# MythraGen: Two-Stage Retrieval Augmented Art Generation Framework

Quang-Khai Le [*,1,2], Cong-Long Nguyen [*,1,2], Minh-Triet Tran[1,2], and Trung-Nghia Le [**,1,2]

[1] University of Science, Ho Chi Minh city, Vietnam
[2] Vietnam National University, Ho Chi Minh city, Vietnam
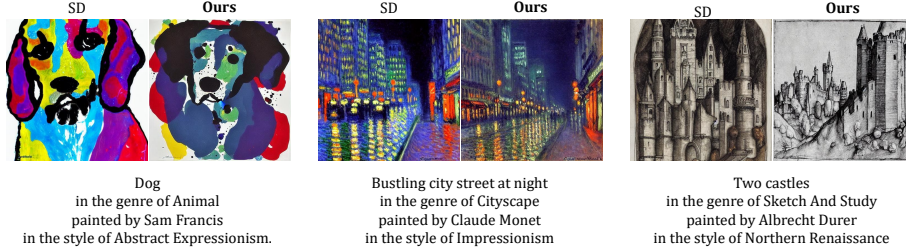{22120148,22120191}@student.hcmus.edu.vn,
{tmtriet,ltnghia}@fit.hcmus.edu.vn

Fig. 1: Examples of images generated by our method against Stable Diffusion.

**Abstract.** Text-to-image generation has seen rapid advancements, especially with the development of generative models. However, challenges remain in achieving high-quality, contextually accurate image outputs that faithfully match the provided textual descriptions, especially in artistic generation. In this paper, we present a simple yet efficient retrieval augmented generation framework, namely MythraGen, for text-to-artistic image generation by integrating an art retrieval mechanism with LoRA-based model fine-tuning. Our method extracts features from a large-scale art dataset, optimizing the generation process by combining artist-specific styles and content. Particularly, retrieved images from an external art database that have the highest similarity to the query prompt are used to finetune Stable Diffusion using LoRA for desired art generation. Experimental results and user studies on the WikiArt dataset show that our proposed method can generate artworks that closely match the user's input, significantly outperforming existing solutions.

**Keywords:** Text-to-image generation · Art retrieval · Art generation

---

[*] These authors contributed equally to this research.

[**] Corresponding author.

# 1   Introduction

In art analysis, content and style are two fundamental elements. Content describes the concepts depicted in the image, such as objects, people, or locations. Style, on the other hand, describes the visual appearance of the artwork, including its color, composition, and shape. Furthermore, each artist expresses his/her own style, creating unique features in his/her works. Through the unique combination of content and the artist's individual style that makes each piece of art special.

Recent advances in deep learning have facilitated powerful breakthroughs in text-to-image generation (T2I)[6,14,1,19,23,27]. T2I methods can incorporate a specific style into generated images by taking a textual description of the style as a prompt. However, conveying artistic style through text descriptions has limitations. These descriptions are often less expressive and informative than visual representations of style, so the style features of T2I outputs are often rough and lack details.

Recent fine-tuning techniques such as Dream-booth [21], Textual Inversion [8], and Low-Rank Adaptation (LoRA) [11] can enhance adaptability to specific T2I generation tasks, and show convincing capability in creating images with unique content and style. Among these methods, LoRA stands out and gains extensive adoption among art designers and T2I enthusiasts, due to its advantages of low-cost an computational efficiency, making it user-friendly and suitable for consumer devices. However, an artist can paint in many different styles. When extending this to hundreds of artists, the need to fine-tune or retrain the model for each artist's style becomes impractical. This process requires a vast amount of computational resources and time, making these methods unrealistic for large-scale application.

To address these issues, we propose MythraGen, a simple yet efficient retrieval augmented art generation framework. First, we employ a retrieval technique to search for paintings from an external database that have the highest similarity in content, genre, and style to the artist described in the prompt. Our art retrieval technique leverages BLIP-2 to encode both the visual features of each image and its associated metadata, including captions, genre, style, and artist information. These encoded features are combined into a comprehensive feature vector, which is indexed using FAISS [7] to facilitate the retrieval of relevant images. Then, we utilize LoRA algorithm [11] for finetuning Stable Diffusion on the identified paintings. This two-stage framework allows for the flexible combination of different styles from various artists and content, optimizing image generation while ensuring the quality of the created images.

