Language-Guided Artistic Style Transfer Using the Latent Space of DALL-E

Anonymous authors
Paper under double-blind review

Abstract

Despite the progress made in the style transfer task, most of the previous work focuses on transferring only relatively simple features like color or texture, while missing other more abstract and creative concepts such as the specific artistic trait of the painter or the overall feeling of the scene. However, these more abstract concepts can be captured by the semantics of the latent space of models like DALL-E or CLIP, which have been trained using huge datasets of images and textual documents. In this paper, we propose a style transfer method that exploits both of these models and uses the natural language to describe abstract artistic styles. Specifically, we formulate the language-guided style transfer task as a non-autoregressive token sequence translation in the DALL-E discrete latent space. Moreover, we propose a textual-prompt-based Reinforcement Learning strategy to incorporate style-specific information in the translation network using the CLIP space as the only guidance. Our empirical results show that we can transfer artistic styles using language instructions at different granularities on content images that are not restricted to a specific domain. Our code will be publicly available.

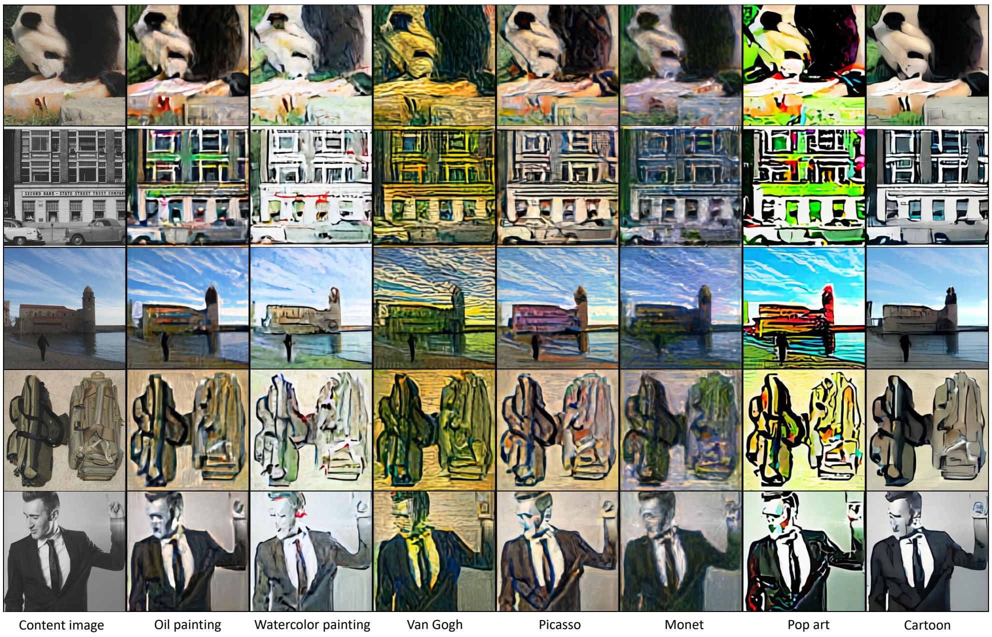

Figure 1: Using language as the guidance, `StylerDALL-E` can transfer different artistic styles, including abstract concepts such as the specific artistic trait of the painter.

