Kandinsky: an Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion

Anton Razzhigaev[1,2], Arseniy Shakhmatov[3], Anastasia Maltseva[3], Vladimir Arkhipkin[3], Igor Pavlov[3], Ilya Ryabov[3], Angelina Kuts [3], Alexander Panchenko[2,1], Andrey Kuznetsov[3,1], and Denis Dimitrov[3,1]

[1]AIRI, [2]Skoltech, [3]Sber AI

{razzhigaev, kuznetsov, dimitrov}@airi.net

Abstract

Text-to-image generation is a significant domain in modern computer vision and has achieved substantial improvements through the evolution of generative architectures. Among these, there are diffusion-based models that have demonstrated essential quality enhancements. These models are generally split into two categories: pixel-level and latent-level approaches. We present Kandinsky[1], a novel exploration of latent diffusion architecture, combining the principles of the image prior models with latent diffusion techniques. The image prior model is trained separately to map text embeddings to image embeddings of CLIP. Another distinct feature of the proposed model is the modified MoVQ implementation, which serves as the image autoencoder component. Overall, the designed model contains 3.3B parameters. We also deployed a user-friendly demo system that supports diverse generative modes such as text-to-image generation, image fusion, text and image fusion, image variations generation, and text-guided inpainting/outpainting. Additionally, we released the source code and checkpoints for the Kandinsky models. Experimental evaluations demonstrate a FID score of 8.03 on the COCO-30K dataset, marking our model as the top open-source performer in terms of measurable image generation quality.

1 Introduction

In quite a short period of time, generative abilities of text-to-image models have improved substantially, providing users with photorealistic quality, near real-time inference speed, a great number of applications and features, including simple easy-to-use web-based platforms and sophisticated AI graphics editors.

This paper presents our unique investigation of latent diffusion architecture design, offering a fresh and innovative perspective on this dynamic field of study. First, we describe the new architecture of Kandinsky and its details. The demo system with implemented features of the model is also described. Second, we show the experiments, carried out in terms of image generation quality and come up with the highest FID score among existing open-source models. Additionally, we present the rigorous ablation study of prior setups that we conducted, enabling us to carefully analyze and evaluate various configurations to arrive at the most effective and refined model design.

Our **contributions** are as follows:

- We present the first text-to-image architecture designed using a combination of image prior and latent diffusion.
- We demonstrate experimental results comparable to the state-of-the-art (SotA) models such as Stable Diffusion, IF, and DALL-E 2, in terms of FID metric and achieve the SotA score among all existing open source models.
- We provide a software implementation of the proposed state-of-the-art method for text-to-image generation, and release pre-trained models, which is unique among the top-performing methods. Apache 2.0 license makes it possible to use the model for both non-commercial and commercial purposes.[2] [3]
- We create a web image editor application that can be used for interactive generation of images by text prompts (English and Russian languages are supported) on the basis of the proposed method, and provides inpainting/outpainting functionality.[4] The video demonstration is available on YouTube.[5]

[1]The system is named after Wassily Kandinsky, a famous painter and an art theorist.