In this paper, the WikiArt dataset [24], consisting of 80,000 unique images from more than 1,100 artists across 27 styles, was used as the external art database. We also leveraged a zero-shot classifier based on Visual Question Answering (VQA) to annotate genres for round 16,000 images missing labels. Extensive experiments and user study showcase an impressive performance our MythraGen, outperforming existing existing open sources and commercial image generation methods in all evaluation metrics.

Our contributions can be summarized as follows:

– We introduce a simple yet retrieval augmented art generation framework called MythraGen to create artistic images with desired content, genre, and style from the query prompt.
– We train LoRA models on various artistic styles and then mix them to create superior composite results.
– We employ Visual Question Answering (VQA) based zero-shot classification to automatically label the genre for 16,452 unclassified images in the WikiArt dataset.
– We conduct a comprehensive evaluation of our proposed framework in reproducing different artistic styles. Experimental results and user study demonstrate the superiority of our proposed method against existing open sources and commercial image generation methods.

## 2   Related Work

### 2.1   Art Retrieval

Similarity search algorithms have played a crucial role in many artificial intelligence applications, with K-Nearest Neighbors (KNN) and Approximate Nearest Neighbors (ANN) being commonly used methods. KNN works by finding the closest points in the dataset to a query point based on a specified distance metric, making it useful for classification and regression tasks. ANN, on the other hand, provides faster search results by approximating the nearest neighbors, making it suitable for high-dimensional data where exact searches are computationally expensive.

Recently, the Facebook AI Similarity Search (FAISS) library [7] is a powerful tool for efficient similarity search and clustering of high-dimensional data, enabling developers to quickly find similar items in large datasets. It is particularly useful for tasks such as image retrieval, recommendation systems, and natural language processing, where finding similar items in large datasets is crucial. Therefore, this paper utilized FAISS due to its ability to quickly and accurately search for similar embedding vectors in latent space. This is especially useful when working with large datasets like WikiArt (around 80k images), allowing us to rapidly gather relevant images to support image generation processes.

To ensure accurate retrieval, the alignment of image and text embeddings is crucial for effective cooperation. Inspired by the vision-language pre-training model BLIP-2 [12], which produces high-quality text-aligned visual representations, we adapted it to extract text-aligned subject representations, improving the model's ability to understand and generate content across modalities.

By integrating BLIP-2 [12] with FAISS, we harness the strengths of advanced vision-language pre-training models and efficient similarity search algorithms. This combination allows us to improve the accuracy of image-text retrieval, providing a more precise and comprehensive dataset for further applications.

{ "content" : "People walking along a tree-lined street",
 "artist" : "Berthe Morisot",
 "genre" : "Genre Painting",
 "style" : "Impressionism",
 "genre-style_scale" : 0.7,
 "artist-style_scale" : 0.4                                    }

**Input Prompt**

**Retrieved Arts**

**External Database**
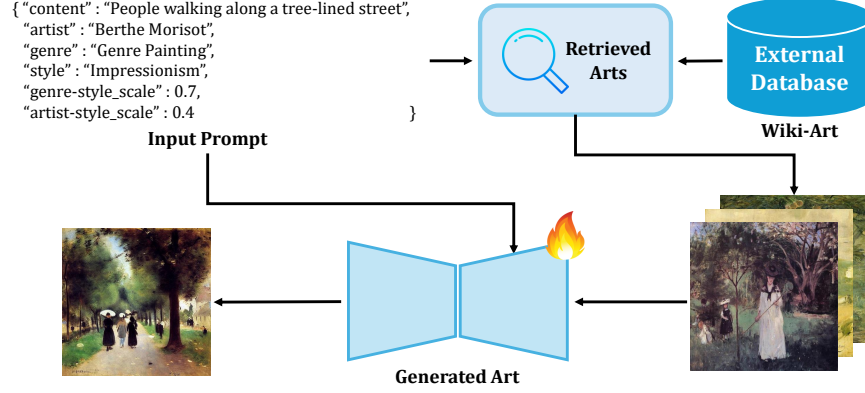
**Wiki-Art**

**Generated Art**

Fig. 2: Overview of the proposed MythraGen framework, with two main stages: (a) the Art Retrieval retrieves images to enhance the image generation process and (b) the Art Generation generates images based on the user's input text combined with the images provided by the Art Retrieval module.

