Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack

Xiaoliang Dai*, Ji Hou*, Chih-Yao Ma*, Sam Tsai*, Jialiang Wang*, Rui Wang*, Peizhao Zhang*,
Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic,
Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song,
Roshan Sumbaly†, Vignesh Ramanathan†, Zijian He†, Peter Vajda†, Devi Parikh†

GenAI, Meta

{xiaoliangdai, jihou, cyma, sstsai, jialiangw, ruiw, stzpz}@meta.com

Figure 1. With quality-tuning, Emu generates *highly* aesthetic images. Prompts: (top) a glass of orange juice; a woman in an apron works at a local bar; a coffee mug; (bottom) an egg and a bird made of wheat bread; a corgi; a shake is next to a cake.

Abstract

Training text-to-image models with web scale image-text pairs enables the generation of a wide range of visual concepts from text. However, these pre-trained models often face challenges when it comes to generating highly aesthetic images. This creates the need for aesthetic alignment post pre-training. In this paper, we propose quality-tuning to effectively guide a pre-trained model to exclusively generate highly visually appealing images, while maintaining generality across visual concepts. Our key insight is that supervised fine-tuning with a set of surprisingly small but extremely visually appealing images can significantly improve the generation quality. We pre-train a latent diffusion model on 1.1 *billion image-text pairs and fine-tune it with only a few thousand carefully selected high-quality images. The resulting model, Emu, achieves a win rate of* 82.9% *compared with its pre-trained only counterpart. Compared to the state-of-the-art SDXLv1.0, Emu is preferred* 68.4% *and* 71.3% *of the time on visual appeal on the standard PartiPrompts and our Open User Input benchmark based on the real-world usage of text-to-image models. In addition, we show that quality-tuning is a generic approach that is also effective for other architectures, including pixel diffusion and masked generative transformer models.*

* Core contributors: equal contribution, alphabetical order.
† Equal last authors.

1. Introduction

Recent advances in generative models have enabled them to generate various high-quality content, such as text [2,33], image [21, 25], music [16], video [32], and even 3D scenes [19, 22, 34], which has fundamentally revolutionized generative artificial intelligence (AI). In this paper, we present a recipe we have found to be effective for training *highly* aesthetic text-to-image models. It involves two stages: a knowledge learning stage, where the goal is to acquire the ability to generate virtually anything from text, which typically involves pre-training on hundreds of millions of image-text pairs; and a quality learning stage, which is necessary to restrict the output to a high-quality and aesthetically pleasing domain. We refer to the process of fine-tuning for the purpose of improving quality and promoting aesthetic alignment as *quality-tuning* for short.

Our key insight is that to effectively perform quality-tuning, a surprisingly small amount – a couple of thousand – exceptionally high-quality images and associated text is enough to make a *significant* impact on the aesthetics of the generated images *without* compromising the generality of the model in terms of visual concepts that can be generated. Although having more data while maintaining the same level of quality may be helpful, any attempts to prioritize quantity over quality may result in a compromise of the quality of generated images.

This is an interesting finding in the broader landscape of fine-tuning generative models. Quality-tuning for visual generative models can be thought of as analogous to instruction-tuning for large language models (LLMs) in terms of improving generation quality. First, before instruction-tuning, language models are more prone to generating low-quality text, which may be inconsistent in tone, overly verbose or concise, or simply unhelpful or even toxic [6, 7, 13, 20, 35]; ChatGPT-level [2] performance is achieved with effective instruction-tuning [20]. Similarly, we find that quality-tuning significantly improves the generation quality of text-to-image models. Second, the recipe for effectively performing instruction-tuning and quality-tuning is similar: use high-quality data, even if the quantity has to be small to maintain quality. Llama2 [33] has been fine-tuned on 27K high-quality prompts, which can be considered a very small quantity compared to the billions or trillions of pre-training tokens. Similarly, we find that strong text-to-image performance can be achieved by fine-tuning with even less data – a few thousand carefully selected images. Lastly, the knowledge obtained from pre-training is mostly retained after both instruction-tuning and quality-tuning. Like instruction-tuned LLMs, quality-tuned text-to-image models retain their generality in terms of the visual concepts that can be generated. These post pre-training stages align the knowledge to downstream user value – improving text quality and following instructions in the case of LLMs, and promoting aesthetic alignment in the case of text-to-image models.