[2]https://github.com/ai-forever/Kandinsky-2

[3]https://huggingface.co/kandinsky-community

[4]https://fusionbrain.ai/en/editor

[5]https://www.youtube.com/watch?v=c7zHPc59cWU

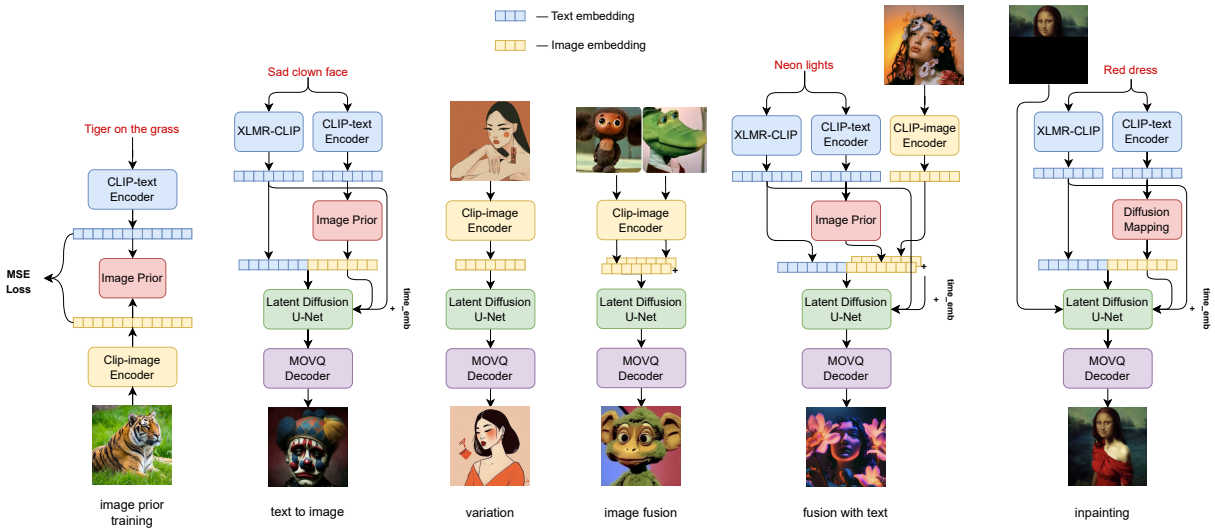

Figure 1: Image prior scheme and inference regimes of the Kandinsky model.

2 Related Work

Early text-to-image generative models, such as DALL-E (Ramesh et al., 2021) and CogView (Ding et al., 2021), or later Parti (Yu et al., 2022) employed autoregressive approaches but often suffered from significant content-level artifacts. This led to the development of a new breed of models that utilized the diffusion process to enhance image quality. Diffusion-based models, such as DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022b), and Stable Diffusion[6], have since become cornerstones in this domain. These models are typically divided into pixel-level (Ramesh et al., 2022; Saharia et al., 2022b) and latent-level (Rombach et al., 2022) approaches.

This surge of interest has led to the design of innovative approaches and architectures, paving the way for numerous applications based on open-source generative models, such as DreamBooth (Ruiz et al., 2023) and DreamPose (Karras et al., 2023). These applications exploit image generation techniques to offer remarkable features, further fueling the popularity and the rapid development of diffusion-based image generation approaches.

This enabled a wide array of applications like 3D object synthesis (Poole et al., 2023; Tang et al., 2023; Lin et al., 2022; Chen et al., 2023), video generation (Ho et al., 2022b; Luo et al., 2023; Ho et al., 2022a; Singer et al., 2023; Blattmann et al., 2023; Esser et al., 2023), controllable image editing (Hertz et al., 2023; Parmar et al., 2023; Liew et al., 2022; Mou et al., 2023; Lu et al., 2023), and more, which are now at the forefront of this domain.

Diffusion models achieve state-of-the-art results in image generation task both unconditional (Ho et al., 2020; Nichol and Dhariwal, 2021) and conditional (Peebles and Xie, 2022). They beat GANs (Goodfellow et al., 2014) by generating images with better scores of fidelity and diversity without adversarial training (Dhariwal and Nichol, 2021). Diffusion models also show the best performance in various image processing tasks like inpainting, outpainting, and super-resolution (Batzolis et al., 2021; Saharia et al., 2022a).

Text-to-image diffusion models have become a popular research direction due to the high performance of diffusion models and the ability to simply integrate text conditions with the classifier-free guidance algorithm (Ho and Salimans, 2022). Early models like GLIDE (Nichol et al., 2022), Imagen (Saharia et al., 2022b), DALL-E 2 (Ramesh et al., 2022) and eDiff-I (Balaji et al., 2022) generate low-resolution image in pixel space and then upsample it with another super-resolution diffusion models. They are also using different text encoders, large language model T5 (Raffel et al., 2020) in Imagen, CLIP (Radford et al., 2021) in GLIDE and DALL-E 2.

3 Demo System

We implemented a set of user-oriented solutions where Kandinsky model is embedded as a core imaging service. It has been done due to a variety of inference regimes, some of which need specific front-end features to perform properly. Overall, we implemented two main inference resources: Tele-

