Fantasy: Transformer Meets Transformer in Text-to-Image Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present Fantasy, an efficient text-to-image generation model marrying the decoder-only Large Language Models (LLMs) and transformer-based masked image modeling (MIM). While diffusion models are currently in a leading position in this task, we demonstrate that with appropriate training strategies and high-quality data, MIM can also achieve comparable performance. By incorporating pre-trained decoder-only LLMs as the text encoder, we observe a significant improvement in text fidelity compared to the widely used CLIP text encoder, enhancing the text-image alignment. Our training approach involves two stages: 1) large-scale concept alignment pre-training, and 2) fine-tuning with high-quality instruction-image data. Evaluations on FID, HPSv2 benchmarks, and human feedback demonstrate the competitive performance of Fantasy against state-of-the-art diffusion and autoregressive models.

1 Introduction

Recent advances in text-to-image (T2I) models [3, 5, 12] have become focal points within the computer vision field. Most advances in T2I models, focused on generating high-quality images based on relatively short descriptions, struggle with intricate long-text semantic alignment due to inherent structure constraints and data limitations. Text encoders used for T2I fall into three categories: CLIP [30], encoder-decoder LLMs, and decoder-only LLMs. Models using encoder-decoder LLMs like T5-XXL [31] have shown improved text-image alignment over CLIP by exploiting enhanced text understanding, increasing token capacity, yet without delving into the semantic alignment for longer texts. ParaDiffusion [43] indicates that directly aligning text embeddings with visual features without prior image-text knowledge is not the most effective approach. Previous works [38, 45] have highlighted shortcomings in existing text-image datasets [37], including image-text mismatches, a lack of informative content, and a pronounced long-tail effect. These deficiencies notably impair training efficiency for T2I models and restrict their ability to learn complex semantic alignment.

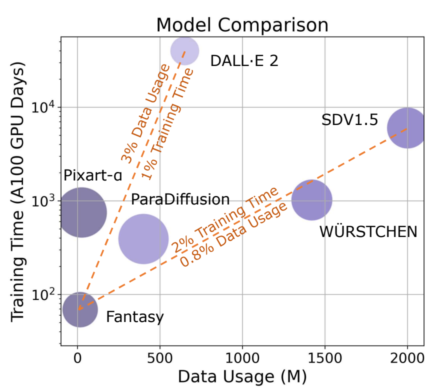


Figure 1: Comparison of data usage, training time and image quality. Colors from dark to light represent parameters increasing in size, and circles from small to large indicate improvements in image quality.

Existing diffusion-based T2I models [33, 5, 9, 26] have achieved unprecedented quality. However, as detailed in Fig. 1, these advanced models come with significant computational demands. The

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

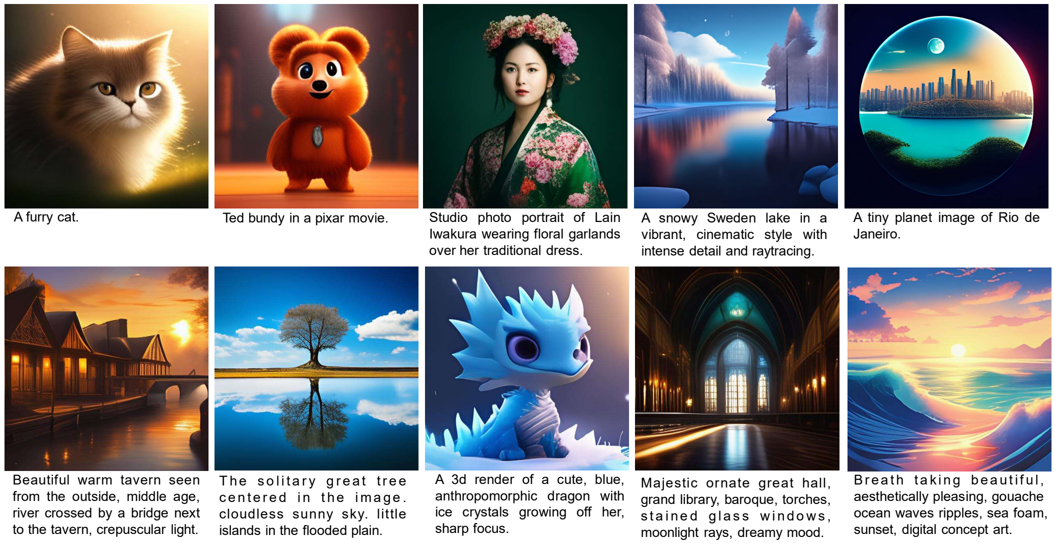


Figure 2: Samples produced by Fantasy (512×512). Each image, generated in 1.26 seconds (without super-resolution models), is accompanied by a descriptive caption showcasing diverse styles and comprehension.

considerable expenses of these models create significant barriers for researchers and entrepreneurs. Meanwhile, economical text-to-image models [25, 15, 48] compromise on image quality, yielding lower resolution and diminished aesthetic appeal.

