-4o model: *Given the prompt '{prompt}', which image do you prefer, Image A or Image B, considering factors like image details, quality, realism, and aesthetics? Respond with 'A' or 'B' or 'none' if neither is preferred.* For each comparison, we also shuffle the order of images to remove any ordering bias. We evaluate the performance on two prompt databases: DrawBench [61] and PartiPrompts [87]. The y-axis in the bar plots indicates the percentage of comparisons in which image from a model is preferred. We breakdown the comparison across individual image category in each prompt database.

and aesthetics? Respond with 'A' or 'B' or 'none' if neither is preferred. For each comparison, we also shuffle the order of images to remove any ordering bias. We generate samples using DrawBench [61] and PartiPrompts (P2) [87], two commonly used prompt databases (Figure 12). On the P2 dataset, images from the combined data model are preferred in 63% of comparisons while images from the real data model are only preferred 21% of the time (16% of comparisons resulted in no preference). For the DrawBench dataset, the combined data model is preferred in 40% of comparisons while the real data model is only preferred in 21% of comparisons. Overall, using a human preference-centric metric clearly demonstrates the benefit of additional synthetic data in improving overall image quality.

(a) a photograph of an astronaut riding a horse; An astronaut riding a pig, highly realistic DSLR photo, cinematic shot; Panda mad scientist mixing sparkling chemicals, artstation

(b) A tiger made of white lego sitting in a realistic, natural field

(c) Portrait of President Obama in style of Vincent Van Gogh starry night; A bird in the style of Claude Monet Lilies painting; A sea otter in the style of girl with pearl earring painting by Johannes Vermeer.

(d) 'A elephant dressed in a suit posing for a photo in ___ style'. Styles are Origami, Pixel art, and Line art

(e) A giant cobra snake made from corn.; A giant cobra snake made from peas.; A giant cobra snake made from sushi.

(f) four Yellow apples lying on a red table; A sports car painted with vibrant colors; A natural scene with 1000 colors

Figure 13. **Comparing generations from our three micro-budget models.** We compare synthesized images from our large-scale models trained with different configurations of the training dataset but with identical computational costs.

(a) a photo of a cow

(b) a photo of a cake and a zebra

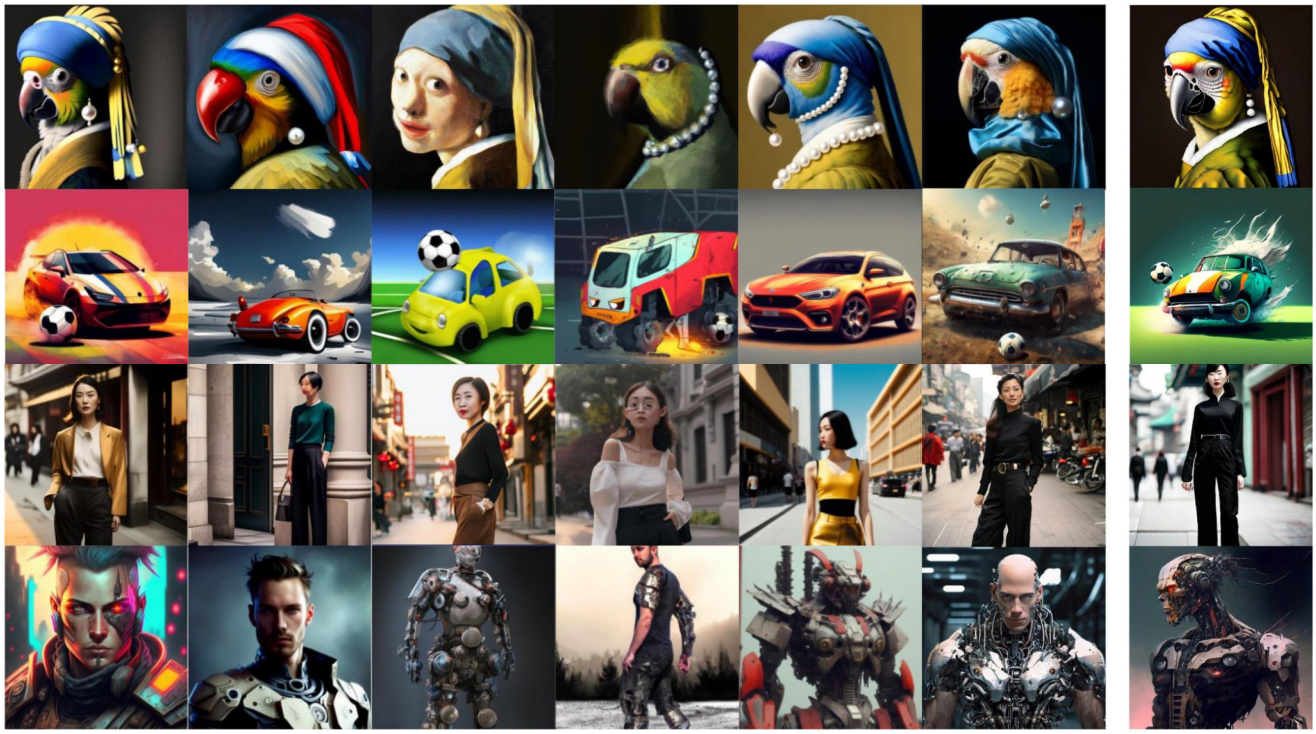
(c) a photo of three buses

(d) a photo of a red potted plant

(e) a photo of an elephant below a surfboard

(f) a photo of an orange cow and a purple sandwich

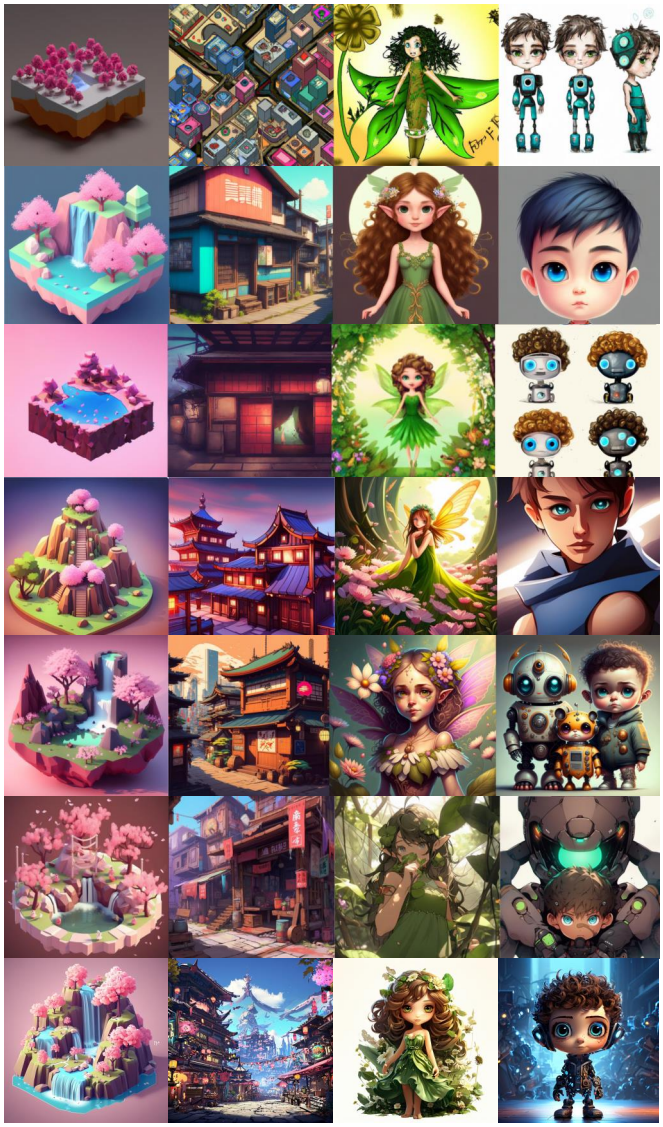
Figure 14. **Comparing generations from our first two micro-budget models.** Generation from a model trained only on real data (on left) and combined real and synthetic data (on right) on GenEval benchmark prompts. Both models use identical random seed for generation.