1 INTRODUCTION

In the last few years, a lot of work has focused on the style transfer task using a *reference image* as a representative of the target style, where the goal is to transfer the style of the reference to a content image (Gatys et al., 2016; Huang & Belongie, 2017; Park & Lee, 2019; Chen et al., 2021b; Lin et al., 2021; Deng et al., 2022a;b; Wang et al., 2022). Recent improvements in this field include: reducing the artifacts (Chen et al., 2021a; Liu et al., 2021b; Cheng et al., 2021; Wang et al., 2022), modeling the style-content relationship (Park & Lee, 2019; Yao et al., 2019a; Liu et al., 2021a), increasing the generation diversity (Li et al., 2017; Ulyanov et al., 2017; Zhang et al., 2019) and many others. However, artistic styles are usually abstract concepts, such as, e.g., oil painting, fauvism or the Van Gogh style. To transfer these abstract artistic styles to a content image, the low-level features (e.g., textures and colors), which are commonly extracted from a single reference image, are not enough. A possible solution is to collect a set of reference images, which can be used, e.g., to train an artist-specific Generative Adversarial Network (GAN) style transfer (Sanakoyeu et al., 2018; Zhu et al., 2017; Kotovenko et al., 2019). The disadvantage of this *set-based representation* is the effort required to collect sufficiently large style-specific data for training. In this paper we follow a different direction, and we model the target style *using natural language*, which is a very powerful tool to describe abstract concepts, especially if used jointly with the semantics embedded in visual-language models pre-trained using huge datasets. Specifically, we use the DALL-E (Ramesh et al., 2021) discrete latent space, which is given by a vocabulary of visual tokens, extracted using a dVAE (Razavi et al., 2019). However, manipulating a discrete token space is drastically different from most of the previous work in style transfer, which operates on the pixel space or in a continuous latent space, and poses some problems to be solved. The first problem is that we cannot use typical style transfer architectures such as, for instance, the U-net (Ronneberger et al., 2015), which have been proved to be very effective to locally modify the pixel appearance while preserving the original content. The second problem is how to train the network without using additional supervision.

As for the first problem, we propose a solution which is inspired by natural language translation, where a (discrete) sequence of tokens in a source language is translated into a (discrete) sequence of tokens in the target language. More specifically, we use a Non-Autoregressive Transformer (NAT) (Gu et al., 2018) network which learns to translate a token-based representation of a low-resolution image into a full-resolution representation in which the final token sequence contains low-level appearance details specialized for the target style. Additionally, the use of a non-autoregressive paradigm offers important advantages with respect to the more common autoregressive (AR) transformer-based generation used, for instance, in DALL-E. In fact, AR models are known to be very computationally expensive because the inference-time sampling process is conditioned on the previously generated tokens. Moreover, they suffer from the exposure bias problem (Ranzato et al., 2016; Schmidt, 2019; Rennie et al., 2017; Bengio et al., 2015), which is due to the inference-time accumulation of errors in the AR sampling mechanism, since the latter differs from the ground-truth conditioned generation used during training. Both problems are largely alleviated in a NAT-like paradigm (Gu et al., 2018) which allows an inference-time *parallel* token generation.

The second problem can be illustrated using the analogy with natural language translation, where people usually use paired training samples of textual documents in the source (e.g., English) and the target (e.g., French) language as supervision. However, this would require collecting target-style images for each specific style, thus being labor extensive. Conversely, we use as (weak) supervision the similarity between the generated image and the textual style description in the CLIP space (Radford et al., 2021), which has been widely recognized to effectively represent a joint vision-language semantic space. A similar solution has been very recently proposed in CLIPStyler (Kwon & Ye, 2022). However, maximizing the CLIP similarity with respect to a textual style description is not enough for a style transfer task, which should also *preserve* the content of the source image. To do so, Kwon & Ye (2022) introduce an additional content loss, measured in an external pre-trained VGG network. Conversely, we propose an alternative direction, which avoids the need to tune the relative contributions (weights) of different loss functions. Specifically, we introduce a two-stage training paradigm: in the first (which is style independent), the network learns to add semantically coherent image details to a low-resolution image using an upscaling pretext task. In the second, the network is fine-tuned with a specific style. Since the fine-tuning phase is built on top of the first stage, the translator is able to keep the semantic consistency with respect to the input image as it learned during pre-training. Moreover, we create a textual prompt by concatenating both the style and the textual description of the image content (i.e., its caption). This way, the prompt simultaneously models both the target appearance (i.e., the abstract style description) and the image content which should be preserved by the translation process. Finally, despite the continuous CLIP space, since our translator's output is discrete, we cannot directly backpropagate the CLIP similarity values through our network. To solve this problem, we introduce a Reinforcement Learning (RL) approach to fine-tune the translator using a reward based on the CLIP similarity between the stylized image and textual prompt.

We call our network `StylerDALL-E` and we empirically show that it can generate stylized images driven by different types of language guidance. Compared with previous language-guided and reference image-based transfer methods, our generated images are less inclined to produce artifacts or semantic errors (e.g. images with many suns or unrealistic content, Sec. 5.2). Moreover, they can capture abstract concepts related to the target style (e.g., the typical brushstrokes of the artist) besides low-level features like texture and colors. Finally, by leveraging the pre-trained CLIP and DALL-E spaces, `StylerDALL-E` is largely domain-independent, and it can be used with basically any type of content image (e.g., animals, indoor/outdoor images, etc.).