Concretely, we pre-train a latent diffusion model (LDM) on 1.1 billion image-text pairs and quality-tune the model on a few thousand hand-picked exceptionally high-quality images selected from a large corpus of images. By its nature, the selection criterion is subjective and culturally dependent. We follow some common principles in photography, including but not limited to composition, lighting, color, effective resolution, focus, and storytelling to guide the selection process. With a few optimizations to the latent diffusion architecture, we start with a strong pre-trained model and dramatically improve the visual appeal of our generated images through quality-tuning. In fact, it significantly outperform a state-of-the-art publicly available model SDXLv1.0 [21] on visual appeal. We call our quality-tuned LDM Emu. We show example generations from Emu in Figure 1 and Figure 2.

Furthermore, we show that quality-tuning is a generic approach that is also effective for pixel diffusion and masked generative transformer models.

Our main contributions are:

- We build Emu, a quality-tuned latent diffusion model that significantly outperforms a publicly available state-of-the-art model SDXLv1.0 on visual appeal.
- To the best of our knowledge, this is the first work to emphasize the importance of a good fine-tuning recipe for aesthetic alignment of text-to-image models. We provide insights and recommendations for a key ingredient of this recipe – supervised fine-tuning with a surprisingly small amount of exceptionally high-quality data can have a significant impact on the quality of the generated images. Image quality should always be prioritized over quantity.
- We show that quality-tuning is a generic approach that also works well for other popular model architectures besides LDM, including pixel diffusion and masked generative transformer models.

2. Related Work

Text-to-Image Models. Generating an image from a textual description has been explored using various approaches. Diffusion-based methods learn a denoising process to gradually generate images from pure Gaussian noise [14]. The denoising process can occur either in pixel space or latent space, resulting in pixel diffusion models [5, 25, 30] or latent diffusion models [21, 27], which feature higher efficiency by reducing the size of spatial features. Generative transformer methods usually train autoregressive [4,10,11,26,37,38] or non-autoregressive (masked) [9] transformers in discrete token space to model the generation

Figure 2. **Selected Examples.** Selected images generated by our quality-tuned model, Emu.

process. Generative adversarial network [17, 31] also show remarkable capability of generating realistic images. While these models exhibit unprecedented image generation ability, they do not *always* generate *highly* aesthetic images.

Fine-Tuning Text-to-Image Models. Given a pre-trained text-to-image model, different methods have been developed to enable specific tasks. A number of techniques have been developed to personalize or adapt text-to-image models to a new subject or style [12, 15, 28]. ControlNet [39] provides additional control to the generation process by additionally conditioning on pose, sketch, edges, depth, etc. InstructPix2Pix [8] makes text-to-image models follow editing instructions by fine-tuning them on a set of generated image editing examples. To the best of our knowledge, this is the first work highlighting fine-tuning for generically promoting aesthetic alignment for a wide range of visual domains.

Fine-Tuning Language Models. Fine-tuning has become a critical step in building high-quality LLMs [1, 20, 33]. It generically improves output quality while enabling instruction-following capabilities. Effective fine-tuning of LLMs can be achieved with a relatively small but high-quality fine-tuning dataset, *e.g.,* using 27K prompts in [33]. In this work, we show that effective fine-tuning of text-to-image models can be also achieved with a *small* but *high-quality* fine-tuning dataset. This finding shows an interesting connection between fine-tuning vision and language models in generative AI.

3. Approach

As discussed earlier, our approach involves a knowledge learning stage followed by a quality-tuning stage. This may seem like a well-known recipe when mapped to pre-training and fine-tuning. That said, the key insights here are: (i) the fine-tuning dataset can be surprisingly small, on the order of a couple of thousand images, (ii) the quality of the dataset needs to be very high, making it difficult to fully automate data curation, requiring manual annotation, and (iii) even with a small fine-tuning dataset, quality-tuning not only significantly improves the aesthetics of the generated images, but does so without sacrificing generality as measured by faithfulness to the input prompt. Note that the stronger the base pre-trained model, the higher the quality of the generated images after quality-tuning. To this end, we made several modifications to the latent diffusion architecture [27] to facilitate high-quality generation. That said, quality-tuning is general enough and can be applied to a variety of architectures.

In this section, we first introduce the latent diffusion architecture we use. Then, we discuss the pre-training stage, followed by the protocol for collecting the high-quality fine-tuning dataset, and finally the quality-tuning stage. Later in Section 4, we demonstrate that quality-tuning is not limited to latent diffusion models but also improve other models such as pixel diffusion [30] and masked generative transformer [9] models.