RAPHAEL *Stable Diffusion XL* *DeepFloyd* *DALL-E-2* *ERNIE-ViLG 2.0* *PixArt-α* ***Ours***

- A parrot with a pearl earring, Vermeer style.
- A car playing soccer, digital art.
- Street shot of a fashionable Chinese lady in Shanghai, wearing black high-waisted trousers.
- Half human, half robot, repaired human, human flesh warrior, mech display, man in mech, cyberpunk.

Figure 15. Comparison with previous works. Figure adapted from Chen et al. [8].



(a) *Previous works*: RAPHAEL, Stable Diffusion XL, DeepFloyd, DALL-E-2, ERNIE-ViLG 2.0, PixArt-α

(b) Our work

- A cute little matte low poly isometric cherry blossom forest island, waterfalls, lighting, soft shadows, trending on Artstation, 3d render, monument valley, fez video game.
- A shanty version of Tokyo, new rustic style, bold colors with all colors palette, video game, genshin, tribe, fantasy, overwatch.
- Cartoon characters, mini characters, figures, illustrations, flower fairy, green dress, brown hair,curly long hair, elf-like wings,many flowers and leaves, natural scenery, golden eyes, detailed light and shadow , a high degree of detail.
- Cartoon characters, mini characters, hand-made, illustrations, robotkids, color expressions, boy, short brown hair, curly hair,blue eyes, technological age, cyberpunk, big eyes, cute, mini, detailed light and shadow, high detail.

Figure 16. Comparison with previous works. Figure adapted from Chen et al. [8].

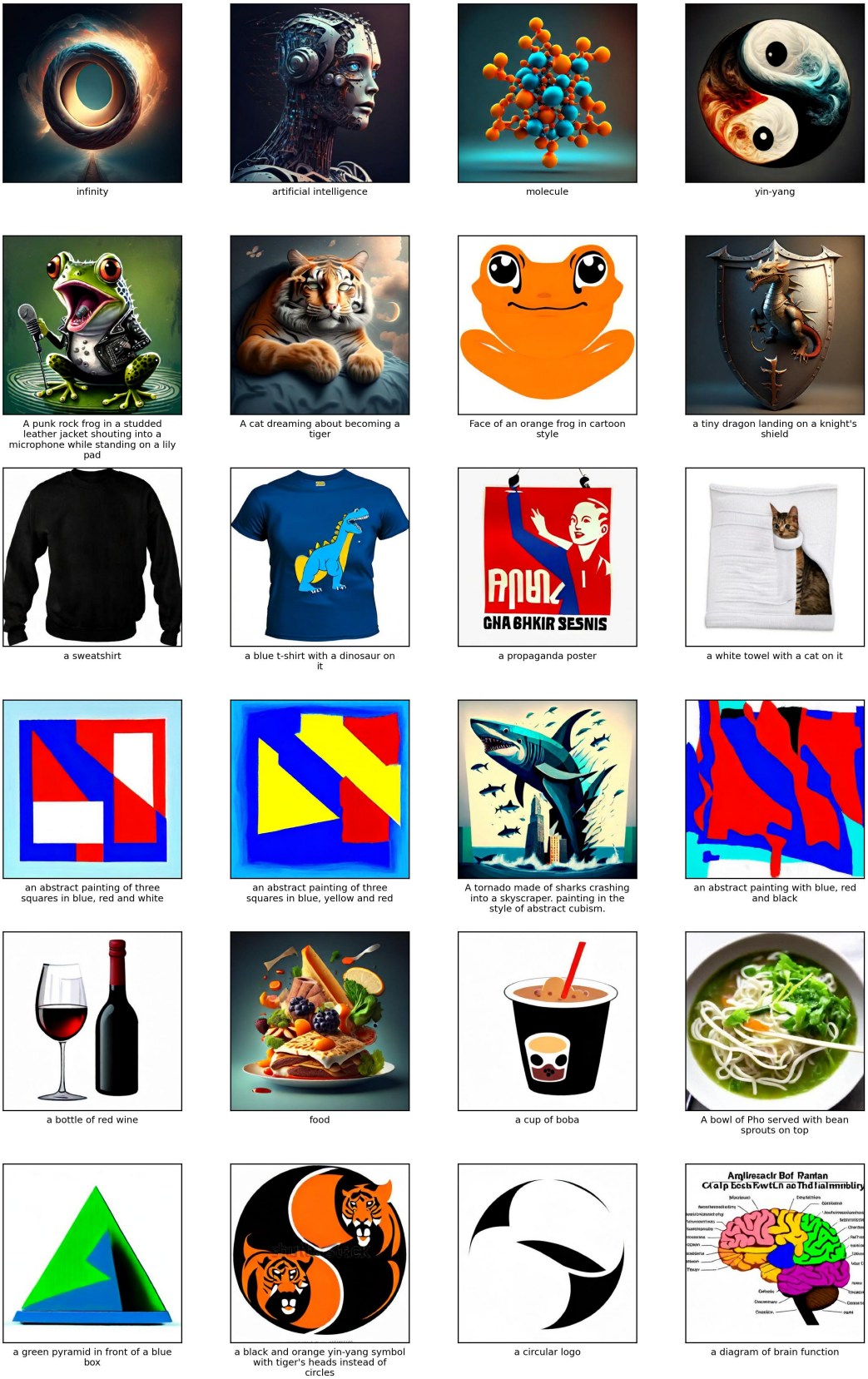


Figure 17. **Evaluating on PartiPrompts.** Synthesized images by our model using randomly selected prompts from PartiPrompts [87]. Rows correspond to following categories: Abstract, Animals, Artifacts, Arts, Food & Beverage, and Illustrations.