[6] https://github.com/CompVis/stable-diffusion

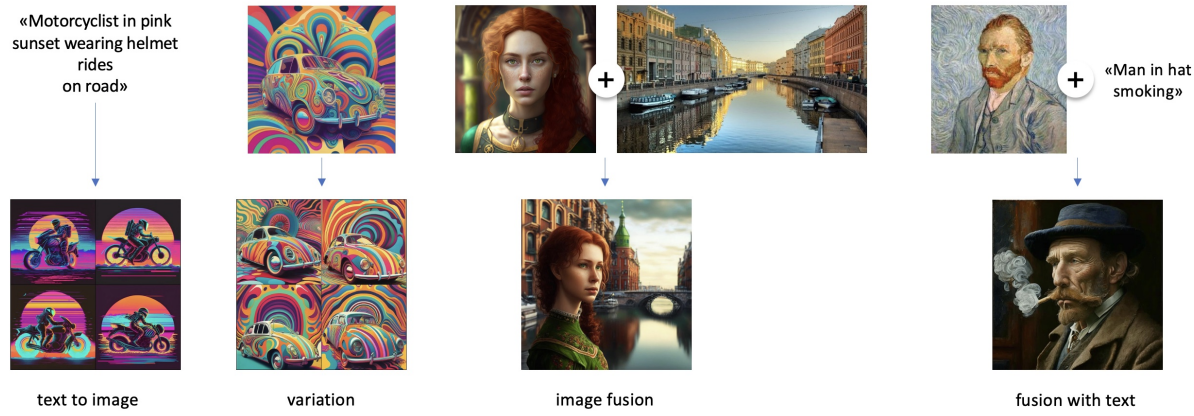

Figure 2: Examples of inference regimes using Kandinsky model.

gram bot and FusionBrain website.

FusionBrain represents a web-based image editor with such features as loading and saving images, sliding location window, erasing tools, zooming in/out, various styles selector, etc. (cf. Figure 3). In terms of image generation, the three following options are implemented on this side:

- text-to-image generation – user inputs a text prompt in Russian or English, then selects an aspect-ratio from the list (9:16, 2:3, 1:1, 16:9, 3:2), and the system generates an image;
- inpainting – using the specific erasing tool, user can remove any arbitrary input image part and fill it, guided by a text prompt or without any guidance;
- outpainting – input image can be extended with a sliding window that can be used as a mask for the following generation (if the window intersects any imaged area, then the empty window part is generated with or without text prompt guidance).

Inpainting and outpainting options are the main image editing features of the model. Architectural details about these generation types can also be found in Figure 1.

Telegram bot contains the following image generation features (cf. Figure 2):

- text-to-image generation;
- image and text fusion – user inputs an image and a text prompt to create a new image guided by this prompt;
- image fusion – user inputs an image as the main one and another 'guiding' image, and the system generates their fusion;
- image variations – user inputs an image, and the system generates several new images similar to the input one.

4 Kandinsky Architecture

In our work, we opted to deliver state-of-the-art text-to-image synthesis. In the initial stages of our research, we experimented with multilingual text encoders, such as mT5 (Xue et al., 2021), XLMR (Conneau et al., 2020), XLMR-CLIP[7], to facilitate robust multilingual text-to-image generation. However, we discovered that using the CLIP-image embeddings instead of standalone text encoders resulted in improved image quality. As a result, we adopted an image prior approach, utilizing diffusion and linear mappings between text and image embedding spaces of CLIP, while keeping additional conditioning with XLMR text embeddings. That is why Kandinsky uses two text encoders: CLIP-text with image prior mapping and XLMR. We have set these encoders to be frozen during the training phase.

The significant factor that influenced our design choice was the efficiency of training latent diffusion models, as compared to pixel-level diffusion models (Rombach et al., 2022). This led us to focus our efforts on the latent diffusion architecture. Our model essentially comprises three stages: text encoding, embedding mapping (image prior), and latent diffusion.

The construction of our model involves three primary steps: text encoding, embedding mapping (image prior), and latent diffusion. At the embedding mapping step, which we also refer to as the