Figure 3. **Autoencoder.** The visual quality of the reconstructed images for autoencoders with different channel sizes. While keeping all other architecture layers the same, we only change the latent channel size. We show that the original 4-channel autoencoder design [27] is unable to reconstruct fine details. Increasing channel size leads to much better reconstructions. We choose to use a 16-channel autoencoder in our latent diffusion model.

3.1. Latent Diffusion Architecture

We design a latent diffusion model that outputs 1024×1024 resolution images. Following standard latent diffusion architecture design, our model has an autoencoder (AE) to encode an image to latent embeddings and a U-Net to learn the denoising process.

We find that the commonly used 4-channel autoencoder (AE-4) architecture often results in a loss of details in the reconstructed images due to its high compression rate. The issue is especially noticeable in small objects. Intuitively, it compresses the image resolution by $64\times$ with three 2×2 downsampling blocks but increases the channel size only

from 3 (RGB) to 4 (latent channels). We find that increasing the channel size to 16 significantly improves reconstruction quality (see Table 1). To further improve the reconstruction performance, we use an adversarial loss and apply a non-learnable pre-processing step to RGB images using a *Fourier Feature Transform* to lift the input channel dimension from 3 (RGB) to a higher dimension to better capture fine structures. See Figure 3 for qualitative results of autoencoders of different channel size.

model	channel	SSIM	PSNR	FID
AE	4	0.80	28.64	0.35
	8	0.86	30.95	0.19
	16	0.92	34.00	0.06
Fourier-AE	16	0.93	34.19	0.04

Table 1. While keeping all other architecture design choices fixed, we first only change the latent channel size and report their reconstruction metrics on ImageNet [29]. We see that AE-16 significantly improves over AE-4 on all reconstruction metrics. Adding a Fourier Feature Transform and an adversarial loss further improves the reconstruction performance.

We use a large U-Net with 2.8B trainable parameters. We increase the channel size and number of stacked residual blocks in each stage for larger model capacity. We use text embeddings from both CLIP ViT-L [23] and T5-XXL [24] as the text conditions.

3.2. Pre-training

We curate a large internal pre-training dataset consisting of 1.1 billion images to train our model. The model is trained with progressively increasing resolutions, similar to [21]. This progressive approach allows the model to efficiently learn high-level semantics at lower resolutions and improve finer details at the highest resolutions We also use a noise-offset [3] of 0.02 in the final stage of pre-training. This facilitates high-contrast generation, which contributes to the aesthetics of the generated images.

3.3. High-Quality Alignment Data

As discussed before, in order to align the model towards highly aesthetic generations – quality matters significantly more than quantity in the fine-tuning dataset (see Section 4.3 for an ablation study on quality vs quantity). As also discussed, the notion of aesthetics is highly subjective. Here we discuss in detail what aesthetics we chose and how we curated our fine-tuning dataset by combining both automated filtering and manual filtering. The general quality-tuning strategy will likely apply to other aesthetics as well.

Automatic Filtering. Starting from an initial pool of billions of images, we first use a series of automatic filters to reduce the pool to a few hundreds of millions. These filters include but are not limited to offensive content removal, aesthetic score filter, optical character recognition (OCR) word count filter to eliminate images with too much overlaying text on them, and CLIP score filter to eliminate samples with poor image-text alignment, which are standard pre-filtering steps for sourcing large datasets. We then perform additional automated filtering via image size and aspect ratio. Lastly, to balance images from various domains and categories, we leverage visual concept classification [36] to source images from specific domains (*e.g.,* portrait, food, animal, landscape, car, etc). Finally, with additional quality filtering based on proprietary signals (*e.g.*, number of likes), we can further reduce the data to 200K.

Figure 4. **Visual Appealing Data**. Examples of visually appealing data that can meet our human filtering criterion.

Human Filtering. Next, we perform a two-stage human filtering process to only retain highly aesthetic images. In the first stage, we train *generalist* annotators to downselect the image pool to 20K images. Our primary goal during this stage is to optimize recall, ensuring the exclusion of medium and low quality that may have passed through the automatic filtering. In the second stage, we engage *specialist* annotators who have a good understanding of a set of photography principles. Their task is to filter and select images of the highest aesthetic quality (see Figure 4 for examples). During this stage, we focus on optimizing precision, meaning we aim to select only the very best images. A brief annotation guideline for photorealistic images is as follows. Our hypothesis is that following basic principles of high quality photography leads to generically more aesthetic images across a variety of styles, which is validated

via human evaluation.