Given these challenges, a pivotal question arises: *Can we develop a **resource-efficient**, **high-quality** image generator for **long** instructions?* In this paper, we present Fantasy, significantly reducing training demands while maintaining the capability of instruction understanding and competitive image generation quality, as shown in Fig. 2. To achieve this, we propose three core designs:

Efficient T2I netwrok. To leverage the powerful understanding ability of a decoder-only LLM, we choose the lightweight Phi-2 [24] as our text encoder. We derive discrete image tokens from a pre-trained VQGAN [27], and employ Transformer-based masked image modeling (MIM) as our T2I architecture. We also utilize the pre-trained VQGAN decoder [27] for pixel space restoration.

Hierarchical Training strategy. We propose a thoughtfully two-stage training strategy to address the high computational demands of current leading models while maintaining competitive performance: (1) large-scale concept alignment pre-training, (2) high-quality instruction-image fine-tuning. To facilitate a coarse image-text alignment, we initially train the T2I model from scratch using relatively lower-quality data. We then fine-tune the pre-trained T2I model and LLM on text-image pair data rich in information density with superior aesthetic quality.

High-quality data. To achieve rough alignment while pre-training, we select the large-scale dataset LAION-2B [37] and employ the filtering strategy proposed by DataComp [14]. We collect long-text prompts and corresponding high-quality synthesized images for instruction tuning, including DiffusionDB [42] and JourneyDB [39]. We further filter and discard texts with special characters and data containing violence or pornography, retaining only instructions exceeding 30 words.

Our main contributions are summarized as follows:

1. We present Fantasy, a novel framework that is the first to integrate a lightweight decoder-only LLM and a Transformer-based MIM for text-to-image synthesis, allowing for long-form text alignment.
2. We show that our two-stage training strategy with high-quality data enables MIM to achieve comparable performance at a significantly reduced training cost.
3. We provide comprehensive validation of the model's efficacy based on automated metrics and human feedback for visual appeal and text faithfulness.

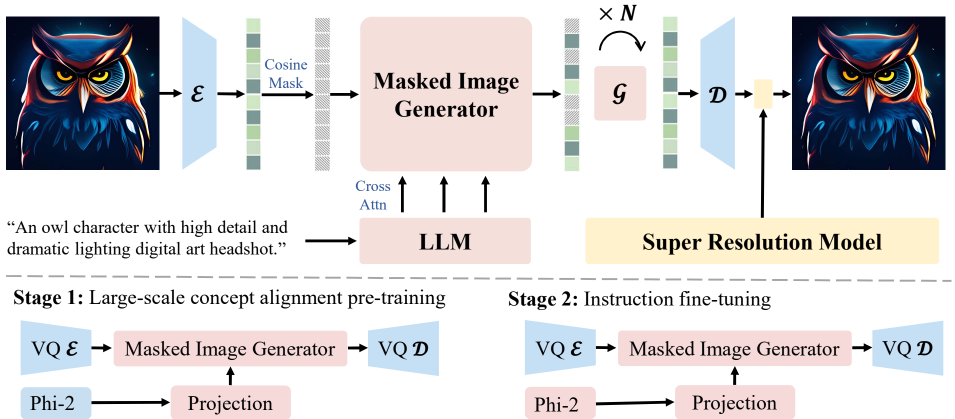


Figure 3: **(Up)** Overview of Fantasy featuring text encoder, VQGAN (encoder $\mathcal{E}$ and decoder $\mathcal{D}$), masked image generator $\mathcal{G}$, and super-resolution model. (**Down**) Our training pipeline involves two stages. The red parts are trainable and the blue parts are frozen; the yellow part is optionally utilized during inference.

2 Method

2.1 Problem Formulation

As depicted in Fig. 3, Fantasy consists of a pre-trained text encoder $\mathcal{T}$, a transformer-based masked image generator $\mathcal{G}$, a sampler $\mathcal{S}$, a frozen VQGAN, and a pre-trained super-resolution model. $\mathcal{T}$ maps a text prompt t to a continuous embedding space. $\mathcal{G}$ processes a text embedding e to generate logits l for the visual token sequence. $\mathcal{S}$ draws a sequence of visual tokens v from logits via iterative decoding [4], which runs N steps of inference conditioned on the text embeddings e and visual tokens decoded from previous steps. Finally, $\mathcal{D}$ maps the sequence of discrete tokens to pixel space Z. To summarize, given a text prompt t, an image $\hat{x}$ is synthesized as follows:

$$\hat{x} = \mathcal{D}(\mathcal{S}(\mathcal{G}, \mathcal{T}(t))), \quad l_n = \mathcal{G}(v_n, \mathcal{T}(t)), \quad v_n = \mathcal{M}(\mathcal{E}(x)) \tag{1}$$

where n is the synthesis step, and l_n are logits, from which the next set of visual tokens v_{n+1} are sampled. $\mathcal{M}$ denotes the masking operator that applies masks to the token in v_n. We refer to [4, 3] for details on the iterative decoding process. The Phi-2 [24] for $\mathcal{T}$ and VQGAN [8] for encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are used. $\mathcal{G}$ is trained on a large text-image pairs D using masked visual token modeling loss:

$$\mathcal{L} = \mathbb{E}_{(x,t)\sim D} \left[CE \left(l_N, \mathcal{E}(x) \right) \right], \tag{2}$$

where CE is a weighted cross-entropy calculated by summing only over the unmasked tokens.

