Draw Your Art Dream: Diverse Digital Art Synthesis with Multimodal Guided Diffusion

Nisha Huang
huangnisha2021@ia.ac.cn
School of Artificial Intelligence, UCAS
NLPR, Institute of Automation, CAS
China

Fan Tang
tfan.108@gmail.com
School of Artificial Intelligence, Jilin University
China

Weiming Dong*
weiming.dong@ia.ac.cn
NLPR, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
China

Changsheng Xu
csxu@nlpr.ia.ac.cn
NLPR, Institute of Automation, CAS
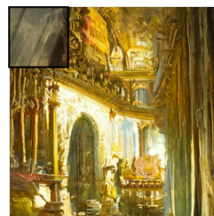School of Artificial Intelligence, UCAS
China

This is a colorful painting of a landscape in the style of cubism drawn by David Burliuk.
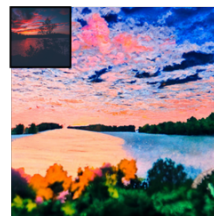
Picasso's painting of a cityscape in cubism style.

A romantic dream-like oil painting by Ivan Aivazovsky of cloudscape.

The beautiful painting depicts the interior of a golden and regal palace drawn by artist Karlsimon.

This is a gorgeous natural landscape painting created by the well-known artist Henri Matisse in pointillism and neo-impressionism styles.

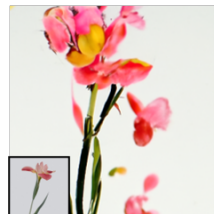A painting of cityscape drawn by Bernardo Bellotto in rococo style.

An oil painting of the autumn on the Seine is drawn by Claude Monet in Impressionism style.

This is a colorful painting of a landscape in the style of cubism drawn by David Burliuk.

The painting shows white waves rolling over a blue sea drawn by Richard Parkes Bonington and John Constable.

This is a Chinese ink painting with flowers.

This fantastic painting depicts the advanced appearance of a future world city 3,000 years from now, full of sense of technology by Theo Prins and Nicolas Ferrand.

An oil painting of a dog in realism style.

Figure 1: Digital art paintings generated by the proposed multimodal guided artwork diffusion (MGAD) model.

ABSTRACT

Digital art synthesis is receiving increasing attention in the multimedia community because of engaging the public with art effectively. Current digital art synthesis methods usually use single-modality inputs as guidance, thereby limiting the expressiveness of the model and the diversity of generated results. To solve this problem, we propose the multimodal guided artwork diffusion (MGAD) model, which is a diffusion-based digital artwork generation approach that utilizes multimodal prompts as guidance to control the classifier-free diffusion model. Additionally, the contrastive language-image pretraining (CLIP) model is used to unify text and image modalities. Extensive experimental results on the quality and quantity of the generated digital art paintings confirm the effectiveness of the combination of the diffusion model and multimodal guidance. Code is available at https://github.com/haha-lisa/MGAD-multimodal-guided-artwork-diffusion.

*Corresponding author.

CCS CONCEPTS

• **Computing methodologies** → **Image manipulation**.

KEYWORDS

diffusion model, digital art, multimodal guidance

1 INTRODUCTION

Many people like to appreciate paintings, but not everyone has the expertise and labor time to create an ideal artwork. Therefore, a tool that can create paintings with high quality and a wide variety from simple inputs would be helpful for novice people to experience art creation easily. In recent years, research on this topic has drawn extensive attention because of its scientific and artistic values.

Existing works use computational algorithms [3, 67, 74] or style transfer approaches [13, 27, 31, 42] for art creation. The task of style transfer involves transferring the style of an image or a collection of images (a pictorial artwork or the artworks of a painter) to another image (usually a photograph) or video clip. Style transfer [11, 14, 69, 76] can generate high-quality art images/videos but does not closely resemble the actual paintings. The resulting content depends on the content of the input photograph, hence limiting the controllability of the creation process and the diversity of the results. Image-to-image translation scheme was also used for artistic image generation [37, 41, 78]. Nevertheless, these methods can only generate results with the style of one or a limited number of artists. Diversified style transfer methods [6, 7, 68] were proposed to generate multiple results from the same input, but all the contents of the results are still the same as the input content image. Recently, some new works used text to control the generation of paintings [20, 30, 57], based on dual language-image encoders, such as CLIP [49]. However, the quality and diversity of the paintings produced by the text guidance still need to be improved urgently. Therefore, we strive to create user-controlled, realistic, high-quality, and diverse artworks.

Meanwhile, outstanding works in image generation have been produced [29, 51, 56, 70, 71, 77]. However, few of them were employed to generate digital paintings. Diffusion models that emerged lately [17, 23, 54] achieve state-of-the-art quality and outstanding diversity, and have the potential to create excellent digital artworks. Restricted by the guidance conditions of existing diffusion models [25, 60], generating digital artworks freely is challenging. For example, ADM-G [17] can only generate natural images of a certain category based on the category label. Therefore, CLIP [49], which enables multimodal prompts as guidance conditions to broaden the application of diffusion models to generate digital artworks, was used. A freshly proposed form of guidance known as classifier-free guidance [26] could produce similar results without the use of a separate classifier. Therefore, we combine the classifier-free diffusion model [26] and CLIP [49] for digital art generation, which has the advantage of high quality and superb diversity.
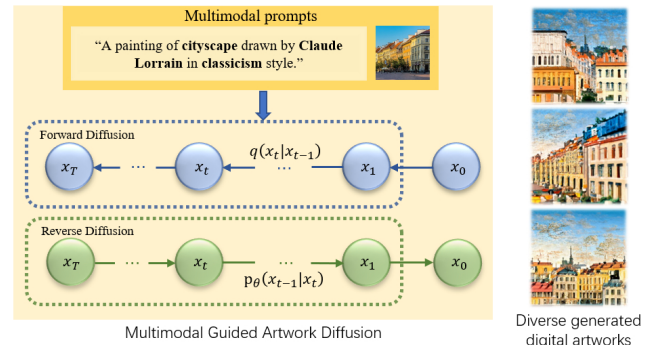

Figure 2: Insight into our work. The diverse generated digital artworks are guided by CLIP loss via the classifier-free diffusion model.

