Style-Frame: A Foundational Framework for Artistic Style Driven Applications

Pavan Gajula, Abhishek Dangeti, Vivek Srivastava, Vikram Jamwal
TCS Research, India
{pavanbhargav.gajula, abhishek.dangeti, srivastava.vivek2, vikram.jamwal}@tcs.com

Abstract

'Style' is an essential element of an artwork. Generative models have opened up new opportunities in content creation, customization, and curation applications involving the style of an artwork. These opportunities include generating images in a particular style, personalizing a given artwork into different styles, and classifying and clustering given artworks in a museum or exhibition based on style. However, in the absence of foundational design frameworks and with a plethora of research outputs and competing technologies, creating robust co-creative technology solutions or platforms that purposefully exploit different aspects of style in content creation is difficult. In this paper, we introduce a framework that aims to cut the clutter on the technological aspects of artistic style. Our proposed framework, *Style-Frame*, aims to synthesize and communicate core concepts in the technology fields related to *'style aspects of an artwork'*. We survey, explore, and evaluate the existing technologies, specifically AI systems, for generating and customizing new artworks in the artistic styles of an artwork or an artist, shedding light on the style concepts, capabilities, and challenges in applying these systems and techniques. We present the various aspects of our framework through experimental case studies based on the artworks archived in the MUNCH Museum in Oslo, Norway.

Introduction

Style can be defined as the organizing principles by which something is achieved or constructed [Knight1994]. Style is a universal concept and applicable to all creative and artistic aspects of our life such as architecture(e.g. neo-futurism, Gothic, Renaissance), visual art (e.g., impressionism, cubism, hyper-realism), music(e.g. classical, jazz, rock), and fashion(casual, formal chick). While the extent to which 'style' and 'style transfer' satisfy different computational creativity desiderata is open to discussion [Brown and Jordanous2022], nevertheless, the authors believe that style forms an essential element of co-creative art systems.

In the visual art domain, there has been a rapid advancement in image-generative AI in recent years. The ability to learn and apply artistic styles through machine learning models has opened new application avenues and business opportunities in creative content creation. Since it is a rapidly evolving field, with a constant stream of innovations, it becomes quite challenging to design a technology solution or a co-creative platform that can effectively utilize the various aspects of new style-transfer capabilities.

To facilitate the design and implementation of the technology solutions, we need to survey, analyze, and critically evaluate the technology landscape for *'artistic style'* in the context of application design. We synthesize this knowledge in the form of an application design framework called *Style-Frame* which has two components - a conceptual model and a process model (section **Style-Frame: Components**). The five dimensions of the conceptual model cover different aspects of knowledge, viz., style specification, style transformation, Gen-AI technologies, process evaluation, and artifact quality evaluation that facilitate appropriate design choices for a technology solution concerning a particular application. While these dimensions cover the static aspect of the framework, the true value of the framework is in applying this knowledge through a process. We further present the dynamic aspect of the framework in the form of a process model.

We illustrate the application of this framework through a creative application solution - *image generation in a particular style* in the context of the MUNCH Museum in Oslo, Norway (section **Case-study**). The case study captures the critical evaluation of different technology design options by considering comparative solutions in light of this framework. We conclude with our experience and by pointing to the areas of future research in the advancement of the framework (section **Conclusion**). Our **main contributions** are:

- We provide a novel design framework that elicits the various dimensions along which style design decisions in a typical co-creative visual art application.
- We point out the various decision choices available for these dimensions in the light of a survey of current technologies.
- We provide the process model for applying this framework.
- We present a real-life case study of the framework usage.

Related Work

AI for Art

The advent of modern AI tools such as Generative Adversarial Networks [Goodfellow et al.2014], transformers

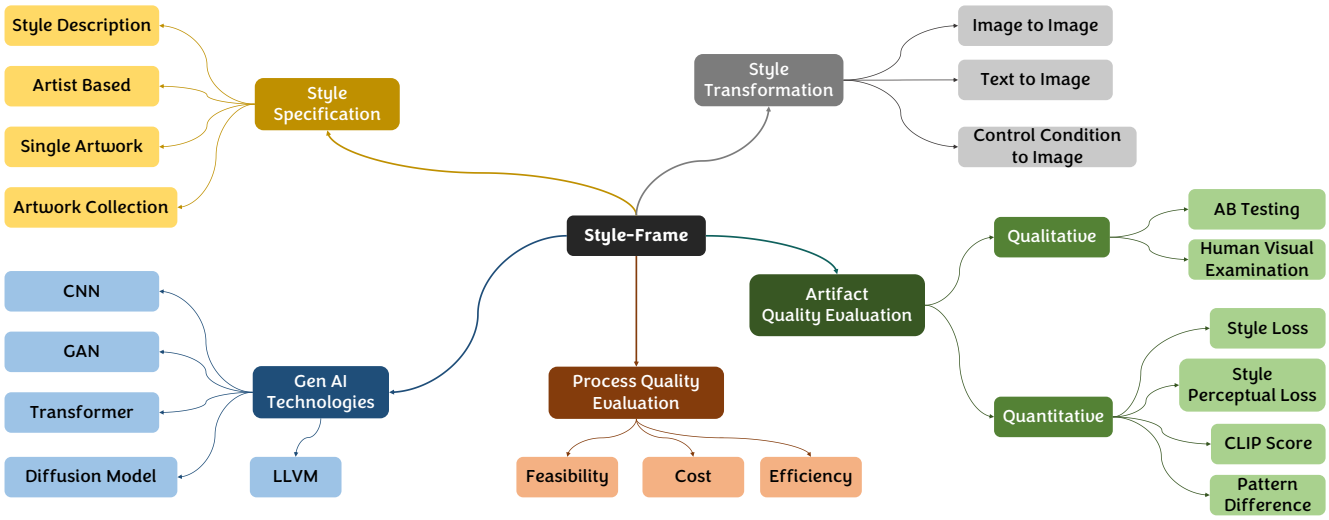

Figure 1: The Conceptual Model of the Style-Frame framework.