In summary, our main contributions are:

- We propose a language-guided style transfer method that manipulates the *discrete* DALL-E latent space using a token sequence translation approach.
- We propose a non-autoregressive translation network which translates a low-resolution content image into a full-resolution image with style-specific details.
- We propose a two-stage training procedure and an RL strategy based on prompting to learn a style-specific mapping function using the CLIP space.
- We show that `StylerDALL-E` can transfer abstract style concepts which go beyond simple texture and color features while simultaneously preserving the semantic content of the translated scene.

2 Related Work

Reference Image-Based Style Transfer. Style and texture transfer tasks were studied since the beginning of the 21st century (Efros & Freeman, 2001; Hertzmann et al., 2001). In 2016, Gatys et al. (2016) propose a neural style transfer method in which a pre-trained CNN is used to extract content and style information from images, and to transfer the latter from an image to another. Following this pioneering work, this research field has attracted a lot of interest, with different methods focusing on different aspects of the topic, such as, e.g., arbitrary style transfer (Huang & Belongie, 2017), diversified style transfer (Ulyanov et al., 2017; Wang et al., 2020), or attention mechanisms to fuse style and content (Yao et al., 2019b; Park & Lee, 2019; Liu et al., 2021a). A specific line of work focuses on artistic style transfer. For instance, Chen et al. (2021b) propose to use internal-external learning and contrastive learning with GANs to bridge the gap between human artworks and AI-created artworks. Wang et al. (2022) introduce an aesthetic discriminator trained with a large corpus of human-created artworks. Other works train GANs using an artist-specific collection of images (Sanakoyeu et al., 2018; Kotovenko et al., 2019; Chen et al., 2021c). In contrast, we use the generic visual and language semantics embedded in DALL-E and CLIP to avoid collecting style or artist specific datasets and we describe a style simply using a textual sentence.

Language-Guided Style Transfer. Similarly to our language-guided style transfer approach, very recently a few works have proposed transferring methods conditioned on a textual description of the style. For instance, Fu et al. (2022) use contrastive learning to train a GAN for artistic style transfer, but the adopted language instructions describe features like textures and colors rather than more abstract styles concepts. Gal et al. (2022) use the CLIP space for a domain adaptation of a pre-trained StyleGAN (Karras et al., 2020). The method which is the closest to our approach is CLIPStyler (Kwon & Ye, 2022), where a patch-wise CLIP loss is used to train a U-Net (Ronneberger et al., 2015). However, to simultaneously condition the style change while preserving the image content, CLIPStyler needs to combine different loss functions and rejection thresholds, which presumably makes this approach sensitive to hyperparameter tuning. In contrast, our method does not have critical hyperparameters and can simultaneously change the style and preserve the content by jointly using the DALL-E and the CLIP space.

Large-scale Text-to-Image Generation Models. Recently, text-to-image models trained with large or very large scale datasets (Ramesh et al., 2021; 2022; Saharia et al., 2022; Yu et al., 2022) have attracted tremendous attention because of their excellent performance in generating realistic images starting from a textual query. However, these methods are not style-transfer methods, and the adaptation to a style transfer task, which should transform a specific content image according to a specific target style, is not trivial. In this paper, we study how to use a pre-trained large-scale text-to-image model (DALL-E) for the style transfer task. Specifically, we use the open-sourced dVAE of DALL-E (Ramesh et al., 2021), although our approach can be applied to other discrete token-based models which can encode and decode an input image.

Non-Autoregresisve Image Generation. There is a very recent growing interest in non-autoregressive image generation methods which try to alleviate AR problems like the large computational costs and the exposure bias (Sec. 1). However, differently from our approach, which is based on NAT (Gu et al., 2018), most of the previous work focuses on a mask-and-predict paradigm (Zhang et al., 2021; Chang et al., 2022; Lezama et al., 2022). For instance, Gu et al. (2022) propose a discrete diffusion process based on a mask-and-replace diffusion strategy to model a VQ-VAE latent space in parallel. Ding et al. (2022) propose a sophisticated region-based masking strategy, in which tokens outside the masked region can attend to all the other tokens, while tokens inside the region have a causal attention. As far as we know, this is the first work proposing a NAT-like architecture for non-autoregressive image generation which does not rely on a masking strategy.

