

# Supplementary Material

The appendix is structured as follows:

- We introduce the related work of Fantasy(A.1).
- We first outline the configuration of our masked image generator for the largest scale of Fantasy (A.2).
- We then explain the necessity of considering LLMs as the text encoder (A.3).
- We illustrate an example of the scheduled parallel decoding process while image generation (A.4).
- We provide cases on adopting Styledrop [39] to Fantasy (A.5).
- We further provide examples on implementing Fantasy for the image inpainting task (A.6).
- We demonstrate failure cases and analyze potential solutions (A.7).
- We show the source distribution for the test sets used in the human evaluation (A.8).
- We visualize the image quality of different sampling steps during inference and provide detailed quantitative results of HPSv2 on various sampling steps (B.1).
- We provide both quantitative and qualitative results, comparing the images upscaled by different super-resolution models (B.2).
- We further show qualitative results, delving into the effect of language model fine-tuning (B.3).
- We discuss the impact of VQGAN fine-tuning (B.4).
- We provide more samples produced by Fantasy (C.1).
- We indicate more visual comparison with leading T2I models (C.2).

## A Implementation Details

### A.1 Related Work

#### A.1.1 Text-to-image Generation

Fueled by the large-scale text-image datasets [38, 10], significant progress has been made in text-to-image generation models, which can be primarily classified into two categories: diffusion-based methods and Transformer-based methods.

**Diffusion-based methods.** A large amount of diffusion-based models [35, 48, 54] approach the text-to-image generation task, which have showcased unprecedented levels of diversity and fidelity. Imagen [37] first incorporates a pre-trained large transformer language model to encode textual input for image generation. However, these advanced models invariably demand significant computational resources for training. For instance, Stable Diffusion v1.5 [31], one of the most notable models in the field, necessitates 6K A100 GPU days. Paradiffusion [46], the recent paragraph-to-image generation diffusion model, approximately costs 392 A100 GPU days for entire training. Additionally, Pixart- $\alpha$  [7] requires 753 A100 GPU days and WÜSTCHEN [30] needs 1025 A100 GPU days. The emergence of a series of methods represented by diffusion models become the dominant method for image [3, 2], video [4], and 3D [28, 9] generation.

**Non-diffusion-based methods.** VQ-VAE [43] and GANs [15] have shown excellent text-to-image generation performance with variants proposed for both convolutional and Transformer architectures. Cogview2 [11] utilizes hierarchical transformers and local parallel auto-regressive generation to speed up and simplify text-to-image models for high-resolution images. StyleSwin [53] leverages Swin transformer-based [23] GAN as the basic building block. NUWA-LIP [26] incorporates a defect-free VQGAN with multi-perspective sequence to sequence. Muse [5] trains two VQGAN models to generate images with high resolution.

Table 1: Configuration and training hyperparameters for masked image generator.

Configuration	Value
Number of Transformer Layers	22
Number of Attention Heads	16
Transformer Hidden Dimension	2560
Transformer MLP Dimension	8192
Layer Normalization Epsilon	$1e - 6$
Optimizer	AdamW
Weight Decay	0.01
Optimizer Momentum	$\beta_1=0.9, \beta_2=0.999$
Accumulation Steps	2
Learning Rate Schedule	cosine decay
Warmup Steps	2000

### 45 A.1.2 Masked Image Modeling

46 Masked image modeling (MIM) significantly advances computer vision training through unsupervised  
 47 learning. BEiT [1] first demonstrates that unsupervised pre-training in computer vision can achieve  
 48 equal or even better results than supervised pre-training by incorporating the Masked Language  
 49 Modeling (MLM). MAE [18] adheres to the spirit of raw pixel restoration, demonstrating for the  
 50 first time that masking a high proportion of the input images can yield a non-trivial and meaningful  
 51 self-supervisory task. MIM has been applied to multiple downstream tasks in computer vision,  
 52 including object detection [13, 19, 40], medical image processing [55, 33, 47], video representation  
 53 [41, 16, 50], 3D point cloud [52, 22, 32] and so on. In this paper, building upon the prior works  
 54 [20, 17, 51], we further explore the application of MIM in image generation.

### 55 A.2 Masked Image Generator Configurations

56 Our masked image generator configuration for our largest model of size 0.6B parameters is given in  
 57 Tab. 1.

### 58 A.3 Necessary of LLMs Using

59 To adhere to prompt instructions, various existing models employ CLIP [45] as their text encoder,  
 60 which is trained on images with predominantly short text pairs. However, these models encounter  
 61 difficulties in handling comprehend dense prompts, particularly when the text describes multiple  
 62 objects, detailed attributes, complex relationships, long-text alignment, etc. Some models [7, 46]  
 63 investigate the incorporation of powerful Large Language Models (LLMs), such as T5 [25] and  
 64 LLaMA-2 [42], to achieve a deeper level of language understanding in text-to-image generation.  
 65 Imagen [37] first demonstrates that text features from LLMs pre-trained on text-only datasets are  
 66 remarkably effective in enhancing text alignment for text-to-image synthesis. Nonetheless, current  
 67 models [7, 46, 5] that employ LLM as a text encoder necessitate the full training of the image generator,  
 68 and ParaDiffusion [46] even fine-tunes the pre-trained LLaMA-2 [42]. Aside from consuming vast  
 69 computational resources, these models are difficult to integrate with the burgeoning community  
 70 models and downstream tools. To leverage the language understanding capability of LLM and the  
 71 image generation potential of Transformer-based models and bridge them effectively, we consider the  
 72 pre-trained Phi-2 [24] as the text encoder, which provides the comprehensive text feature extracted  
 73 from the last hidden state to condition image generation.