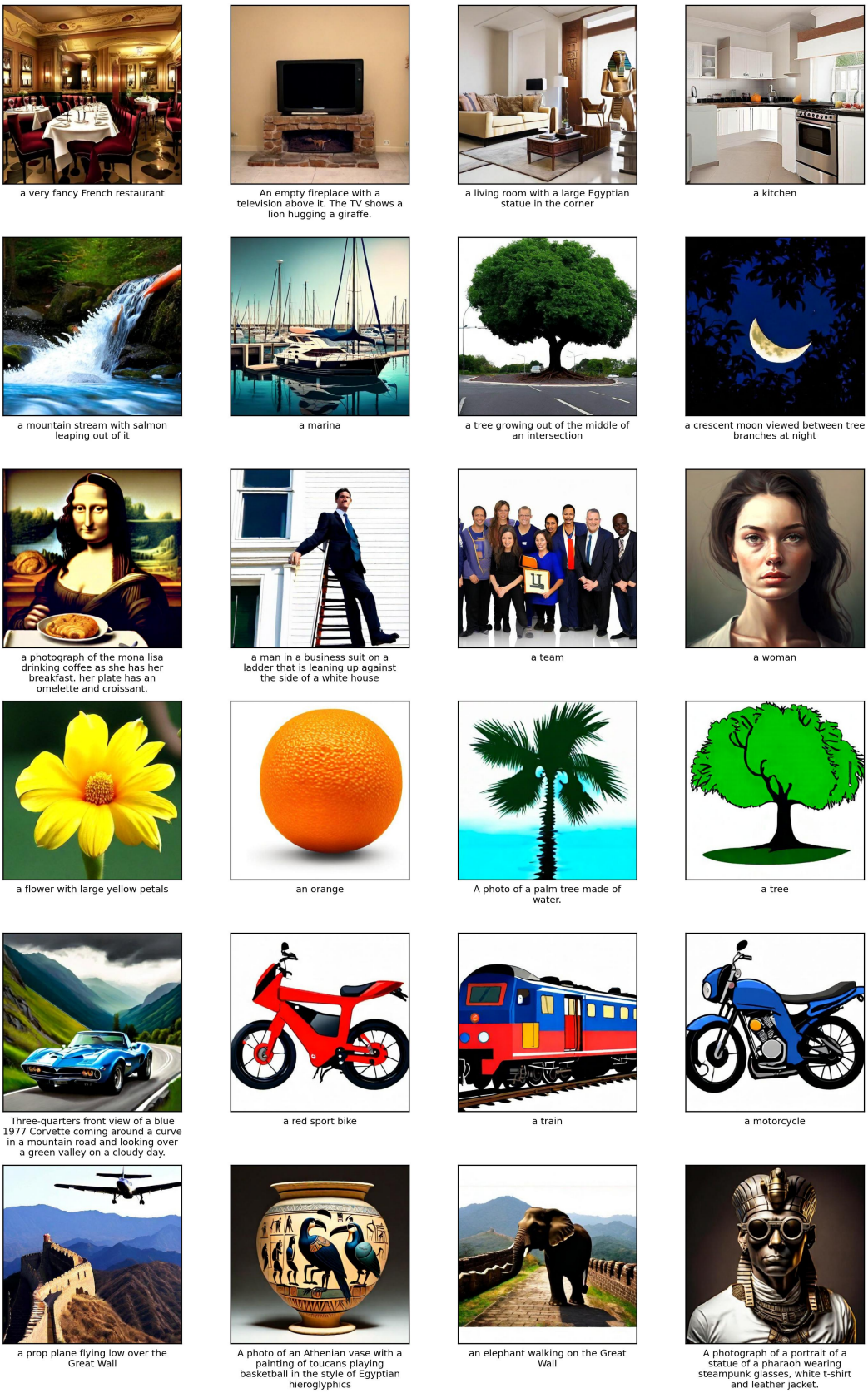
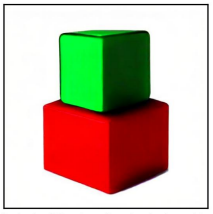


Figure 18. **Evaluating on PartiPrompts.** Synthesized images by our model using randomly selected prompts from PartiPrompts [87]. Rows correspond to following categories: Indoor scenes, Outdoor scenes, People, Produce & Plants, Vehicles, and World knowledge.



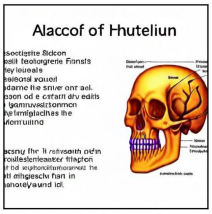
A stack of 3 cubes. A red cube is on the top, sitting on a red cube. The red cube is in the middle, sitting on a green cube. The green cube is on the bottom.

An emoji of a baby panda wearing a red hat, green gloves, red shirt, and green pants.

New York Skyline with 'Hello World' written with fireworks on the sky.

A keyboard made of water, the water is made of light, the light is turned off.

A small blue book sitting on a large red book.



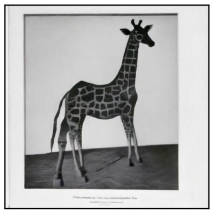
An organ of soft nervous tissue contained in the skull of vertebrates, functioning as the coordinating center of sensation and intellectual and nervous activity.

A carrot on the left of a broccoli.

A giraffe underneath a microwave.

A storefront with 'Text to Image' written on it.

An umbrella on top of a spoon.



A photocopy of a photograph of a painting of a sculpture of a giraffe.

A fluffy baby sloth with a knitted hat trying to figure out a laptop, close up, highly detailed, studio lighting, screen reflecting in its eyes.

Painting of the orange cat Otto von Garfield, Count of Bismarck-Schönhausen, Duke of Lauenburg, Minister-President of Prussia. Depicted wearing a Prussian Pickelhaube and eating his favorite meal - lasagna.

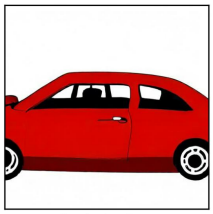
A medieval painting of the wifi not working.

A sign that says 'Deep Learning'.

A black colored car.

A bird scaring a scarecrow.

A yellow and black bus cruising through the rainforest.



A red colored car.

Abraham Lincoln touches his toes while George Washington does chin-ups. Lincoln is barefoot. Washington is wearing boots.

Three dogs on the street.

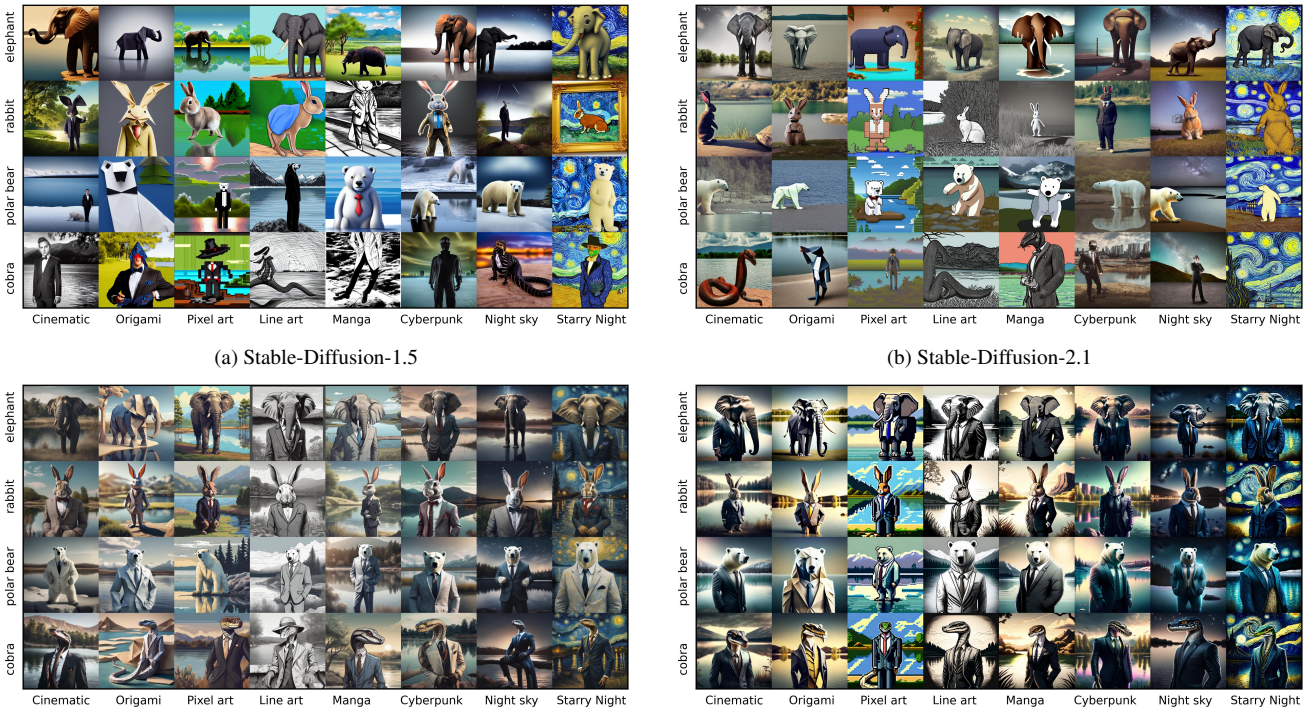
A cat on the right of a tennis racket.