[Vaswani et al.2017], and diffusion models [Rombach et al.2022a] has given a boost to human-machine co-creativity.

In the space of artwork creation and manipulation, AI plays a fundamental role in applications such as style transfer and personalization. Style transfer has been an active area of interest in computer vision and allied fields where it is usually studied as a problem of texture synthesis [Jing et al.2019], which is to extract and transfer the texture from the style reference image to the target image [Efros and Freeman2023], [Drori, Cohen-Or, and Yeshurun2003]. Traditional style transfer methods used handcrafted features to match the patches between the content image and the style reference image [Zhang et al.2013, Resales, Achan, and Frey2003]. In recent years, several advanced technologies such as deep convolutional neural networks have been used to capture and transfer style patterns [Gatys, Ecker, and Bethge2016a]. Apart from the widely popular image-to-image style transfer task, recent works on text-to-image generation also facilitate generating novel images in a desired style controlled through the text prompt [Kwon and Ye2022, Sohn et al.2024, Liu et al.2023]. These AI technologies empower modern artists to create high quality artworks. It also encourages many non-artists to participate in the creative process of artwork generation. It further underscores the importance of frameworks such as *Style-Frame* to effectively build AI applications for these use cases.

Frameworks for AI applications

Applications powered with AI has emerged as a dominant mechanism to build products and deliver services in multiple domains such as retail [Oosthuizen et al.2021, Anica-Popa et al.2021], healthcare [Osman Andersen et al.2021], and e-commerce [Bawack et al.2022]. The traditional software development life-cycle (SDLC) includes five stages: software requirements, software design, software implementation, software testing, and software maintenance. These stages help to design and build high-quality software in a cost-effective and time-efficient manner. Several frameworks (such as waterfall, spiral, and agile) have been proposed in the past to effectively manage and deliver the project. In modern AI application development, we observe a dramatic change in the activities performed in each stage of the SDLC [Ishikawa and Yoshioka2019]. Owing to such paradigm shift, the frameworks for AI applications is the need of the hour [Smith and Eckroth2017]. We observe an increase in the interest in building such frameworks for AI-driven applications centered around healthcare [Soenksen et al.2022], manufacturing [Kaymakci, Wenninger, and Sauer2021], marketing [Huang and Rust2021], etc.

Style-Frame: Components

As discussed in the previous sections, the design of artistic style-driven AI applications involves making several key decisions. In such a scenario, it is essential to have a framework that can educate on the different concepts and design decisions involved in style-driven applications and help effectively navigate through different stages of the application design workflow.

Some key questions are: From where do we draw the style reference? What is the input to which we apply style guidance? What is the required output? What kind of technology is available for style transformation? What kind of artifact and production quality is of interest, and how do we evaluate such quality? To address these questions and facilitate the design of artistic style-driven applications, we present a novel framework, viz., *Style-Frame*. The framework has two aspects: (A) The conceptual model, which covers the structural or knowledge aspects of a style-driven application, and (B) The Process model, which covers the dynamic or the application aspects of applying the knowledge to achieve the desired outcomes in style-driven application design.

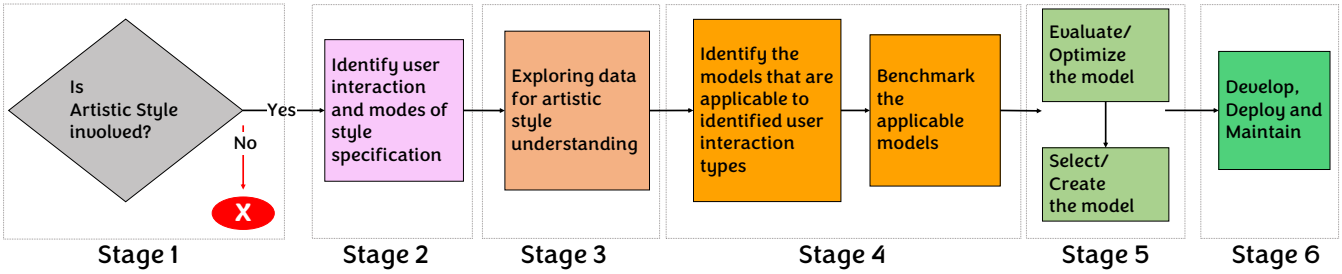

Figure 2: The Process Model of the Style Frame Framework.

The Conceptual Model of Style-Frame

At the top level, Style-Frame has five conceptual components that capture the different dimensions of knowledge required in the design of Style-driven applications (refer to Figure 1). These are:

1. Style specification: The fundamental requirement in building a style-driven application is to identify how a user would provide the style reference. A user can specify the style through diverse techniques such as providing a single artwork or a collection of artworks. A user can also specify the style through verbal description such as by providing the name of the artist (e.g. "in the style of Edvard Munch") or through a more detailed style description such as "*in the style of The Scream by Edvard Munch*", "*in the style of watercolor paintings by Vincent van Gogh*", etc.

2. Style transformation: Based on the style reference specification by the user, the next step is to identify the mechanism to apply the style transformation. Given an image as the style reference, the *image-to-image* style transformation technique allows us to transform an image in the style of the style reference image. On the other hand, the *text-to-image* style transformation technique lets us generate new image based on the style description provided in the text prompt. Whereas the *control-condition based* style transformation facilitates us to use additional control conditions such as line drawing and hand-drawn sketch and further refine them in a certain style with style transfer models.