Taking advantage of the capabilities of guided diffusion models to generate images and the abilities of text-to-image or image-to-image models to handle prompts, guided diffusion is applied to the issue of multimodal-conditional digital artwork synthesis. In this paper, we propose a multimodal guided artwork diffusion (**MGAD**) model, which adopts **CLIP** to assist multimodal prompts for generating digital artworks guided by classifier-free diffusion models (Figure 2). We begin by utilizing a fine-tuned unconditional model with 512×512 resolution for the diffusion model from OpenAI's class-conditional ImageNet diffusion model [17]. Then, we employed a secondary diffusion model [10] that was previously trained on the Yahoo Flickr Creative Commons 100 million [1] dataset to produce higher-quality outputs. Findings show that the samples (Figure 1) from our model generated pleasing and artistic results. The main contributions of this work are summarized as follows:

- We propose MGAD, a method used to refine each transition in the generative process by matching each latent variable with the given multimodal guidance.
- We enable the user to control the semantic content of the prompts with respect to the similarity of the results by using CLIP and the classifier-free diffusion model.
- The aggregate experimental results show that MGAD outperforms the baseline approaches and achieves excellent results in terms of diversity and quality of the generated digital artworks.

2 RELATED WORK

Guided Image Synthesis. Previous research has focused on the combination of natural language and image, including tasks, such as text-guided [35, 45] or image-guided synthesis [9, 66]. Consequently, some prominent works for vision and language– representations [8, 16, 40] have been studied in depth to integrate image–text embedding. The pre-trained text–image embedding model CLIP [49] can be efficiently transferred non–trivially to most tasks, generally without any specific dataset training compared with fully supervised baselines. Its representations have

been proven robust and comprehensive enough to perform zero-shot classification and various vision-language tasks on different datasets. The combination of CLIP and GANs, which utilize CLIP to guide the optimization of a latent code for a required image manipulated or generated, has emerged in subsequent studies. For instance, StyleCLIP [48] used CLIP embedding vectors to tune the latent codes. StyleGAN-NADA [21] utilized CLIP frame to adjust the zero-shot domain. Our goal is to recognize these descriptions automatically. CLIPstyler [38] proposed patchCLIP for transferring semantic texture information on text conditions.

GLIDE [47] and DALL-E 2 [50] focus on open domain image synthesis. Both of them are implemented with the idea of integrating image generators and joint text-image encoders into their architectures. They all contain pre-trained models with large-scale datasets of numerous text-image pairs. In contrast, we do not intend to spend such expensive training resources and time. We aim to only use CLIP to measure the similarity between the prompts and the generated results to guide the reverse direction. In addition, the above CLIP-based image editing methods only allowed the users to supply a textual description as the style condition. To improve controllability, we attempt to input multiple modality information as control conditions for art image synthesis.

Diffusion Models. Diffusion models [58], which consist of one forward process (signal to noise) and one reverse process (noise to signal), have been newly demonstrated to produce high-quality images [17, 25, 59, 61]. Denoising diffusion probabilistic models (DDPM) [25] and score-based generative models [61, 62], have recently gained remarkable success in the field of image generation [25, 32, 60, 62]. Ablated diffusion model (ADM) [17] has demonstrated higher image synthesis quality than variational autoencoders (VAEs) [52], flow-based models [36], auto-regressive models [46], and GANs [22, 33, 34]. The generative power of these models [17, 25, 62] stems from a natural adaptation to the inductive biases of image-like data when their underlying neural skeleton is implemented as a U-Net [55]. Recently, diffusion models have been explored for conditional generation, such as class-conditional generation [9], image-guided synthesis [17], text-guided synthesis [23, 35], semantics-guided synthesis [44], and super-resolution [54].

Art Painting Synthesis. Artwork analysis [12, 15, 24] and synthesis [28, 65] are the most challenging tasks that enable effective engagement of the public with art, balancing between sophisticated computational/engineering techniques and artistic purposes. Tan et al. [63, 64] proposed ArtGAN, where the label information was propagated back to the generator for more efficient learning. More recently, CLIPDraw [20] and StyleCLIPDraw [57] began generating paintings from randomized Bézier curves that fit a given text and style. In contrast, these two models mainly capture larger features, such as shapes or outlines, rather than fine-grained textures. Yi et al. [72] explored the generation of fine art painting that used diffusion models without any guidance.

Inspired by the aforementioned works, we propose a novel MGAD model, which concentrates on synthesizing drawings rather than realistic pictures. We explore whether ADM [17] can be guided by text and image prompts to synthesize high-quality and fantasy art paintings rather than only using text prompts, which makes generating results less controllable.
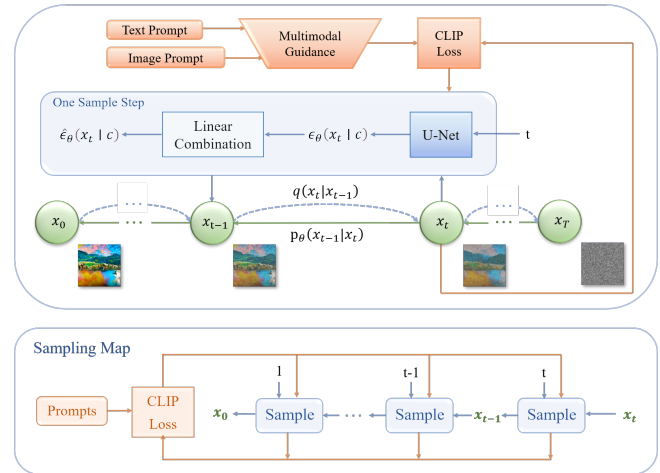

Figure 3: Overall framework of our method. One sample step consists of a U-Net and linear combination. Every sample step is guided by multimodal guidance.