A blue cup and a green cell phone.

An ancient Egyptian painting depicting an argument over whose turn it is to take out the trash.

Figure 19. **Evaluating on DrawBench.** Synthesized images by our model using randomly selected prompts from DrawBench [61].

Figure 20. **Evaluating on GenEval.** Synthesized images by our model using randomly selected prompts from GenEval [23].



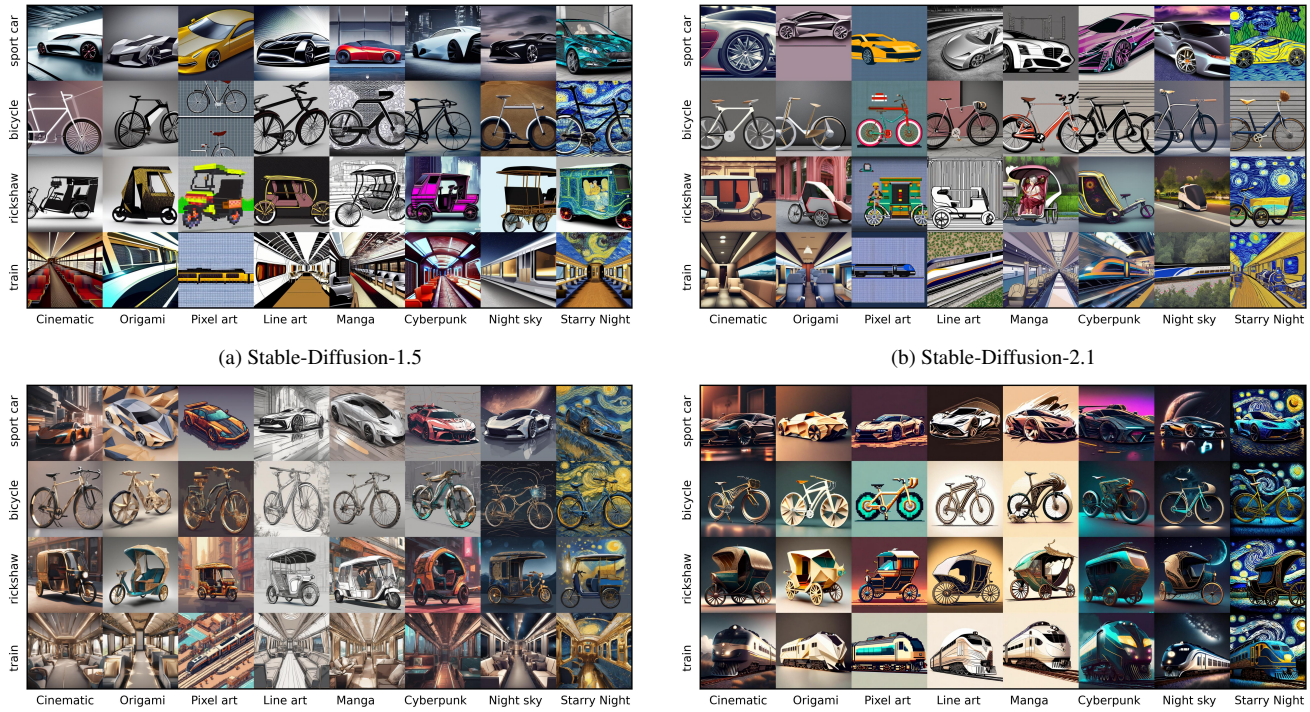
(a) Stable-Diffusion-1.5

(b) Stable-Diffusion-2.1

(c) Stable-Diffusion-XL

(d) Ours

Figure 21. **Control and diversity in image styles.** Evaluating the ability to generate diverse styles. *Prompt*: A ___ dressed in a suit posing for a photo in ___ style. Natural lake landscape in background, detailed light and shadow, high detail.



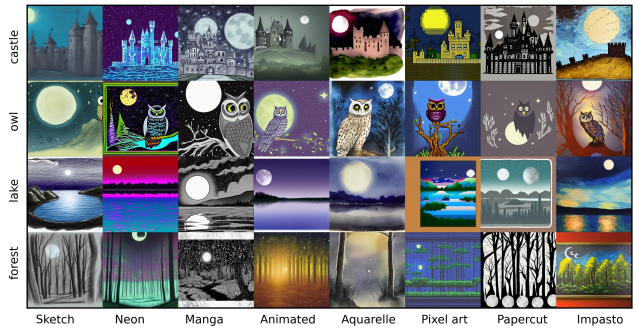
(a) Stable-Diffusion-1.5

(b) Stable-Diffusion-2.1

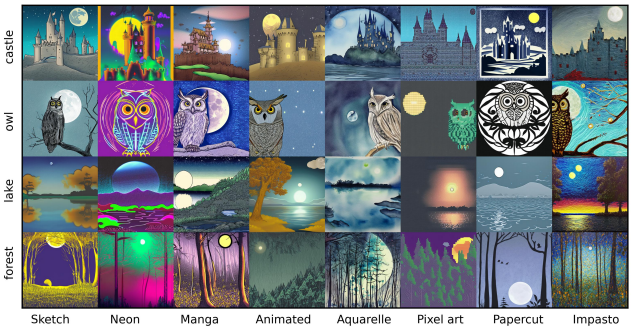
(c) Stable-Diffusion-XL

(d) Ours

Figure 22. **Control and diversity in image styles.** Evaluating the ability to generate diverse styles. *Prompt*: A luxurious ___ with a next-generation modern design illustrated in ___ style.