3. Generative AI technologies: The choice of generative AI technology influences key decisions in application development. Also, the style reference specification by the user and the type of style transformation would influence the choice of AI technology for the application. For instance, if a user wishes to specify the style through verbal description in a text prompt, the diffusion model-based techniques that are widely popular for text-to-image generation would be a suitable choice. With the rapid pace of development in the AI landscape, we observe several prominent fundamental technologies available at our disposal with their own merits and limitations such as the inference speed, cost of computation, etc. Some of the AI technologies frequently used in the style transfer literature include diffusion models, transformers [Ramesh et al.2021, Chang et al.2023a], GANs [Goodfellow et al.2014, Richardson et al.2021], convolutional neural networks, etc.

4. Artifact quality evaluation: While training (and post-training) the AI models for the desired style transformation, we need to evaluate the quality of the artifact produced by these models. The quality of the artifact reflects the AI model's capability on the given style transformation task. For instance, in the case of the image-to-image style transformation task, we evaluate the artifact quality by measuring the style similarity between the style reference image and the stylized image. We evaluate the quality of the artifact (and the model) with both qualitative and quantitative measures. Qualitative evaluation generally involves assessing the quality based on human visual perception. Whereas the quantitative evaluation involves metrics such as style loss [Gatys, Ecker, and Bethge2016b], perceptual loss [Johnson, Alahi, and Fei-Fei2016], prompt fidelity, etc.

5. Process evaluation: Furthermore, we need to evaluate the different processes involved in the entire workflow such as data curation, bench-marking existing models, artifact quality evaluation, etc. The idea behind the process evaluation is to identify and troubleshoot the bottleneck in the involved processes. We evaluate the processes on criteria such as feasibility, cost, and efficiency. Through different mechanisms, we identify if the processes are easy to set up and execute while being cost-efficient and satisfying the project requirements.

The Process Model of Style-Frame

Apart from knowledge explication in a style-driven application design, the real value of the framework is in applying it. The process model describes the process or the steps of applying the *Style-Frame* framework to a style driven application design (refer to Figure 2). Essentially it consists of following steps:

1. Checking if the problem involves artistic style
2. Identifying the user interaction and modes of style specification
3. Data exploration for artistic style understanding
4. Identifying and benchmarking applicable models
5. Developing Gen AI technology solution and evaluation
6. Deploying and maintaining the application

We shall present the various aspects of applying the framework through experimental application design cases-study based on the artworks in MUNCH Museum Image Archive.

Case Study

MUNCH Museum in Oslo, Norway, is one of the largest art museums in the world dedicated to a single artist. It is dedicated to the life and works of the Norwegian artist Edvard Munch.

Problem Statement We wanted to understand and evaluate how good are the modern AI technologies in capturing the style of an artist (Edvard Munch in this case) and utilizing that learned style in creating new creative content. In this regard, we explore two different problem statements in this case study:

- *P1*: Stylizing a given image in the style of Edvard Munch.
- *P2*: Generating new images in the style of Edvard Munch with user-defined controls.

For these tasks, we utilize the MUNCH Museum's image archive of artworks [Mun] comprising paintings and sketches. For the problem P2, we consider the text prompt as the principal user-defined conditional control but the usage can be extended to other conditional controls such as line-drawings, scribbles, depth-maps, and pose-markers

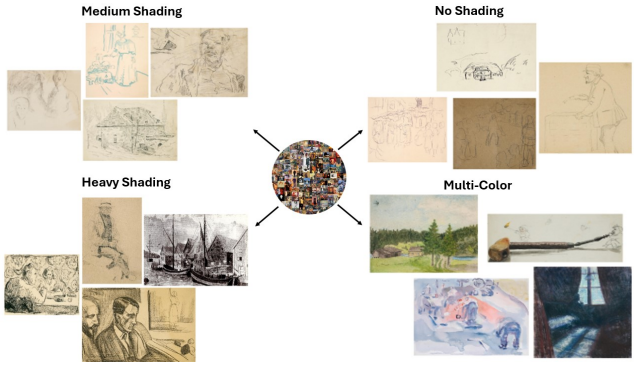

Figure 3: Artwork categories from the MUNCH museum.

Our Exploration

We demonstrate the usage of *Style-Frame* framework by applying it to both the problems and show its effectiveness in helping us make critical design and technology choices. We discuss in detail the various decision choices, technological capabilities, and roadblocks throughout these stages. The six stages (refer to Figure 2) are:

1. Check if the problem involves artistic style:

We first examine the problem statement to identify if it involves artistic style requirements. Problems such as image generation and artwork personalization with AI techniques are popularly known to include artistic style transformations at different stages.

In both our problems (*P1* and *P2*), we need to capture and transfer the artistic style from the existing artworks of Edvard Munch available in the MUNCH Museum's Image archive. Hence, both the problems involve artistic style and satisfy the first stage requirements of the Style-Frame's process model. Therefore, we successfully move ahead with the subsequent stages of the framework.

2. Identify the user-interaction and the modes of style specification:

A fundamental task in this stage is to understand how the user will interact with the application and specify the style. As discussed, the *Style Specification* component of the *Style-Frame* framework allows us to categorize the style specification into four broad categories. We also assess the feasibility and ease of specifying the style reference by the user.

In the context of *P1*, the user specifies the intended style through a single artwork of Edvard Munch or a collection of artworks of similar style. Leveraging this style specification from the user, the model (refer to Stages 4 and 5 for a detailed discussion) will attempt to transfer the style to the user-provided target content image. For *P2*, the user typically specifies the intended artistic style through the text prompt by either providing the style description (e.g. "watercolor painting"), the name of the artist (e.g. "in the style of Edvard Munch"), or both. In addition, we can fine-tune the current text-to-image generation models on a single artwork or a collection of artworks along with the artwork caption as the text prompt.