### 74 A.4 Scheduled Parallel Decoding with Masked Generative Imgae Transformer

75 Autoregressive Transformers [12, 34] flatten an image into a 1D sequence of tokens following a raster  
 76 scan ordering. MaskGIT [6] adopts a novel non-autoregressive decoding method to synthesize an  
 77 image in constant number of steps. As we employ MaskGIT as the backbone of the masked image  
 78 generator, Fantasy generates an image and the predictions within each step are parallelizable. Fig. 2  
 79 illustrates an example of our decoding process in  $T = 32$  iterations.

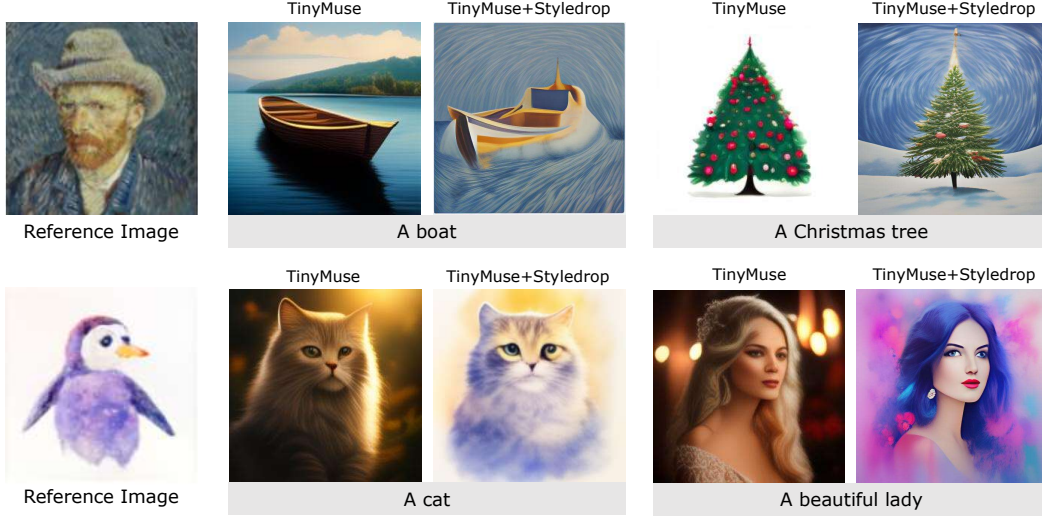


Figure 1: Examples of Styledrop with Fantasy.

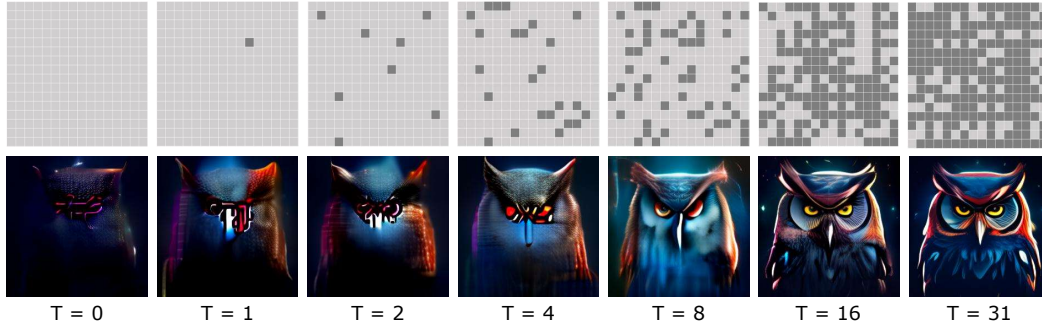


Figure 2: Fantasy's scheduled parallel decoding. Rows 1 is the input latent masks at each iteration, and row 2 are the samples generated at that iteration. Our decoding starts with all unknown codes (marked in lighter gray), and gradually fills up the latent representation with more and more scattered predictions in parallel (marked in darker gray), where the number of predicted tokens increases sharply over iterations.

## 80 A.5 Exploration in Styledrop

81 Styledrop [39] introduces an effective method for effective single example image style adoption,  
 82 optionally enhanced by generating extra training samples for dataset expansion. As illustrated in  
 83 Fig. 1, we achieve good results with fine-tuning on a single image and not generating any additional  
 84 training samples. Styledrop can cheaply fine-tune Fantasy in as few as 150-200 training steps.

## 85 A.6 Implement in Image Inpainting Task

86 As shown in Fig. 3, the sampling procedure of Fantasy enables us to perform zero-shot text-guided  
 87 image inpainting. We convert an input image into a set of tokens, mask out the tokens corresponding  
 88 to a local region, and then sample the masked tokens conditioned on unmasked tokens and a text  
 89 prompt.

## 90 A.7 Failure Cases for Fantasy

91 While Fantasy demonstrates proficiency in generating images of visual appeal and text-image align-  
 92 ment, it still encounters persistent issues. As illustrated in Fig. 4, Fantasy has a significant shortcoming  
 93 in text rendering. Regardless of whether text prompts explicitly request the display of characters,



Figure 3: Examples of zero-shot text-guided image inpainting using Fantasy.



(a) Explicitly request to generate images with text included.



(b) Implicitly request to generate images with text included.

Figure 4: Failure Cases for Fantasy.

94 the generated images fail to display English characters properly. Other state-of-the-art open-source  
 95 generative models also fail to accomplish this task, but the performance of Stable Diffusion XL [36]



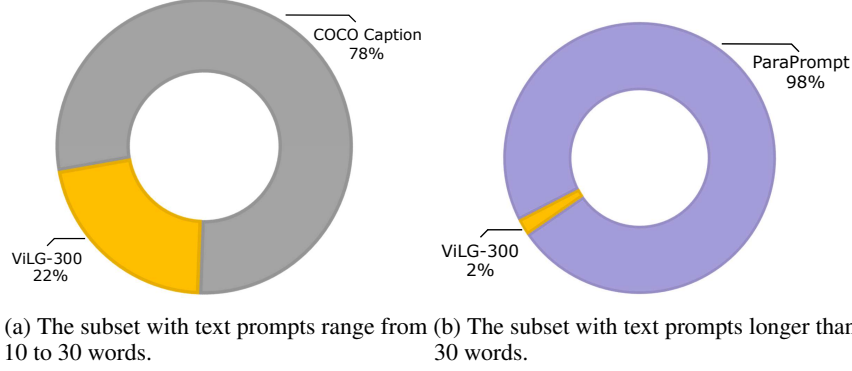


Figure 5: Distribution of the Sources of the Test Prompts.