(a) Stable-Diffusion-1.5

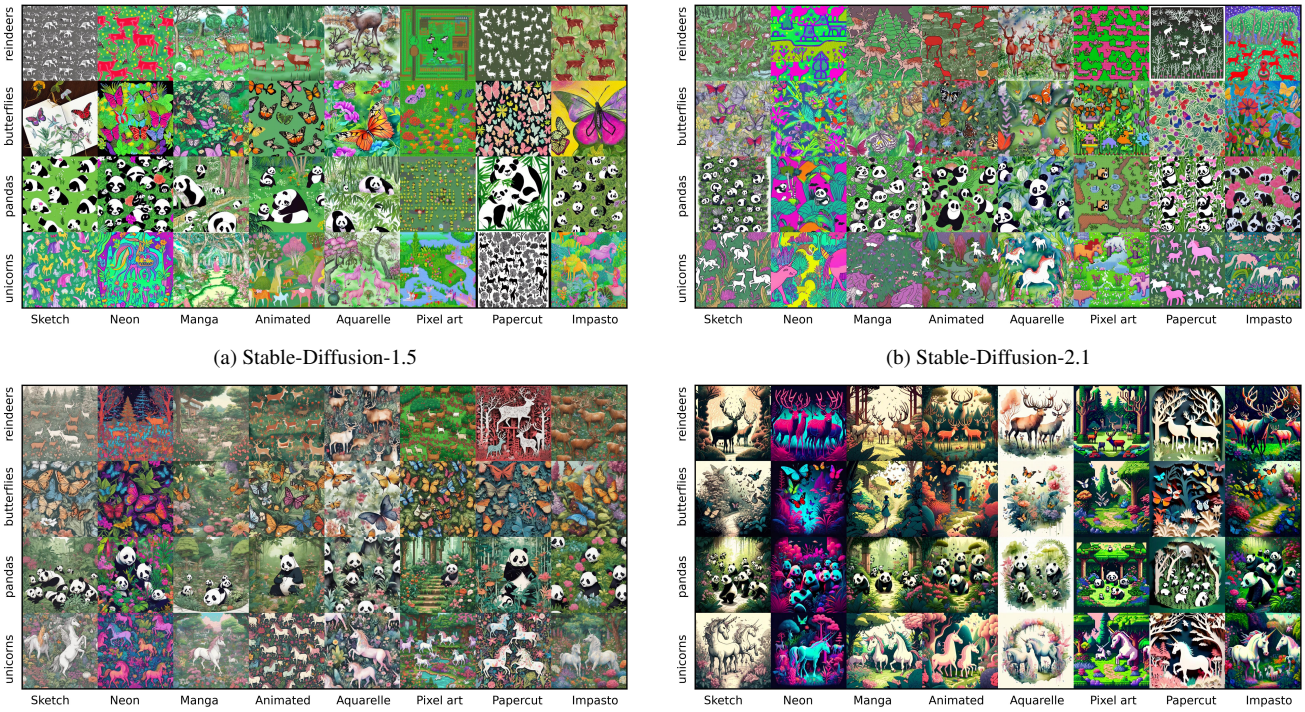


(b) Stable-Diffusion-2.1

(c) Stable-Diffusion-XL

(d) Ours

Figure 23. **Control and diversity in image styles.** Evaluating the ability to generate diverse styles. *Prompt*: A moonlit night over a ___, mysteriously rendered in ___ style.



(a) Stable-Diffusion-1.5

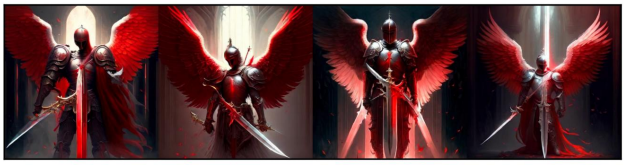
(b) Stable-Diffusion-2.1

(c) Stable-Diffusion-XL

(d) Ours

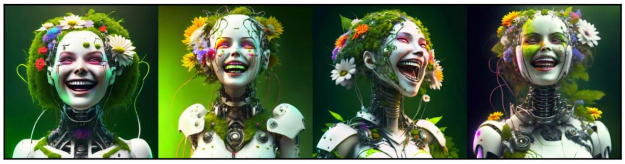
Figure 24. **Control and diversity in image styles.** Evaluating the ability to generate diverse styles. *Prompt*: A lush garden full of ___, vividly illustrated in ___ style.

flat vector, in a circle,Niagara Falls, Canada & USA, 8k, high resolution, adobe illustration color, no words



The Knight's Red Sword Cuts the Angel's Shining Wings,

Lo-fi style anime manga pirate and pirate ship, kawaii sticker, vaporwave, sparklecore, vibrant holographic gradient, iridescent highlights, flat vector, minimalist, □❤□,



gorgeous girl white glossy female android covered in moss and wires in the style of irobot, laughing facial expression, colorful flowers, solarpunk, detailed portrait, glamor, hi-resolution 4k, sharp focus, 8k, octane render

photo portrait of a blonde female racing driver in a black and purple race suit. She has a very determined expression on her face. Foggy background, dramatic lighting

The god of dreams, endovelicus, endovelico, with a beard and white robes with a crown of palm leaves, peering into a dark abyss with his hand outstretched behind him, beckoning you to hold his hand and peer into the stygian blackness, celtic, gallic, gaulish, extremely detailed etching, celtiberian

a 2D professional logo for a podcast, blue background, microphone, no text

shrew feeling wonderful and blissed out, calm, stoned, sleepy, curled up

time machine, cinematic lighting, fine-art photography, photorealistic, Annie Leibovitz style. aesthetic, dynamic lighting, digital illustration , cinematic lighting, 8k,

symmetrical ancient woven fabric, minimal,animals circling the sun in the center

rose garden, lamp, photorealistic, insane_details, cinematic lighting, canon 5d mk4, EF200mm F2L IS USM, iso 200

Two criminals, one male and one female, stand in silhouette with their backs to camera as they watch the city of London burns behind them + ultra realistic + photography + Full Perspective + crime + cinematography by Roger Deakins + award winning realism + The image is shot in a dark style + full body perspective + The composition is balanced and dynamic + The image is shot with a Arri Alexa, 18 mm lens 8k

Figure 25. **Evaluating on JourneyDB.** Synthesized images by our model using randomly selected prompts from the test set of JourneyDB dataset [23].

first day of work in a big administration with a lot of scary workers, 70s, horror style, cinematic lighting

"A hijab woman is petting a dog in front of her house. The woman has green eyes and is wearing glasses. The dog has a black head and a white body."::: charcoal style::1

journalism photo of a futuristic alien ufo ring in the sky dramatic sky over a valley cinematic lighting realistic flying saucer detailed minimal ring, historic, breaking news, dramatic, fog, cinematic, old kodak photo, grainy, film scratches, news archive photo, poloriod photo

The mystical gate of Valhalla. The door is almost half open. Worn stairs leading to the door. Viking Runes glowing blue in the door frames. hyper detailed. Hyper quality. Hyper reality. 8k

concept art of a male hispanic mech pilot

bearded sorcerer Benedict Wong playing magical guitar with colorful lights in a dark tavern

8k, The most realistic photo of the Sahara desert terrain, with a black hole in the middle

2d game monster design, wizardry style, beautiful female werewolf, full body drawing, nendoroid, anime, black background ,

Fantasy concept art of an anthropomorphic bear shaman

anime concept art□ a male field explorer

computer virus and hacker,

mob boss, mafia, red tie, cyberpunk, epic, dark, mysterious

Figure 26. **Evaluating on JourneyDB.** Synthesized images by our model using randomly selected prompts from the test set of JourneyDB dataset [23].

hyperrealistic detailed mk2 vw golf gti in metalic blue, manchester city in background, , caption, watermark, logo, or signature 8k

Jesus Christ Jehovah strong peaceful happy

Glass Cityscape in a recursive intricacy , sunrise

Full-length view, oil painting, woman morphing into colorful bird, gossamer wings, flowing robes, feminine, exquisite, delicate, ethereal, mystical, fantastical, ar3:4