1. **Composition.** The image should adhere to certain principles of professional photography composition, including the "Rule Of Thirds", "Depth and Layering", and more. Negative examples may include imbalance in visual weight, such as when all focal subjects are concentrated on one side of the frame, subjects captured from less flattering angles, or instances where the primary subject is obscured, or surrounding unimportant objects are distracting from the subject.

2. **Lighting.** We are looking for dynamic lighting with balanced exposure that enhances the image, for example, lighting that originates from an angle, casting highlights on select areas of the background and subject(s). We try to avoid artificial or lackluster lighting, as well as excessively dim or overexposed light.

3. **Color and Contrast.** We prefer images with vibrant colors and strong color contrast. We avoid monochromatic images or those where a single color dominates the entire frame.

4. **Subject and Background.** The image should have a sense of depth between the foreground and background elements. The background should be uncluttered but not overly simplistic or dull. The focused subjects must be intentionally placed within the frame, ensuring that all critical details are clearly visible without compromise. For instance, in a portrait, the primary subject of image should not extend beyond the frame or be obstructed. Furthermore, the level of detail on the foreground subject is extremely important.

5. **Additional Subjective Assessments.** Furthermore, we request annotators to provide their subjective assessments to ensure that only images of *exceptionally* aesthetic quality are retained by answering a couple of questions, such as: (i) Does this image convey a compelling story? (ii) Could it have been captured significantly better? (iii) Is this among the best photos you've ever seen for this particular content?

Through this filtering process, we retained a total of 2000 exceptionally high-quality images. Subsequently, we composed ground-truth captions for each of them. Note that some of these handpicked images are below our target resolution of 1024×1024. We trained a pixel diffusion upsampler inspired by the architecture proposed in [30] to upsample these images when necessary.

3.4. Quality-Tuning

We can think of the visually stunning images (like the 2000 images we collected) as a subset of all images that share some common statistics. Our hypothesis is that a strongly pre-trained model is already capable of generating highly aesthetic images, but the generation process is not properly guided towards always producing images with these statistics. Quality-tuning effectively restricts outputs to a high-quality subset.

We fine-tune the pre-trained model with a small batch size of 64. We use a noise-offset of 0.1 at this stage. Note that early stopping is important here as fine-tuning on a small dataset for too long will result in significant overfitting and degradation in generality of visual concepts. We fine-tune for no more than 15K iterations despite the loss still decreasing. This total iteration number is determined empirically.

4. Experiments

We compare our quality-tuned model to our pre-trained model to demonstrate the effectiveness of quality-tuning. To place the visual appeal of our generated images in context with a current state-of-the-art model, we compare our model to SDXLv1.0 [21]. Due to lack of access to training data of SDXL and their underlying model, we leveraged their corresponding APIs for our comparison. Note that unlike SDXL, we use a single stage architecture, and do not use a subsequent refinement stage. As stated earlier, we also show that quality-tuning is not specific to LDMs, and can be applied to other architectures – pixel diffusion and masked generative transformer models.

4.1. Evaluation Setting

Prompts. We evaluate on two large sets of prompts: 1600 PartiPrompts [37] which is commonly used for text-to-image generation benchmarking, and our 2100 Open User Input (OUI) Prompts. The OUI prompt set is based on real-world user prompts. It captures prompts that are popular with text-to-image models, reflecting visual concepts that are relevant to real-world use cases (Figure 5), paraphrased by LLMs to be closer to how regular users might input prompts (as opposed to being highly prompt engineered). The overall motivation was to capture the creativity and intricacies of popular prompts for text-to-image models so we are pushing the capabilities of models, while also being grounded in likely real world use cases.

Metrics. We use two separate evaluation metrics: visual appeal and text faithfulness. Visual appeal refers to the overall aesthetic quality of a generated image. It combines various visual elements such as color, shape, texture, and composition that creates a pleasing and engaging look. The concept of visual appeal is subjective and varies from person to person, as what may be aesthetically pleasing to one person may not be to another. Therefore, we ask five annotators to rate each sample. Concretely, we show annotators

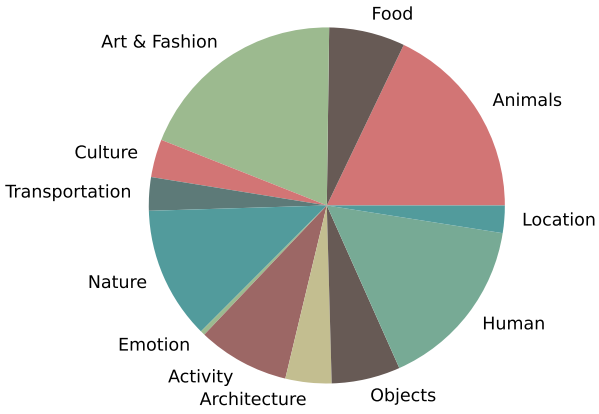

Figure 5. **Prompt distributions**. The distribution of different concepts in our Open User Input prompts. We cover a comprehensive list of common concepts people typically use to generate images.

two images A and B, side-by-side, each generated by a different model using the same caption. The text captions are not displayed. Annotators choose which image is more visually appealing by selecting "A", "B" or "Tie".