Figure 6: Visual appeal comparison between images generated from different sampling steps.

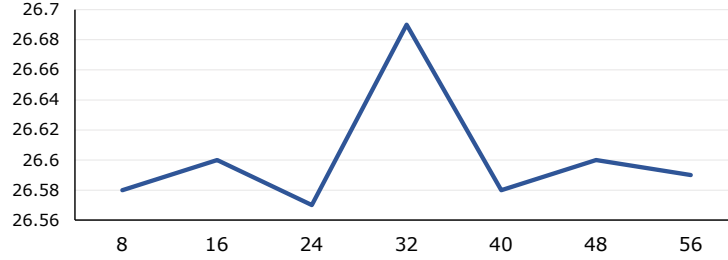


Figure 7: The relationship between HPSv2 (represented by metrics on Drawbench) and sampling steps.

is notably superior. We speculate that this may be related to a lack of relevant training data. This assumption warrants further investigation and remediation in our forthcoming research endeavors.

#### A.8 Source Distribution for the Test Sets in User Study

We collect 600 text prompts from ParaPrompt [46], ViLG-300 [14], and COCO Caption [8], and divide them into two subsets based on the length of the prompts. The distribution of the sources of these two subsets are depicted in Fig. 5.

## B Ablation Study

### B.1 Influence of Sampling Steps during Inference

We visualize the evolution of masked tokens over the sequence of steps for the masked image generator in Fig. 6. Additionally combining with Fig. 7, we set sampling steps to 32 during inference.

### B.2 Differences between Super-resolution Models

The pre-trained VQGAN decoder reconstructs generated masked image representation to pixel space with a resolution of  $256 \times 256$ . Our trials with various super-resolution models (e.g., SwinIR [21], StableSR [44], and SUPIR [49]) to upscale images to  $512 \times 512$  resolution reveal minor metric

Table 2: Ablation study on various super-resolution models with the best bolded. ‘Base’ refers to the generated images without upscaling.

Model	Animation	Concept-art	Painting	Photo	DrawBench [37]
Base	$26.61 \pm 0.147$	$26.37 \pm 0.132$	$26.40 \pm 0.169$	$26.35 \pm 0.186$	$26.48 \pm 0.57$
SwinIR [21]	$26.61 \pm 0.208$	<b><math>26.84 \pm 0.129</math></b>	$26.52 \pm 0.166$	$26.47 \pm 0.102$	$26.76 \pm 0.59$
StableSR [44]	$27.03 \pm 0.131$	$26.66 \pm 0.117$	<b><math>26.72 \pm 0.176</math></b>	$26.80 \pm 0.174$	$26.78 \pm 0.52$
SUPIR [49]	<b><math>27.10 \pm 0.150</math></b>	$26.63 \pm 0.116$	$26.67 \pm 0.181$	<b><math>26.95 \pm 0.181</math></b>	<b><math>26.85 \pm 0.44</math></b>

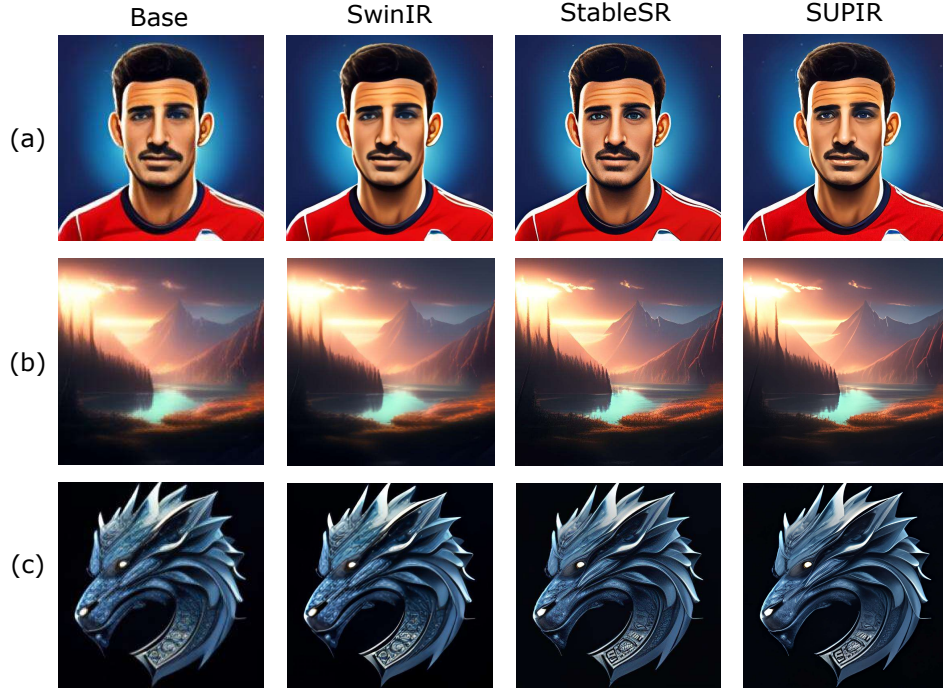


Figure 8: Comparison between Images Upscaled by Different Super-resolution Models. (a) *Lionel Messi portrayed as a sitcom character*. (b) *Concept art of a highly detailed landscape, centered and utilizing rule of thirds, with dynamic lighting for a cinematic effect*. (c) *Metal dragon head badge with detailed relief, displaying a digital painting of concept art, illustrated by Giger, Rutkowski, Shimoda, Leighton, and Bowater*.

fluctuations across models as detailed in Tab. 2, yet the overall performance remains comparable to DALL-E 2 [27]. Different super-resolution models retain the low-resolution images’ object attributes and the relationships. As shown in Fig. 8, different super-resolution models retain the low-resolution images’ object attributes and the relationships while marginally enhance visual appeal. This stability in maintaining text-image alignment, unequivocally highlights Fantasy’s intrinsic strengths in generating high-quality images.