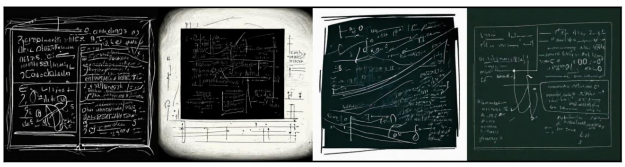
The Starry night composed by cosmetics

queen of tooth fairies character, solid white background

boxes of liquor chocolates, smashed, 1930s, prohibition, police officers, candy store

a traditional Chinese watercolor painting of a Chinese village with several bamboo houses in the sub-tropical mountains of Yunnan with a tea plantation in the background, traditional chinese dressed people, cobblestone path, in a round frame, three black chinese characters on a white background q2

a flatlay of a single Japanese watch face, 1943, no strap, against white background , writing implements, paper, hands



abstract, mathematical, chalkboard annotations a professor would use to illustrate complex financial derivatives in front of a classroom, clean vector lines, in the style of cy twombly

watercolore,3d render,cute moth fairy doll

photorealistic, HD, steampunk, owl with open wings, celtic designs on wings

Figure 27. **Evaluating on JourneyDB.** Synthesized images by our model using randomly selected prompts from the test set of JourneyDB dataset [23].

coffee cup roast heart love wall art, Brian Allen, flat, sticker, black and white svg vector

steampunk percolator coffee pot sign which says steampunk, steampunk style, neon lights, smoky, transparent background

Stunning 4K Realistic Fog Bright Light neon pink blue fairy fantasy beach

corgi dressed as queen of england, in royal attire, pen and ink

Inscrutable wraith ghost hyperborean resistance research athaeneum, ubiquitous darkness enigma diffused moonlight

liquid girl, graffiti style,

T-shirt with short sleeves, printed in a climate of extreme sports, beautiful graphics

**Victorian candle, dark roses and peonies arrangement, surrounded by red floating fairy lights, hyper realistic, hyper detailed, low light, dark, gothic, witchy, 8d,

the Parthenon, Photo realistic, unreal engine, cinematic lighting, .5

black evil goblin, red eyes

young person from the 1920's, with a antique PC from 1920's, in their home shopping online in the 1920's , hyper realistic

The Multiverse:: Max Detailed Foreground:: In A Multi-Perspective Surreal Aesthetic, DMT, Psychedelic Style:: With a Cinematic Imaginative Background That Cuts Into The Foreground With Realistic Light Reflections and Perfect Scene Lighting:: Insanely detail, photomanipulation, photorealistic, C4D, Octane, Blender::

Figure 28. **Evaluating on JourneyDB.** Synthesized images by our model using randomly selected prompts from the test set of JourneyDB dataset [23].

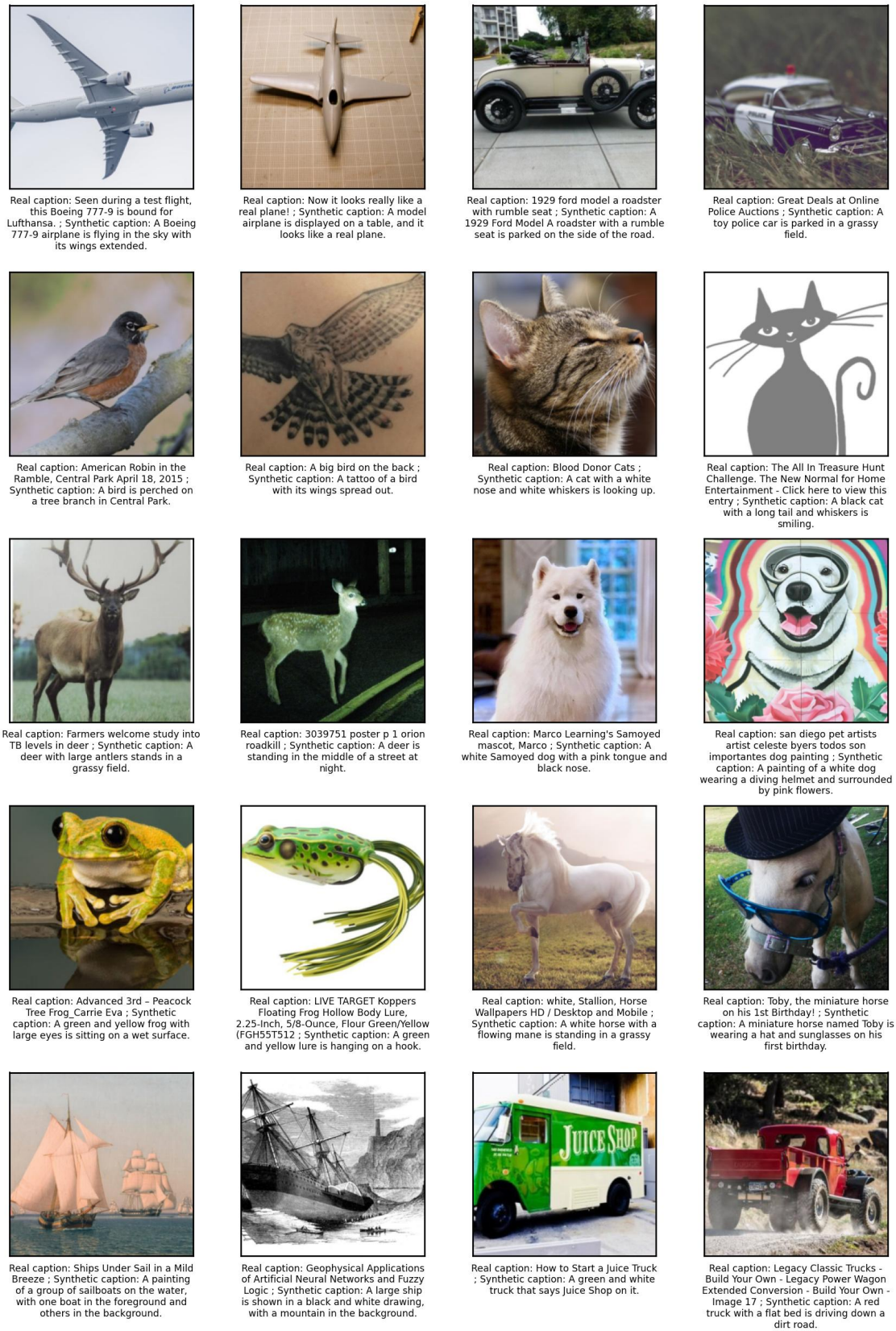


Figure 29. **Samples from our cifar-captions dataset.** Selected images with corresponding real and synthetic captions from the cifar-captions dataset. We created the cifar-captions dataset, imitating the widely used CIFAR-10 dataset [41], to enable small scale experimentation on the text-to-image generative models.

List of extended DrawBench [61] prompts used in human-centric evaluation.