Text faithfulness refers to the degree of similarity between a generated image and a text caption. In this task, we again display two generated images A and B, side-by-side, but with the caption alongside the images. The annotators are asked to ignore the visual appeal of the images and choose which ones best describe the caption with choices "A ", "B", "Both", and "Neither", where "Both" and "Neither" are considered as "Tie". In this task, we have three annotators to annotate each sample pair.

We do not report "standard" metrics such as FID scores. As argued in many recent papers (*e.g.,* [18, 21]), FID scores do not correlate well with human assessment of the performance of generative models.

4.2. Results

Effectiveness of Quality-Tuning. First, we compare our quality-tuned model, Emu, with the pre-trained model. See Figure 7 for random (not cherry-picked) qualitative examples before and after quality-tuning. Note the highly aesthetic non-photorealistic image as well, validating our hypothesis that following certain photography principles in curating the quality-tuning dataset leads to improved aesthetics for a broad set of styles. We show more examples of generated images using Emu in Figure 8 and Figure 9.

Quantitatively, as shown in Figure 6 (top), after quality-tuning, Emu is preferred in both visual appeal and text faithfulness by a significant margin. Specifically, Emu is preferred 82.9% and 91.2% of the time for visual appeal, and 36.7% and 47.9% of the time for text faithfulness on PartiPrompts and OUI Prompts, respectively. In contrast, the

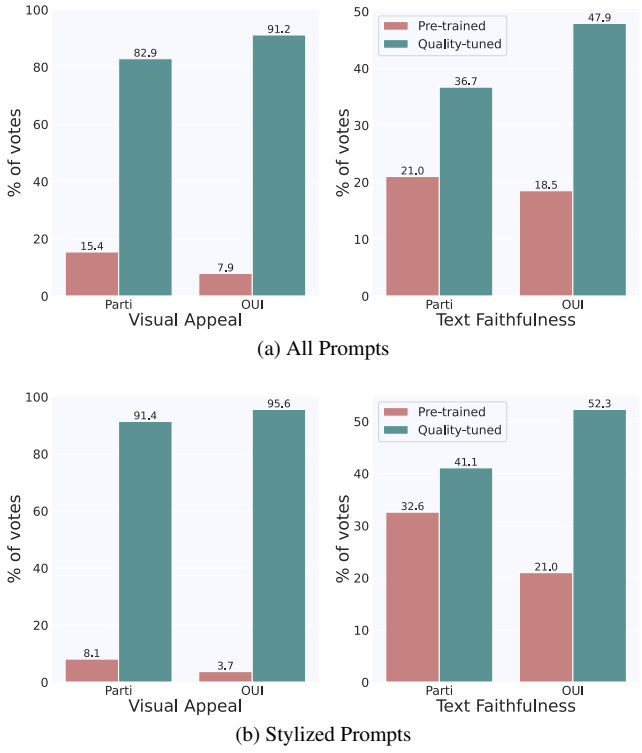

Figure 6. **Quality-Tuning vs Pre-training**. Human evaluation on both the PartiPrompts and Open User Input prompts shows that our quality-tuned model, Emu, significantly outperforms the pre-trained model on visual appeal, without loss of generality of visual concepts or styles that can be generated.

pre-trained model is preferred only 15.4% and 7.9% of the time for visual appeal, and 21.0% and 18.5% of the time for text faithfulness on PartiPrompts and OUI Prompts. The remaining cases result in ties. From the two large sets of evaluation data that covers various domains and categories, we did not observe degradation of generality across visual concepts. In fact, as seen, text-faithfulness also improved. We hypothesize this is because the captions of the 2000 quality-tuning images were manually written, while the captions in the pre-training dataset tend to be noisy. Finally, we analyze the results on non-photorealistic stylized prompts (*e.g.,* sketches, cartoons, etc.). We find that the improvements broadly apply to a variety of styles, see Figure 6 (bottom).