3. Data exploration for artistic style understanding: In this stage, we leverage the available data and the metadata to explore different artistic styles present in the artworks. An artist's style evolves over time and as a result, we may have multiple styles representations in the artwork collection. The metadata such as the genre, motifs, caption, color, and shading information can help one put the artworks in different style clusters.

In this study, we work with the dataset of 7411 Edvard Munch's artworks provided by the MUNCH Museum Image Archive. In the earlier exploration with this art corpus [Sivertsen et al.2023], the museum categorized the dataset into 5 style categories based on the shading and the color information: *heavy shading* (230 artworks), *medium shading* (1353 artworks), *no shading* (3529) artworks, *multi-color* (894 artworks), and *no label* (1405 artworks). Representative artworks from each category are shown in Figure 3. We will use this style categorized dataset in the subsequent stages to benchmark, train, and evaluate AI technologies for both the problems.

4. Identifying and benchmarking applicable Gen AI models:

After exploring different artistic styles present in the dataset, the next step is to identify and benchmark different *AI technologies* available for the given problem. This stage is crucial to understand and estimate the computational requirements, technological limitations, and the capabilities of the AI technologies available at our disposal. We benchmark different AI technologies along multiple dimensions such as model configuration, model training, model inference, and model performance.

In our case, for problem *P1*, the technologies such as transformer models, GANs, and CNNs seems a good fit among the available options at the moment as they give state-of-the-art performance. Transformers are good at modelling relationships among different visual entities, GANs can learn the input data distributions and CNNs, on the other hand, capture visual features across different levels of gran-

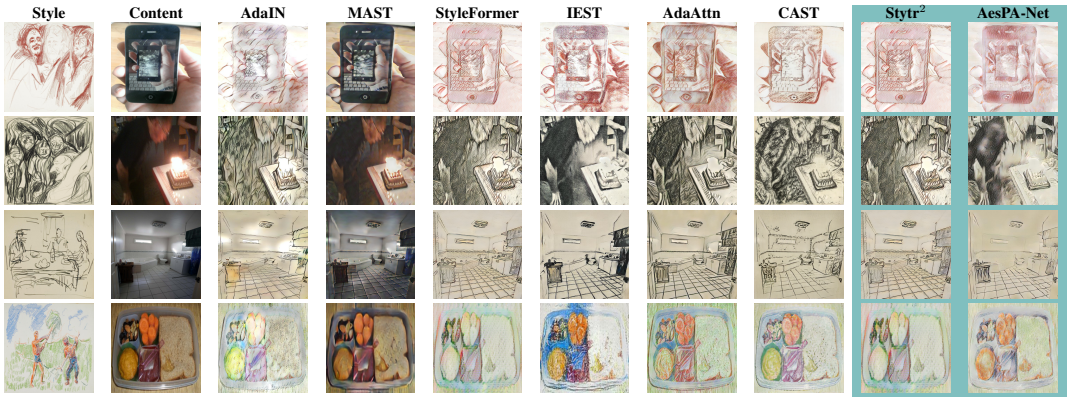

Table 1: Qualitative comparison of different image-to-image style transfer models. Manual inspection of the output from these models suggests that the Stytr^2 and AesPA-NET outperforms the other models. The output from these two models has higher tendency to retain the content from the content image while simultaneously transferring the style from the style image.

Model	Model Configuration					Model Training				Model Inference		Model Performance	
	Open-source	License	Para-meters	Disk space	Compute	Dataset	AI Technology	Reported Time/Steps	Cost	Time	Cost	Style Loss (↓)	Content Loss (↓)
AdaIN	Yes	MIT License	7.01M	92MB	Pascal Titan X	WA (80k), MS (80k)	CNN	-	-	1.28s	$0.30	0.0013	22.59
MAST	Yes	-	17.1 M	152MB	RTX 2080 Ti	WA, MS	CNN	-	-	0.35s	$0.08	0.015	13.54
StyleFormer	Yes	Apache License 2.0	19.9 M	621 MB	Tesla V100	WA (80k), MS (80K)	Transformer	5 days / 800K steps	∼$ 148 (GCP)	0.019s	$0.01	0.0013	17.178
IEST	Yes	MIT License	3.5 M	107MB	A100	WA, MS	GAN	160K steps	-	2.4s	$0.58	0.0048	17.89
AdaAttN	Yes	Apache License 2.0	13.2M	120MB	Tesla P40	WA, MS	CNN, Attention	50K steps	-	0.62s	$0.15	0.0028	38.32
CAST	Yes	Apache License 2.0	10.5 M	40MB	RTX 3090	WA (20k), MS (20k)	GAN	18 hrs / 800K steps	$ 4.90 (Salad)	0.25s	$0.06	0.0014	17.938
Stytr^2	Yes	-	48.2 M	210MB	2 Tesla P100, 2 RTX 3090	WA, MS	Transformer	24 hrs / 160K steps	$ 12.96 (Salad)	1.84s	$0.44	0.00076	10.753
AesPA-Net	Yes	-	24.19 M	50MB	GTX 3090 Ti	WA (80k), MS (120k)	Transformer	-	-	0.34s	$0.08	0.00309	19.994

Table 2: Benchmarking different image-to-image style transfer models (AdaIN [Huang and Belongie2017], Mast [Huo et al.2021], StyleFormer [Wu et al.2021], IEST [Chen et al.2021], AdaAttN [Liu et al.2021], CAST [Zhang et al.2022], Stytr^2 [Deng et al.2022], and AesPA-Net [Hong et al.2023]). WA: WikiArt dataset [Saleh and Elgammal2015], MS: MSCOCO dataset [Lin et al.2015]. Inference time is reported on an A100 GPU. Inference cost is calculated to stylize 1000 images on an A100 GPU on GCP. We identify Stytr^2 as the best available option owing to its performance and the requirements for the problem *P1*.