Sleek red car reflects cityscape, admired by passersby.
Black car cruises rainy streets, reflecting vibrant neon lights.
Glossy pink car cruising a sunny highway, surrounded by lush greenery, chrome gleaming, catching admiring glances.
Majestic black dog stands alert in a meadow, wildflowers sway, sunlight glinting on its sleek coat.
Curly red dog chases a butterfly in a windy, flower-filled meadow, frolicking playfully in the sunshine.
Vibrant blue dog prances in a sunny field, tail wagging, butterflies fluttering nearby.
A green banana contrasts with purple and yellow fruit, sunlight streaming in.
A red banana sits on a wooden table, surrounded by sunlight filtering through the leaves.
A black banana lies on white marble, hinting at overripe mystery beneath its dark peel.
White bread sandwich filled with cream cheese, cucumber, and turkey on a pristine plate, bathed in soft, warm light.
Black sandwich with charcoal bread, lettuce peeking out, set on a white plate beside lemonade.
Orange sandwich of roasted pumpkin and cheddar on whole wheat bread, sitting on a rustic table.
Pink giraffe in a wildflower field, plucking flowers with its long neck, butterflies swirling around.
Bright yellow giraffe nibbles leaves from tall trees under the warm savannah sun.
Brown giraffe strides across a sunlit savannah, its unique pattern glowing in the vibrant landscape.
Red car parked near a white sheep on a peaceful country road, hills rolling in the distance.
Blue bird perched on a brown bear's shoulder in a wildflower meadow, sharing a serene moment.
Green apple atop a black backpack on a park bench, with trees and children playing in the background.
Green cup of steaming coffee sits beside a blue cell phone on a rustic table in warm sunlight.
Yellow book and red vase filled with flowers sit on a sunlit wooden table, radiating elegance.
White car parked near a red sheep in a lush meadow, under a blue sky.
Brown bird sings from a cherry blossom branch while a blue bear gazes up, surrounded by flowers.
Black apple beside a green backpack on a rustic table, sunlight streaming through a window.
Blue ceramic cup filled with hot coffee sits next to a green cell phone, casting steam into the air.
Red book and yellow vase sit together on an antique table, illuminated by sunlight.
A horse wearing a spacesuit stands triumphantly on an astronaut's shoulders on a moon-like surface.
Pepperoni pizza bakes in a brick oven, surrounded by dancing flames and delicious aromas.
Cardinal startles a scarecrow in a cornfield, sending straw-filled arms flailing comically.
A blue pizza topped with blueberries, blackberries, and edible flowers sits on a rustic table.
A hovering cow abducts aliens with a tractor beam, while a spaceship glows in the night sky.
Panda making latte art in a cozy café, skillfully shaping a bamboo leaf in a cup of espresso.
A great white shark swims through desert dunes, its fin slicing through the sand like water.
An elephant swims under the sea, wearing a bubble-shaped helmet, surrounded by coral reefs and fish.
Rainbow-colored penguin waddles joyfully across a vibrant arctic landscape, reflecting the sky's colors.
A large fish leaps from the ocean to devour a startled pelican mid-flight during a storm.
One red sports car parked on a quiet cobblestone street in a picturesque European village.
Two vintage cars, one red and one blue, race down a charming cobblestone street.
Three colorful vintage cars line up along a cobblestone street, bathed in the setting sun's glow.
Four vintage cars in vibrant colors cruise down a bustling cobblestone street in retro style.
Five unique cars, from vintage to sleek sports models, line a lively city street.
A golden retriever trots along a cobblestone street, wearing a red bandana, as the sun sets.
Two dogs, a retriever and a beagle, walk together in the city, proudly sporting red and green bandanas.
Three dogs—a retriever, terrier, and dalmatian—walk down a lively street, leashes loosely held.
Four dogs—retriever, poodle, dachshund, and bulldog—strut down a bustling street in colorful bandanas.
Five playful dogs of various breeds chase each other on a cobblestone street, wagging tails happily.
A dog and a cat sit together on a grassy hill, gazing at the sunset over a lake.
A cat and two dogs sit peacefully on a grassy patch in a sunlit park, surrounded by flowers.
A regal cat sits with three cheerful dogs in a meadow, wearing a tiny gold crown.
Two cats and a dog relax on the grass as the sunset bathes them in warm colors.
Two cats and two dogs sit on a grassy hill, basking in the sunlight and gentle breeze.
Two cats and three dogs lounge together under a tree, enjoying the sunny meadow.
Three cats and a dog sit peacefully in the grass, enjoying each other's company.

Three cats and two dogs share a sunlit meadow, lounging gracefully on the soft grass.
Three cats and three dogs form a harmonious group, lounging together on a grassy hill.
Purple triangular flower pot brimming with green plants sits on a sunlit windowsill.
Orange triangular picture frame holds a serene landscape painting, set against a pale blue background.
A bright pink triangular stop sign stands at the edge of a forest road, glowing in the sunlight.
A denim-textured cube sits on a wooden table, surrounded by sewing needles and spools of thread.
A sphere made of kitchen tiles reflects sunlight in a mesmerizing pattern.
A large cube made of red bricks sits in a grassy field, vines creeping up its side.
A collection of nails neatly arranged on a wooden table in a cozy workshop.
A grand brass clock sits on an antique wooden table in a warmly lit room.
Two ornate crystal glasses sit on a polished wooden table, casting gentle shadows.
A red elephant sits atop a small blue mouse in a whimsical, flower-filled field.
A green elephant stands behind a large red mouse in a bright, grassy meadow.
A small blue book sits on top of a large red book in an elegant library.
Three stacked plates—two blue, one green—sit neatly on a rustic wooden table.
Three stacked cubes—two red, one green—rest on a wooden surface in vibrant colors.
A stack of three books—green, red, and blue—sits on a polished table in a sunlit room.
Baby panda emoji in a red hat and green pants grins adorably.
Baby panda wearing a red hat and blue gloves sits happily in its colorful outfit.
A turtle sits in a lush forest, captured through a fisheye lens, with the trees curving around.
A majestic owl perches in a wildflower field, its feathers ruffled by the gentle breeze.
A detailed cross-section of a brain showcases intricate structures and pathways.
A vintage bicycle rests against a brick wall, flowers in its wicker basket.
A modern bus cruises through a vibrant city, picking up passengers.
A small wooden boat glides peacefully across a serene lake, leaving a gentle wake.
A red fire hydrant stands at a busy city intersection as firefighters connect hoses.
A sleek parking meter stands beside a bustling city street, displaying the time left.
A vintage umbrella with lace patterns protects from a gentle rain in a sunlit park.
A vintage wooden chair with intricate carvings sits alone in a softly lit room.
An old-fashioned icebox sits in a cozy rustic kitchen, filled with fresh produce.
A steampunk-inspired clock with exposed gears sits on an ornate table, pendulum swinging.
A pair of ornate scissors rests on a patterned tablecloth, ready to cut delicate fabrics.
A majestic chestnut horse grazes in a sunlit meadow, its mane blowing in the breeze.
A bunch of ripe, yellow bananas hangs from a tree in a vibrant jungle scene.
A sleepy calico cat lounges on a windowsill, bathed in sunlight.
A well-groomed dog with a long snout stands proudly in a sunlit backyard, sniffing the air.
A colorful human brain sits within a transparent skull, pulses of energy flowing through it.
A futuristic office of a multinational tech company buzzing with innovation.
A grand piano with intricate carvings stands in a sunlit room, strings vibrating with music.
A golden coin symbolizes the decentralized nature of cryptocurrency, floating above a computer.
A thick-skinned hippopotamus stands at the edge of a calm river, basking in the sun.
A humanoid robot mimics a scientist's movements in a futuristic lab filled with machinery.
Customer pays for a tiny pizza with a giant quarter, both grinning in surprise.
An elegant couple in formal wear caught in a downpour, sharing a tender moment.
Pint cartons of milk sit neatly on the top shelf of a grocery store refrigerator.
A man stands in the shadow of a maple tree on a crisp January afternoon in New England.
An elephant hides behind a tree, trunk visible on one side, back legs on the other.
A tomato sits atop a pumpkin on a kitchen stool, a fork stuck in the side.
A pear cut into seven even pieces is arranged in a ring on a rustic table.
A donkey and octopus play tug-of-war while a cat leaps over the rope at sunset.
Supreme Court justices face off against FBI agents in a friendly game of baseball.
Abraham Lincoln touches his toes while George Washington does chin-ups in a meadow.
Tennis racket rests against a wooden bench on a sunlit clay court.