## 2.2   Art Generation

Early text-to-image models [17,25,26,28] made significant progress by utilizing Generative Adversarial Networks (GANs) [8] trained on large paired image-caption datasets, which can lead to model collapse issues [2,5,10]. Recently, diffusion models [9,18,15,23] have become powerful in text-to-image (T2I) tasks due to their ability to generate high-quality images and their flexibility in adapting images to the context of the text. GLIDE [15] and Imagen [22] employ classifier-free guidance by replacing the label in the class-conditioned diffusion model with text descriptions of the images. Stable Diffusion [20] utilizes VQ-GAN for the latent representation of images, enhancing photorealism through an adversarial objective. DALL-E2 [16] employs a multimodal latent space where image and text embeddings are aligned to generate images reflecting a deeper level of language understanding. However, these models often struggle when handling prompts containing less common genres or styles related to artists. Instead of generating a suitable image, they tend to either create nonexistent genres or styles or use similar but more popular ones, which does not align with the user's intent, leading to a mismatch between the original prompt and the final image.

Unifying both image retrieval and generation processes, Re-Imagen [4] addressed augmenting rare entities to improve image quality. However, our goal is to use retrieval to enhance the less common drawing styles of various artists, while ensuring that when these styles are applied, the main content of the prompt remains preserved. Additionally, Re-Imagen trains its image generation process on the cascaded diffusion model [10], while our method uses LoRA [11] to fine-tune the Stable Diffusion model [12] for reducing required resources and speed up the training process.
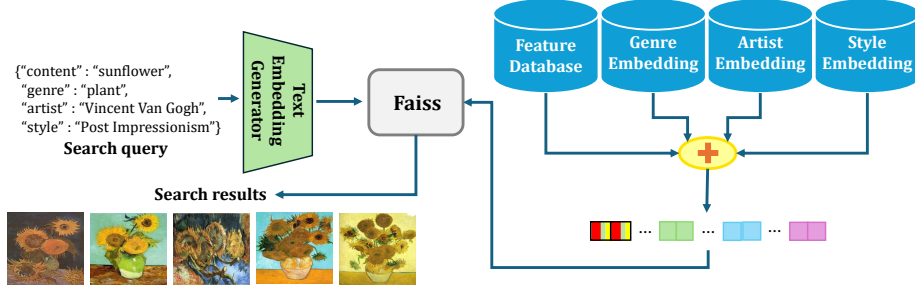
Fig. 3: Pipeline of the Art Retrieval module.

## 3    Proposed Method

### 3.1    Overview

Figure 2 illustrates an overview of our proposed two-stage retrieval augmented art generation framework. The input prompt is fed into the Art Retrieval module to search for images that are similar to the input text content. In the Art Generation module, we use the images obtained from the Art Retrieval module to enrich the data during the image generation process, which is also controlled by the text embedding vector. Finally, the Art Generation module produces a target image that has the highest similarity to the user's input text.

### 3.2    Art Retrieval

In this section, we present our approach for retrieval from an external database to find images with the highest similarity in content, genre, and style to the artist described in the input query. Figure 3 provides a visual representation of the Art Retrieval module, built around the Bootstrapping Language Image Pre-training (BLIP-2 [12]) architecture. First, each 256-dimensional vector from the feature, genre, artist, and style databases, which are related to each other, are concatenated into a single 1024-dimensional vector. This vector is then fed to the FAISS system for indexing to support the retrieval process. When the input query is processed by the Text Embedding Generator, generated by the Q-Former component of BLIP-2 [12], the resulting vector is sent to the FAISS system to return a set of images with the highest similarity in content, genre, and style to the artist described in the input query.

The Image Embedding Generator consists of two main components: a frozen pre-trained image encoder and a multimodal encoder called Q-Former (see Fig. 4). The process begins with the input image, which is passed through the image encoder to extract visual features. These features are then combined with learnable query tokens and fed into the Q-Former. The output of the Q-Former is then passed through a fully connected layer and normalized to produce the final image feature vector. Similarly, the Text Embedding Generator tokenizes the input text and processes it through the Q-Former. The output from the Q-Former is
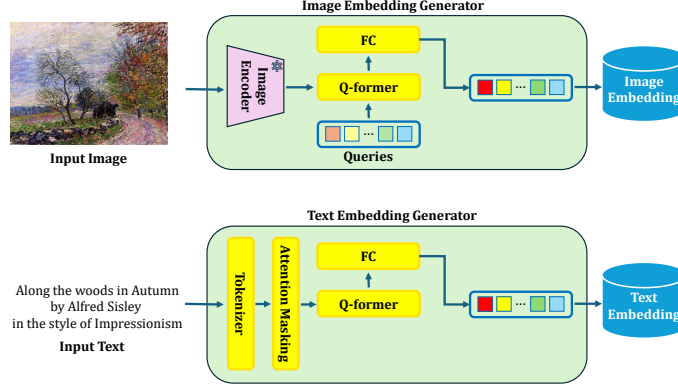
Fig. 4: Multimodal representation, where image and text embeddings processes rely primarily on the Q-former.