3 METHODS

3.1 Overview

Figure 3 shows the overall framework of the proposed MGAD for art painting synthesis. MGAD is a new unified framework that incorporates different modality guidance into a pre-trained classifier-free diffusion model. Our goal is to generate the required art paintings according to one or both images and text prompts to take advantage of multimodal guidance. The multimodal guidance enables the controllable art painting synthesis and realizes complementarity between modes. First, we utilize a pre-trained diffusion model ϵ_θ to transform the noise image x_0 into latent $x_{t_0}(\theta)$. Second, the target y_{tar} leads the diffusion model to generate samples guided by the CLIP loss at the reverse process. Third, the deterministic forward processes are based on DDPM [25]. For translation among unseen domains, the image generation is also performed by combining multimodal guidance and diffusion models [17, 26].

3.2 Multimodal Guided Diffusion

Diffusion models [58] are inspired by non-equilibrium thermodynamics. They define a Markov chain of diffusion steps to add random noise to data slowly and then learn to reverse the diffusion process to construct the desired data samples from noise. Unlike VAEs [52] or flow-based [36] models, diffusion models are learned using a fixed procedure, and the latent variable has high dimensionality (same as the original data).

A diffusion model includes one forward and one reverse diffusion process. The forward process is fixed to a Markov chain trained using variational inference, which gradually adds noise to the data. The data distribution is defined as $x_0 \sim q(x_0)$ and a Markov chain forward process named as q. In addition, the data generates noised samples from x_1 to x_T. At each step of the forward process, Gaussian noise is added to the data accordingly by a variance schedule

$\beta_1, ..., \beta_T$:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), \tag{1}$$

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right). \tag{2}$$

The training goal is to optimize the negative log likelihood of the usual variational bound:

$$
\begin{aligned}
L &:= \mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right) - \sum_{t \geq 1} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right], \\
&= \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right], \\
&\leq \mathbb{E}\left[-\log p_\theta\left(\mathbf{x}_0\right)\right].
\end{aligned}
\tag{3}
$$

Forward process samples $\mathbf{x}_t$ at time step t:

$$q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}, \left(1-\alpha_t\right), \mathcal{I}\right), \tag{4}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$. The variance of the noise for an arbitrary time step is defined as $1 - \bar{\alpha}_t$ to determine the noise schedule. The form of the diffusion models [58] can be represented by $p_\theta\left(\mathbf{x}_0\right) := \int p_\theta\left(\mathbf{x}_{0:T}\right) d\mathbf{x}_{1:T}$. $p_\theta\left(\mathbf{x}_{0:T}\right)$ is the expression of the reverse process that is defined as a learning Gaussian distribution Markov chain that initiates at $p\left(\mathbf{x}_T\right) = \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$.

If the magnitude $1 - \alpha_t$ of the noise added at each step is small enough, then the posterior $q\left(x_{t-1} \mid x_t\right)$ could be well-approximated by a diagonal Gaussian. Furthermore, if the magnitude $1 - \alpha_1 \ldots \alpha_T$ of the total noise added throughout the chain is large enough, then x_T could be well-approximated by $\mathcal{N}(0, \mathcal{I})$. These properties suggest learning a model $p_\theta\left(x_{t-1} \mid x_t\right)$ to approximate the true posterior:

$$p_\theta\left(x_{t-1} \mid x_t\right) := \mathcal{N}\left(\mu_\theta\left(x_t\right), \Sigma_\theta\left(x_t\right)\right), \tag{5}$$

which can be used to produce samples $x_0 \sim p_\theta\left(x_0\right)$ by starting with Gaussian noise $x_T \sim \mathcal{N}(0, \mathcal{I})$ and gradually reducing the noise in a sequence of steps $x_{T-1}, x_{T-2}, \ldots, x_0$. To compute this surrogate objective, Ho et al. [25] found that predicting ϵ worked best, especially when combined with a reweighted loss function:

$$L_{simple} := E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta\left(x_t, t\right)\|^2\right]. \tag{6}$$

The DDPM model [25] shows how to derive $\mu_\theta\left(x_t\right)$ from $\epsilon_\theta\left(x_t, t\right)$, and fix Σ_θ to a constant. Results from the DDPM show that they can rapidly sample and achieve better log-likelihoods used parameterization and simplified training objective to improve the learning of Σ_θ.

Guided Diffusion. To explicitly incorporate class information into the diffusion process, Dhariwal et al. [17] trained a classifier $f_\phi\left(y \mid \mathbf{x}_t, t\right)$ on noisy image x_t and use gradients $\nabla_{x_t} \log p_\phi\left(y \mid x_t\right)$ to guide the diffusion sampling process toward the target class label y. The new resulting perturbed mean $\hat{\mu}_\theta\left(x_t \mid y\right)$ is given by

$$\hat{\mu}_\theta\left(x_t \mid y\right) = \mu_\theta\left(x_t \mid y\right) + s \cdot \Sigma_\theta\left(x_t \mid y\right) \nabla_{x_t} \log p_\phi\left(y \mid x_t\right), \tag{7}$$

where mean $\mu_\theta\left(x_t \mid y\right)$ and variance $\Sigma_\theta\left(x_t \mid y\right)$ is perturbed additively by the gradient of the log-probability $\log p_\phi\left(y \mid x_t\right)$ of a target class y predicted by a classifier. The ADM and the one with additional classifier guidance (ADM-G) can achieve results that are better than those of state-of-the-art generative models [4].

Algorithm 1 Multimodal guided diffusion sampling, given a diffusion model $\left(\mu_\theta\left(x_t\right), \Sigma_\theta\left(x_t\right)\right)$, guidance function $F\left(x_t, t\right)$ and *CLIP* model

Input: Multimodal guidance c, gradient scale s, diffusion steps T.
Output: generated image $\hat{x}_0$ according to multimodal guidance c.
1: $t = T$;
2: $x_t \longleftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$;
3: **repeat**
4: $\quad t \leftarrow t - 1$;
5: $\quad \mu, \Sigma \leftarrow \mu_\theta\left(x_t\right), \Sigma_\theta\left(x_t\right)$;
6: $\quad \hat{\epsilon}_\theta\left(x_t \mid c\right) \leftarrow (1-s) \cdot \epsilon_\theta\left(x_t \mid \emptyset\right) + s \cdot \epsilon_\theta\left(x_t \mid c\right)$;
7: $\quad x_{t-1} \leftarrow$ sample from $\mathcal{N}\left(\mu + s\Sigma\nabla_{x_t} F\left(x_t, t\right)\right)$;
8: **until** $t < 0$