ularity. The architectures of these technologies enable the capturing of styles in the input data better. For problem P2, the diffusion models and transformers based text-to-image generation models give state-of-the-art performance and are the popular choices for several similar applications. Diffusion models and transformers are highly capable of modelling complex input data distributions in both unconditional and conditional settings. Next, we identify different models that are based on these technologies for both the problems and benchmark them:

Benchmarking image-to-image style transformation models for *P1*:

We study and benchmark multiple state-of-the-art image-to-image style transfer models build with the identified AI technologies for P1 i.e. transformer models, GANs, and CNNs. The pre-trained versions of these models draw style reference from the WikiArt dataset which is a collection of artworks from multiple artists across different genres such as landscape, figurative, self-portrait, etc. To benchmark the model's off-the-shelf style transfer capability, we use the pre-trained model weights and test on the test dataset created with 50 pairs of style and content images from the Edvard Munch's archive and MS-COCO dataset respectively.

The first step is to **qualitatively compare** different models for style and content image pairs. For example, in Table 1, we observe that Stytr^2 and AesPA-Net are the best performing models with respect to the visual quality of the output images. In the second step, we further **quantitatively benchmark** these models on several criteria to get sharper insights and objective comparison (see Table 2). For instance, StyleFormer takes the least inference time and cost whereas IEST is the most expensive model for inference. We also evaluate the model's quantitative performance with: *style loss* and *content loss* [Gatys, Ecker, and Bethge2016b].

Table 3: Qualitative comparison of different text-to-image generation models with respect to their performance on textual prompts. Here, we compare these models based on their capability to reproduce some of the original artworks from Edvard Munch through text instructions. Even though these models fail to perfectly reproduce the original artworks, we observe that all the three variants of stable diffusion model are able to generate the content provided in the text instruction while bringing in different elements of style from Edvard Munch.

Model	Model Configuration					Model Training				Model Inference		Model Performance	
	Open Source	Licence	Parameters	Disk Space	Compute	Dataset	AI Technology	Reported Time/Steps	Cost	Time	Cost	Prompt Fidelity (↑)	Image Similarity (↑)
Imagen	No	-	2B	-	TPU V4	Imagen Dataset, LAION-400M	Diffusion Model	2.5M Steps	-	-	-	-	-
VQGAN-CLIP	Yes	MIT License	227M	934MB	Tesla V100	-	GAN	-	-	239s	$19.25	0.27566	48.85
Stable Diffusion 1.5	Yes	MIT License, CreativeML Open RAIL-M License	860M	4.27GB	256 Tesla A100	LAION-2B	Diffusion Model	595K Steps	-	9.54s	$0.70	0.33301	53.5934
Stable Diffusion 2.1	Yes	MIT License, CreativeML Open RAIL++-M License	865M	5.21GB	Tesla A100	LAION-5B	Diffusion Model	55K Steps	-	16.57s	$1.33	0.32863	53.3886
Stable Diffusion XL	Yes	MIT License, CreativeML Open RAIL++-M License	2.6B	6.94GB	Tesla A100	Internal Dataset	Diffusion Model	800K Steps	-	43.36s	$3.49	0.3503	56.2288
Muse	No	-	3B	-	TPU V4	Imagen Dataset	Transformer	1M Steps, 1 Week	-	-	-	-	-
Rich Text-to-Image	Yes	MIT License, CreativeML Open RAIL++-M License	2.6B	6.94GB	RTX A6000	-	Diffusion Model	-	-	55.667s	$4.47	0.28235	37.85625
DALL-E 3	No	-	-	-	-	-	Diffusion Model	500K Steps	-	-	-	-	-

Table 4: Benchmarking text-to-image generation models (Imagen [Saharia et al.2022], VQGAN-CLIP [Crowson et al.2022], Stable Diffusion 1.5 [Rombach et al.2022b], Stable Diffusion 2.1 [Rombach et al.2022b], Stable Diffusion XL [Podell et al.2023], Muse [Chang et al.2023b], Rich Text-to-Image [Ge et al.2023], and DALL-E 3 [Betker et al.2023]). Inference time is reported on Tesla T4 GPU. Inference cost is calculated to generate 1000 images on a Tesla T4 GPU. We identify SD1.5 as the best available option owing to its competitive performance to the other high performing models and the requirements for the problem $P2$.

	Heavy Shading		Medium Shading		No Shading		Multi Color	
	SL	CL	SL	CL	SL	CL	SL	CL
PT	0.01	30.29	0.0072	28.98	0.006	27.46	0.005	30.17
FT	0.01	30.65	**0.0064**	**28.69**	0.006	27.55	**0.0047**	30.54

Table 5: Style loss (SL) and content loss (CL) of pretrained (PT) and fine-tuned (FT) Stytr^2 models for different style categories. We observe that the SL either decreases or remains same for all the style categories.

Style loss quantifies the differences in the style between style reference image and the stylized image while content loss quantifies the differences in the content between the content image and the stylized image. A lower score is preferred for both the metrics. We observe that Stytr^2 outperforms all the other models on both the metrics while giving a strong competition to AesPA-Net on other benchmarking criteria. The quantitative analysis further reinforces our observation that the image stylization capability of Stytr^2 is superior to the other models including AesPA-Net for our purposes.

