MMCP Team at ImageCLEFmed MEDVQA-GI 2024 Task: Diffusion Models for Text-to-Image Generation of Colonoscopy Images

Mikhail Chaichuk[1,*], Elena Tutubalina[1,2,3]

[1]National Research University Higher School of Economics (HSE University), Russia
[2]Kazan Federal University, Russia
[3]Artificial Intelligence Research Institute, Russia

Abstract

This paper introduces models developed for the ImageCLEFmed 2024 MEDVQA-GI task, aimed at leveraging text-to-image generative models to create a comprehensive dataset of artificial colonoscopy images from textual prompts. The task's complexity arises from the novel and relatively uncharted nature of the provided training dataset, its limited size, and the specificity required in the generated images. We explore multiple approaches, including the efficient fine-tuning of large generative models such as Kandinsky and the modification of conditional latent Diffusion Probabilistic Models (DDPMs) tailored to text prompts. Our model achieved first place, with a Frechet Inception Distance (FID) score close to 0.1 on the official test set, reflecting the high quality and realism of the generated images.

Keywords

Neural networks, Generative networks, colonoscopy images, Kandinsky model, Multimodal architectures, Diffusion models, LoRa, PEFT, ImageCLEF

1. Introduction

The development of artificial neural networks for healthcare applications has recently become one of the most promising research directions within the field of artificial intelligence. Neural networks are widely used in medicine for various tasks, such as medical image segmentation [1], disease and anomaly detection, and diagnosis [2, 3, 4]. However, a major challenge in this area is the need for large volumes of annotated data to train the models. The acquisition of such data is complicated by patient confidentiality concerns and the necessity of involving qualified specialists for annotation.

Research in this field has been hindered for a long time by the lack of sufficient training data. Recently, there has been an increasing interest in the idea of creating synthetic medical images using generative neural networks. This represents an important area of study and has been explored in several papers [5, 6, 7]. The ImageCLEFmed [8] MEDVQA-GI 2024 Image Synthesis challenge [9] aims to utilize text-to-image generative models for creating medical images based on textual prompts. This involves generating images of various medical conditions from provided text descriptions. For example, if given a text description such as "An early-stage colorectal polyp", models are expected to produce an image that accurately reflects the given description.

This study addressed the ImageCLEFmed MEDVQA-GI 2024 task by investigating several generative models. Our work focuses primarily on fine-tuning large generative models using parameter-efficient fine-tuning techniques and modifying conditional diffusion models to better accommodate textual prompts. The fine-tuning of the Kandinsky model [10] was divided into three stages: initial training of the decoder model with a LoRa [11] rank of 32, followed by experiments with the Prior model using various LoRa [11] ranks, and finally fine-tuning the decoder with the best-performing Prior model. As an alternative approach, we proposed a model named MSDM based on Stable Diffusion [12] with

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
mishac22@yandex.ru (M. Chaichuk); tutubalinaev@gmail.com (E. Tutubalina)
0009-0009-8900-2195 (M. Chaichuk); 0000-0001-7936-0284 (E. Tutubalina)

two components — VAE [13] and U-Net [1] - and Text Encoder with CrossAttention within the U-Net blocks. In addition, we explored several image and text augmentation techniques to enhance the dataset provided by the organizers.

This article is organized as follows: Section 2 reviews relevant research, Section 3 details the research problem and data, Section 4 outlines the models and architectures used, Section 5 presents the experimental design and results, and Section 6 summarizes the findings and discusses future directions.

2. Related Work

For a significant period, generative adversarial networks (GANs) dominated this field, demonstrating promising results [14, 15, 16, 17, 18]. However, the emergence of diffusion models [19, 20, 12, 1] has significantly changed the landscape. Recent studies suggest that diffusion models can outperform GANs in terms of quality or exhibit similar performance in most tasks, including medical image generation [21, 22, 23, 24, 25, 26, 27]. Moreover, Savchenko et al. [28] demonstrate how leveraging textual clues can significantly enhance the interpretation of images. Müller-Franzes et al. [29] summarizes various studies comparing GANs and diffusion models in the context of medical image generation. It concludes that diffusion models generally outperform GANS both in terms of image quality and addressing issues associated with GANs, such as mode collapse and instability during training.

Recently, one of the most popular directions in this field has been the generation of medical images through fine-tuning large generative models. Although these models, in their original configurations, do not always perform well on highly specific tasks, they possess significant general generative capabilities that can be leveraged to create higher-quality images.

One of the most popular and effective fine-tuning techniques is the LoRa (Low-Rank Adaptation) method, which was first proposed by Hu et al. [11]. Rather than training all parameters in a model, the LoRa technique involves adding low-rank matrices to the weights of the model. This allows us to approximate changes in the weights using a smaller number of parameters, which has been shown to be highly efficient and has become widely popular.

