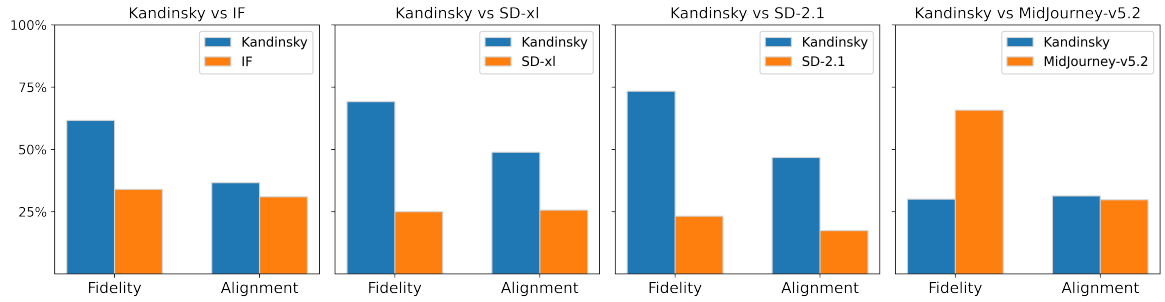

Figure 6: Human evaluation: competitors vs Kandinsky with diffusion prior on Drawbench. The total count of votes is 5000.

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.