then passed through a fully connected layer and normalized to produce the final text feature vector.

To improve the retrieval performance, we combine image embedding with caption and genre embeddings. Let $E_i$ with $i \in \{image, caption, genre\}$ represents the image, caption, and genre embeddings. We compute the weighted sum of the embeddings as follows:

$$V = \sum_i W_i * E_i, \tag{1}$$

where $W_i$ denotes the corresponding weight of embedding $E_i$. The resulting vector $V$ becomes the final feature vector used for retrieval. By combining information from images, captions, and genres, a stronger feature vector is created for better retrieval performance for genre.

The feature vector may not perform well in retrieving style and artist information because it mainly captures the content and genre of the artwork. Moreover, many art styles share significant similarities in technique, such as *Action Painting* and *Abstract Expressionism*, or between movements like *Cubism*, *Analytical Cubism* and *Synthetic Cubism* or *Impressionism* and *Post-Impressionism*, etc. This makes it difficult to distinguish between closely related styles or specific artists. To address this, we concatenate additional style embedding and artist embedding to enrich the data for the vector before the indexing phase (See Fig. 3). By incorporating these additional embeddings, the model can better capture the nuances of style and artist characteristics, thereby improving retrieval performance in these specific areas. Moreover, despite the increased complexity from the added embeddings, our optimized retrieval system maintain an impressive speed, achieving a retrieval time of just 0.1 second per query. This efficiency demonstrates the robustness of our approach, ensuring that even with enriched data, the system remains highly responsive.
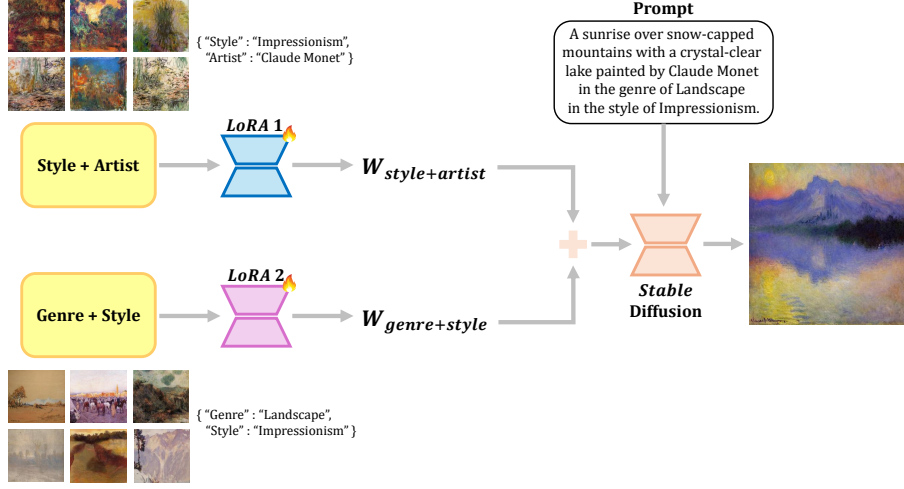
Fig. 5: LoRA combination in Art Generation module. The first LoRA model is fine-tuned for the artist and style ($W_{style+artist}$), while the second LoRA model is fine-tuned for the genre and style ($W_{genre+style}$). Both are combined and applied to the Stable Diffusion model to generate an image that faithfully reflects the input prompt, incorporating the specific style, artist, and genre.