Another popular method for adapting large models for specialized tasks is the Textual Inversion approach, which was presented by Gal et al. [30]. This approach aims to train a model to associate new concepts with specific textual tokens. The idea is to introduce “virtual“ tokens that represent specific visual concepts or styles and then train these tokens such that they can be used in text prompts to generate images with desired characteristics.

De Wilde et al. [31], Chambon et al. [32] applied the aforementioned techniques to fine-tune the Stable Diffusion model [12] on small specialized datasets. Both techniques demonstrated high flexibility and efficiency. However, ultimately, the authors concluded that direct fine-tuning of the model is preferable compared to using the Textual Inversion method.

3. Task and Dataset Description

3.1. Task

In this study, we address a task in the domain of developing innovative approaches for creating artificial medical images — ImageCLEFmedical Visual Question Answering for Colonoscopy Images 2024 (ImageCLEFmed MEDVQA-GI) [9] , proposed at the international ImageCLEF 2024 task [8].

Our focus is on the first of two subtasks in this competition — ImageCLEFmed MEDVQA-GI 2024 Image Synthesis. This subtask involves developing a machine learning algorithm that transforms textual descriptions into pictures to create a set of artificial medical images simulating the results of procedures such as colonoscopy and gastroscopy.

Participants are required to provide a model capable of generating images from textual queries that closely resemble real images of the human digestive system (esophagus, stomach, intestines), both pathological changes and normal conditions.

3.2. Dataset

(a) Generate an image with 1 polyp.

(b) Generate an image with polyps of size 5-10mm and 11-20mm.

(c) Generate an image with an abnormality located in the upper-right, lower-right, and center-right.

Figure 1: An example of images (with the original proportions preserved) and their corresponding text queries from the Development Dataset.

The participants were provided with two datasets for this task: the development and testing datasets.

The development dataset consists of 20,241 pairs of "text-images". The texts in these pairs represent examples of queries that the final model should be able to process, and they generally reflect the content of the corresponding images. It is important to note that not all data in the dataset are unique. There are a total of 2,000 different images and 483 unique text queries in the dataset. Accordingly, each pair represents a diverse combination of these elements, with each image having an average of approximately 10 different suitable descriptions (ranging from 7 to 14). Examples of images and their descriptions are presented in Figure 1. We divided the development dataset into a training and a validation set. The validation set consists of 200 images with approximately 2,000 associated descriptions, while the training subset includes the remaining 1,800 images, along with approximately 18,000 related descriptions. We used the train set to train models and the validation set to correctly evaluate their quality in terms of the loss function.

The testing dataset, in contrast, only contains text queries that will be used for image generation by the final model. This dataset contains 5,000 queries, of which 260 are unique. Each of these test queries also appears in the training dataset, ensuring that the model has seen them before.

4. Methodology

In this study, we tested two different approaches to solving the task.

4.1. Fine-tuning Large Models

The initial hypothesis of this research posited that pre-trained large generative models would be the most effective in generating the required images. These models exhibit a high capability for image creation due to their training on diverse and extensive datasets, which enables them to capture complex patterns and structures. To adapt such models to this specific task, it was necessary to fine-tune them on the available data. However, large generative models typically possess a large number of trainable parameters and require substantial computational resources for complete retraining. Therefore, the most feasible approach was to employ efficient fine-tuning tools, with the most advanced method being LoRA (Low-Rank Adaptation) [11].

For fine-tuning, the Kandinsky model family [10] was selected. This choice was driven by several factors. Firstly, Kandinsky is one of the most powerful image generation models, ranking among the best in the world [33]. Secondly, its weights have been made publicly available by the developers, significantly simplifying the process of working with the model.

4.1.1. Kandinsky 2.2

Initially, the Kandinsky 2.2 model [10] was tested, as it demonstrated high-quality results while requiring fewer training resources compared to the more recent Kandinsky 3.0 model [34]. The architecture of Kandinsky 2.2 consists of the following components:

1. **Image Prior Model**: The DiffusionMapping [35] model, which generates a visual embedding using CLIP [36] from a given text prompt or CLIP text embedding, all while remaining within the paradigm of latent visual space. The CLIP model used is CLIP-ViT-G.
2. **Image Decoder**: A U-Net [1] diffusion model employed for the generation of images.
3. **Sber-MoVQGAN** [37]: A modified version of VQGAN [38] developed by the authors, which has demonstrated high quality in experiments.