2.2 Model Architecture

2.2.1 VQGAN as Image Processor

VQGAN [8] is capable of transforming each image into discrete tokens with higher-level semantic information from a learned codebook, while ignoring low level noise. The autoregressive tokens prediction of VQGAN shares the same form as text tokens generated by LLMs. Prior research [46] has shown that unifying vision and language by the same token space could enhance the coherency for vision-text alignment. Furthermore, compared with RGB pixels, the visual token representation has proven to reduce disk storage and improve the capability of robustness and generalization.

To reduce the computational burden, we initially compress an RGB image $v \in \mathbb{R}^{H\times W\times 3}$ into a diminished representation with a resolution of $h \times w \times 3$, where $h = H/f$ and $w = W/f$, with f denoting the downsampling factor. We then employ a pre-trained $f16$ VQGAN [27] encoder $\mathcal{E}$ to quantizate images $x \in \mathbb{R}^{3\times 256\times 256}$ into discrete tokens of spatial dimensions 16×16 from a pre-trained codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K}$ consisting of $K = 8192$ vectors, resulting in the quantized representation $z = \mathcal{E}(x, \mathcal{Z})$.

2.2.2 LLM as Text Encoder

Recent studies [10, 5, 3] tend to use encoder-decoder LLMs [31] for text encoding over CLIP [30], which is adept at handling tasks that involve complex mappings between input and output sequences. Due to the tremendous success of ChatGPT, attention has been drawn to models that consist solely of a decoder. Also, [43] presents an insight that efficiently fine-tuning a more powerful decoder-only LLM can yield stronger performance in long-text alignment. Consequently, to capitalize on the enhanced semantic comprehension and generalization potential of LLMs while simultaneously reducing the training burden, we employ Phi-2 [24], a state-of-the-art, lightweight LLM, as the text encoder.

Given the text prompt t, Fantasy first passes it through Phi-2, extracting the text embedding from the last hidden layer L. However, typically, decoder-only architectures are not adept at feature extraction and mapping tasks. [23] proposes that the conceptual representations learned by LLM's are roughly linearly mappable to those learned by models trained on vision tasks. Therefore, the embedding vectors are linearly projected to the hidden size of the image generator $\mathcal{G}$:

$$c = \mathcal{P}(\mathcal{T}_L(t)) \tag{3}$$

where $\mathcal{T}(\cdot)$ denotes the decoder-only Phi-2 and L is the index of the last hidden layer. $\mathcal{P}$ represents the projection from text space to visual space, and c is the text feature suitable for the image generator.

2.2.3 MIM as Image Generator

MIM narrows the gap between its modeling and the extensively studied area of language modeling, making it straightforward to leverage the findings of the LLMs research community. Therefore, we adopt a masked transformer as the image generator backbone of Fantasy [46].

During training, we leave the projected text embeddings c unmasked and the image tokens z are masked at a variable masking rate based on a Cosine scheduling $\mathcal{M}$ as [4, 3]. Specifically, for each training example, we sample a masking rate r from $[0, 1]$ from a truncated $arccos$ distribution with density function $p(r) = \frac{2}{\pi}(1 - r^2)^{-\frac{1}{2}}$. While autoregressive methods learn fixed-order token distributions $P(z_i|z_{<i})$, random masking with variable ratios enables learning $P(z_i|z_{\neq i})$ for any token subset, crucial for our parallel sampling scheme. The sampling of a new state s_{n+1} at each successive step is conditioned on the previous state and the specified text condition c:

$$P(s \mid c) = \int P(s_N \mid s_{N-1}, c) \prod_{n=1}^{N-1} P(s_n \mid s_{n-1}, c)\, ds_1 \dots ds_{N-1} \tag{4}$$

For each training example, the most confidently predicted tokens are revealed at each step n, maintaining $\cos\left(\frac{n}{N} \cdot \frac{\pi}{2}\right)$ masked until reaching N total steps.

For the base model, we use a variant of MaskGiT [4], a masked image generative Transformer to predict randomly masked tokens by attending to tokens in all directions. Leveraging the multi-layered structure of the Transformer, we have developed scalable image generators with varying layer counts, ranging in size from 257M parameters to 611M parameters (for the image generator; the Phi-2 model has an additional 2.7B parameters). We first employ a series of Cross Attention blocks to optimize text-driven feature extraction, before passing through O layers of the masked image generator. Each layer o of the Transformer is again formed by Multi-Head Self-Attentuib(MSA), LayerNorm (LN), Cross Attention (CA) and Multi-Layer Perceptron (MLP) blocks:

$$Y_o = \mathrm{MSA}(\mathrm{LN}(Z_o)), \quad Z_{o+1} = \mathrm{MLP}(\mathrm{CA}((\mathrm{LN}(Y_o), c))). \tag{5}$$

At the output layer, to reduce the training burden, ConvMLP [18] is utilized to transform masked image embeddings into logits sets, aligning with the VQGAN codebook dimensions. Eventually, the reconstructed lower-resolution tokens are restored with the pre-trained 256×256 resolution VQGAN decoder to the pixel space, resulting in the generated image $\hat{x}$:

$$\hat{x} = \mathcal{D}(\mathrm{ConvMLP}(Z_O), \mathcal{Z}) \tag{6}$$

2.3 Training Strategy

Fig. 3 illustrates Fantasy's two-stage training approach. Following prior works[43, 35, 9], we employ large-scale pre-training to achieve general text-image concept alignment, and simultaneous fine-tuning of Phi-2 [24] and the masked image generator using high-quality instruction-image pairs.