### 3.3    Art Generation

We utilize the LoRA technique [11] to fine-tune two distinct LoRA models, each is specifically trained on different types of input data sourced from the Art Retrieval module (see Fig. 5). The fine-tuning process allows us to tailor each LoRA model to accurately reflect particular artistic attributes, such as genre, style, and artist characteristics. After fine-tuning, each LoRA model is assigned its own unique weight, carefully calibrated to balance the influence of each model on the final output. These weighted models are then seamlessly mixed together, in conjunction with the given prompt, to generate the final target image. This process allows to create highly nuanced and accurate representations of artistic styles. For example, as shown in Fig. 5, by combining the *Landscape genre* with the *Impressionism style* and integrating the influence of *the artist Claude Monet*, the technique produces an image of a sunrise over snow-capped mountains with a crystal-clear lake. The resulting image not only depicts the specified scene but also strongly embodies Monet's distinctive Impressionist style, showcasing the effectiveness of our approach in capturing complex artistic nuances.

## 4    Experiments

### 4.1    Implementations

In our experiments, we leveraged the PyTorch deep learning framework on a computer equipped with an NVIDIA RTX A4500 GPU with 24 GB. For the Art
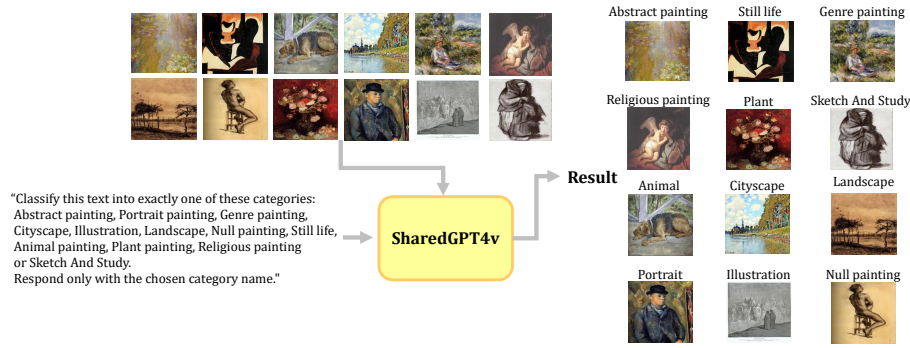
Fig. 6: Visual Question Answering (VQA) model used to classify the genre of an image when the genre is unknown.

Retrieval module, we used BLIP-2 [12] pre-trained with ViT-L/14 to embed the image and its related metadata. Weights in Eq. 1 is set $W_{\text{image}} = 1.0$, $W_{\text{caption}} = 0.9$, and $W_{\text{genre}} = 0.75$.

Regarding the Art Generation module, we employed Stable Diffusion V1.5 [19] as the backbone. We utilized two LoRA models for our experiments: one fine-tuned for genre + style and another fine-tuned for artist + style. For both LoRA, we set max_train_steps to 4095, using AdamW8bit as the optimizer to save memory, with xformers enabled for memory optimization and mixed_precision = "bf16" for faster computation without compromising quality. The learning rates were set to $1 \times 10^{-4}$ for the U-Net and $5 \times 10^{-5}$ for the text encoder. The LoRA network was configured with network_dim = 32 and network_alpha = 1, adjusting the rank and scaling of the LoRA layers for efficient fine-tuning.

## 4.2   Dataset

We used WikiArt as the external database for all experiments in this paper. WikiArt has complete labels for artists and styles, but over 16,452 images are missing genre labels. Therefore, we employed a Visual Question Answering (VQA) model (i.e., ShareGPT4V [3]) to label these 16,452 images according to 12 corresponding genres (genre painting, illustration, landscape, etc.) as described in Fig. 6, where the model analyzes each image and assigns the appropriate genre label based on visual content.

We performed retrieval on the WikiArt dataset using three separate categories: artist, styles, and genres, as well as a combined category that includes all three.

## 4.3   Art Retrieval

Table 1 compares the performance of three BLIP-2 versions in retrieval tasks across genres, artists, and styles. All models get a perfect mAP@5 of 1.0, showing

Table 1: Results of different BLIP-2 versions on retrieval across characteristics.