The Image Prior model synthesizes a visual embedding from a given text, which is subsequently used in the image decoder's training process. This way, the reverse diffusion mechanism learns to reconstruct the latent representation of an image not only from textual input but also from visual embeddings, thereby enhancing the overall quality. During the course of working with the model, checkpoints [39, 40] provided by the authors on the Hugging Face [41] platform were utilized.

4.1.2. Kandinsky 3.0

Kandinsky 3.0 [34] is a newer version of the Kandinsky model. It is significantly larger in size and has been trained on a much larger dataset than previous versions. Its developers claim that this model can achieve better results when used with LoRa than its predecessors. Therefore, we have decided to include it in this study.

In Kandinsky 3.0, the developers have abandoned the two-stage generation process that was used in Kandinsky 2.2 and adopted a more traditional approach by directly feeding text into the model. This change was made possible by the development of new large language models that have a better understanding of text than the CLIP text encoder used in previous versions. As a result, the new architecture consists of three main components:

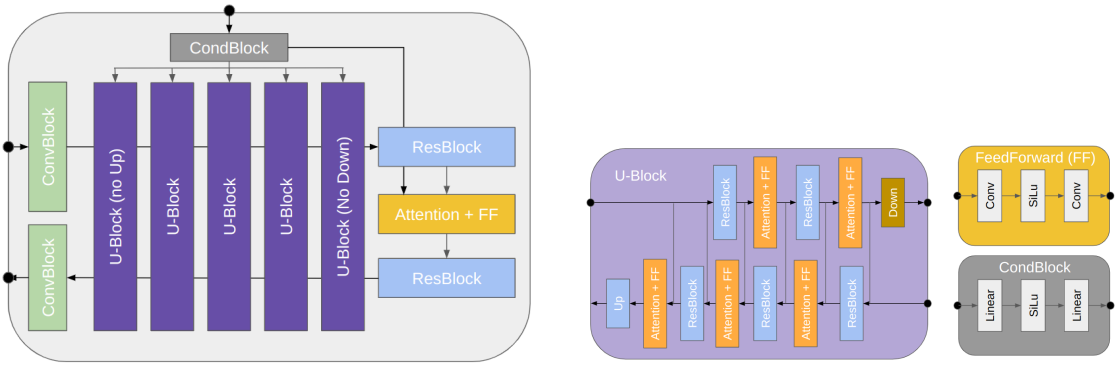
1. **FLAN-UL2** [42]: A large language model based on the T5 architecture [43]. In Kandinsky 3.0, only the Encoder of this model is utilized.
2. **U-Net with a modified architecture**: A U-Net model, which primarily consists of BigGAN-deep blocks [44], allowing for a doubling of the architecture's depth while maintaining the same overall number of parameters.
3. **Sber-MoVQGAN**: The same decoder as used in version 2.2.

These architectural modifications significantly simplified the training and fine-tuning processes. Only the U-Net requires training, while all other models are used in a frozen state. As with the previous version, for fine-tuning Kandinsky 3.0, the weights [45] provided by the authors on Hugging Face [41] were used.

4.2. MSDM: Medical Synthesis with Diffusion Models

4.2.1. MSDM Architecture

In addition to efficient training of large models, we propose a denoising diffusion probabilistic model (DDPM) designed for text-guided medical image generation. Our proposed model named MSDM is based on the Medfusion model [29], which is based on **Stable Diffusion** [12] and comprises two components — **VAE** [13] and **U-Net** [1], which need to be trained sequentially. In its initial configuration, the Medfusion model was not designed for text-to-image generation, as Müller-Franzes et al. [29] worked with class labels. Therefore, we extend this architecture with the following modifications (Figure 2):



(a) MSDM U-Net architecture

(b) MSDM U-Net blocks

Figure 2: MSDM architecture

1. Integration of **CrossAttention layers** within the U-Net blocks, which facilitates the extraction of information from text embeddings during training;
2. **FeedForward layer** after each CrossAttention layer.

For text encoding, the **CLIP ViT-L/14** model [46] was employed. This model was chosen because it demonstrates good performance while maintaining a relatively small size and is recommended as a standard solution by the Hugging Face community [41]. Despite the trend towards replacing CLIP with larger language models, the uniformity and simplicity of texts in the utilized dataset render the application of such models excessive for this study.

4.2.2. Text Diversity Enhancement with Paraphrasing

Insufficient text diversity in the training dataset might lead to model overfitting, causing the generation of homogeneous images for each given text. To address this issue, we utilize two potential approaches:

1. Implementing the context dropout technique, which randomly drops descriptions during training with a certain probability, thus replacing conditional generation with unconditional generation.
2. Increasing the diversity of the existing set of texts using a generative language model.

We used the GPT Turbo 3.5 model [47] to generate new texts. Using this model, we prepared a set of paraphrased original texts, increasing the number of unique elements from 483 to over 11,000.