Pre-training Stage. To perform general text-image concept alignment, the VQGAN and LLM weights are frozen, and only the image generator is pre-trained on deduplicated LAION-2B [37] with images above a 4.5 aesthetic score. We exclusively preserve prompts in English, filter out images above a 50% watermark probability or above a 45% NSFW probability, yielding a final set of 9 million images. Since the computational cost of upsampling is much lower than training a super-resolution model, Fantasy is started with training at a resolution of 256×256. Note that the pre-training only needs approximate image-text alignment, substantially lowering the training costs.

Fine-tuning Stage. [43] has proven that LLMs trained solely on text data lack prior image-text knowledge, and that merely aligning their text embeddings with visual features might not be optimal. Therefore, in the second stage, we gather an internal dataset of 7 million high-quality instruction-image pairs to fine-tune both the Phi-2 model and the image generator of Fantasy, which ensures enhanced compatibility of text embeddings within the text-image pair space, facilitating the use of decoder-only LLMs in text-to-image generation tasks and harnessing their inherent advantages. To prevent catastrophic forgetting in LLMs and preserve their understanding abilities during training, we select questions from BIG-bench [2] and monitor the common sense question-answering ability of Phi-2 in real-time throughout the training process. We construct our training dataset for the fine-tuning stage by incorporating JourneyDB [39] and an internal synthetic dataset to enhance the aesthetic quality of generated images beyond realistic photographs. To facilitate instruction-image alignment learning, we retain only data with descriptions exceeding 30 words, as these provide enough detailed insights into the image objects, including attributes and spatial relations.

With this approach, Fantasy trains a 0.6B parameter T2I model in about 69 A100 GPU days, significantly reducing computation compared to existing diffusion-based methods, while maintaining comparable visual and numerical fidelity. Throughout this paper, we present a comprehensive evaluation of Fantasy's efficacy, showcasing the potential in training high-quality transformer-based image synthesis models compared to diffusion-based models in future.

2.4 High-quality Data Collection

To ensure rough alignment in the pre-training phase, we utilize the large-scale dataset LAION-2B [37] and apply the filtering strategy developed by DataComp [14]. Furthermore, we gather long-text prompts and corresponding high-quality images to achieve finer-grained text-image alignment through instruction tuning. CapsFusion [47] employs a fine-tuned LLaMA [40] for recaptioning LAION-2B [37] and LAION-COCO [1]. However, this approach still results in suboptimal image quality and occasional mismatches between images and text. SAM-LLAVA [5] utilizes LLaVA [20] to recaption the SAM dataset [17], which leads to images with blurred faces, a consequence of the dataset's inherent face-blurring. Therefore, we shift focus to synthesize images, mainly including DiffusionDB [42] and JourneyDB [39], produced by Stable Diffusion and MidJourney, respectively. To augment the diversity of the images, we minimize the use of datasets from specific domains, such as gaming and anime. Furthermore, we implement filtering to discard texts with special characters and data containing violence or pornography, retaining only instructions exceeding 30 words.

3 Experiments

In this section, we outline detailed training, inference, and evaluation protocols, followed by comprehensive comparisons across three key metrics.

3.1 Implementation Details

Training Details. Different from the prior works [9, 43, 32, 34], we used a lightweight but powerful decoder-only large language model Phi-2 [24] as the text encoder. Diverging from prior approaches that extract a standard and fixed short text tokens, we extend the extraction to 256 tokens to master long-term instruction-image alignment, ensuring precise alignment for more fine-grained prompts. For the entire training process, we train Fantasy on $4\times$A100 80G GPUs and set the accumulation step to 2. At different stages, we employ varying learning rate strategies with single-cycle cosine annealing decay. Furthermore, the AdamW optimizer [22] is utilized with a weight decay of 0.01. Fantasy trains a 0.6B parameter T2I model in about 84.5 A100 GPU days, significantly reducing computation compared to existing diffusion-based methods as shown in Fig. 1.

Table 1: Evaluation of diffusion (upper) and transformer (down) models on HPSv2. We underline the highest value and color the first above Fantasy in blue.