| Model | Genres | | | | |
|---|---|---|---|---|---|
| | mAP@5 | mAP@15 | mAP@25 | mAP@40 | mAP@50 |
| BLIP2 using ViT-g/14 (EVA-CLIP) | 100% | 98.4% | 98.4% | 98.6% | 98.6% |
| BLIP2 using ViT-L/14 (CLIP) | 100% | **99.8%** | **99.7%** | **99.6%** | **99.5%** |
| BLIP2 finetuned on COCO [13] | 100% | 98.8% | 98.3% | 97.8% | 97.7% |

| Model | Artists, Styles | | | | |
|---|---|---|---|---|---|
| | mAP@5 | mAP@15 | mAP@25 | mAP@40 | mAP@50 |
| BLIP2 using ViT-g/14 (EVA-CLIP) | 100% | 100% | 100% | 100% | 100% |
| BLIP2 using ViT-L/14 (CLIP) | 100% | 100% | 100% | 100% | 100% |
| BLIP2 finetuned on COCO [13] | 100% | 100% | 100% | 100% | 100% |

| Model | Genres + Artists + Styles | | | | |
|---|---|---|---|---|---|
| | mAP@5 | mAP@15 | mAP@25 | mAP@40 | mAP@50 |
| BLIP2 using ViT-g/14 (EVA-CLIP) | 90.5% | 92.1% | 92.4% | 92.4% | 92.6% |
| BLIP2 using ViT-L/14 (CLIP) | 92.8% | **94.8%** | **94.4%** | **93.8%** | **94.2%** |
| BLIP2 finetuned on COCO [13] | **93.4%** | 94% | 93.7% | 93.7% | 93.7% |

Table 2: Results of comparing the performance between using a standalone image embedding and combining it with caption and genre information across different versions of BLIP-2.

| Embedding vector | BLIP2 using ViT-g/14 (EVA-CLIP) | | | | |
|---|---|---|---|---|---|
| | mAP@5 | mAP@15 | mAP@25 | mAP@40 | mAP@50 |
| Image embedding | 82.7% | 78.5% | 77.0% | 75.7% | 75.2% |
| Image + Caption + Genre embedding | **100%** | **98.4%** | **98.4%** | **98.6%** | **98.6%** |

| Embedding vector | BLIP2 using ViT-L/14 (CLIP) | | | | |
|---|---|---|---|---|---|
| | mAP@5 | mAP@15 | mAP@25 | mAP@40 | mAP@50 |
| Image embedding | 89.0% | 81.3% | 78.6% | 75.3% | 74.1% |
| Image + Caption + Genre embedding | **100%** | **99.8%** | **99.7%** | **99.6%** | **99.5%** |

| Embedding vector | BLIP2 finetuned on COCO [13] | | | | |
|---|---|---|---|---|---|
| | mAP@5 | mAP@15 | mAP@25 | mAP@40 | mAP@50 |
| Image embedding | 84.6% | 79.8% | 77.0% | 75.5% | 74.9% |
| Image + Caption + Genre embedding | **100%** | **98.8%** | **98.3%** | **97.8%** | **97.7%** |

high accuracy for the top 5 results. For genres, BLIP2 with ViT-L/14 (CLIP) scores the highest mAP@50 at 99.5%, followed by ViT-g/14 (EVA-CLIP) at 98.6%, and the COCO-finetuned model at 97.7%. In artist-based and style-based retrieval, all models reach a perfect mAP of 1.0 at all levels, showing equal skill in retrieving artist and style information. However, when combining genre, artist, and style , the COCO-finetuned model performs best at mAP@5 with 93.4%, but ViT-L/14 (CLIP) performs better as the number of queries increases, reaching the highest mAP@50 at 94.2%. Overall, ViT-L/14 (CLIP) performs best in more complex tasks, while ViT-g/14 (EVA-CLIP) has slightly lower results.

Table 2 shows the comparison between using only image embedding and combining it with caption and genre information across different versions of BLIP-2. The results use metrics like mAP@5, mAP@15, mAP@25, mAP@40, and mAP@50 to measure accuracy. Using only image embedding gives good results, such as 82.7% with BLIP-2 using ViT-g/14 (EVA-CLIP) and 89.0% with BLIP-2 using ViT-L/14 (CLIP) at mAP@5. However, when adding caption and

Table 3: Comparison of methods in terms of CLIP-T, CLIP-I, and FID metrics.

| Metrics | Methods | | | |
|---|---|---|---|---|
| | **MythraGen** | **BingAI** | **Midjourney** | **SD** |
| CLIP-T ↑ | **30.68** | 27.77 | 30.13 | 29.61 |
| CLIP-I ↑ | **79.84** | 66.82 | 65.19 | 75.29 |
| FID ↓ | **322.9** | 373.85 | 329.22 | 325.79 |

genre embeddings, the performance improves significantly, reaching over 98% for most cases.