A well-worn baseball glove lies on a patch of grass, ready for the next catch.
A retro red refrigerator stands in a cozy kitchen filled with vintage décor.
A polished dining table set with fine china and silverware, beneath a glowing chandelier.
A vintage parking meter stands on a bustling street, surrounded by classic cars.
A small boat propelled by oars floats peacefully on a lake at sunset.
A fluffy cat rests in a cozy living room, intently watching a dangling toy mouse.
A pair of elegant stainless steel scissors with engraved handles rests on a patterned table.
A happy, well-groomed dog stands in a grassy backyard, sniffing the air.
A grand piano with wooden carvings stands in a sunlit room, with strings and hammers visible.
A red steam locomotive rides atop a surfboard, slicing through ocean waves at sunset.
A golden retriever balances a wine glass on its head, sitting proudly by a warm fireplace.
A vintage bicycle rests on top of a small wooden boat floating near the shore.
A delicate umbrella balances on top of a polished silver spoon on a wooden table.
A fluffy teddy bear supports a sleek laptop, its kind eyes peeking over the top.
A curious giraffe stands beneath an oversized microwave hanging from a tree branch.
A pink frosted donut lies beneath a white porcelain toilet in a whimsical bathroom scene.
A hair dryer blows warm air beneath a sheep standing calmly in a grassy field.
A tennis racket rests beneath a traffic light in a bustling city street.
A zebra stands under a gigantic broccoli tree, its stripes contrasting with the green leaves.
A banana rests on a wooden table to the left of a shiny red apple.
A red velvet couch sits beside a vintage leather chair in a sunlit living room.
A red sports car cruises beside a double-decker bus on a busy city street.
A sleek black cat lounges next to a tennis racket on a sunlit tennis court.
A stop sign leans against a refrigerator in a cozy, vintage-inspired kitchen.
A fluffy white sheep stands beside a wine glass filled with red wine in a grassy meadow.
A zebra stands next to a red fire hydrant in a bustling city street.
Acersecomicke, a majestic creature, flies through a vibrant sky, searching for adventure.
A family gathers for a jentacular feast in a warm kitchen, sharing laughter and food.
Matutinal sun rises over a peaceful valley, casting golden light on the dewy grass.
Pigeons socialize on a red-bricked rooftop in a village, creating a peristeronic scene.
The mystical artophagous feasts on colorful paintings, absorbing their creative energy.
An abandoned backlot of a film studio, overgrown and filled with forgotten treasures.
The octothorpe, a massive metallic creature, roams through an abandoned city.
Stained glass windows of a church depict a hamburger and fries, casting colorful rays inside.
Otto von Garfield, Duke of Lauenburg, eats lasagna in an elegant painting.
A baby fennec fox sneezes onto a strawberry, backlit ears glowing in detail.
A confused grizzly bear attends calculus class, staring at the chalkboard in puzzlement.
Egyptian painting shows two figures arguing over who should take out the trash.
A baby sloth in a knitted hat stares at a laptop, trying to figure it out.
A tiger in a lab coat works a science machine in a 1980s Miami-style laboratory.
Animals dressed as humans pose for a 1960s yearbook photo in vintage clothing.
A Lego version of Arnold Schwarzenegger stands confidently, holding a mini gun.
A yellow and black bus cruises through the dense foliage of a rainforest.
Medieval scholars gather around a broken Wi-Fi router, trying to restore the connection.
An IT guy struggles with tangled cables like Laocoön while fixing a PC tower.
A handful of colorful Skittles scattered across a smooth surface.
A gothic-style McDonald's church with stained glass windows stands proudly in a village.
An athletic cat addresses a scandal at a press conference, surrounded by journalists.
A marble statue shows a man tripping over a surprised cat in a lush garden.
A 1920s airship shaped like a pig floats over a golden wheat field.
A tuxedo cat sings in a barbershop quartet, wearing a straw boater hat and bow tie.
Astronaut couple poses in American Gothic style, holding a flag and a space helmet.
The regal Burger King poses with a Whopper in an opulent oil painting.

A keyboard made of water shimmers as the light is turned off, blending into darkness.
Mona Lisa viewed from behind, her hair cascading down, gazing at a Tuscan landscape.
A hyper-realistic photo of an abandoned industrial site during a storm.
An iOS app screen shows different types of milk available for ordering.
Super Mario leaps through a bustling city in a realistic 8K Ultra HD photograph.
Cats climb the Eiffel Tower in a futuristic cyberpunk coloring page.
A mega Lego space station towers inside a child's galaxy-themed bedroom.
A spider with a moustache greets a gentlemanly grasshopper as they cross paths.
A framed photo of a photocopy of a painting of a giraffe statue.
Bird's-eye view of a bridge connecting Europe and North America across the Atlantic Ocean.
A maglev train plunges vertically downward at high speed in New York City.
A magnifying glass reveals a page from a 1950s Batman comic, highlighting a dramatic scene.
A futuristic car plays soccer on a digital field, sending a holographic ball flying.
Darth Vader plays with a raccoon on Mars at sunset, red sky glowing around them.
A 1960s poster warns against climate change with vibrant, psychedelic imagery.
A mouse uses a mushroom as an umbrella during a gentle rain, ripples forming in puddles.
A Pomeranian dressed as a 1980s wrestler strikes a pose in neon wrestling tights.
A pyramid of falafel stands in the desert under a partial solar eclipse.
A storefront displays "Hello World" in large, welcoming letters across the glass window.
A charming storefront with "Diffusion" written elegantly across the window.
A storefront with "Text to Image" written above the door, surrounded by framed pictures.
A sleek storefront with "NeurIPS" etched on the glass, reflecting the busy street outside.
A quaint storefront with "Deep Learning" written in gold script on the window.
A modern storefront with "Google Brain Toronto" etched above the glass doors.
A rustic storefront with "Google Research Pizza Cafe" warmly inviting passersby inside.
A wooden sign in a meadow says "Hello World," surrounded by colorful wildflowers.
A weathered wooden sign reading "Diffusion" sways gently in the breeze.
A sign in a grassy field says "Text to Image," with butterflies fluttering around.
A large conference center sign displays "NeurIPS" in bold letters, lit by string lights.
A wooden sign engraved with "Deep Learning" stands in a meadow, glowing in the sunset.
A sleek metal sign says "Google Brain Toronto," standing in front of a glass building.
A colorful sign reading "Google Research Pizza Cafe" hangs outside a charming café.
The New York skyline glows as fireworks spell out "Hello World" in the evening sky.
Fireworks over the New York skyline spell out "Diffusion" against the night sky.
The New York skyline at dusk with "Text to Image" written in dazzling fireworks.
Fireworks over New York spell out "NeurIPS," illuminating the cityscape below.
"Deep Learning" is written in fireworks above the New York skyline, reflected in the water.
"Google Brain Toronto" written with fireworks above the New York skyline at night.
The New York skyline shines as fireworks spell "Google Research Pizza Cafe."
"Hello World" spelled out in brilliant fireworks over New York's famous skyline.
"Diffusion" written with fireworks above the New York skyline, creating a dazzling display.
The New York skyline lights up with "Text to Image" written in vibrant fireworks.