Model	Type	Params	Animation	Concept-art	Painting	Photo	DrawBench [36]
GLIDE [25]	Diff	5.0B	23.34 ± 0.198	23.08 ± 0.174	23.27 ± 0.178	24.50 ± 0.290	25.05 ± 0.84
VQ-Diffusion [15]	Diff	0.37B	24.97 ± 0.186	24.70 ± 0.149	25.01 ± 0.145	25.71 ± 0.222	25.44 ± 0.83
Latent Diffusion [34]	Diff	1.45B	25.73 ± 0.125	25.15 ± 0.140	25.25 ± 0.178	26.97 ± 0.183	26.17 ± 0.85
DALL·E 2 [26]	Diff	6.5B	27.34 ± 0.175	26.54 ± 0.127	26.68 ± 0.156	27.24 ± 0.198	27.16 ± 0.64
Stable Diffusion v1.4 [33]	Diff	0.8B	27.26 ± 0.156	26.61 ± 0.082	26.66 ± 0.143	27.27 ± 0.226	27.23 ± 0.57
Stable Diffusion v2.0 [33]	Diff	0.8B	27.48 ± 0.174	26.89 ± 0.076	26.86 ± 0.120	27.46 ± 0.198	27.31 ± 0.68
DeepFloyd-XL [11]	Diff	4.3B	27.64 ± 0.108	26.83 ± 0.137	26.86 ± 0.131	27.75 ± 0.171	27.64 ± 0.72
LAFITE [48]	Trans	0.075B	24.63 ± 0.101	24.38 ± 0.087	24.43 ± 0.155	25.81 ± 0.213	25.23 ± 0.72
FuseDream [21]	Trans	-	25.26 ± 0.125	25.15 ± 0.107	25.13 ± 0.183	25.57 ± 0.248	25.72 ± 0.71
DALL·E mini [7]	Trans	0.4B	26.10 ± 0.132	25.56 ± 0.137	25.56 ± 0.112	26.12 ± 0.233	26.34 ± 0.76
VQGAN + CLIP [8]	Trans	0.2B	26.44 ± 0.152	26.53 ± 0.075	26.47 ± 0.111	26.12 ± 0.210	26.38 ± 0.43
CogView2 [12]	Trans	6B	26.50 ± 0.129	26.59 ± 0.119	26.33 ± 0.100	26.44 ± 0.271	26.17 ± 0.74
Fantasy (ours)	Trans	0.6B	**27.03±0.131**	**26.66±0.117**	**26.72±0.176**	**26.80±0.174**	**26.78±0.523**

Table 2: Comparison with recent T2I models. 'Trained' indicates the model develops a text encoder from scratch, foregoing a pre-trained one.

Method	**Type**	**Text Encoder**	**#Params**	**#Images**	**FID-30K (↓)**
LDM [34]	Diff	Trained	1.4B	400M	12.64
GLIDE [25]	Diff	Trained	5.0B	-	12.24
DALL·E 2 [26]	Diff	CLIP	6.5B	650M	10.39
Stable Diffusion v1.5 [33]	Diff	CLIP	0.9B	2000M	9.62
SD XL [29]	Diff	CLIP	2.6B	-	>18
Würstchen [28]	Diff	CLIP	0.99B	1420M	23.6
ParaDiffusion [43]	Diff	LLaMA V2	1.3B	>300M	9.64
Pixart-α [5]	Diff	T5	0.6B	-	5.51
Cogview2 [12]	Trans	CogLM	6B	35M	24.0
Muse [3]	Trans	T5-XXL	3B	460M	7.88
Fantasy	Trans	Phi-2	0.6B	16M	23.4

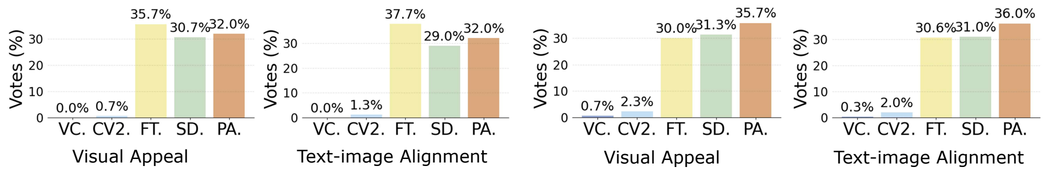
Inference Details. We use $N = 32$ sampling steps in all of our evaluation experiments. Since Fantasy is trained at a resolution of 256×256, we employ the pre-trained diffusion-based super-resolution model StableSR [41] to upscale images to 512×512.

Evaluation Metrics. We comprehensively evaluate Fantasy via four primary metrics, i.e., alignment on HPSv2 [44], FID [16] on MSCOCO dataset [19] and human evaluation on a collected dataset.

3.2 Performance Comparisons and Analysis

Results on HPSv2. We utilize HPSv2 [44] as our primary automated metric, a preference prediction model which can be used to compare images generated with the same prompt across five categories: anime, concept art, paintings, photography, and DrawBench [36]. We present the results of HPSv2 between Fantasy and other state-of-the-art generative models in Tab. 1. Fantasy exhibited outstanding performance across all key aspects among previous Transformer-based methods like CogView2 [12], which is expected. The results also reveal its competitive performance compared to prior diffusion-based methods, especially in concept-art and painting, demonstrating similar performance to DALL·E 2 [26]. This remarkable performance is primarily attributed to the text-image alignment learning in fine-tuning stage, where high-quality text-image pairs were leveraged to achieve superior alignment capabilities. In comparison, DeepFloyd-XL and other diffusion-based models achieve better scores, while utilizing larger models with significantly higher compute budget.

Results on FID. We employ FID [16] to evaluate our models on COCO-30K [19]. To allow for a fair comparison, all images are downsampled to 256×256 pixels. The comparison between our method and other methods in FID, and their training time is summarized in Tab. 2. We observe that the FID of Fantasy is substantially higher compared to other state-of-the-art models. Visual inspections reveal that images generated by Fantasy are smoother than those from other leading T2I models. This discrepancy is most noticeable in real-world images like COCO, on which we compute the FID-metric. Although the state-of-the-art models [43, 11, 29] exhibit lower FID, it relies on unaffordable resources. Furthermore, prior studies [29, 5, 11] have demonstrated that FID may not



(a) User study on long prompts. (b) User study on short prompts.