### 4.4   Art Generation

**Compared Methods.**  We compared our method with state-of-the-art (SOTA) methods, including Stable Diffusion (SD) version 2.0 [19] and two well-known commercial products: BingAI [3] and Midjourney [4].

**Dataset.**  The comparison was conducted on a customized dataset extracted from WikiArt. The dataset consists of 50 randomly selected (genre, artist, style) combinations from WikiArt, along with 50 prompts based on scenes that we generated. Each prompt was carefully crafted to evaluate the methods' ability to accurately generate images that adhere to both the style and content described. This customized dataset allowed for a comprehensive comparison of the models' performance in generating stylistically consistent and content-accurate images.

**Experimental Results.**  Table 7 shows that our MythraGen outperforms the other methods across all three metrics. In terms of CLIP-T, which measures the textual similarity between the prompt and the generated image, MythraGen achieves the highest score at 30.68, surpassing Midjourney (30.13), SD (29.61), and BingAI (27.77). For CLIP-I, which measures the style similarity, MythraGen also lead with a score of 79.84, followed by SD (75.29), BingAI (66.82), and Midjourney (65.19). Finally, for FID, which measures the quality of the images, MythraGen obtains the lowest score (indicating better performance) at 322.9, compared to SD (325.79), Midjourney (329.22), and BingAI (373.85).

Fig. 7 illustrates the visual results of these methods. While SD [19] struggles to balance style and content due to poor representation of the text extracted from the reference image, the content of the images generated by BingAI, and Midjourney is relatively closer to the prompt, but their style differs from the style reference. In contrast, our method produces images that are more faithful to the style of the reference image, especially regarding brushstrokes, lines, etc. This demonstrates that our MythraGen method achieves a better balance between content similarity, style similarity, and generated quality according to objective metrics.

---

[3] https://www.bing.com/images/create/

[4] https://www.midjourneyfree.ai/

Glassmakers street in Vilno 1906 in the genre of Cityscape painted by Mstislav Dobuzhinsky in the style of Symbolism



A vibrant garden scene with blurred flowers in the genre of Plant painted by Pierre Auguste Renoir in the style of Impressionism.

Reference    **MythraGen**    BingAI    Midjourney    Stable Diffusion 2.0

Fig. 7: Qualitative comparison with SOTA methods based on style reference image. Methods like BingAI, Midjourney, and SD 2.0 lack specific stylistic information drawn from the original artist, leading to difficulties in balancing content and style. In contrast, MythraGen performs better in generating both style and content as intended.

### 4.5    Human Evaluation

**Setup.**    We also conducted a human evaluation of MythraGen against three SOTA methods as in Section 4.4, on the same dataset as in Section 4.4. The user study focused on two key criteria:

- **Faithfulness** assesses the extent to which the generated image is true to both the entity's appearance and the text description. It considers factors such as object fidelity and text relevance, ensuring that the image accurately represents the caption provided
- **Naturalness** evaluates the visual quality of the generated image. It encompasses aspects of style similarity to the original image, ensuring that the image adheres faithfully to the style described in the prompt without noticeable artifacts or discrepancies that detract from its overall aesthetic appeal.

The user study involves 32 participants (48.4% of whom are male) with age between 11 and 60 (most of them are from 11 to 20). Each participant was asked to score the outputs from different methods on a scale of 1 (worst) to 5 (best) based on two primary criteria. Each participant evaluated a total of 30 images per method.

**Quantitative Results.**    Fig. 8 shows the quantitative results of the human evaluation, where MythraGen achieves the highest scores for both faithfulness and naturalness. For faithfulness, MythraGen gets an average score of 3.09, clearly improving over SD 2.0 (2.88), Midjourney (2.97), and BingAI (2.53). This result shows that MythraGen is better at generating images that accurately match the descriptions, both visually and textually. Similarly, for naturalness, MythraGen also outperforms other methods with an average score of 3.28. Participants find
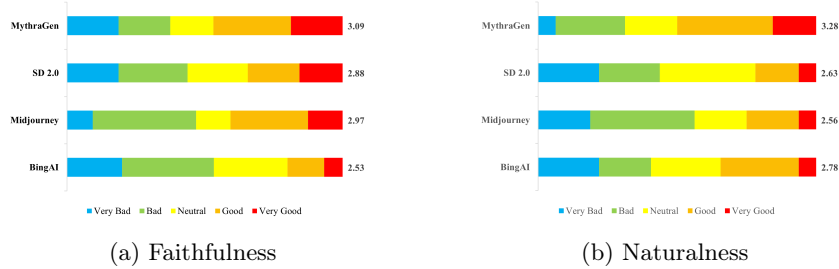
(a) Faithfulness

(b) Naturalness

Fig. 8: Humans evaluate the methods based on two criteria: Faithfulness and Naturalness.