Classifier-free guidance. The disadvantage of classifier guidance is that it needs an additional classifier model, thereby complicating the training process. Ho and Salimans [26] presented classifier-free guidance, a methodology for guiding diffusion models that do not demand the training of a separate classifier model. Throughout the training, the tag y in a class-conditional diffusion model $\epsilon_\theta\left(x_t \mid y\right)$ is substituted with a null tag $\emptyset$ with a defined likelihood for classifier-free guidance. The output of the model is further extended in the direction of $\epsilon_\theta\left(x_t \mid y\right)$ and away from $\epsilon_\theta\left(x_t \mid \emptyset\right)$ during sampling:

$$\hat{\epsilon}_\theta\left(x_t \mid y\right) = \epsilon_\theta\left(x_t \mid \emptyset\right) + s \cdot \left(\epsilon_\theta\left(x_t \mid y\right) - \epsilon_\theta\left(x_t \mid \emptyset\right)\right). \tag{8}$$

The guidance scale is $s \geq 1$. The latent classifier inspired this equation.

$$p^i\left(y \mid x_t\right) \propto \frac{p\left(x_t \mid y\right)}{p\left(x_t\right)}, \tag{9}$$

where gradient is expressed as a function of the true scores ϵ^*

$$
\begin{aligned}
\nabla_{x_t} \log p^i\left(x_t \mid y\right) &\propto \nabla_{x_t} \log p\left(x_t \mid y\right) - \nabla_{x_t} \log p\left(x_t\right), \\
&\propto \epsilon^*\left(x_t \mid y\right) - \epsilon^*\left(x_t\right).
\end{aligned}
\tag{10}
$$

As shown in Algorithm 1, the amended prediction $\hat{\epsilon}$ is then used to steer us towards the multimodal prompts c:

$$\hat{\epsilon}_\theta\left(x_t \mid c\right) = \epsilon_\theta\left(x_t \mid \emptyset\right) + s \cdot \left(\epsilon_\theta\left(x_t \mid c\right) - \epsilon_\theta\left(x_t \mid \emptyset\right)\right). \tag{11}$$

Overall, classifier-free guidance has two advantages. First, rather than depending on the information of a separate (and possibly smaller) classification model, it enables a single model to exploit its expertise during guiding. Second, when conditioned on information that is difficult to anticipate using a classifier, it simplifies guiding.

3.3 CLIP-based Multimodal Guidance

CLIP [49] was presented to acquire visual concepts with natural language supervision and can provide the similarity scores between texts and images. Several works have used CLIP to steer generative models, such as GANs [21, 38, 48], toward user-defined text prompts. In this paper, we leverage a pre-trained CLIP model for text-driven and image-driven art paintings synthesis. Both prompts are formulated as the cosine similarity of the created paintings' features to control the semantic content of the generated digital art paintings.

Text prompt l and image prompt x are embedded into the joint embedding space. The image encoder E_I is time-dependent and
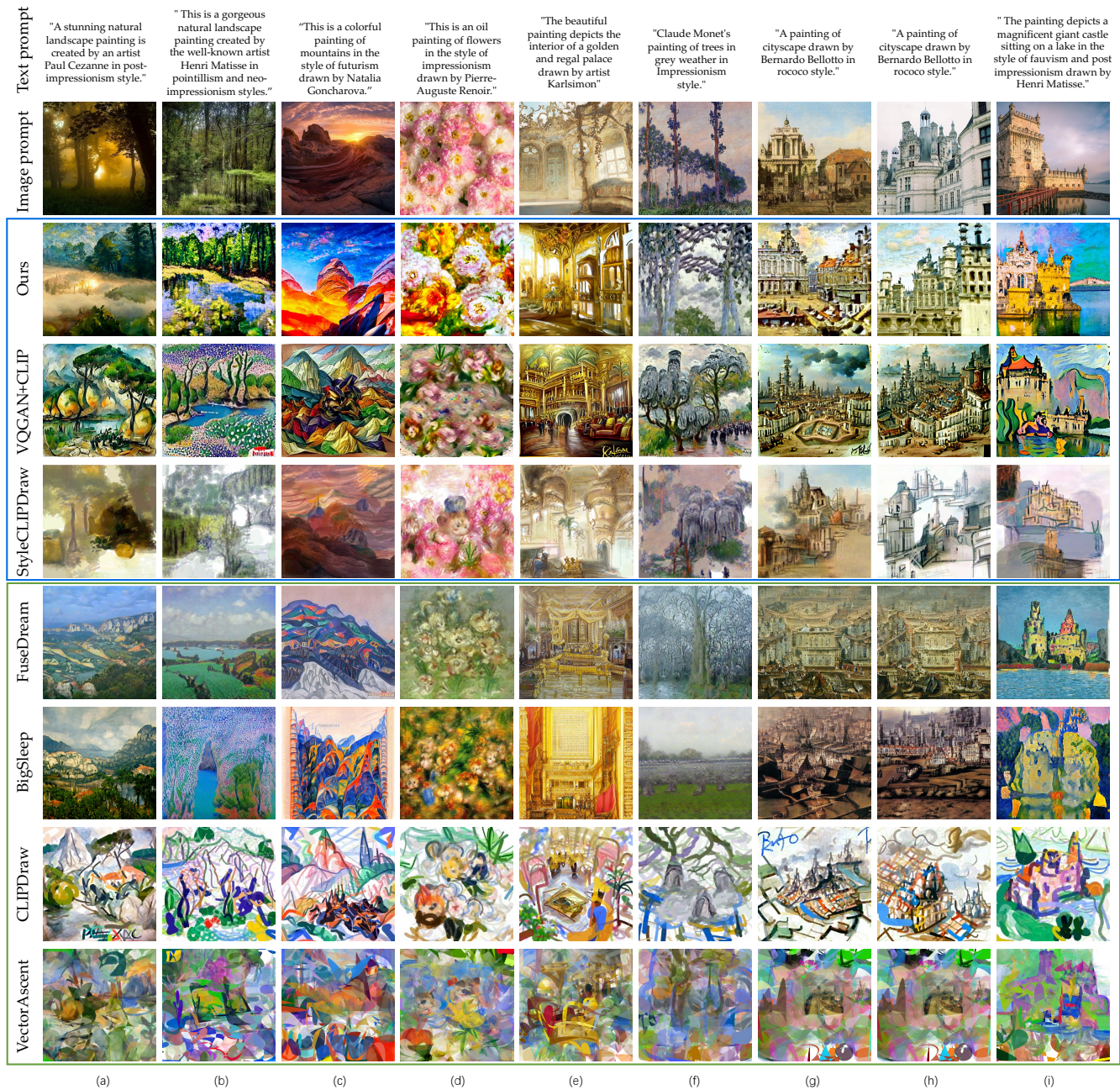
Figure 4: Comparisons of digital art synthesis results using different methods. Methods framed in blue are guided by text and image prompts while methods framed in green are guided by text prompts only.