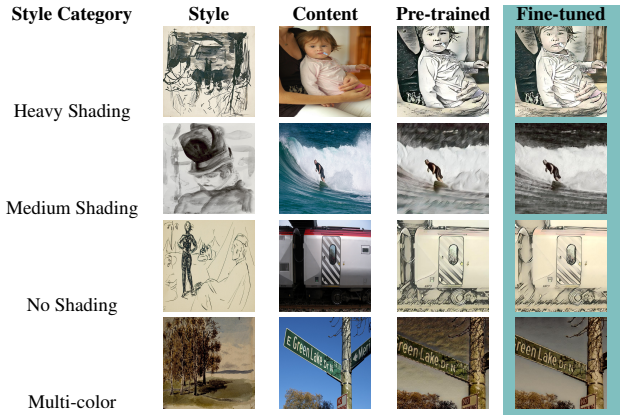

Table 6: Qualitative results of pre-trained and fine-tuned Stytr^2 models on different style categories.

Models	Heavy Shading		Medium Shading		No Shading		Multi Color	
	PF	IS	PF	IS	PF	IS	PF	IS
SD 1.5	0.322	49.185	0.320	47.761	0.319	45.599	0.311	45.919
FMFT	0.302	**51.363**	0.304	48.375	0.295	**49.994**	0.299	**47.288**
LoRA	0.326	45.436	0.325	45.932	0.323	44.130	**0.332**	42.613
DreamLoRA	**0.329**	50.721	**0.331**	**51.119**	**0.327**	46.376	0.331	45.234

Table 7: Prompt fidelity (PF) and Image similarity (IS) for pre-trained model and fine-tuned SD1.5 models (bottom-3 rows) for different style categories.

Benchmarking text-to-image generation models for *P2*: We benchmark multiple state-of-the-art text-to-image generation models build with the identified AI technologies for *P2* i.e. diffusion models and transformers. To benchmark these models, we leverage the pre-trained versions of these open-source models and provide style reference in the prompt by specifying the artist's name as a post-fix (e.g. *a drawing of a man in the style of Edvard Munch*).

We select 50 artworks from the Edvard Munch's archive and generate the caption for these artworks using the BLIP model [Li et al.2022]. We manually refine these captions and correct the errors, if any. We use these refined captions as the prompts to create images with different models. We first **qualitatively compare** different models (refer to Table 3), to evaluate their capability in style based generation through text prompts. We find that all the three versions of the stable diffusion model show good performance in capturing different elements of Edvard Munch's style such as the similarity in drawing human faces and the choice of colors, lines, and texture. Furthermore, we **quantitatively benchmark** these models on different criteria (Table 4 presents a summary). We observe that among all the variants of Stable Diffusion models, SD 1.5 is the most efficient model in terms of inference time and cost. Both VQGAN-CLIP and Rich Text-to-Image are costlier than the three versions of stable diffusion model. To further evaluate the model performance, we leverage two evaluation metrics: prompt fidelity and image similarity. *Prompt fidelity* gives the similarity between the prompt and the generated image while *Image similarity* gives the similarity between the generated image and the ground truth artwork. We use CLIP [Radford et al.2021] embedding and CLIPScore [Hessel et al.2022] to get the similarity scores. We observe that SDXL outperforms all the other open-source models on both the metrics. However, both SD1.5 and SD2.1 achieve competitive performance to that of SDXL with fewer parameters and less inference time and cost. For the next stage, we choose SD1.5 model to develop custom models for P2 due to its relatively inexpensive inference and competitive performance to that of SD2.1 and SDXL.

5. Gen AI technology creation and evaluation: In this stage, we further develop and refine the identified AI technology from the previous stage. We can also build a novel AI technology depending on the requirements such as faster inference, lightweight model, and artifact quality.

For our problems, we leverage the style categorized dataset from the MUNCH Museum's image archive for developing custom models. Here, we create style-specific models by fine-tuning the pre-trained Stytr^2 (for *P1*) and SD1.5 (for *P2*) models for each style category in the dataset.

Image-to-image style transfer for *P1*: To develop the custom style-transfer models for P1, we fine-tune the Stytr^2 model on artworks from different style categories in the dataset. For *style reference*, we sample 50 artworks from each style category, and for *input content images* we sample 11400 content images from MSCOCO training set. We use this style-content paired data as the training set for our experiments. We fine-tune Stytr^2 pre-trained model weights separately on each of the style categories for 55000 iterations. The time taken for fine-tuning each model is about 10 hours on an A100 GPU. To evaluate the performance, we test the models on the style-content pairs from the remaining artworks in each category and the 50 content images used earlier for benchmarking. We compute the style and content losses and report our findings in the Table 5.

We observe that the style loss either decreases or remains same for all the style categories. Conversely, we see a decline in the performance on content loss metric post fine-tuning across all the categories except medium shading. This could possibly indicate that the models are altering the finer details of the content image in an attempt to capture the elements of artistic style from the reference artwork. Qualitative examination of the results (see Table 6) indicates visual difference in the stylized outputs of the pre-trained and fine-tuned models across all the categories. A closer inspection of the output in Table 6 reveals that the fine-tuned models can effectively transfer the fine-grained style attributes such as line, color and texture from the style reference image while simultaneously reducing noise artifacts in the output.