We experimented with several strategies:

1. Adding paraphrased descriptions to the original texts.
2. Randomly replacing some of the original texts with paraphrased ones.
3. Completely replacing all original texts with paraphrased versions.

Our experiments on the validation set show that the best result was achieved in the first experiment by adding paraphrased texts to the originals (Section 5.1.4). In the other two cases, the model's performance on original texts deteriorated.

5. Experiments and Results

5.1. Experiments on the Development Dataset

5.1.1. Experimental Setup

To evaluate the performance of the model, we considered several metrics. First, we examined the Fréchet Inception Distance (FID) score [48], which was mentioned by the organizers as a key metric for

this task. In addition, we assessed the quality and variety of images using Precision and Recall metrics adapted for image evaluation [49, 50], as well as F1-score metrics that combine these measures.

To calculate the metrics, we used all 2,000 images from the development dataset as a set of real data, as the FID metric heavily relies on the number of images under evaluation. We generated artificial images for evaluation within 150 generation steps, as this number of iterations is optimal in terms of balance between time spent and metric quality, as shown in the paper [29] and confirmed through our own experiments. The TorchMetrics module [51] was used for the metric calculation.

During the training and validation of the models, images of size 256×256 pixels were utilized, as this size was found to be optimal in terms of striking a balance between memory requirements for calculations and the final image quality. Additionally, the model architectures employed are well-suited for working with this resolution. For the training and fine-tuning of each model, the AdamW optimizer [52] was employed with default parameters unless otherwise specified. In the course of each of the experiments described, the selection of the optimal value for the learning rate parameter (lr) was carried out. However, in all cases, the value of lr=10^{-4} was found to be optimal. Therefore, it can be assumed, by default, that this value of the parameter is used unless specified otherwise.

Each of the Kandinsky models was trained for a default of 100 epochs, with the model demonstrating the best performance during this training period being selected for further evaluation. In contrast, the MSDM model underwent an extended training regime of 500 epochs, and similarly, the model exhibiting superior performance post-training was selected for subsequent analysis.

The experiments were conducted on the HSE University cHARISMa supercomputer cluster [53]. For all experiments, we used 1 to 4 NVIDIA A100 GPUs.

5.1.2. Kandinsky 2.2

The main challenge in fine-tuning the Kandinsky 2.2 model was the need to simultaneously fine-tune both its components in order to achieve optimal results and accurate metric calculations. Therefore, the fine-tuning process was divided into three stages.

In the first stage, the decoder model with a LoRa rank of 32 was trained for 40 epochs. During this training, metrics were not calculated. In the second stage, we used this already fine-tuned decoder model to conduct experiments with the Prior model. The primary parameter adjusted during these experiments was the LoRa rank. The rank value varied, starting from 4, then from 8 to 64 in steps of 8. Additionally, a rank of 128 was considered. The results of these experiments are presented in Table 1. Overall, we observed a consistent improvement in the metrics as the rank increased. However, there was a relatively minor change in the FID metric when we transitioned from a rank of 64 to 128, which was interpreted as indicating that further increases in the rank would not significantly improve the results.

The next step involved fine-tuning the Decoder model using the best-performing Prior model with a rank of 128. The rank selection for the Decoder model was carried out in a similar manner. The results are presented in Table 2. In this case, the effect of the rank value on the final image quality was less significant. Based on the obtained data, it can be inferred that the prior model has the greatest influence on the quality of the generated images.

Additionally, experiments were conducted to select the batch size and the alpha parameter value for LoRa and LoRa modifications such as DoRa [54] and rsLoRa [55]. However, these experiments resulted in only a slight quality improvement.

5.1.3. Kandinsky 3.0

As with Kandinsky 2.2, initial experiments were conducted by varying the rank values for LoRa during training. Analysis of the resulting images showed that they became much more diverse compared to Kandinsky 2.2, as confirmed by an increase in the Recall metric to values approaching 0.5 across all experiments. However, a visual analysis also revealed shortcomings in the model's photorealism, as the images had an unusual texture resembling drawings on paper or those created in a 3D editor. Examples

Table 1
Kandinsky 2.2 Image Prior model fine-tuning with LoRa results on the development dataset

LoRa Rank	Number of parameters	FID (↓)	Precision (↑)	Recall (↑)	F1 (↑)
4	1,310,720	112.543	0.442	0.248	0.317
8	2,621,440	106.061	0.479	0.271	0.346
16	5,242,880	91.256	0.486	0.263	0.341
24	7,864,320	88.971	0.504	0.280	0.360
32	10,485,760	87.266	0.543	0.283	0.372
40	13,107,200	85.732	0.572	0.298	0.391
48	15,728,649	81.563	0.609	0.315	0.415
56	18,350,080	78.145	0.628	0.359	0.456
64	20,971,520	76.988	0.677	0.362	0.471
128	41,943,040	**75.835**	**0.701**	**0.385**	**0.497**