### B.3 Effect of Language Model Fine-tuning

Fig. 9 showcases the ablation study concerning the language model adaptation in the fine-tuning stage. It is obvious that after the joint fine-tuning of MIM and Phi-2, the generated images not only become visually more appealing but also demonstrate a significant enhancement in text faithfulness, aligning closer to finer text-image alignment.

### B.4 Discussion of VQGAN Fine-tuning

To reconstruct more sharper details without re-training any of the other model components, Muse [5] increase the capacity of the VQGAN decoder by the addition of more residual layers and channels

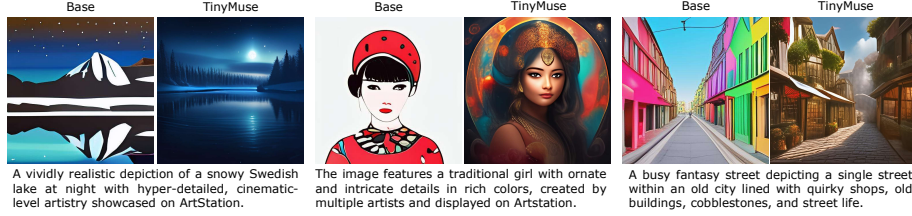


Figure 9: Ablation Study for Language Model Fine-tuning.



Figure 10: Visual example of the difference between the VQGAN reconstruction and the reconstruction with a finetuned decoder. We can see especially that fine details are worse preserved in the finetuned decoder. (a) *A piece of fine art art photography titled The Dawn of Li River by Yan Zhang.* (b) *A portrait of the photographer’s uncle looking at electrical equipment.* (c) *Infinity Dining Table.* (d) *Mansion in rich residential area.* (e) *This cave in Vietnam is so big it can fit a complete city.* (f) *Summer in the Dolomites of the lake’s most affluent and many colors.*

while keeping the encoder capacity fixed. To improve the reconstruction of high-resolution images, aMused [29] further fine-tunes the VQGAN decoder on a dataset of images greater than  $1024 \times 1024$  resolution. Therefore, to enhance the visual appeal of images generated by Fantasy, we explore the implementation of a third stage to fine-tune both the masked image generator and the VQGAN decoder on a synthetic dataset containing images with heightened aesthetic quality. However, as illustrated in Fig. 10, after fine-tuning for 14K steps, we observe that while the content and structure of the generated images remain largely unchanged, the edges of objects within the images significantly blur, consequently diminishing their visual appeal. On the contrary, we believe that solely fine-tuning the VQGAN decoder on the data with high-quality images is insufficient to achieve significant improvements in image quality, and it is necessary to fine-tune the VQGAN encoder simultaneously.

## C Case Study

### C.1 More Samples Produced by Fantasy

We provide more samples produced by Fantasy as illustrated in Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, and Fig. 17.

### C.2 More Visual Comparison with Existing T2I Models

To gain a more intuitive understanding of Fantasy, we supplement more examples comparing Fantasy with other leading models in Fig. 18.



A capybara wearing sunglasses.



An image portraying Barry Lyndon.



A building in a landscape.



New York Skyline with fireworks on the sky.



There is a old black motorcycle inside of a garage.



A bouquet of bright flowers grows in a vase.



Kurdish soldier, highly detailed, digital painting, award winning art, sharp focus .



Fruit in a jar filled with liquid sitting on a wooden table.



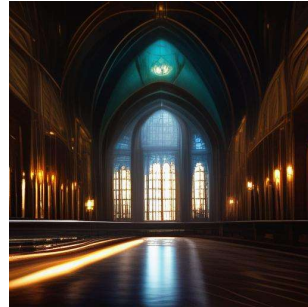
A bald general with an angry expression in an intricately detailed digital painting.



Portrait of Beautiful blonde Slavic woman in her early 30's, league of legends, fantasy, digital painting, artstation, concept art, sharp focus.



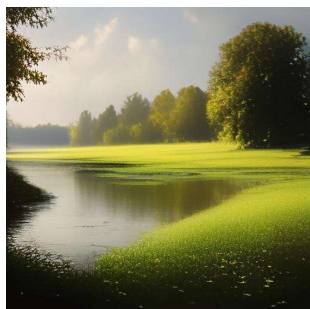
A photorealistic portrait of a colorful fantasy landscape with a hyper-realistic river, mountains, trees, and bright blue sky.



Majestic ornate great hall, grand library, baroque, torches, stained glass windows, moonlight rays, dreamy mood.

Figure 11: Samples Produced by Fantasy.

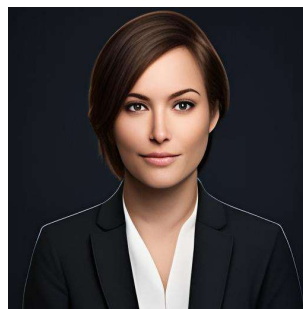




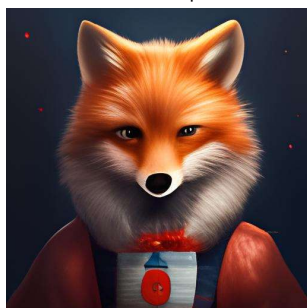
A scene of a scenic place.