[7] https://github.com/FreddeFrallan/Multilingual-CLIP

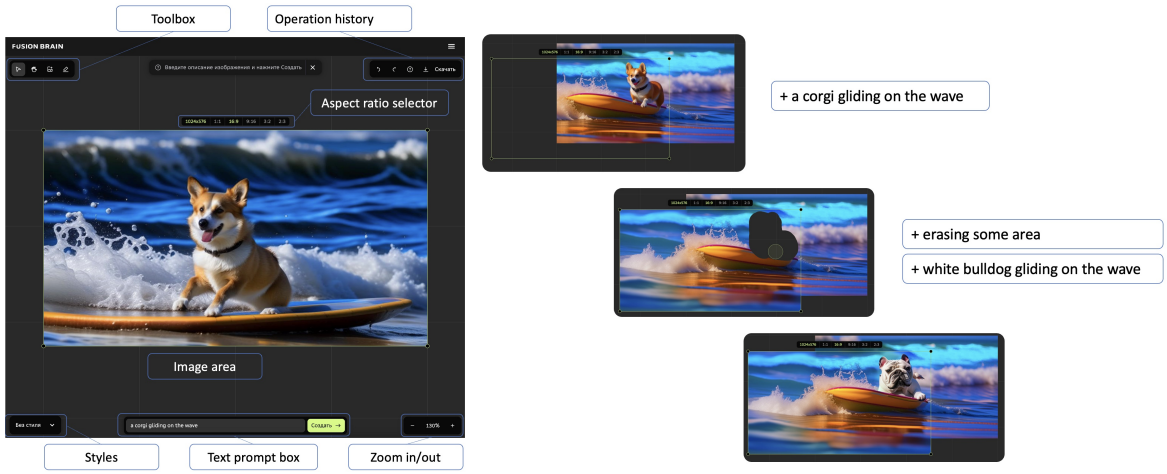

Figure 3: Kandinsky web interface for “a corgi gliding on the wave”: generation (left) and in/outpainting (right).

Table 1: Proposed architecture comparison by FID on COCO-30K validation set on 256×256 resolution. * For the IF model we reported reproduced results on COCO-30K, but authors provide FID of 7.19.

Model	FID-30K
Open Sourced Techologies	
Kandinsky (Ours)	**8.03**
Stable Diffusion 2.1 (2022) [8]	8.59
GLIDE [8] (Nichol et al., 2022)	12.24
IF* (2023) [12]	15.10
Kandinsky 1.0 (2022) [9]	15.40
ruDALL-E Malevich (2022) [9]	20.00
GLIGEN [10] (Li et al., 2023)	21.04
Proprietary Technologies	
eDiff-I (Balaji et al., 2022)	**6.95**
Imagen (Saharia et al., 2022b)	7.27
GigaGAN (Kang et al., 2023)	9.09
DALL-E 2 (Ramesh et al., 2022)	10.39
DALL-E (Ramesh et al., 2021)	17.89

image prior, we use the transformer-encoder model. This model was trained from scratch with a diffusion process on text and image embeddings provided by the CLIP-ViT-L14 model. A noteworthy feature in our training process is the use of element-wise normalization of visual embeddings. This normalization is based on full-dataset statistics and leads to faster convergence of the diffusion process. We implemented inverse normalization to revert to the original CLIP-image embedding space in the inference stage.

The image prior model is trained on text and image embeddings, provided by the CLIP models.

[8] https://github.com/Stability-AI/stablediffusion

[9] https://github.com/ai-forever/ru-dalle

[10] https://github.com/gligen/GLIGEN

We conducted a series of experiments and ablation studies on the specific architecture design of the image prior model (Table 3, Figure 6). The model with the best human evaluation score is based on a 1D-diffusion and standard transformer-encoder with the following parameters: num_layers=20, num_heads=32, and hidden_size=2048.

The latent diffusion part employs a UNet model along with a custom pre-trained autoencoder. Our diffusion model uses a combination of multiple condition signals: CLIP-image embeddings, CLIP-text embeddings, and XLMR-CLIP text embeddings. CLIP-image and XLMR-CLIP embeddings are merged and utilized as an input to the latent diffusion process. Also, we conditioned the diffusion process on these embeddings by adding all of them to the time-embedding. Notably, we did not skip the quantization step of the autoencoder during diffusion inference as it leads to an increase in the diversity and the quality of generated images (cf. Figure 4). In total, our model comprises 3.3 B parameters (Table 2).

Table 2: Kandinsky model parameters.

Architecture part	Params	Freeze
Diffusion Mapping	1B	False
CLIP image encoder (ViT-L14)	427M	True
CLIP text encoder	340M	True
Text encoder (XLM-R-L)	560M	True
Latent Diffusion UNet	1.22B	False
MoVQ image autoencoder	67M	True

We observed that the image decoding was our main bottleneck in terms of generated image quality; hence, we developed a Sber-MoVQGAN, our custom implementation of MoVQGAN (Zheng

Table 3: Ablation study: FID on COCO-30K validation set on 256×256 resolution.