trained on noisy pictures. E'_I is the label given to a time-dependent image encoder for noisy pictures. The text guidance function can be defined as:

$$F\left(x_t, l, t\right) = E'_I\left(x_t, t\right) \cdot E_L(l). \tag{12}$$

To define the image guiding function, we employ an image encoder finetuned using denoised painting synthesis, similar to how text

guidance is generated. The guidance signal at time step t is:

$$F\left(x_t, x'_t, t\right) = E'_I\left(x_t, t\right) \cdot E'_I\left(x'_t, t\right). \tag{13}$$

The ability to unify image and text guidance simultaneously increases user control flexibility and controllability. Both can be readily included in our pipeline (Fig. 3).

$$F\left(x_t, t\right) = w_1 F\left(x_t, l, t\right) + w_2 F\left(x_t, x'_t, t\right). \tag{14}$$

A stunning natural landscape painting is created by an artist Paul Cezanne in post-impressionism style

A cubism cityscape painting by Picasso.

This cityscape is painted by Henri Matisse in a neo-impressionism style.

A fauvism cityscape painting by Henri Matisse.

A baroque cityscape painting by Adriaen van de Veide.

A painting of cityscape drawn by Albrecht Durer in northern renaissance style.

Figure 5: The results are obtained by fixing one mode while the other mode is changed. The first line of the figure fixes the image prompt, and the text prompts are different. The second line fixes the text prompt, and the image prompts are distinct.

Users may modify the balance between the two by adjusting each modality's weighting and scale factors.

We replace the classifier with a CLIP model. As illustrated in Algorithm 1, the gradient of the dot product of the multimodal prompts and generated image encoding affect the reverse-process mean:

$$\hat{\mu}_\theta\left(x_t \mid c\right) = \mu_\theta\left(x_t \mid c\right) + s \cdot \Sigma_\theta\left(x_t \mid c\right) \nabla_{x_t} F\left(x_t, t\right). \quad (15)$$

For MGAD, we employ noised CLIP model, which has been specifically trained to be noise-aware.

4 EXPERIMENTS

4.1 Implementation Details

For the diffusion model, we used an unconditional model of resolution 512×512 fine-tuned from OpenAI's class-conditional ImageNet diffusion model [17]. Additionally, we use a secondary diffusion model [10] pre-trained on Yahoo Flickr Creative Commons 100 Million (YFCC100m) [1] dataset to achieve a better performance. For the CLIP model, we used ViT-B/32 released by OpenAI for the Vision Transformer [18]. The output size of MGAD is 512×512. For sampling, we set w_1, w_2 and s to 1.0, 1.0 and 5, 000, respectively. We utilize the U-Net [55] architecture based on Wide-ResNet [73]. The model has seven downsampling and seven upsampling layers. The 4×4 feature is generated from the 512×512 input image via one input convolution and fine Resblocks. From the 32×32 to 4×4 resolution, self-attention blocks are added to the Resblocks. The clip model for generating is ViT-B/32. To ensure the quality of the results and to maintain the consistency of the parameters, the diffusion step and the time step used for the experiments in this work are set to 2, 000. Remarkable results are generated when the time step is set to 500 or more. One upsampling step takes approximately 63 milliseconds on a single A40 GPU. For more hyperparameter settings, please refer to the Supplementary Material.

4.2 Qualitative Evaluation

Baselines. To demonstrate the performance of our method, we compare MGAD with SOTA digital painting synthesis works, including VQGAN-CLIP [53], FuseDream [43], BigSleep [2], Style-CLIPDraw [57], CLIPDraw [20], and VectorAscent [30]. Figure 4 shows the digital art generation results. VQGAN-CLIP [19] and StyleCLIPDraw [57] use two modal prompts. The others are only able to use text prompts. VQGAN-CLIP [19], FuseDream [43], and BIGGAN-CLIP [2] utilize CLIP-loss between the prompts and the generated results to realize latent code optimization. Still, problems are encountered in generating paintings with similar brush strokes and structures. Their results have similar repeated textures in different image locations. VQGAN-CLIP [19] can make use of the two modal prompts. However, it may not mimic the painter's strokes well (e.g., the 4th row in Figures 4a and 4d-4i) during the generation process, and the generated results do not resemble the image prompt very much. FuseDream and BigSleep use pre-trained BigGANs [4] and CLIP to achieve high-resolution text-to-image generation. The weights of the generator are frozen; only the latent Z vectors are optimized, so they often generate repetitive patterns in the results (e.g., the 6th and 7th rows in Figures 4g-4h). CLIPDraw [20] and VectorAscent [30] rely on CLIP because of its aligned text and image encoders and diffvg[39], which is a differentiable vector graphics rasterizer used to generate raster images from vector paths. By using gradient ascent, they can optimize for a vector graphic whose rasterization has high similarity with a user-provided caption, backpropagating through CLIP [49] and diffvg [39] to the vector graphic parameters. Their results have resemblances to the text prompt. However, this type of method can only produce discontinuous outlines that resemble the touch of a watercolor brush or blocks of color, unlike human paintings (e.g., the 5th, 8th and 9th rows in Figure 4). The StyleCLIPDraw [57] adds a stylized module to the CLIPDraw [20], thereby making the

Table 1: Quantitative comparisons for digital art generation. We compute the average learned perceptual image patch similarity (LPIPS) and difference Hash (dHash) values of results to measure diversity. The best results are presented in bold.