A fursona, furry art.



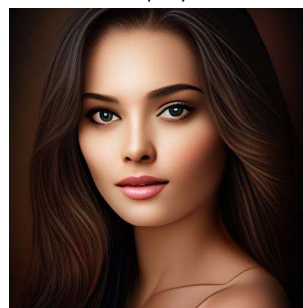
A woman company CEO.



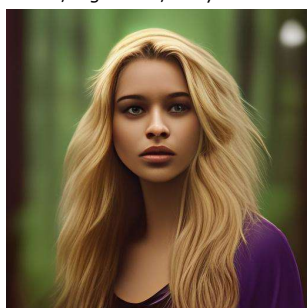
A fox fursona wearing a maid outfit, digital art, furry art .



A beautiful painting tyrannosaurus rex.



A beautiful picture of beautiful lady.



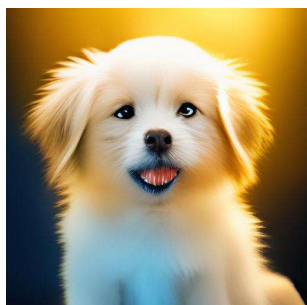
A breathtakingly beautiful young man with flowing blonde hair and amethyst eyes standing in a forest.



A 3 d render of a cute, blue, anthropomorphic dragon with ice crystals growing off her, sharp focus, unreal engine.



A detailed portrait of a frog, digital art, realistic painting, character design.



A beautiful portrait, golden background, gorgeous fantasy cute puppy, professionally retouched, soft lighting, realistic, long white hair, cute ears, wide angle, sharp focus on the eyes.



Pencil drawing of a beautiful greek goddess aphrodite wearing a laurel wreath and arrowhead earrings, beautiful confident eyes, beautiful flowing hair, glowing god eyes, hyper realistic face.



A memorial grove of trees of various sizes dedicated to missing people, metal plaques, solemn, brooding, somber tone, surreal dream landscape.

Figure 12: Samples Produced by Fantasy.





An avocado armchair.



Ted bundy in a pixar movie.



Wolf geometric wall art.



A portrait of a medieval anthropomorphic foxdigital art, backlighting.



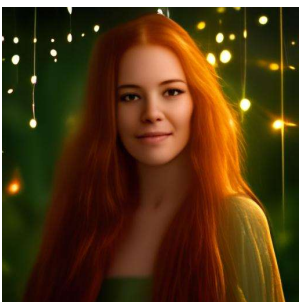
An anthropomorphic jackal anubis wearing sunglasses and a leather jacket, furry art.



A serene landscape with a singular building in the style of Anton Pevzne.



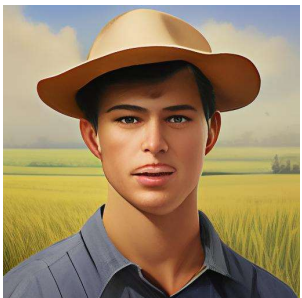
A mushroom house in a tall mushroom, small door and windows in the mushroom, warm light coming from the windows, in a dark forest, macro, cool tones.



Adorable woman, serene smile surrounded by golden firefly lights, amidst nature fully covered by a intricate detailed dress, long red hair, precise linework, smooth oval shape face, empathic, expressive .



Cinematic shot epic portrait an male survivor wearing a brown jacket and a white dirty shirt, dirty clothes, beard, short hair, serious, broad light, ambient occlusion, volumetric light effect.



Portrait of a handsome young ohio farmer.

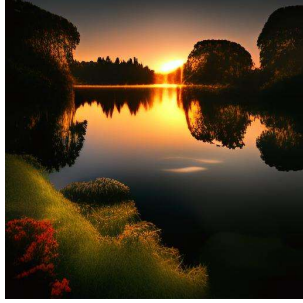


Portrait of a beautiful french empress royalty.



Diana the huntress, portrait, ultrarealistic.

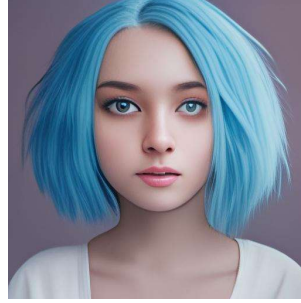
Figure 13: Samples Produced by Fantasy.



A beautiful landscape of the garden of eden. lake reflections in the foreground, sunset, dramatic lighting.



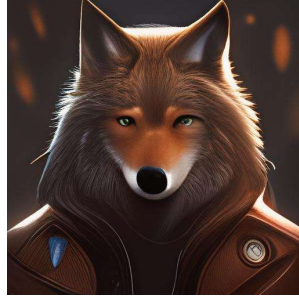
A lomographic photo of old lada standing in typical soviet yard in small town, hrushevka on background.



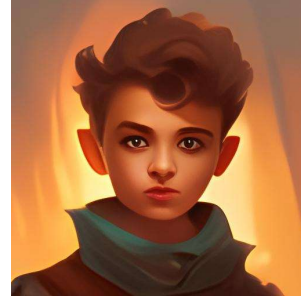
A pale girl with blue hair, soft facial features, round face, looking directly at the camera, neutral expression.



An older elf black-haired, archer with leather, armor and long bow, torso concept character art.



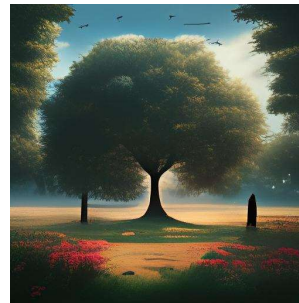
Athro female wolf wearing a leather jacket, Mark Edward Fischbach, intricate, furry digital painting.



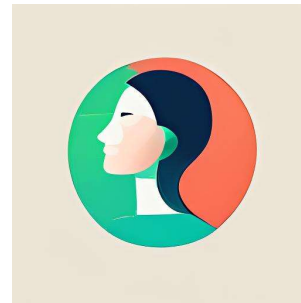
Cute little anthropomorphic rick astley cute and adorable, pretty, beautiful, dnd character art portrait, matte fantasy painting.