3 Background

Non-Autoregressive Language Translation. Gu et al. (2018) propose a non-autoregressive architecture (NAT) for natural language translation, which consists of an encoder and a decoder. The encoder takes as input a source sentence $X = \{x_1, ..., x_{N'}\}$ of N' tokens and outputs a distribution over possible output sentences $Y = \{y_1, ..., y_N\}$, where Y is the translation of X in the target language and, usually, $N \neq N'$. The main novelty of NAT with respect to AR translation networks is that, during training, NAT uses a *conditional independent* factorization for the target sentence and the following log-likelihood:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log\ p(y_n | x_{1:N'}; \theta), \tag{1}$$

which differs from the common AR factorization in which the prediction of the n-th token (y_n) depends on the previously predicted tokens ($p(y_n | y_{0:n-1}, x_{1:N'}; \theta)$). This conditional independence assumption makes it possible a *parallel* generation of Y at inference time, largely accelerating the translation time with respect to AR models. Importantly, to make parallel generation possible, the encoder input (X) is provided as input to the decoder as well, and individual tokens ($x_n \in X$) can be copied zero or more times, with the number of times each input is copied depending on a specific "fertility" value (predicted by the encoder). As we will see in Sec. 4.1, we do not need to predict fertilities because, in our case, the cardinality of the input copies is fixed and determined by the upsampling task we use for the image translation.

Image Tokenization. DALL-E (Ramesh et al., 2021) use a dVAE (Razavi et al., 2019) to produce a discrete representation of the images. In more detail, an image I is transformed into a $k \times k$ grid of tokens $X = \{x_{i,j}\}_{i,j=1,...k}$, using an encoder $E(\cdot)$. Each token $x_{i,j} \in X = E(I)$ is an index of a codebook of embeddings ($C = \{\boldsymbol{e}_1, ...\boldsymbol{e}_M\}$, $1 \leq x_{i,j} \leq M$), built during the dVAE training, and corresponds to a patch in I. A decoder $G(\cdot)$ takes as input a grid of embeddings and reconstructs the original image: $\hat{I} = G(\{\boldsymbol{e}_{x_{i,j}}\}_{i,j=1,...k})$. Training in DALL-E is split in two stages. The first stage is dedicated to train the dVAE, while in the second phase an AR model is used to learn a prior distribution over the text and the image tokens. In `StylerDALL-E` (Sec. 4) we use only the dVAE (i.e., $E(\cdot)$, $G(\cdot)$ and C) pre-trained in DALL-E, discarding its AR model.

4 Method

The language-guided style transfer task can be described as follows. Given an image I, we want to generate a new image I^s which preserves the semantic content of I but changes its appearance according to a style description provided by a textual sentence t_s (e.g., $t_s =$ *"cartoon"*). At training time, we are given a dataset of image-caption pairs $D = \{(I, t_a)\}$, where t_a is a textual annotation describing the content in I (e.g., $t_a =$ *"A man's hand is adjusting his black tie."*). Note that the annotations in D do *not* contain style information and D does *not* contain samples of the target style. This way, we can use for training, e.g., a generic image captioning dataset (specifically, we used COCO (Lin et al., 2014)). At inference time, the annotation (t_a) is not needed.

In `StylerDALL-E`, we formulate this task as a *visual-token based translation* problem (Fig. 2). Specifically, given a content image I we first downsample I to get a half-resolution image I'. Then, I' is fed to the tokenization encoder (Sec. 3) which extracts a discrete grid of $k \times k$ source tokens $X' = E(I')$. X' can now be "translated" into a target (discrete) representation $\hat{Y}$, where $\hat{Y}$ is a grid at the original resolution ($2k \times 2k$) and $\hat{Y} = f(X')$, being $f(\cdot)$ the translation function implemented by an non-autoregressive network. Finally, $\hat{Y}$ is projected-back into the pixel space obtaining $I^s = G(\hat{Y})$ (see Fig. 2). The reason why we use downsampled versions of the content image is that "style" is commonly assumed to be mostly involved in the low-level image details, such as colors, texture, painting strokes, etc. X', which in our formulation represents I at a lower resolution, presumably keeps most of the content in I discarding some details, this way facilitating the style translation process. Preliminary experiments in which we fed the encoder with (tokenized) full-resolution images lead to poor results, demonstrating that the different cardinality between the source and the target sequence is an important component in this translation process. In the following subsections, we describe the architecture of $f(\cdot)$ and the way in which it is trained.