Visual Appeal in the Context of SoTA. To place the visual appeal of our generated images in the context of current state-of-the-art, we compare Emu with SDXLv1.0 [21]. As shown in Table 2, our model is preferred over SDXLv1.0 in visual appeal by a significant margin – including on stylized (non-photorealistic) prompts.

Figure 7. **Qualitative Comparison**. a comparison of images generated by the pre-trained and quality-tuned model.

Quality-Tuning Other Architectures. Next, we show our quality-tuning can also improve other popular architectures, such as pixel diffusion and masked generative transformer models. Specifically, we re-implement and re-train from scratch a pixel diffusion model, similar to Imagen [30] architecture, and a masked generative transformer model, similar to Muse [9] architecture, and then quality-tune them on 2000 images. We evaluate both quality-tuned models on $1/3$ randomly sampled PartiPrompts. As shown in Figure 10, both architectures show significant improvement after quality-tuning on both visual appeal and text faithfulness metrics.

Eval data	win (%)	tie (%)	lose (%)
Parti (All)	68.4	2.1	29.5
OUI (All)	71.3	1.2	27.5
Parti (Stylized)	81.7	1.9	16.3
OUI (Stylized)	75.5	1.4	23.1

Table 2. **Emu vs SDXL on Visual Appeal**. Our model is preferred over SDXL by a large margin, including on stylized, non-photorealistic prompts.

4.3. Ablation Studies

We do ablation studies on the fine-tuning dataset with a focus on visual appeal. We first investigate the impact of the dataset size. We report results of quality-tuning on randomly sampled subsets of sizes 100, 1000 and 2000 in Table 3. With even just 100 fine-tuning images, the model can already be guided towards generating visually-pleasing images, jumping from a win rate of 24.8% to 60% compared with SDXL.

fine-tune data	win (%)	tie (%)	lose (%)
w/o quality-tuning	24.8	1.4	73.9
100	60.3	1.5	38.2
1000	63.2	1.9	35.0
2000	67.0	2.6	30.4

Table 3. **Visual Appeal by Fine-Tuning Dataset Size.** All the numbers are against SDXL as baseline. With merely 100 high-quality images as fine-tuning data, our quality-tuned model can already outperform SDXL in visual appeal. Our model's visual appeal further improves as more images are used.

5. Limitation

Limitation of Human Evaluation. Relative to most published works, the prompt sets we evaluate on are reasonably large (1600 Parti prompts and our 2100 Open User Input prompts), under a multi-review annotation setting. Even then, the evaluation results may not fully reflect real-world usage of the models. In addition, human evaluation of text-to-image models, especially when it comes to aesthetics, is inherently subjective and noisy. As a result, evaluation on a different set of prompts or with different annotators and guidelines may lead to different results.

Limitations of Small-Scale Fine-Tuning. The role of quality-tuning is to restrict the output distribution to a high-quality domain. However, issues rooted from pre-training