MythraGen's images more visually appealing and better aligned with the described styles in the input prompt. These results demonstrate that MythraGen not only excels in generating images that are faithful to their descriptions but also produces outputs that accurately reflect the requested artistic style, proving its effectiveness over SOTA methods in both accuracy and quality.

Our experimental results show that MythraGen is highly effective at generating images that faithfully represent both the text prompt and the desired artistic style. By using a retrieval-augmented approach, MythraGen leverages existing artwork to fine-tune the generation process, producing high-quality outputs that closely match user expectations. Additionally, the use of LoRA for efficient fine-tuning helps reduce computational costs while maintaining impressive performance, making the model accessible even on less powerful hardware.

## 5    Conclusion

In this paper, we present MythraGen, which represents the advancement in the field of text-to-artistic image generation by effectively combining retrieval-based techniques with LoRA fine-tuning. Our work not only enhances the quality and contextual accuracy of generated artworks but also enables the incorporation of diverse artistic styles that meet the user's expectations. Experimental results demonstrate that MythraGen outperforms existing methods in generating images that faithfully reflect text descriptions and are highly natural, as evidenced by user studies. We further demonstrate that our model is particularly effective at generating images from text that requires a greater diversity of artistic genres and periods. We believe that our work can inspire further innovations in the intersection of art and artificial intelligence, fostering deeper engagement with both creators and audiences.

# References

1. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Trans. Graph. **42**(6), 243 (2023). https://doi.org/10.1145/3618322, https://doi.org/10.1145/3618322

2. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv (2018), https://api.semanticscholar.org/CorpusID:52889459

3. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv (2023), https://arxiv.org/abs/2311.12793

4. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. arXiv (2022), https://api.semanticscholar.org/CorpusID:252596087

5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

6. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., Tang, J.: Cogview: Mastering text-to-image generation via transformers. In: NeurIPS. pp. 19822–19835 (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/a4d92e2cd541fca87e4620aba658316d-Paper.pdf

7. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024), https://arxiv.org/abs/2401.08281

8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022), https://arxiv.org/abs/2208.01618

9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv (2020), https://api.semanticscholar.org/CorpusID:219955663

10. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR **23**(47), 1–33 (2022), http://jmlr.org/papers/v23/21-0635.html

11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022), https://openreview.net/forum?id=nZeVKeeFYf9

12. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023), https://api.semanticscholar.org/CorpusID:256390509

13. Lin, T.Y., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312

14. Liu, D., Fan, H., Liu, J.: Expogenius: Robust personalized human image generation using diffusion model for exposure variation and pose transfer. In: ICMR. pp. 239–247 (2024). https://doi.org/10.1145/3652583.3658071, https://doi.org/10.1145/3652583.3658071

15. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models (2022), https://arxiv.org/abs/2112.10741

16. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv (2022), https://arxiv.org/abs/2204.06125

17. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML. Proceedings of Machine Learning Research, vol. 48, pp. 1060–1069 (2016), https://proceedings.mlr.press/v48/reed16.html

18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685 (2021), https://api.semanticscholar.org/CorpusID:245335280

19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

21. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv (2022), https://arxiv.org/abs/2208.12242

22. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS **35**, 36479–36494 (2022)

23. Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS. p. 2643 (2024). https://doi.org/10.5555/3600270.3602913

24. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE Trans. Image Process. **28**(1), 394–409 (2019). https://doi.org/10.1109/TIP.2018.2866698, https://doi.org/10.1109/TIP.2018.2866698

25. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018)

26. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. TPAMI **41**(8), 1947–1962 (2019). https://doi.org/10.1109/TPAMI.2018.2856256

27. Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., Sun, T.: Towards language-free training for text-to-image generation. In: CVPR. pp. 17886–17896 (2022). https://doi.org/10.1109/CVPR52688.2022.01738

28. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR. pp. 5795–5803 (2019), https://api.semanticscholar.org/CorpusID:91183909