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.

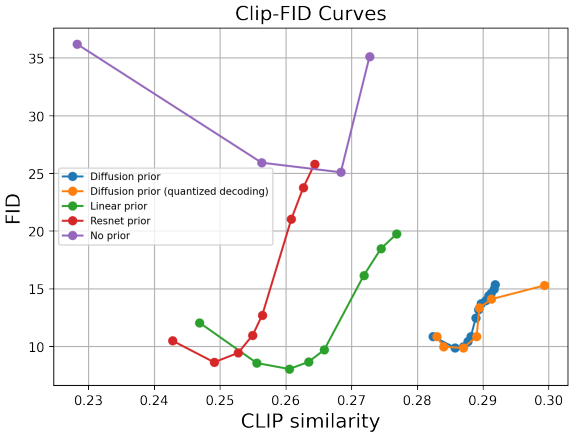

Figure 4: CLIP-FID curves for different setups.

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