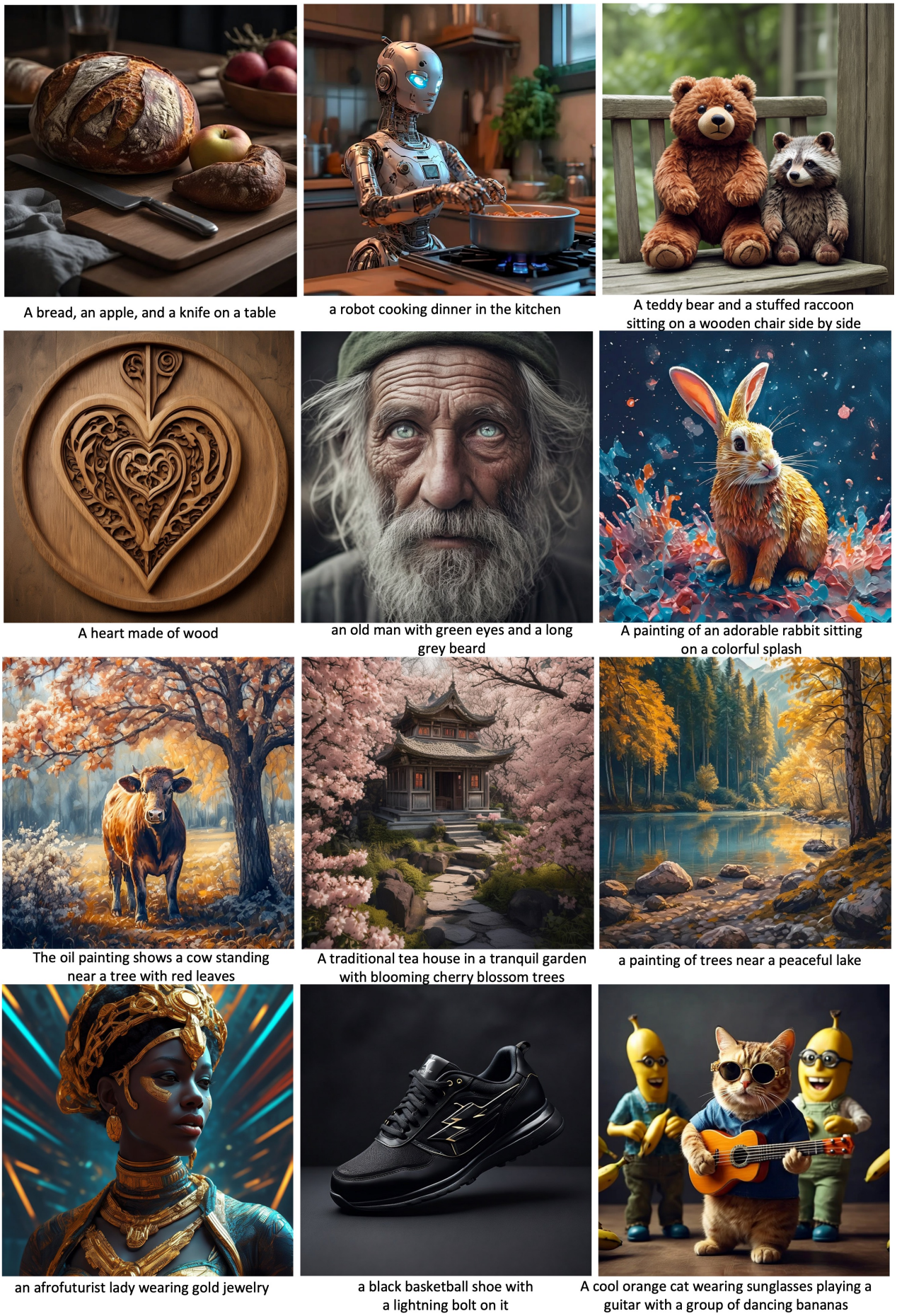

Figure 8. **Generated Examples.** Images generated by our quality-tuned model, Emu.

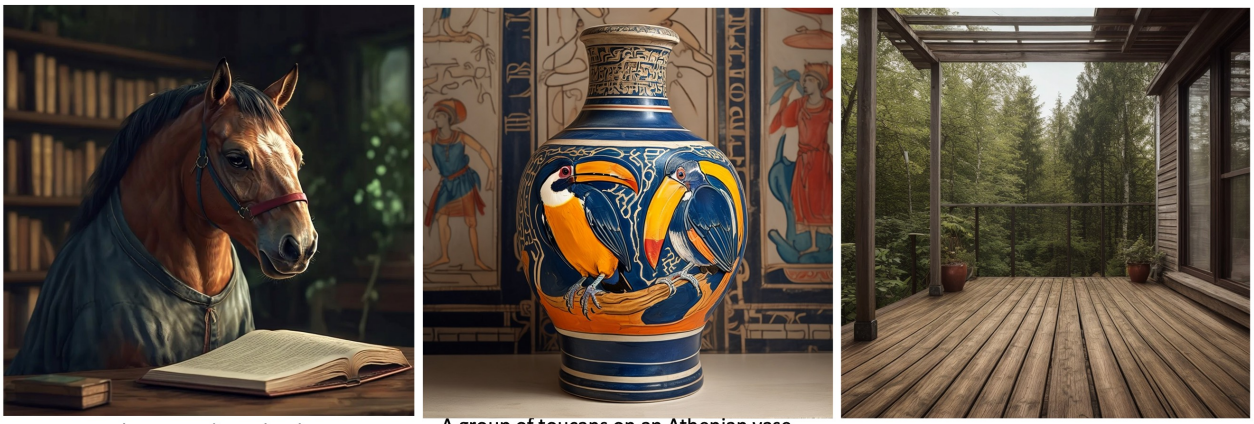

Figure 9. **More Examples.** Images generated by our quality-tuned model, Emu.