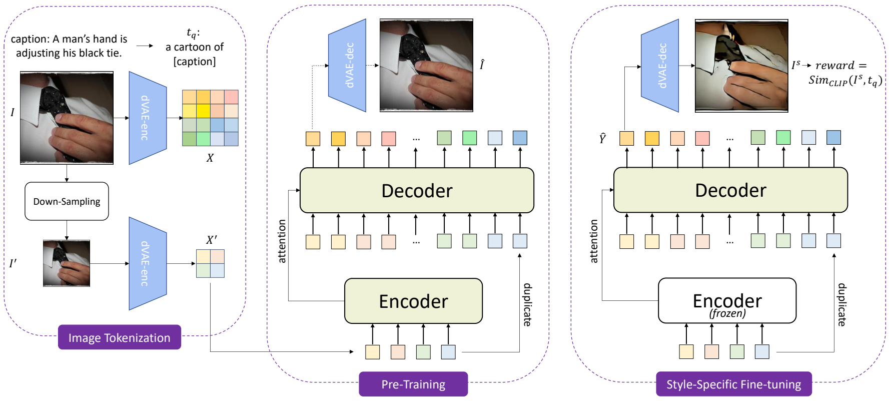

Figure 2: A schematic illustration of `StylerDALL-E`.

4.1 Architecture and Self-Supervised Pre-Training

For our translation network $f(\cdot)$ we use a NAT architecture (Gu et al., 2018) (Sec. 3) which we train from scratch using a self-supervised learning pretext task consisting in predicting the image at full resolution. Specifically, given the downsampled image I' (Sec. 4) and its corresponding grid of tokens $X' = E(I')$, we use the indexes in X' to extract the corresponding embeddings from C (Sec. 3). For each $x_{i,j} \in X'$, let $\boldsymbol{e}_{x_{i,j}}$ be the corresponding embedding in C and let $N' = k^2$. The so obtained set of embeddings is flattened into a sequence, and, for each element $\boldsymbol{e}_n$ of the sequence ($1 \leq n \leq N'$), we add an absolute positional embedding (Vaswani et al., 2017) $\boldsymbol{p}_n$, where $\boldsymbol{p}_n$ has the same dimension as $\boldsymbol{e}_n$: $\boldsymbol{v}_n = \boldsymbol{e}_n + \boldsymbol{p}_n$. The final sequence $V' = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_n, ..., \boldsymbol{v}_{N'}\}$ is input to the encoder of $f(\cdot)$. Note that an alternative solution is to directly fed $f(\cdot)$ with (a flattened version of) X' and let $f(\cdot)$ learn its own initial token embedding. However, using the embeddings in C has the advantage of exploiting the image representation of the dVAE pre-trained in DALL-E (also) as the initial representation for the tokens of $f(\cdot)$. Moreover, from the original image I we extract the ground truth $Y = E(I)$, which is flattened in a sequence of $N = 4k^2$ tokens.

Finally, following (Gu et al., 2018), we build a second sequence of input embeddings V, with cardinality N, which is fed to the decoder of $f(\cdot)$ (Fig. 2). As mentioned in Sec. 3, differently from NAT, we do not predict fertility values. Instead, we upsample X' and we get X by simply replicating each element $x_{i,j} \in X'$ into $x_{i,j}, x_{i+1,j}, x_{i+1,j+1}, x_{i,j+1}$. Then, X is used to extract the initial embeddings from C, which are flattened and added with a new positional encoding (computed over the new sequence of N elements). The rationale behind this choice is that $f(\cdot)$ is trained to predict the full-resolution image, and each encoder input ($\boldsymbol{e}_{x_{i,j}}$) corresponds to a patch in the subsampled image I' and to 4 patches in the full-resolution image I. Thus, initializing the decoder with 4 replicas of each source-image patch initial embedding provides a coarse-grained signal for the upsampling task. Note that $X \neq Y$, and we cannot use Y to build V because this would make the upsampling task of the decoder a trivial identity mapping operation.