A portrait of a bear, full medium shot, abreast, volumetric, baroque, rococo, tarot card color scheme, cinematic lights, artistically realistic.



A memorial grove of trees dedicated to missing people, plaques, solemn, flowers game art matte painting hyperdetailed, surreal dream landscape.



Attention filter, memory, minimalist logo without text for a research lab that studies human cognition, vector art, minimalism.



Badger eating popcorn, professional photography.



Beautiful ink sketch, a close-up byzantine princess.



Face icon stylized minimalist samurai warrior.

Figure 14: Samples Produced by Fantasy.





Awesome cute mongrel dog portrait intricate artwork , beautiful, very coherent symmetrical artwork, cinematic, hyper realism, vibrant colors.



Beautiful aesthetic inspirational masterful professional ink pen and watercolor sketch of an occult mystic flower, ultra detailed.



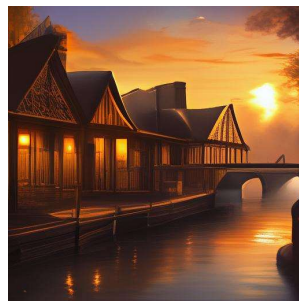
Combat android, gun for a head, katana arms, skeletal face, military prototype, photography, desert, photorealism, mercury metal.



Beautiful aesthetic inspirational masterful professional ink pen and watercolor sketch of a fantasy gazebo, ultra detailed.



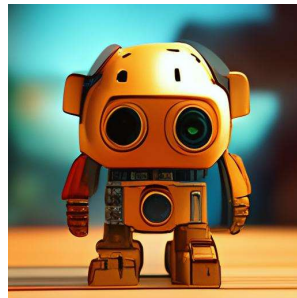
Beautiful aesthetic inspirational masterful professional ink pen and watercolor sketch of a palace, ultra detailed.



Beautiful warm tavern seen from the outside, middle age, river crossed by a bridge next to the tavern, crepuscular light.



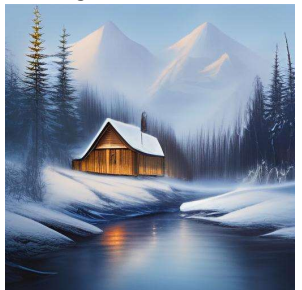
Close up photo of a fennec fox wearing round lense glasses with a golden frame .



Cute funny figurine wooden, retrorobot, brtualism, concept art, digital art.



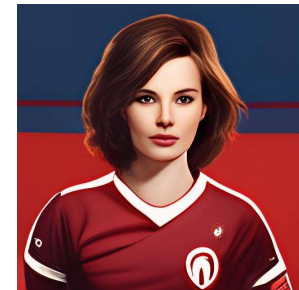
A person staring into a lucid dream world with an adventure waiting.



Snowy cozy home by a small river in a forest in canadian mountains.



Still portrait of darth avder as the joker in the new star wars movie.



Emma watson in new England patriots football uniform fanart, digital art.

Figure 15: Samples Produced by Fantasy.



Awesome cute mongrel dog portrait intricate, very coherent symmetrical artwork, caramel, cinematic, hyper realism, vibrant colors.



Complex 3d render hyper detailed ultra sharp beautiful futuristic stunning biomechanical humanoid woman with porcelain ivory face, close-up, filigree lace.



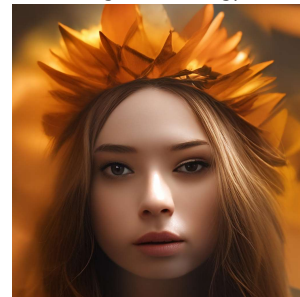
The image is a vibrant and intricate illustration of a man, with a focus on his shoulder and head, created using inkpen and Unreal Engine technology.



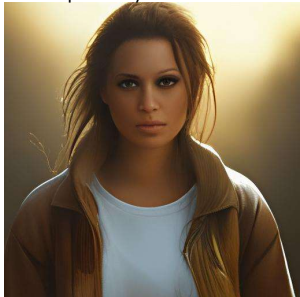
Imogen Poots portrayed as a D&D Paladin in a fantasy concept art by Tomer Hanuka.



A psychedelic shaman with celtic tattoos in an ancient temple.



A close-up portrait of a beautiful girl with an autumn leaves headdress and melting wax.



Cinematic shot epic portrait an female survivor wearing a brown jacket and a white T-shirt, dirty clothes, shiny skin, tied hair, broad light, ambient occlusion, volumetric light effect.



Clouds and waves, an aesthetically pleasing, energetic, lively, complex, intricate, detailed, fine art of a beach, ripples, waves, sea foam, light and shadow, overlaid with aizome patterns.

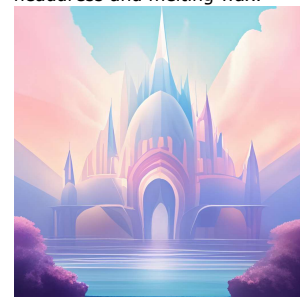
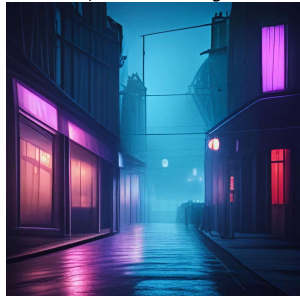


Image featuring a crystal palace with a dream-like guide line composition and a soft Monet-inspired tintal effect, created by multiple artists and trending on Artstation.



A cyberpunk street scene in Saint-Petersburg.



A drawing of a female warrior in full body pose.



Wonderful princess with fair skin, accent lighting, dramatic light.

Figure 16: Samples Produced by Fantasy.