Text-to-image generation for *P2*: To develop the custom style-transfer models for *P2*, we fine-tune SD1.5 on artworks from different style categories in the dataset. We pick 50 artworks from each category and manually refine the captions generated by BLIP [Li et al.2022] for these artworks. We fine-tune SD1.5 based on the following three techniques: (i) Full Model Fine-Tuning (FMFT), (ii) LoRA based Fine-Tuning [Hu et al.2021], and (iii) Dreambooth LoRA Fine-Tuning [Ruiz et al.2023]. We test the performance of these style-specific fine-tuned models by re-creating the remaining artworks in that style category and report the results with

Style Category	Prompt	Original Artwork	SD1.5	SD2.1	SDXL	FMFT	LoRA	DreamLoRA
Heavy Shading	Drawing of a man with a mustache and a suit							
Medium Shading	Drawing of a city with a church and a river in the foreground							
No Shading	Drawing of a man sitting on a cart with a pipe							
Multi-Color	Painting of a couple standing in front of a tree with blue and green leaves							

Table 8: Qualitative comparison of pre-trained and fine-tuned text-to-image generation models. We append the style category and the artist's name to the prompt as post-fix (e.g. "Drawing of a man with a mustache and a suit, *heavy shaded style, in the style of Edvard Munch*") for the pre-trained models to provide the style information [Mun, Sivertsen et al.2023]. SDXL and DreamLoRA outputs are more closer to Edvard Munch's style highlighting their ability in capturing intricate artistic styles better even when there is no explicit mention of artist name/style in the prompt

prompt fidelity and image similarity (see Table 7) metrics.

We observe that DreamLoRA fine-tuned SD1.5 models consistently outperform the pre-trained model for all style categories. It highlights the importance of dreambooth fine-tuning to bind the rare token with an artistic style shown visually through the artwork. The FMFT and LoRA techniques boost the performance of SD1.5 on only one of the two evaluation metrics suggesting the need to study and provide more sophisticated prompts to recreate artworks in a given style. In Table 8, we present a qualitative comparison of the artworks recreated with different models. We observe that the fine-tuning of the model helps to better capture the various elements of the artistic style without explicitly mentioning the style information in the prompt.

6. Deploying and maintaining the application

Although not a part of our core framework, this stage builds on the outputs of the designed solution and is critical for delivering a reliable and efficient experience to the end-users. The developed AI technology is integrated into a software application and the application is further deployed to serve the end-users. During deployment, the application is installed on the target systems (e.g., on a cloud), and the environments are configured according to the application requirements. Post-deployment, the applications are to be regularly monitored and maintained for performance.

The solution designs that were produced as a result of earlier stage problems *P1* and *P2* would be used later for further downstream tasks such as co-drawing experiences with Munch and AI-driven learning of the core styles of the artist Edvard Munch.

Scope and Limitations While the framework has been developed in the context of the visual art domain in this paper, the framework can be easily extended to include other domains that employ style. The hierarchical structure of Style-Frame ensures that the top levels are more resilient to extension, while the lower levels would change with domain specifics. Similarly, the modular structure of Style-Frame ensures that the new changes can be easily incorporated. The present scope of the framework (in this paper) is around style relevancy in generative AI visual art scenarios.

Conclusion

We propose a framework, Style-Frame, that synthesizes and communicates core knowledge concepts in the design of creative applications dealing with *style aspects of an artwork*. Framework is mainly aimed at creative solution architects, and because of its applied nature, we believe, it is more useful than blindly chasing the SOTA technologies or any isolated literature review, benchmarking, or tutorials on the topic. We applied Style-Frame in the experimental work for a future interactive drawing experience for visitors at the MUNCH Museum in Norway. The framework helped us successfully navigate a very complex technology landscape. It enabled us to systematically approach the design, understand the design choices, and make more educated trade-off decisions on technology options.

Our work on style raises an interesting question: Can modern AI understand and replicate the style of an artist? Some of our experiments revealed that an average person with some exposure to the artwork finds it difficult to distinguish between the artist's and AI-created images in terms of attribution. However for art experts, style can be a much deeper question than only the visual style statement - it might involve the cognitive, experiential, creative, and expressive elements of an artist's craft; more research is required in computational creativity to understand all these systemic aspects well. The proposed framework can also be extended in its dimensions as future AI research sheds more light on these aspects.

Ethical Considerations

In the development of the Style-Frame framework for the MUNCH Museum, we prioritised ethical principles, particularly the importance of maintaining authenticity and integrity within the realm of generative AI. This approach is designed to safeguard the authenticity of artworks and ensure the responsible use of technology, thus protecting the artistic legacy of Edvard Munch whilst expanding the knowledge about his work.

Acknowledgement

We are thankful to Birgitte Aga (Head of Innovation and Research) and Nikita Mathias (Senior Concept Developer), MUNCH Museum for their insights on Edvard Munch's artworks and the help in the understanding museum's human-technology interaction landscape, and ethical sensitivities in AI-Art explorations. We also would like to thank Shirish Karande, Sankha Som, Shishir Dahake, Akash Mohan, Anirban Gupta, Tamanna Desai, Anandita, Anuj Sharma, Rajan Maheshwari (Country Head Norway, TCS), Gautam Shroff (Head Research, TCS), and Harrick Vin (CTO, TCS) for their inputs, guidance, and support of the project.

References

[Anica-Popa et al.2021] Anica-Popa, I.; Anica-Popa, L.; Rădulescu, C.; and Vrîncianu, M. 2021. The integration of artificial intelligence in retail: benefits, challenges and a dedicated conceptual framework. *Amfiteatru Economic* 23(56):120–136.

[Bawack et al.2022] Bawack, R. E.; Wamba, S. F.; Carillo, K. D. A.; and Akter, S. 2022. Artificial intelligence in e-commerce: a bibliometric study and literature review. *Electronic markets* 32(1):297–338.

[Betker et al.2023] Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Guo, J. L. Y.; Manassra, W.; Dhariwal, P.; Chu, C.; Jiao, Y.; and Ramesh, A. 2023. Improving image generation with better captions.

[Brown and Jordanous2022] Brown, D. G., and Jordanous, A. K. 2022. Is style reproduction a computational creativity task? In *International Conference on Innovative Computing and Cloud Computing*.

[Chang et al.2023a] Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; Li, Y.; and Krishnan, D. 2023a. Muse: Text-to-image generation via masked generative transformers.