Table 2
Kandinsky 2.2 Image Decoder model fine-tuning with LoRa results on the development dataset

LoRa Rank	Number of parameters	FID (↓)	Precision (↑)	Recall (↑)	F1 (↑)
4	823,296	75.713	0.651	0.371	0.472
8	1,646,592	73.176	0.668	0.373	0.478
16	3,293,184	71.241	0.673	0.382	0.487
24	4,939,776	70.351	0.681	0.391	0.496
32	6,586,368	67.407	0.687	0.397	0.503
40	8,232,960	68.221	0.709	0.413	0.521
48	9,879,552	68.332	0.714	0.429	0.535
56	11,526,144	70.012	0.6985	0.426	0.529
64	13,172,736	66.968	0.731	**0.435**	**0.545**
128	26,345,472	**66.869**	**0.745**	0.421	0.537

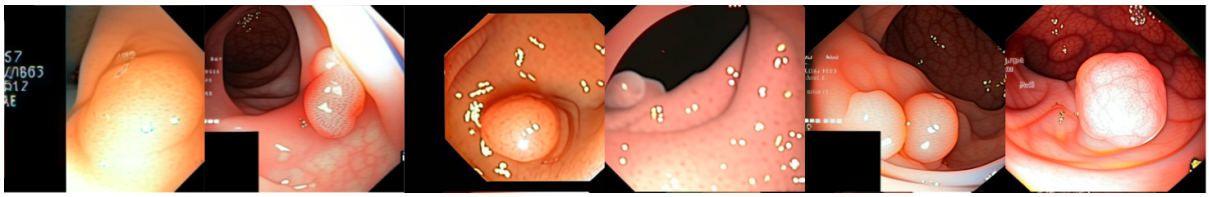


Figure 3: Examples of images obtained after fine-tuning Kandinsky 3.0. As can be observed, these images exhibit an unusual texture reminiscent of 3D graphics or hand-drawn illustrations, which results in a lack of photorealism.

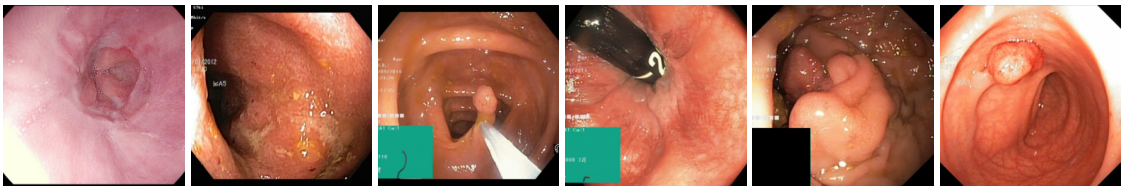
of the generated images and the issues related to their photorealism are presented in Figure 3. This negatively affected the FID metric values, which ranged between 90 and 110, significantly lower than those achieved with the Kandinsky 2.2 model. We experimented with increasing the batch size and reducing the parameter ϵ of the AdamW optimizer from 10^{-8} to a range between 10^{-15} and 10^{-10}. Corresponding experiments were conducted but did not lead to a noticeable improvement in results.

5.1.4. MSDM

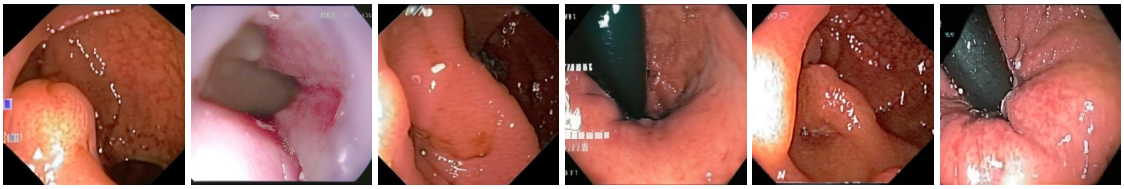
Implementation of MSDM is based on the code provided by the Medfusion developers on GitHub [56]. Weights for CLIP 14 were obtained from Hugging Face [57].

The MSDM model immediately demonstrated significantly improved quality and higher textual comprehension abilities compared to the Kandinsky models. After 100,000 training steps, the FID score was reduced to 35, and the Precision value increased to 0.82. However, the Recall parameter remained relatively low at 0.35, indicating insufficient diversity in the generated images.