	Ours	Vector Ascent	FuseDream	BigSleep	CLIPDraw	StyleCLIPDraw	VQGAN+CLIP
LPIPS ↓	**0.458**	0.614	0.58	0.544	0.735	0.489	0.618
dHash ↑	**0.713**	0.551	0.616	0.502	0.534	0.519	0.516

Table 2: User study results. Each number represents the percentage of votes that the results generated by the other models received when selected against our results.

	VectorAscent	FuseDream	BigSleep	CLIPDraw	StyleCLIPDraw	VQGAN+CLIP
Preference rate	10.9%	29.3%	27.5%	9.5%	29.4%	26.3%

stroke styles more diverse, but it still falls short in approaching the realism of the content described by the text prompts (e.g., the 5th and 8th rows in Figure 4).

Compared with baselines, the proposed MGAD can combine the semantic content of visual and language modalities to generate the requested digital artworks. CLIP [49] provides a good understanding of the different modal prompts at the level of semantic content. Diffusion models [17] can balance high quality and diversity in the generation of images. Therefore, using the MAGD, generating complete and sophisticated artworks with skills that meet the user's requirements for the content and the texture and style that closely resembles the paintings created by the mentioned artist is possible.

Strong Understanding for Wide Prompts. By observing the content of the prompts in Figure 1, we find that MGAD can meet the prompts that present a variety of requirements. There is a good understanding of the painter, painting style, color, and texture, and can show their differences. Image prompt can generate good results regardless of whether it is a real image or a painting. MGAD's ability to blend the two modalities and find commonalities between them is outstanding.

4.3 Quantitative Evaluation

To demonstrate the advantages of our method in terms of the quality and diversity of digital art synthesis, we compare it with the existing methods, namely, VectorAscent [30], FuseDream [43], BigSleep [2], CLIPDraw [20], StyleCLIPDraw [57] and VQGAN-CLIP [53] using LPIPS [75], and dHash [5]. User study is also conducted.

LPIPS for diversity evaluation. LPIPS [75] measures the perceptual similarity between two images. We calculate the LPIPS score between paired images generated from the same prompts for guidance (Table 1). Intuitively, the lower the LPIPS score is, the lesser the similarity of the generated images and the higher the diversity of the generated paintings will be. The results show that our model outperforms baseline models by showing the best LPIPS score and surpassing others by a margin. This finding indicates that the LPIPS score results have almost the same tendencies as the score of the user study.

dHash. dHash [5] is a hash mapping method based on pixel point level to measure the similarity between images. As an implementation, dHash is nearly identical to aHash, but it performs much better. A value of 0 indicates the same hash and likely a similar picture. The comparison between the different model dHash values is shown in the bottom row of Table 1. Our model scores significantly superior to other models on this evaluation metric. This result demonstrates that MGAD is notable in terms of diversity.

User Study. We conduct a user study to further compare our method. The user study was divided into two groups. The first group compares the models using text as guidance, and the second group uses text and image as guidance. VectorAscent [30], FuseDream [43], BigSleep [2], and CLIPDraw [20] are in the first group, and the other models are in the second group. We invited 82 users to participate in the first study and invited 89 users to participate in the second study to evaluate the results of different approaches.

Given the guidance, for each set, we show the result generated by our approach and the output from another randomly selected method for comparison and ask the user to select which digital artwork has better effects. Participants in the first group wrote 27 questions and collected 2, 214 votes. Participants in the second group reported 15 questions and gathered 1, 335 votes. We calculated the percentage of votes where the existing methods are superior to ours and show the statistical results in Table 2. The results show that most people prefer our approach compared with other methods.

4.4 Ablation Study

The Influence of Multimodal Guidance Scale. The MGS decides how strongly the result should match the multimodal prompts. We compared the generated results for different MGS values to verify the impact of multimodal guidance. As shown in Figure 6, the generated results are more creative and uncertain when using a lower MGS. Meanwhile, using a higher MGS brings the generated results closer to the semantic content of the prompts and the user requirements. With the appropriate scale of multimodal guidance, the results can be made to fit the prompt content while exploiting the creativity of the model. Users can adjust the scaling factor to control how diverse they expect the generated images to be.
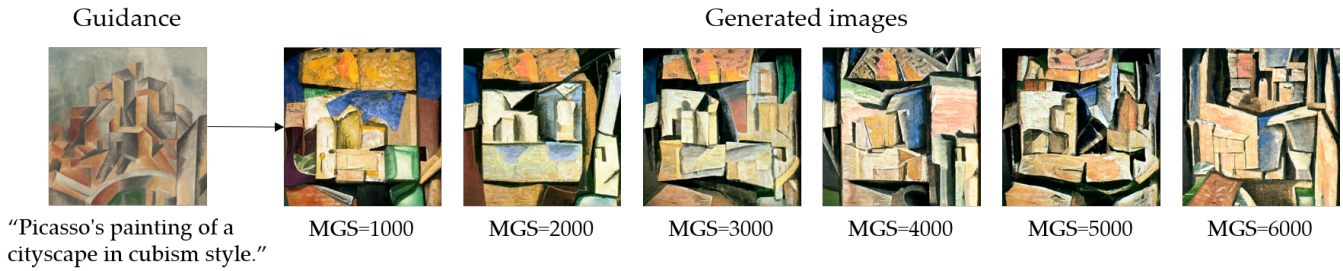
Nisha Huang, Fan Tang, Weiming Dong, & Changsheng Xu

Guidance

Generated images

"Picasso's painting of a
cityscape in cubism style."

MGS=1000   MGS=2000   MGS=3000   MGS=4000   MGS=5000   MGS=6000

Figure 6: Effect of multimodal guidance scale (MGS) on the generated results.