Setup	FID-30K	CLIP
Diffusion prior with quantization	9.86	**0.287**
Diffusion prior w/o quantization	9.87	0.286
Linear prior	**8.03**	0.261
Residual prior	8.61	0.249
No prior	25.92	0.256

et al., 2022) with minor modifications. We trained this autoencoder on the LAION HighRes dataset (Schuhmann et al., 2022), obtaining the SotA results in image reconstruction. We released the weights and code for these models under an open source licence[11]. The comparison of our autoencoder with competitors can be found in Table 4.

5 Experiments

We sought to evaluate and refine the performance of our proposed latent diffusion architecture in our experimental analysis. To this end, we employed automatic metrics, specifically FID-CLIP curves on the COCO-30K dataset, to obtain the optimal guidance-scale value and compare Kandinsky with competitors (cf. Figure 4). Furthermore, we conducted investigations with various image prior setups, exploring the impact of different configurations on the performance. These setups included: no prior, utilizing text embeddings directly; linear prior, implementing one linear layer; ResNet prior, consisting of 18 residual MLP blocks; and transformer diffusion prior.

An essential aspect of our experiments was the exploration of the effect of latent quantization within the MoVQ autoencoder. We examined the outputs with latent quantization, both enabled and disabled, to better comprehend its influence on image generation quality.

To ensure a comprehensive evaluation, we also included an assessment of the IF model [12], which is the closest open-source competitor to our proposed model. For this purpose, we computed FID scores for the IF model [13] (Table 1).

However, we acknowledged the limitations of automatic metrics that become obvious when it comes to capturing user experience nuances. Hence, in addition to the FID-CLIP curves, we conducted a blind human evaluation to obtain insightful feedback and validate the quality of the generated images from the perspective of human perception based on the DrawBench dataset (Saharia et al., 2022b).

The combination of automatic metrics and human evaluation provides a comprehensive assessment of Kandinsky performance, enabling us to make informed decisions about the effectiveness and usability of our proposed image prior to design.

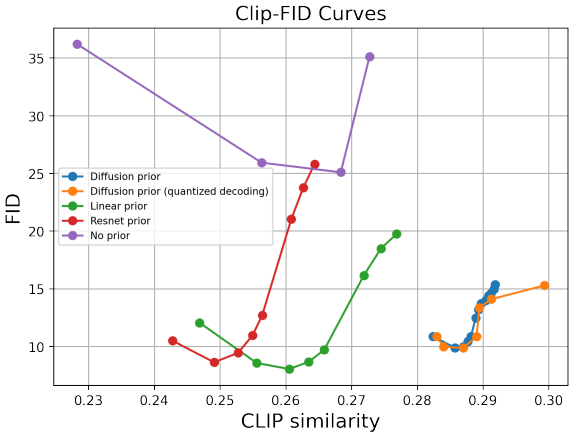

Figure 4: CLIP-FID curves for different setups.

Figure 5: Image generation results with prompt "astronaut riding a horse" for original image prior and linear prior trained on 500 pairs of images with cats.

6 Results

Our experiments and evaluations have showcased the capabilities of Kandinsky architecture in text-to-image synthesis. Kandinsky achieved the FID score of 8.03 on the COCO-30K validation set at a resolution of 256×256, which puts it in close competition with the state-of-the-art models, and among the top performers within open-source systems. Our methodical ablation studies further dissected the performance of different configurations: quantization of latent codes in MoVQ slightly improves

[11] https://github.com/ai-forever/MoVQGAN
[12] https://github.com/deep-floyd/IF
[13] https://github.com/mseitzer/pytorch-fid

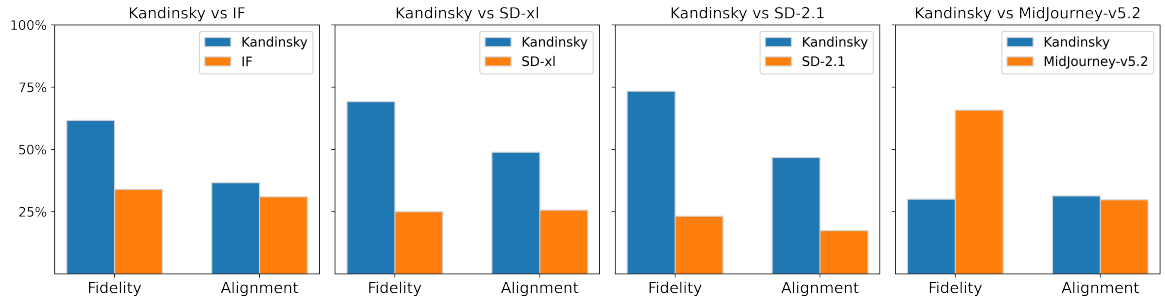

Figure 6: Human evaluation: competitors vs Kandinsky with diffusion prior on Drawbench. The total count of votes is 5000.