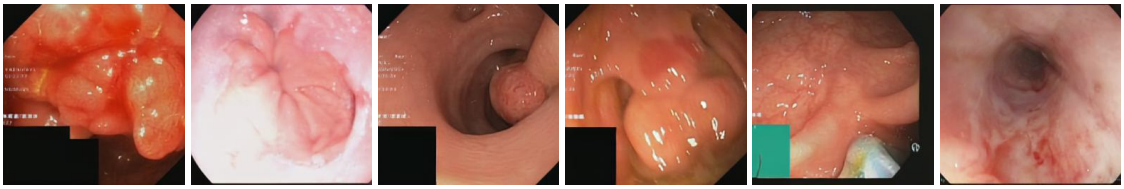
One possible reason for the lack of variability in the generated images could be the limited size of the training dataset and, consequently, insufficient diversity in the training examples. In such cases, additional random data augmentations are often employed. However, for this dataset, the use of



(a) Original images from the Developing Dataset



(b) Kandinsky 2.2 + LoRa, rank 128



(c) MSDM

Figure 4: Examples of images generated using Kandinsky 2.2 and MSDM compared to the original images from the development dataset.

common augmentation methods such as horizontal and vertical flips or changes in color palette proved problematic. This is due to the fact that text prompts contained information about the placement of elements in specific parts of the image and required colors. Altering these parameters could negatively impact the model's text perception. However, we applied augmentations that randomly altered the brightness and contrast of the images.

Another method for increasing the diversity of generated images is the "self-conditioning" technique. This approach allows the model to consider its own results from previous steps in the generation process. Thus, the model receives a more comprehensive context for its work, which contributes to improved consistency, variability, and quality of the final images. The "self-conditioning" technique was integrated into MSDM and tested in combination with other methods.

Additionally, one potential cause of reduced image generation quality could be model overfitting due to a small amount of data. To combat this phenomenon, dropout [58] is employed, which helps reduce the risk of overfitting and improves the model's generalization capability. The optimal dropout probability value from the experiments was found to be 0.1. The combination of all three aforementioned approaches resulted in an increase in the Recall metric to 0.437 without a noticeable loss in quality. Regarding experiments with paraphrased texts, we achieved the following results for the development dataset with the first strategy: FID is 33.58, precision is 0.929, recall is 0.498, and the F1 score is 0.648.

5.2. Official Results on the Testing Dataset

The competition organizers conducted submissions testing using data from both single-center and multi-center sources as sets of real images. We were provided with a list of prompts for which we generated one image per prompt and sent it to the organizers.

As the final solution, we submitted three sets of images, each containing 5,000 images that were generated based on prompts from the test dataset. We used the following three models for the final generation:

- Kandinsky 2.2 model, both parts of which were fine-tuned using LoRa with a rank of 128.
- Two versions of the MSDM model — one trained without adding paraphrased texts, and the other including them.

Table 3
The results of the evaluation of the three final models on the test dataset provided by the competition organizers. The results for the FID and Inception Score (IS) metrics are shown.

Model Name	Dataset Type	FID (↓)	IS (avg) (↑)	IS (med) (↑)
Kandinsky 2.2 + LoRa, rank 128	single	**0.086**	1.624	1.633
	multi-center	**0.064**	1.624	1.633
	both	**0.066**	1.624	1.633
MSDM	single-center	0.120	**1.791**	**1.792**
	multi-center	0.117	**1.791**	**1.792**
	both	0.114	**1.791**	**1.792**
MSDM + paraphrases	single-center	0.125	1.773	1.775
	multi-center	0.121	1.773	1.775
	both	0.119	1.773	1.775

Examples of the images generated by these models, as well as their comparison with the original images, can be seen in Figure 4. Table 3 presents the evaluation results of the models based on the FID and Inception Score metrics. As can be seen, the models demonstrated significantly higher quality scores on the test dataset according to the FID metric compared to the results obtained during development. Interestingly, the Kandinsky 2.2 model (4.6 billion parameters in total, 68 million trainable parameters) ultimately demonstrated better FID results on the test set compared to the MSDM model (463 million parameters in total, 326 million trainable parameters), contrary to the trend observed during the dev phase. At the same time, MSDM turned out to be slightly better in terms of Inception Score.

6. Conclusion

In this paper, we present our solution to the ImageCLEFmed [8] MEDVQA-GI 2024 Image Synthesis task [9]. The task involved developing a generative neural network to create artificial colonoscopy images from textual descriptions. We experimented with several approaches during the research process, each demonstrating their effectiveness. The resulting models showed excellent quality, achieving an FID score of approximately 0.1 on the test dataset and receiving high praise from the organizers of the ImageCLEFmed 2024 competition [8].

The findings of this study could be beneficial for future developments in the field of medical imaging. There is significant potential for further research in this area. Future research could involve experimenting with other architectures, such as large models or autoencoders, on the same dataset or applying the studied methods to different types of medical images, such as MRI, CT scans, or X-rays.