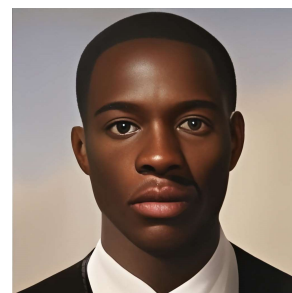




Sharp focus, breath taking beautiful, aesthetically pleasing, gouache ocean waves ripples, sea foam, sunset, digital concept art background.



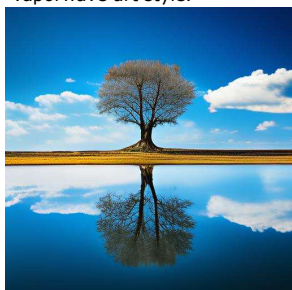
Portrait of a cool and stylish vaporwave anthropomorphic anthro male fox furry fursona, wearing a hoodie, vibrant colors, vaporwave art style.



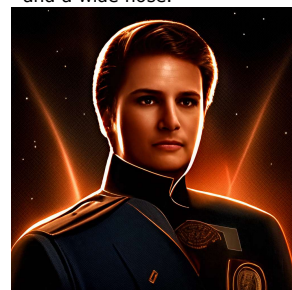
The image is a painting by William-Adolphe Bouguereau of Alfric Overguard, a calm and strong black man with alert eyes and a wide nose.



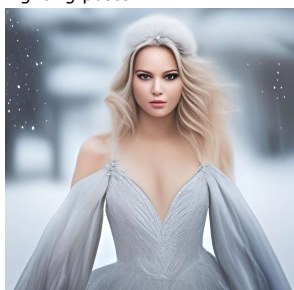
Ezio audiore in Vietnam, realistic shaded, fine details, fine - face, realistic shaded lighting poster.



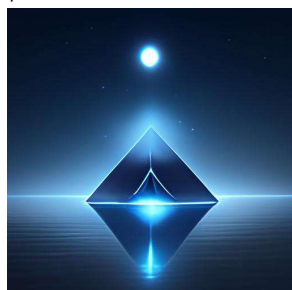
The solitary great tree centered in the image. cloudless sunny sky. little islands in the flooded plain.



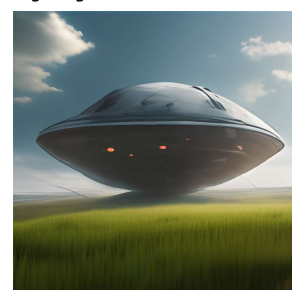
A digital portrait of a young, handsome Captain Kirk with intricate details and dramatic lighting.



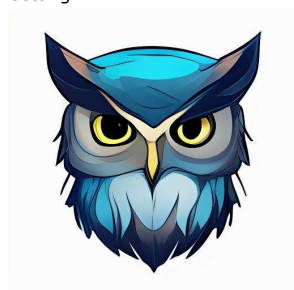
A full-body shot of a beautiful female in an intricate dress, with a sharp focus on her perfect eyes, captured by artist Artgerm in a snowy winter setting.



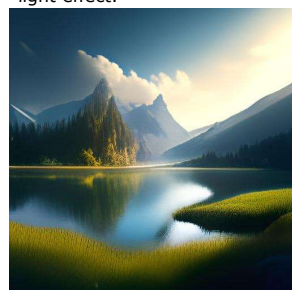
Beyond the horizon, just one gaint very lonely floating magnificent tetrahedron spacecraft, on which mystical symbols glowing, moon night, light effect.



A hyper-realistic landscape from a Neil Blomkamp film featuring a crashed spaceship, detailed grass, and a photorealistic sky.



Cute owl logo manga style.



A realistic beautiful natural landscape.



A digital art depicting a two ai android facing each other.

Figure 17: Samples Produced by Fantasy.