Guidance

Generated images

"The beautiful painting depicts the
interior of a golden and regal palace
drawn by artist Karlsimon"

steps=500   steps=1000   steps=1500   steps=2000   steps=2500   steps=3000
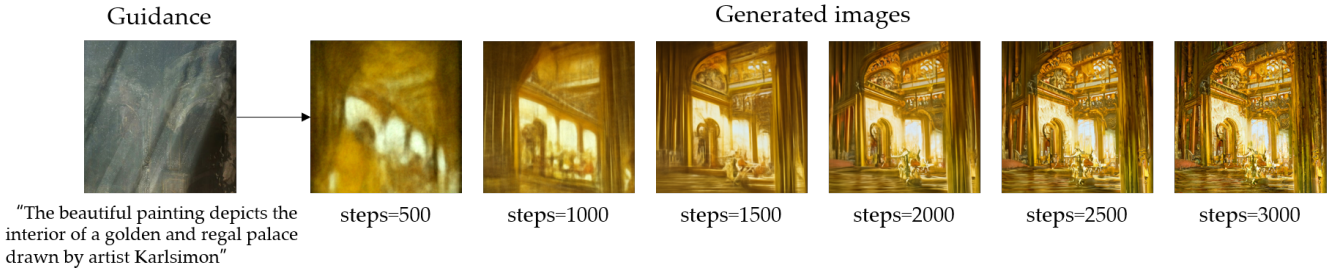
Figure 7: Results of different diffusion steps.

Picasso's painting
of a cityscape in
cubism style.

2 : 0   1.5 : 0.5   1 : 1   0.5 : 1.5   0 : 2

Text prompt ←————————————————————————————————→ Image prompt
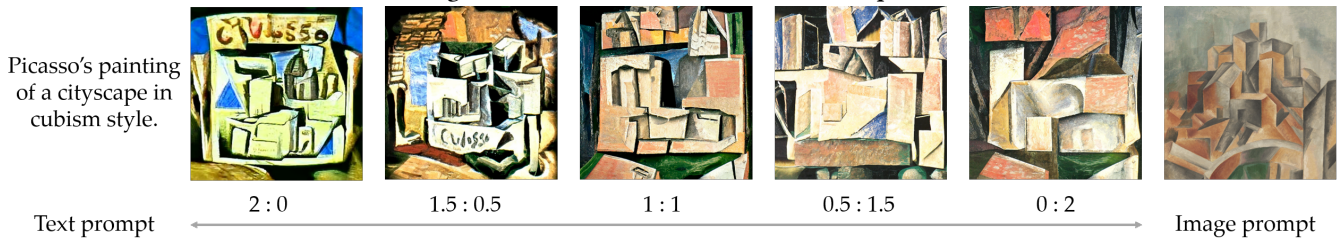
Figure 8: Visualizations of the trade-off between a text prompt and an image prompt.

Impact of Diffusion Steps. To investigate the effect of the number of diffusion steps on the quality and generation speed of the generated images, we tested the generation at different diffusion steps. The experiments show that the sampling process is faster when the number of diffusion steps is small, although the generated images are unclear. After the diffusion steps gradually increase, the image quality improves, and the sampling time increases. However, as shown in Fig. 7, after a certain number of diffusion steps, the image quality no longer changes significantly. In the trade-off between generation time and generation quality, we usually set the number of diffusion steps to 2000.

Balance between Text and Image Prompts. As shown in Figure 8, we can adjust the degree of guidance of both for the final generated results by setting the values of the weight w_1 for the image prompt and the weight w_2 for the text prompt. During our experiments, we found that if the difference between image prompt and text prompt in semantic content is too large, then it will lead to the generation of unsatisfactory results. For this reason, we propose the adaptive prompt discarding module, which calculates the similarity between image prompt and text prompt and then compares the weights of the two to decide whether one of the modalities should be discarded to obtain a better generation result.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose the MGAD model, which is a digital artwork generation model that combines multimodal prompt guidance with classifier-free diffusion model guidance. To enable MGAD to control the generation of digital artwork with the semantic content of multimodal prompts, we propose multimodal guidance loss to constrain the diversity of models and the content similarity with prompts. Adequate experiments show that our approach can achieve a balance between the quality and diversity of digital artwork generation again. In future work, we will focus more on the disentangling and fusion of multimodal prompts and involving art stroke/model-based techniques to make the model more general, creative, and user friendly.

ACKNOWLEDGMENTS

REFERENCES

[1] 2022. *Yahoo Flickr Creative Commons 100 Million (YFCC100m) dataset.* http://projects.dfki.uni-kl.de/yfcc100m/

[2] Adverb. 2022. *The BigSleep: BigGAN+CLIP.* https://colab.research.google.com/drive/1NCceX2mbiKOSlAd_o7IU7nA9UskKN5WR?usp=sharing#scrollTo=WtlDVVMvzMUd

[3] Luis Alvarez, Nelson Monzón, and Jean-Michel Morel. 2021. Interactive Design of Random Aesthetic Abstract Textures by Composition Principles. *Leonardo* 54, 2 (2021), 179–184.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

[5] Johannes Buchner. 2021. *ImageHash:An image hashing library written in Python.* https://github.com/JohannesBuchner/imagehash

[6] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. DualAST: Dual Style-Learning Networks for Artistic Style Transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 872–881.

[7] Haibo Chen, Lei Zhao, Huiming Zhang, Zhizhong Wang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Diverse Image Style Transfer via Invertible Cross-Space Mapping. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 14860–14869.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*. 104–120.

[9] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 14347–14356.

[10] Katherine Crowson and Chainbreakers AI. 2022. *Diffusion 512x512, secondary model method.* https://github.com/crowsonkb/v-diffusion-pytorch

[11] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1210–1217.

[12] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Feiyue Huang, Oliver Deussen, and Changsheng Xu. 2020. Exploring the representativity of art paintings. *IEEE Transactions on Multimedia* 23 (2020), 2794–2805.

[13] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. StyTr^2: Image Style Transfer with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2719–2727.

[15] Yingying Deng, Fan Tang, Weiming Dong, Fuzhang Wu, Oliver Deussen, and Changsheng Xu. 2019. Selective clustering for representative paintings selection. *Multimedia Tools and Applications* 78, 14 (2019), 19305–19323.

[16] Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11162–11173.

[17] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

[19] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12873–12883.