Acknowledgments

The work of E.T. has been supported by the Russian Science Foundation grant # 23-11-00358. We acknowledge the computational resources of HPC facilities at the HSE University. We express our gratitude to the Kandinsky development team for their assistance and consultations during the experiments with their models. We also thank the PhD student of the Faculty of Computer Science at HSE University, Airat Valiev, for his help in generating paraphrased texts.

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[2] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, D. Merhof, Diffusion models in medical imaging: A comprehensive survey, Medical Image Analysis (2023) 102846.

[3] A. S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on mri, Zeitschrift für Medizinische Physik 29 (2019) 102–127. URL: https://www.sciencedirect.com/science/article/pii/S0939388918301181. doi:https://doi.org/10.1016/j.zemedi.2018.11.002, special Issue: Deep Learning in Medical Physics.

[4] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annual Review of Biomedical Engineering 19 (2017) 221–248. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-bioeng-071516-044442. doi:https://doi.org/10.1146/annurev-bioeng-071516-044442.

[5] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, A. Weller, Synthetic data–what, why and how?, arXiv preprint arXiv:2205.03257 (2022).

[6] J. Jordon, A. Wilson, M. van der Schaar, Synthetic data: Opening the data floodgates to enable faster, more directed development of machine learning methods, arXiv preprint arXiv:2012.04580 (2020).

[7] T. Wang, Y. Lei, Y. Fu, J. F. Wynne, W. J. Curran, T. Liu, X. Yang, A review on medical imaging synthesis using deep learning and its clinical applications, Journal of Applied Clinical Medical Physics 22 (2021) 11–36. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13121. doi:https://doi.org/10.1002/acm2.13121. arXiv:https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/acm2.13121.

[8] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[9] S. A. Hicks, A. Storås, P. Halvorsen, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2024 – medical visual question answering for gastrointestinal tract, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[10] A. Razzhigaev, A. Shakhmatov, A. Maltseva, V. Arkhipkin, I. Pavlov, I. Ryabov, A. Kuts, A. Panchenko, A. Kuznetsov, D. Dimitrov, Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023, pp. 286–295.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[13] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[14] G. Kwon, C. Han, D.-s. Kim, Generation of 3d brain mri using auto-encoding generative adversarial networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 118–126.

[15] L. Sun, J. Chen, Y. Xu, M. Gong, K. Yu, K. Batmanghelich, Hierarchical amortized gan for 3d high resolution medical image synthesis, IEEE Journal of Biomedical and Health Informatics 26 (2022) 3966–3975. doi:10.1109/JBHI.2022.3172976.

[16] N. K. Singh, K. Raza, Medical image generation using generative adversarial networks, arXiv preprint arXiv:2005.10687 (2020).

[17] T. Han, S. Nebelung, C. Haarburger, N. Horst, S. Reinartz, D. Merhof, F. Kiessling, V. Schulz, D. Truhn, Breaking medical data sharing boundaries by employing artificial radiographs, bioRxiv

(2019). URL: https://www.biorxiv.org/content/early/2019/11/14/841619. doi:10.1101/841619. arXiv:https://www.biorxiv.org/content/early/2019/11/14/841619.full.pdf.

[18] J. Krause, H. I. Grabsch, M. Kloor, M. Jendrusch, A. Echle, R. D. Buelow, P. Boor, T. Luedde, T. J. Brinker, C. Trautwein, A. T. Pearson, P. Quirke, J. Jenniskens, K. Offermans, P. A. van den Brandt, J. N. Kather, Deep learning detects genetic alterations in cancer histology generated by adversarial networks, The Journal of Pathology 254 (2021) 70–79. URL: https://pathsocjournals.onlinelibrary.wiley.com/doi/abs/10.1002/path.5638. doi:https://doi.org/10.1002/path.5638.

[19] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International conference on machine learning, PMLR, 2015, pp. 2256–2265.

[20] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in neural information processing systems 33 (2020) 6840–6851.

[21] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Advances in neural information processing systems 34 (2021) 8780–8794.

[22] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, M. J. Cardoso, Brain imaging generation with latent diffusion models, in: MICCAI Workshop on Deep Generative Models, Springer, 2022, pp. 117–126.

[23] F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, et al., Medical diffusion: Denoising diffusion probabilistic models for 3d medical image generation, arXiv preprint arXiv:2211.03364 (2022).

[24] Z. Dorjsembe, S. Odonchimed, F. Xiao, Three-dimensional medical image synthesis with denoising diffusion probabilistic models, in: Medical Imaging with Deep Learning, 2022. URL: https://openreview.net/forum?id=Oz7lKWVh45H.