Figure 4: User study on prompts with different length. VC. , CV2. , FT. , SD. , and PA. refer to VQGAN+CLIP [8], CogView2 [12], our Fantasy, Stable Diffusion v2.0 [33], and Pixart-α [5].

be an appropriate metric for image quality evaluation, as a lower score does not necessarily reflect superior image generation, and it is more authoritative to use the evaluation of human users.

3.3 Results on Human Evaluation

Following prior works [5, 43, 28], we also conduct a study with human participants to supplement our evaluation and provide a more intuitive assessment of Fantasy's performance. Participants are asked to select a preference of the images based on the visual appeal of the generated images and the precision of alignments between the text prompts and the corresponding images.

As involving human evaluators can be time-consuming, we choose the top-performing open-source diffusion-based models (e.g., SD XL [33], and Pixart-α [5]) and transformer-based models (e.g., VQGAN+CLIP [8] and CogView2 [12]) as our baseline, which are accessible through APIs and capable of generating images. We randomly select a total of 600 prompts from existing prompt sets (e.g., ParaPrompt [43], ViLG-300 [13], COCO Captions [6]). To comprehensively contrast the capabilities of Fantasy and other models in interpreting text prompts of varying lengths, we allocate one subset to consist of 300 prompts ranging from 10 to 30 characters and another subset comprising 300 prompts exceeding 30 characters. For each model, we use a consistent set to generate images, which are then evaluated by 50 individuals.

Fig. 4a clearly demonstrates that images generated on relatively long text prompts (longer than 30 words) by Fantasy are distinctly favored among the four models in both two perspective, especially for text-image alignment, aligning closely with the intended use case of Fantasy. As illustrated in Fig. 4b, for text prompts shorter than 30 words, our model outperforms existing open-source Transformer-based models in fidelity and alignment for shorter prompts. Our model slightly lags behind diffusion-based models in visual appeal, limited by the 8,192 size of VQGAN's codebook and not targeting visual appeal. Simultaneously, Fantasy lacks a distinct advantage in text-image alignment in the short subset. We hypothesize that this is due to two main reasons: diffusion-based models' ability to handle shorter prompts, and vague prompts generating diverse images that make preferences more subjective, thus biasing outcomes towards aesthetically superior images. In summary, the human preference experiments confirm the observation made in the HPSv2 benchmarks.

3.4 Case Study

Fig. 5 vividly illustrates Fantasy's superior visual appeal and text-image alignment over leading open-source transformer-based T2I models [12, 8] and diffusion-based T2I models [29, 26]. Fantasy significantly surpasses existing transformer-based T2I models, matches the performance of SDXL [29], and qualitatively outperforms Dall·E 2 [26]. Despite being trained on images with a resolution of 256×256, Fantasy ensures generated low-resolution images contain sufficient details, indirectly supporting long prompts. Limited by computing resources, we haven't

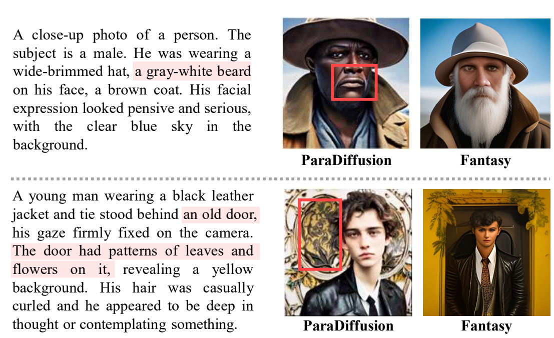


Figure 6: Visual Comparison with ParaDiffusion [43]: Red markings and boxes highlight text misalignments in images generated by ParaDiffusion.