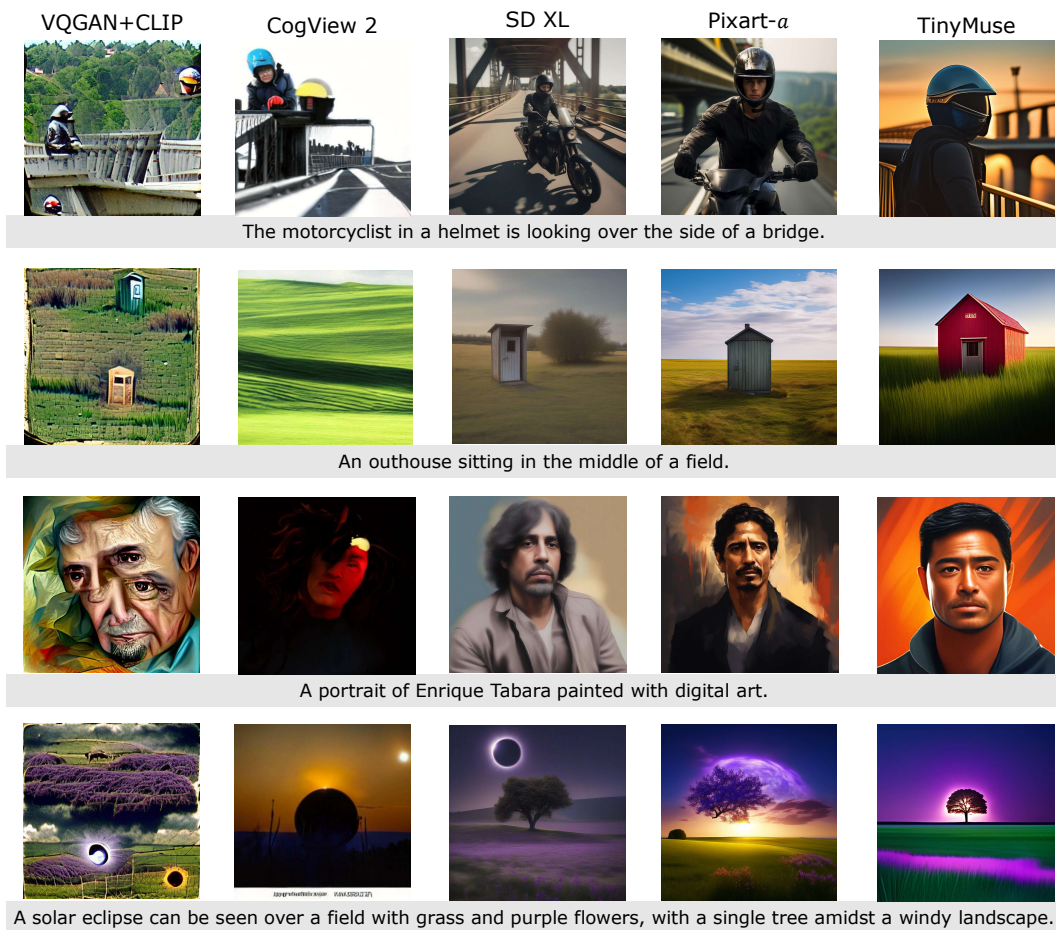


Figure 18: Samples Produced by Fantasy.

## 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](https://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.

## References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [2] Shidong Cao, Wenhao Chai, Shengyu Hao, and Gaoang Wang. Image reference-guided fashion design with structure-aware transfer by diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3524–3528, 2023.
- [3] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Diffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *arXiv preprint arXiv:2302.06826*, 2023.
- [4] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Jie Deng, Wenhao Chai, Jianshu Guo, Qixuan Huang, Wenhao Hu, Jenq-Neng Hwang, and Gaoang Wang. Citygen: Infinite and controllable 3d city layout generation. *arXiv preprint arXiv:2312.01508*, 2023.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [11] 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.
- [12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [13] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6244–6253, 2023.
- [14] 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.
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [16] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.
- [17] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700*, 2023.
- [18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [19] Guoqiang Jin, Fan Yang, Mingshan Sun, Ruyi Zhao, Yakun Liu, Wei Li, Tianpeng Bao, Liwei Wu, Xingyu Zeng, and Rui Zhao. Seqco-detr: Sequence consistency training for self-supervised object detection with transformers. *arXiv preprint arXiv:2303.08481*, 2023.
- [20] Jaewoong Lee, Sangwon Jang, Jaehyeon Jo, Jaehong Yoon, Yunji Kim, Jin-Hwa Kim, Jung-Woo Ha, and Sung Ju Hwang. Text-conditioned sampling framework for text-to-image generation with masked generative models. *arXiv preprint arXiv:2304.01515*, 2023.
- [21] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

- [22] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. Meshmae: Masked autoencoders for 3d mesh data analysis. In *European Conference on Computer Vision*, pages 37–54. Springer, 2022.
- [23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [24] Microsoft. Phi-2. <https://huggingface.co/microsoft/phi-2>, 2023.
- [25] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang, editors. *Sentence-T5: Scaling up Sentence Encoder from Pre-trained Text-to-Text Transfer Transformer*, 2022.
- [26] Minheng Ni, Xiaoming Li, and Wangmeng Zuo. Nuwa-lip: Language-guided image inpainting with defect-free vqgan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14192, 2023.
- [27] OpenAI. Dall-e 2. <https://openai.com/dall-e-2>, 2022.
- [28] Yichen Ouyang, Wenhao Chai, Jiayi Ye, Dapeng Tao, Yibing Zhan, and Gaoang Wang. Chasing consistency in text-to-3d generation from a single image. *arXiv preprint arXiv:2309.03599*, 2023.
- [29] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction. *arXiv preprint arXiv:2401.01808*, 2024.
- [30] 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.
- [31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sd-xl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [32] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023.
- [33] Hao Quan, Xingyu Li, Weixing Chen, Mingchen Zou, Ruijie Yang, Tingting Zheng, Ruiqun Qi, Xinghua Gao, and Xiaoyu Cui. Global contrast masked autoencoders are powerful pathological representation learners. *arXiv preprint arXiv:2205.09048*, 2022.
- [34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.
- [40] Yunjie Tian, Lingxi Xie, Zhaozhi Wang, Longhui Wei, Xiaopeng Zhang, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Integrally pre-trained transformer pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18610–18620, 2023.
- [41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [42] 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.
- [43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [44] 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.



- 602 [45] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a  
603 text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022.
- 604 [46] Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di  
605 Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model.  
606 *arXiv preprint arXiv:2311.14284*, 2023.
- 607 [47] Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. Mrm: Masked relation modeling for medical image  
608 pre-training with genetics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,  
609 pages 21452–21462, 2023.
- 610 [48] Haoxuan You, Mandy Guo, Zhecan Wang, Kai-Wei Chang, Jason Baldridge, and Jiahui Yu. Cobit: A  
611 contrastive bi-directional image-text generation model. *arXiv preprint arXiv:2303.13455*, 2023.
- 612 [49] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and  
613 Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the  
614 wild. *arXiv preprint arXiv:2401.13627*, 2024.
- 615 [50] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann,  
616 Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In  
617 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–  
618 10469, 2023.
- 619 [51] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A Ross,  
620 Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. Spae: Semantic pyramid autoencoder for multimodal  
621 generation with frozen llms. *arXiv preprint arXiv:2306.17842*, 2023.
- 622 [52] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d  
623 point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on*  
624 *Computer Vision and Pattern Recognition*, pages 19313–19322, 2022.
- 625 [53] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining  
626 Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the*  
627 *IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022.
- 628 [54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion  
629 models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847,  
630 2023.
- 631 [55] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-  
632 training with masked autoencoders for medical image classification and segmentation. *arXiv preprint*  
633 *arXiv:2203.05573*, 2022.