[25] K. Packhäuser, L. Folle, F. Thamm, A. Maier, Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 2023, pp. 1–5.

[26] P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani, A. Bashashati, A morphology focused diffusion probabilistic model for synthesis of histopathology images, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1999–2008. doi:10.1109/WACV56688.2023.00204.

[27] H. K. Kim, I. Ryu, J. Y. Choi, T. K. Yoo, Early experience of adopting a generative diffusion model for the synthesis of fundus photographs, 2022. doi:10.21203/rs.3.rs-2183608/v2.

[28] A. Savchenko, A. Alekseev, S. Kwon, E. Tutubalina, E. Myasnikov, S. Nikolenko, Ad lingua: Text classification improves symbolism prediction in image advertisements, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1886–1892.

[29] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J. N. Kather, D. Truhn, A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis, Sci. Rep. 13 (2023) 12098.

[30] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: Personalizing text-to-image generation using textual inversion, arXiv preprint arXiv:2208.01618 (2022).

[31] B. De Wilde, A. Saha, R. P. ten Broek, H. Huisman, Medical diffusion on a budget: textual inversion for medical image generation, arXiv preprint arXiv:2303.13430 (2023).

[32] P. Chambon, C. Bluethgen, C. P. Langlotz, A. Chaudhari, Adapting pretrained vision-language foundational models to medical imaging domains, arXiv preprint arXiv:2210.04133 (2022).

[33] A. Raza, Technologya showdown of creativity: A comparative analysis of proprietary generative ai image models, 2024. URL: https://techbullion.com/a-showdown-of-creativity-a-comparative-analysis-of-proprietary-generative-ai-image-models/.

[34] V. Arkhipkin, A. Filatov, V. Vasilev, A. Maltseva, S. Azizov, I. Pavlov, J. Agafonova, A. Kuznetsov, D. Dimitrov, Kandinsky 3.0 technical report, arXiv preprint arXiv:2312.03511 (2023).

[35] W. Peebles, S. Xie, Scalable diffusion models with transformers, in: Proceedings of the IEEE/CVF

International Conference on Computer Vision, 2023, pp. 4195–4205.
[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[37] github.com, Sber-movqgan, 2024. URL: https://github.com/ai-forever/MoVQGAN.
[38] P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.
[39] huggingface.co, Hugging face hub kandinsky community kandinsky 2.2 prior, 2024. URL: https://huggingface.co/kandinsky-community/kandinsky-2-2-prior.
[40] huggingface.co, Hugging face hub kandinsky community kandinsky 2.2 decoder, 2024. URL: https://huggingface.co/kandinsky-community/kandinsky-2-2-decoder.
[41] huggingface.co, Hugging face hub documentation, 2024. URL: https://huggingface.co/docs/hub/index.
[42] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, et al., Ul2: Unifying language learning paradigms, arXiv preprint arXiv:2205.05131 (2022).
[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.
[44] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, arXiv preprint arXiv:1809.11096 (2018).
[45] huggingface.co, Hugging face hub kandinsky community kandinsky 3, 2024. URL: https://huggingface.co/kandinsky-community/kandinsky-3.
[46] openai.com, Clip: Connecting text and images, 2024. URL: https://openai.com/index/clip/.
[47] openai.com, Gpt-3.5 turbo fine-tuning and api updates, 2024. URL: https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/?ref=mlq.ai.
[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).
[49] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, T. Aila, Improved precision and recall metric for assessing generative models, Advances in neural information processing systems 32 (2019).
[50] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, S. Gelly, Assessing generative models via precision and recall, 2018.
[51] lightning.ai, Torchmetrics module, 2024. URL: https://lightning.ai/docs/torchmetrics/stable/.
[52] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[53] P. S. Kostenetskiy, R. A. Chulkevich, V. I. Kozyrev, Hpc resources of the higher school of economics, Journal of Physics: Conference Series 1740 (2021) 012050. URL: https://dx.doi.org/10.1088/1742-6596/1740/1/012050. doi:10.1088/1742-6596/1740/1/012050.
[54] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen, Dora: Weight-decomposed low-rank adaptation, arXiv preprint arXiv:2402.09353 (2024).
[55] D. Kalajdzievski, A rank stabilization scaling factor for fine-tuning with lora, arXiv preprint arXiv:2312.03732 (2023).
[56] github.com, Medfusion - medical denoising diffusion probabilistic model, 2024. URL: https://github.com/mueller-franzes/medfusion.
[57] huggingface.co, Hugging face hub clip patch 14, 2024. URL: https://huggingface.co/openai/clip-vit-large-patch14.
[58] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, 2014. URL: https://api.semanticscholar.org/CorpusID:6844431.