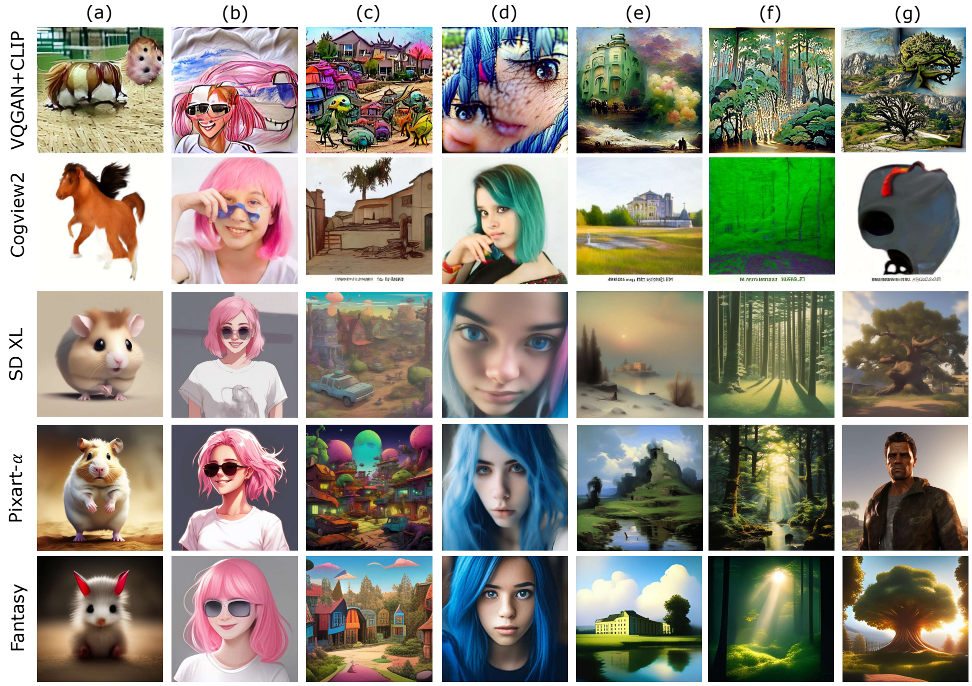


Figure 5: Visual comparison with existing T2I models. (a) *A hamster resembling a horse.* (b) *A frontal portrait of a anime girl with chin length pink hair wearing sunglasses and a white T-shirt smiling.* (c) *A colorful illustration of a suburban neighborhood on an ancient post-apocalyptic planet featuring creatures made by Jim Henson's workshop.* (d) *A blue-haired girl with soft features stares directly at the camera in an extreme close-up Instagram picture.* (e) *A building in a landscape by Ivan Aivazovsky.* (f) *Aoshima's masterpiece depicts a forest illuminated by morning light.* (g) *The image is a highly detailed portrait of an oak in GTA V, created using Unreal Engine and featuring fantasy artwork by various artists.*

Table 3: Ablation study on two stages with the best bolded. 'Base' indicates the model after the pre-training stage.

Model	Training Part	Animation	Concept-art	Painting	Photo	DrawBench [36]
Base	MIM	25.27 ± 0.190	24.20 ± 0.166	24.60 ± 0.146	25.32 ± 0.208	25.49 ± 0.230
Fantasy	MIM+Phi-2	**27.03±0.131**	**26.66±0.117**	**26.72±0.176**	**26.80±0.174**	**26.78±0.521**

trained on higher resolutions like 512×512 but aim to enhance Fantasy by training at higher resolutions in the future.

ParaDiffusion [43] pioneers the use of decoder-only large language models as text encoders in text-to-image generation. As illustrated in Fig. 6, our observations suggest that Fantasy more closely aligns details with prompts than ParaDiffusion [43].

4 Ablation Study

This section analyzes the effects of LLMs fine-tuning, and model scale on Fantasy's performance through ablation studies. More ablation study refers to appendix.

4.1 Effect of Language Model Fine-tuning

To assess the effect of training strategies on the comprehension of complex instructions, we perform a human preference evaluation, as detailed in Sec. 3.3, using a subset of 300 prompts longer than 30 characters. 'Base' denotes general text-image alignment with filtered LAION-2B [1] in the pre-training stage. Compared to the base model, our synergy fine-tuning with Phi-2 demonstrates a notable improvement in all aspects in Tab. 3.

Table 4: Ablation study on models at different scales with the best **bolded**. DB. represents DrawBench [36].

Layers	Param	Animation	Concept-art	Painting	Photo	DB.
6	257M	25.79±0.15	25.84±0.11	25.92±0.19	25.63±0.18	25.18±0.22
12	421M	26.34±0.17	26.29±0.06	26.45±0.17	26.19±0.17	25.68±0.14
22	611M	**27.03±0.13**	**26.66±0.11**	**26.72±0.17**	**26.80±0.17**	**26.78±0.52**

Table 5: Training cost for Fantasy at 3 different scales. BS. denotes batch size and LR. denotes learning rate.

Layers	Pre-training			Fine-tuning		
	Steps (K)	BS.	LR.	Steps (K)	BS.	LR.
6	180	768	1e-4	180	192	1e-4
12	220	768	1e-4	250	192	1e-4
22	370	256	5e-4	280	128	3e-4

4.2 Scale of Image Generator

The hierarchical structure of the Transformer allows us to train image generators with varying numbers of Transformer layers. As shown in Tab. 4, we evaluate models of different sizes on the HPSv2 benchmark. The insight indicates that as trainable parameters increase from 257 million to 611 million, performance consistently improves. Therefore, we set the number of Transformer layers to 22 with 611 million trainable parameters as the optimal setting. Tab. 5 showcases the required resources for models of three different scales. Fig. 7 offers visual comparisons across models of varying scales, illustrating a clear trend: models with fewer parameters underperform on the HPSv2 benchmark, frequently resulting in distorted images and omitted details, yet they may still generate acceptable outcomes. Significantly, the visual quality diverges as model size increases, highlighting the potential for scaling up masked image modeling to enhance instruction-image alignment and elevate generation quality.

Figure 7: Examples generated by models at different scales: 1^{st} column for 6 layers, 2^{nd} column for 12 layers and 3^{rd} column for 22 layers.

5 Limitations and Social Impact

Limitations. Despite Fantasy achieving competitive performance in text-image alignment and visual appeal, it requires improvements in handling complex scenes. We propose two possible strategies to overcome the challenge in future research: Firstly, augmenting the dataset with high-quality images can enhance diversity and refine the model. Secondly, since the scale of the masked image generator affects instruction-image alignment, training an upscale image generator based on higher resolution left further explored.