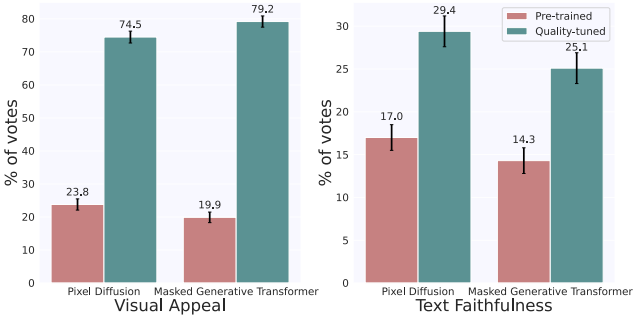

Figure 10. **Quality-Tuning vs Pre-training on Pixel Diffusion and Masked Generative Transformer**. We adapt our quality-tuning to other popular text-to-image model architectures. Our results indicate that the success of quality-tuning can be transferred to other architectures, beyond latent diffusion models.

may still persist. For instance, the models may struggle to generate certain objects that were not sufficiently learned during pre-training.

Limitations of Text-to-Image Models in General. Like other text-to-image models, our models may generate biased, misleading, or offensive outputs. We've invested a significant amount in fairness and safety of our models - starting with balanced dataset construction, creating dedicated evaluation sets for these risk categories and investing multiple hours of redteaming.

6. Conclusion

In this paper, we demonstrated that manually selecting high quality images that are highly aesthetically-pleasing is one of the most important keys to improving the aesthetics of images generated by text-to-image generative models. We showed that with just a few hundred to thousand fine-tuning images, we were able to improve the visual appeal of generated images without compromising on the generality of visual concepts depicted. With this finding, we build Emu, a LDM for high-quality image synthesis. On the commonly used PartiPrompts and our Open User Input Prompts, we carefully evaluate our model against a publicly available state-of-the-art text-to-image model (SDXLv1.0 [21]) as well as our pre-trained LDM. We also show that quality-tuning not only improves LDMs, but also pixel diffusion and masked generative transformer models.

7. Acknowledgement

This work would not have been possible without a large group of collaborators who helped with the underlying infrastructure, data, privacy, and the evaluation framework. We extend our gratitude to the following people for their contributions (alphabetical order): Eric Alamillo, Andrés Alvarado, Giri Anantharaman, Stuart Anderson, Snesha Arumugam, Chris Bray, Matt Butler, Anthony Chen, Lawrence Chen, Jessica Cheng, Lauren Cohen, Jort Gemmeke, Freddy Gottesman, Nader Hamekasi, Zecheng He, Jiabo Hu, Praveen Krishnan, Carolyn Krol, Tianhe Li, Mo Metanat, Vivek Pai, Guan Pang, Albert Pumarola, Ankit Ramchandani, Stephen Roylance, Kalyan Saladi, Artsiom Sanakoyeu, Dev Satpathy, Alex Schneidman, Edgar Schoenfeld, Shubho Sengupta, Hardik Shah, Shivani Shah, Yaser Sheikh, Karthik Sivakumar, Lauren Spencer, Fei Sun, Ali Thabet, Mor Tzur, Mike Wang, Mack Ward, Bichen Wu, Seiji Yamamoto, Licheng Yu, Hector Yuen, Luxin Zhang, Yinan Zhao, and Jessica Zhong.

Finally, thank you Connor Hayes, Manohar Paluri and Ahmad Al-Dahle for your support and leadership.

References

[1] https://cdn.openai.com/papers/gpt-4.pdf.

[2] https://openai.com/blog/chatgpt/.

[3] https://www.crosslabs.org/blog/diffusion-with-offset-noise/.

[4] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. arXiv preprint arXiv:2201.07520, 2022.

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022.

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages 610–623, 2021.

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18392–18402, 2023.

[9] 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.

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems, 34:19822–19835, 2021.

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In

European Conference on Computer Vision, pages 89–106. Springer, 2022.

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

[13] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462, 2020.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[16] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. arXiv preprint arXiv:2302.03917, 2023.

[17] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10124–10134, 2023.

[18] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569, 2023.

[19] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 300–309, 2023.

[20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[21] 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.

[22] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988, 2022.

[23] 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.

[24] 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.

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

[26] 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.

[27] 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.

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115:211–252, 2015.

[30] 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.

[31] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515, 2023.

[32] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[34] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213, 2023.

[35] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359, 2021.

[36] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546, 2019.

[37] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2(3):5, 2022.

[38] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. arXiv preprint arXiv:2309.02591, 2023.

[39] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543, 2023.