[Chang et al.2023b] Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; Li, Y.; and Krishnan, D. 2023b. Muse: Text-to-image generation via masked generative transformers.

[Chen et al.2021] Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021. Artistic style transfer with internal-external learning and contrastive learning. In *Neural Information Processing Systems*.

[Crowson et al.2022] Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance.

[Deng et al.2022] Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. Stytr^2: Image style transfer with transformers.

[Drori, Cohen-Or, and Yeshurun2003] Drori, I.; Cohen-Or, D.; and Yeshurun, H. 2003. Example-based style synthesis. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, II–143. IEEE.

[Efros and Freeman2023] Efros, A. A., and Freeman, W. T. 2023. Image quilting for texture synthesis and transfer. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 571–576.

[Gatys, Ecker, and Bethge2016a] Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016a. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

[Gatys, Ecker, and Bethge2016b] Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016b. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2414–2423.

[Ge et al.2023] Ge, S.; Park, T.; Zhu, J.-Y.; and Huang, J.-B. 2023. Expressive text-to-image generation with rich text. In *IEEE International Conference on Computer Vision (ICCV)*.

[Goodfellow et al.2014] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks.

[Hessel et al.2022] Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2022. Clipscore: A reference-free evaluation metric for image captioning.

[Hong et al.2023] Hong, K.; Jeon, S.; Lee, J.; Ahn, N.; Kim, K.; Lee, P.; Kim, D.; Uh, Y.; and Byun, H. 2023. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22758–22767.

[Hu et al.2021] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models.

[Huang and Belongie2017] Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization.

[Huang and Rust2021] Huang, M.-H., and Rust, R. T. 2021. A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science* 49:30–50.

[Huo et al.2021] Huo, J.; Jin, S.; Li, W.; Wu, J.; Lai, Y.-K.; Shi, Y.; and Gao, Y. 2021. Manifold alignment for semantically aligned style transfer. In *IEEE International Conference on Computer Vision*, 14861–14869.

[Ishikawa and Yoshioka2019] Ishikawa, F., and Yoshioka, N. 2019. How do engineers perceive difficulties in engineering of machine-learning systems?-questionnaire survey. In *2019 IEEE/ACM Joint 7th International WorkshopCESI and 6th International Workshop SER&IP*, 2–9. IEEE.

[Jing et al.2019] Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; and Song, M. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics* 26(11):3365–3385.

[Johnson, Alahi, and Fei-Fei2016] Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution.

[Kaymakci, Wenninger, and Sauer2021] Kaymakci, C.; Wenninger, S.; and Sauer, A. 2021. A holistic framework for ai systems in industrial applications. In *Innovation*

Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues, 78–93. Springer.

[Knight1994] Knight, T. 1994. *Transformations in Design: A Formal Approach to Stylistic Change and Innovation in the Visual Arts*. Cambridge University Press.

[Kwon and Ye2022] Kwon, G., and Ye, J. C. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18062–18071.

[Li et al.2022] Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

[Lin et al.2015] Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft coco: Common objects in context.

[Liu et al.2021] Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*.

[Liu et al.2023] Liu, Z.-S.; Wang, L.-W.; Siu, W.-C.; and Kalogeiton, V. 2023. Name your style: Text-guided artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3529–3533.

[Mun] Munch museum image archive. `https://foto.munchmuseet.no/fotoweb/`. Accessed: 2024-05-08.

[Oosthuizen et al.2021] Oosthuizen, K.; Botha, E.; Robertson, J.; and Montecchi, M. 2021. Artificial intelligence in retail: The ai-enabled value chain. *Australasian Marketing Journal* 29(3):264–273.

[Osman Andersen et al.2021] Osman Andersen, T.; Nunes, F.; Wilcox, L.; Kaziunas, E.; Matthiesen, S.; and Magrabi, F. 2021. Realizing ai in healthcare: challenges appearing in the wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–5.

[Podell et al.2023] Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis.

[Radford et al.2021] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision.

[Ramesh et al.2021] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation.

[Resales, Achan, and Frey2003] Resales; Achan; and Frey. 2003. Unsupervised image translation. In *Proceedings Ninth IEEE International Conference on Computer Vision*, 472–478. IEEE.

[Richardson et al.2021] Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2287–2296.

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

[Rombach et al.2022b] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models.

[Ruiz et al.2023] Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

[Saharia et al.2022] Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

[Saleh and Elgammal2015] Saleh, B., and Elgammal, A. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature.

[Sivertsen et al.2023] Sivertsen, C.; Haas, R.; Jensen, H. H.; and Løvlie, A. S. 2023. Exploring a digital art collection through drawing interactions with a deep generative model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–5.

[Smith and Eckroth2017] Smith, R. G., and Eckroth, J. 2017. Building ai applications: Yesterday, today, and tomorrow. *Ai Magazine* 38(1):6–22.

[Soenksen et al.2022] Soenksen, L. R.; Ma, Y.; Zeng, C.; Boussioux, L.; Villalobos Carballo, K.; Na, L.; Wiberg, H. M.; Li, M. L.; Fuentes, I.; and Bertsimas, D. 2022. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine* 5(1):149.

[Sohn et al.2024] Sohn, K.; Jiang, L.; Barber, J.; Lee, K.; Ruiz, N.; Krishnan, D.; Chang, H.; Li, Y.; Essa, I.; Rubinstein, M.; et al. 2024. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems* 36.

[Vaswani et al.2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

[Wu et al.2021] Wu, X.; Hu, Z.; Sheng, L.; and Xu, D. 2021. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14618–14627.

[Zhang et al.2013] Zhang, W.; Cao, C.; Chen, S.; Liu, J.; and Tang, X. 2013. Style transfer via image component analysis. *IEEE Transactions on multimedia* 15(7):1594–1601.

[Zhang et al.2022] Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH*.