Social Impact. Generative models for media bring both benefits and challenges. They foster creativity and make technology more accessible, yet pose risks by facilitating the creation of manipulated content, spreading misinformation, and exacerbating biases, particularly affecting women with deep fakes. Concerns also include the potential exposure of sensitive training data collected without consent. Despite generative models potentially offering better data representation, the impact of combining adversarial training with likelihood-based objectives on data distortion remains a crucial research area. Ethical considerations of these models are significant and require thorough exploration.

6 Conclusion

In this paper, we introduce Fantasy, a lightweight and efficient text-to-image model that combines Large Language Models (LLMs) with a transformer-based masked image modeling (MIM), effectively transferring semantic understanding capabilities from LLMs to the text-to-image generation. With our proposed two-stage training strategy and high-quality dataset, Fantasy significantly reduces computational requirements while producing high-fidelity images. Extensive experiments demonstrate that Fantasy achieves comparable performance to models trained with significantly more computational resources, illustrating the viability of our approach and suggesting potential efficient scalability to even larger masked image modeling for text-to-image generation.

References

[1] Köpf Andreas, Vencu Richard, Coombes Theo, and Beaumont Romain. Laion coco: 600m synthetic captions from laion2b-en.[eb/ol], 2022.

[2] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arxiv 2015. *arXiv preprint arXiv:1504.00325*, 2015.

[7] Craiyon. Dall·e mini: Generate images from any text prompt. `https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-mini-Generate-images-from-any-text-prompt--VmlldzoyMDE4NDAy`, 2023. Accessed: 2024-02-27.

[8] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[9] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[10] Deepfloyd. Deepfloyd. `https://www.deepfloyd.ai/`, 2023.

[11] DeepFloyd. IF-I-XL-v1.0: A model by deepfloyd on hugging face models. `https://huggingface.co/DeepFloyd/IF-I-XL-v1.0`, 2023. Accessed: 2024-02-28.

[12] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

[13] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023.

[14] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

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

[17] 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. *arXiv preprint arXiv:2304.02643*, 2023.

[18] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6306–6315, 2023.

[19] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[21] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021.

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

[23] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

[24] Microsoft. Phi-2. `https://huggingface.co/microsoft/phi-2`, 2023.

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[26] OpenAI. Dall-e 2. `https://openai.com/dall-e-2`, 2022.

[27] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction. *arXiv preprint arXiv:2401.01808*, 2024.

[28] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[29] 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 preprint arXiv:2307.01952*, 2023.

[30] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

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

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022. *URL https://arxiv. org/abs/2205.11487*, 4.

[36] 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. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[37] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[38] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023.

[39] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[41] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023.

[42] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[43] Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model. *arXiv preprint arXiv:2311.14284*, 2023.

[44] 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.

[45] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024.

[46] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[47] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*, 2023.

[48] Y Zhou, R Zhang, C Chen, C Li, C Tensmeyer, T Yu, J Gu, J Xu, and T Sun. Lafite: Towards language-free training for text-to-image generation. arxiv 2021. *arXiv preprint arXiv:2111.13792*.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly define the paper's contributions, which involve advancements in urban simulation accuracy and computational efficiency. These claims are backed by robust experimental validation detailed in the subsequent sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a comprehensive discussion on limitations, particularly focusing on the scalability of our simulations in extremely large urban environments and potential biases in the modeling processes.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are accompanied by a clear statement of assumptions and are supported by complete proofs provided in the supplementary materials. Each theorem and lemma are properly referenced and numbered for clarity and ease of access.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of the experimental setup, including data splits, hyperparameters, and the type of optimizer used. We also provide access to the source code and datasets in the supplementary materials to ensure full reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper does not propose a benchmark and we will release the code if the paper is accepted. The model depends on non-open-sourced dataset, and the copyright of the checkpoint belongs to the company. Detailed instructions for training our model, including command lines, are provided in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental section of the paper provides comprehensive details about the training and test setups, including the rationale behind choosing specific hyperparameters and the types of optimizers used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experimental results are presented with error bars reflecting the standard deviation across multiple runs. We provide a detailed explanation of how these were calculated and the assumptions underlying our statistical tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper details the computational resources required for each experiment, including the types of GPUs used, the amount of memory, and the execution time. This ensures that other researchers can allocate the appropriate resources to reproduce our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres strictly to the NeurIPS Code of Ethics. We have considered ethical implications, especially regarding the generation of images from text, and have implemented measures to prevent misuse.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes content about broader impacts that discusses both the potential positive applications of our method in educational and creative industries, and potential negative impacts, such as the misuse of generated images. We also suggest mitigation strategies for potential negative uses.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks. If then, we will describe the safeguards implemented in releasing our models, including usage guidelines and limitations to access, ensuring responsible use and mitigating risks of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All third-party assets used in our research are properly credited, and we have explicitly mentioned and complied with the licensing terms. URLs and version numbers of datasets and code are clearly listed in the references.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Any new datasets or models introduced in the paper are accompanied by thorough documentation detailing their creation, intended use, limitations, and licensing information.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research did not involve human subjects, thus no IRB approval was necessary.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.