[20] Kevin Frans, LB Soros, and Olaf Witkowski. 2021. CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders. *arXiv preprint arXiv:2106.14843* (2021).

[21] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. arXiv:2108.00946 [cs.CV]

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Neural Information Processing Systems (NIPS)*.

[23] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2021. Vector Quantized Diffusion Model for Text-to-Image Synthesis. *arXiv preprint arXiv:2111.14822* (2021).

[24] Peter Hall, Hongping Cai, Qi Wu, and Tadeo Corradi. 2015. Cross-depiction problem: Recognition and synthesis of photographs and artwork. *Computational Visual Media* 1, 2 (2015), 91–103.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *arXiv: Learning* (2020).

[26] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[27] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1501–1510.

[28] Zhengyu Huang, Yichen Peng, Tomohiro Hibino, Chunqi Zhao, Haoran Xie, Tsukasa Fukusato, and Kazunori Miyata. 2022. dualface: Two-stage drawing guidance for freehand portrait sketching. *Computational Visual Media* 8, 1 (2022), 63–77.

[29] Yuchi Huo and Sung-eui Yoon. 2021. A survey on deep learning-based Monte Carlo denoising. *Computational Visual Media* 7, 2 (2021), 169–185.

[30] Ajay Jain. 2021. *VectorAscent: Generate vector graphics from a textual description.* https://github.com/ajayjain/VectorAscent

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*. 694–711.

[32] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Ioannis Mitliagkas, and Remi Tachet des Combes. 2021. Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations (ICLR)*.

[33] Tero Karras, Samuli Laine, and Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv: Neural and Evolutionary Computing* (2018).

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8107–8116.

[35] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2021. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. (2021). https://doi.org/10.48550/ARXIV.2110.02711

[36] Diederik P. Kingma and Prafulla Dhariwal. 2018. Glow: Generative Flow with Invertible 1x1 Convolutions. *arXiv: Machine Learning* (2018).

[37] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. 2019. Content and Style Disentanglement for Artistic Style Transfer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 4422–4431.

[38] Gihyun Kwon and Jong Chul Ye. 2022. CLIPstyler: Image Style Transfer with a Single Text Condition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. 2020. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics* 39, 6 (2020), 1–15.

[40] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*. 121–137.

[41] Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Changsheng Xu, and Chongyang Ma. 2021. Distribution Aligned Multimodal and Multi-Domain Image Stylization. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3, Article 96 (2021), 17 pages.

[42] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 6649–6658.

[43] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. 2021. FuseDream: Training-Free Text-to-Image Generation with Improved CLIP+GAN Space Optimization. arXiv:2112.01573 [cs.CV]

[44] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. 2021. More Control for Free! Image Synthesis with Semantic Diffusion Guidance. arXiv:2112.05744 [cs.CV]

[45] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. 2020. Describe What to Change: A Text-Guided Unsupervised Image-to-Image Translation Approach. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, 1357–1365.

[46] Jacob Menick and Nal Kalchbrenner. 2018. Generating High Fidelity Images with Subscale Pixel Networks and Multidimensional Upscaling. *international conference on learning representations (ICLR)* (2018).

[47] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[48] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[51] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*. 8821–8831.

[52] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*.

[53] Nerdy Rodent. 2022. *Source Code of VQGAN-CLIP*. https://github.com/nerdyrodent/VQGAN-CLIP

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*.

[56] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. 2021. DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 13960–13969.

[57] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. 2022. StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Translation. *arXiv preprint arXiv:2202.12362* (2022).

[58] Jascha Sohl-Dickstein, Eric L. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *arXiv: Learning* (2015).

[59] Jascha Sohl-Dickstein, Eric L. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *arXiv: Learning* (2015).

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*.

[61] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019).

[62] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*.

[63] Wei Ren Tan, Chee Seng Chan, Hernán Aguirre, and Kiyoshi Tanaka. 2017. ArtGAN: Artwork Synthesis with Conditional Categorial GANs.

[64] Wei Ren Tan, Chee Seng Chan, Hernán Aguirre, and Kiyoshi Tanaka. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing* 28 (2019), 394–409.

[65] Fan Tang, Weiming Dong, Yiping Meng, Xing Mei, Feiyue Huang, Xiaopeng Zhang, and Oliver Deussen. 2018. Animated Construction of Chinese Brush Paintings. *IEEE Transactions on Visualization and Computer Graphics* 24, 12 (2018), 3019–3031.

[66] Hao Wang, Guosheng Lin, Steven C. H. Hoi, and Chunyan Miao. 2021. Cycle-Consistent Inverse GAN for Text-to-Image Synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 630–638.

[67] Miaoyi Wang, Bin Wang, Yun Fei, Kanglai Qian, Wenping Wang, Jiating Chen, and Jun-Hai Yong. 2014. Towards Photo Watercolorization with Artistic Verisimilitude. *IEEE Transactions on Visualization and Computer Graphics* 20, 10 (2014), 1451–1460.

[68] Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, and Dongming Lu. 2020. Diversified Arbitrary Style Transfer via Deep Feature Perturbation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7786–7795.

[69] Hua-Peng Wei, Ying-Ying Deng, Fan Tang, Xing-Jia Pan, and Wei-Ming Dong. 2022. A Comparative Study of CNN- and Transformer-Based Visual Style Transfer. *Journal of Computer Science and Technology* 37, 3 (2022), 601–614.

[70] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1316–1324.

[71] Yuan Xue, Yuan-Chen Guo, Han Zhang, Tao Xu, Song-Hai Zhang, and Xiaolei Huang. 2022. Deep image synthesis from intuitive user input: A review and perspectives. *Computational Visual Media* 8, 1 (2022), 3–31.

[72] Da Yi, Chao Guo, and Tianxiang Bai. 2021. Exploring Painting Synthesis with Diffusion Models. In *IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI)*. 332–335.

[73] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

[74] Kang Zhang and Jinhui Yu. 2016. Generation of Kandinsky Art. *Leonardo* 49, 1 (2016), 48–54.

[75] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.

[76] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings)*.

[77] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. 2018. Multi-View Image Generation from a Single-View. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea). Association for Computing Machinery, New York, NY, USA, 383–391.

[78] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2223–2232.