Table 4: Sber-MoVQGAN comparison with competitors on ImageNet dataset.

Model	Latent size	Num Z	Train steps	FID ↓	SSIM ↑	PSNR ↑	L1 ↓
ViT-VQGAN*	32x32	8192	500,000	1.28	-	-	-
RQ-VAE*	8x8x16	16384	10 epochs	1.83	-	-	-
Mo-VQGAN*	16x16x4	1024	40 epochs	1.12	0.673	22.42	-
VQ CompVis	32x32	16384	971,043	1.34	0.650	23.85	0.0533
KL CompVis	32x32	-	246,803	0.968	0.692	25.11	0.0474
Sber-VQGAN	32x32	8192	1 epoch	1.44	0.682	24.31	0.0503
Sber-MoVQGAN 67M	32x32	1024	5,000,000	1.34	0.704	25.68	0.0451
Sber-MoVQGAN 67M	32x32	16384	2,000,000	0.965	0.725	26.45	0.0415
Sber-MoVQGAN 102M	32x32	16384	2,360,000	0.776	0.737	26.89	0.0398
Sber-MoVQGAN 270M	32x32	16384	1,330,000	**0.686**	**0.741**	**27.04**	**0.0393**

the quality of images (FID 9.86 vs 9.87). The best CLIP score and human-eval score are obtained by diffusion prior.

The best FID score is achieved using Linear Prior. This configuration stands out with the best FID score of 8.03. It is an intriguing outcome: the simplest linear mapping showcased the best FID, suggesting that there might exist a linear relationship between visual and textual embedding vector spaces. To further scrutinize this hypothesis, we trained a linear mapping on a subset of 500 cat images and termed it the "cat prior". Astonishingly, this mapping displayed high proficiency (cf. Figure 5).

7 Conclusion

We presented Kandinsky, a system for various image generation and processing tasks based on a novel latent diffusion model. Our model yielded the SotA results among open-sourced systems. Additionally, we provided an extensive ablation study of an image prior to design choices. Our system is equipped with free-to-use interfaces in the form of Web application and Telegram messenger bot. The pre-trained models are available on Hugging Face, and the source code is released under a permissive license enabling various, including commercial, applications of the developed technology.

In future research, our goal is to investigate the potential of the latest image encoders. We plan to explore the development of more efficient UNet architectures for text-to-image tasks and focus on improving the understanding of textual prompts. Additionally, we aim to experiment with generating images at higher resolutions and to investigate new features extending the model: local image editing by a text prompt, attention reweighting, physics-based generation control, etc. The robustness against generating abusive content remains a crucial concern, warranting the exploration of real-time moderation layers or robust classifiers to mitigate undesirable, e.g. toxic or abusive, outputs.

8 Limitations

The current system produces images that appear natural, however, additional research can be conducted to (1) enhance the semantic coherence between the input text and the generated image, and (2) to improve the absolute values of FID and image quality based on human evaluations.

9 Ethical Considerations

We performed multiple efforts to ensure that the generated images do not contain harmful, offensive, or abusive content by (1) cleansing the training dataset from samples that were marked to be harmful/offensive/abusive, and (2) detecting abusive textual prompts.

While obvious queries, according to our tests, almost never generate abusive content, technically it is not guaranteed that certain carefully engineered prompts may not yield undesirable content. We, therefore, recommend using an additional layer of classifiers, depending on the application, which would filter out the undesired content and/or use image/representation transformation methods tailored to a given application.

Acknowledgements

As usual, we would like to thank the anonymous reviewers for their useful comments. We would also like to thank Sergey Markov and his team for helpful feedback and discussions, for collaboration in multimodal dataset collecting, labelling and processing.

References

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. 2021. Conditional image generation with score-based diffusion models. *CoRR*, abs/2111.13606.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. *CoRR*, abs/2304.08818.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *CoRR*, abs/2303.13873.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19822–19835.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. *CoRR*, abs/2302.03011.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022a. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. volume abs/2207.12598.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022b. Video diffusion models. In *NeurIPS*.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up gans for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver,*