Both the encoder and the decoder have self-attention layers and no causal masking is used. However, following (Gu et al., 2018), in the decoder, we mask out each query position (n) only from attending to itself. Using V' and V, $f(\cdot)$ generates N *parallel* posterior distributions over the dVAE vocabulary ($\{1, ..., M\}$): $P = f_\theta(V', V)$, where P is a $N \times M$ matrix, $P_n \in [0,1]^M$ and $P_n[y] = p_\theta(Y_n = y|V', V)$. Using $Y = \{y_1, ..., y_n, ..., y_N\}$, $f(\cdot)$ is trained to maximize:

$$\mathcal{L}_{pre-train}(\theta) = \sum_{n=1}^{N} \log P_n[y_n]. \tag{2}$$

This pre-training stage is independent of the target style and it can be shared over different styles. After this initialization stage, $f(\cdot)$ is able to generate realistic low-level details (which are missing in I') by manipulating the DALL-E tokens. In the following section, we describe how a specific style is incorporated in $f(\cdot)$ using a fine-tuning phase.

4.2 STYLE-SPECIFIC FINE-TUNING

Given a style description provided with a textual sentence t_s (Sec. 4), the goal is to fine-tune the pre-trained translator $f(\cdot)$ (Sec. 4.1) to make it generate image details in the style of t_s. We fine-tune only the decoder of $f(\cdot)$, keeping frozen the encoder. As the fine-tuning guidance we use the CLIP space and we maximize the cosine similarity between the image generated by $f(\cdot)$ and the projection of t_s into this space. However, merely maximizing this similarity does not take into account content preservation, so there is a risk of losing content information. To avoid this, at fine-tuning time we introduce additional language supervision, i.e., a textual description of the source image *content*, corresponding to the annotation t_a associated with I in D (Sec. 4). Specifically we (automatically) concatenate t_s and t_a to get a joint content-and-style description of the desired image I^s, obtaining a *prompt* (Radford et al., 2021) sentence t_q. For instance, given $t_a =$ *"A man's hand is adjusting his black tie"* and $t_s =$ *"cartoon"*, we obtain $t_q =$ *"a cartoon of a man's hand is adjusting his black tie*. On the other hand, in order to represent the image generated by $f(\cdot)$, we first need to sample the distributions in P (Sec. 4.1), and we do so using a plain argmax operation:

$$\hat{Y}_n = \arg \max_{y \in \{1,...,M\}} P_n[y] \quad \forall n \in \{1, ..., N\}. \tag{3}$$

The sampled sequence $\hat{Y}$ is reshaped to a $2k \times 2k$ grid and fed to the dVAE decoder to get the final image $I^s = G(\hat{Y})$. Finally, using the CLIP visual and textual encoders we compute the cosine similarity on the CLIP space:

$$r = Sim_{CLIP}(I^s, t_q). \tag{4}$$

However, directly using Eq. 4 as the fine-tuning objective function is not possible because Eq. 3 is not differentiable. Hence, we use an RL approach and the REINFORCE algorithm (Williams, 1992), which updates the parameters of $f_\theta(\cdot)$ using r as the reward. This leads to the gradient estimate:

$$\nabla_{\theta|_d} \mathcal{L}_{fine-tune}(\theta|_d) = \sum_{n=1}^{N} r \nabla_{\theta|_d} \log P_n[y_n], \tag{5}$$

where $\theta|_d$ indicates the parameters of the decoder only. Note that by maximizing Eq. 5 we encourage $f(\cdot)$ to generate images which have both the content (t_a) and the style (t_s) of the prompt t_q.

5 Experiments

In this section, we show the stylized results of StylerDALL-E and a comparison with the results of two types of style transfer methods, i.e., language-guided methods and standard reference image-based methods. Moreover, our ablation study, additional experimental results and implementation details, are shown in Appendix A.1, A.2 and A.3, respectively.

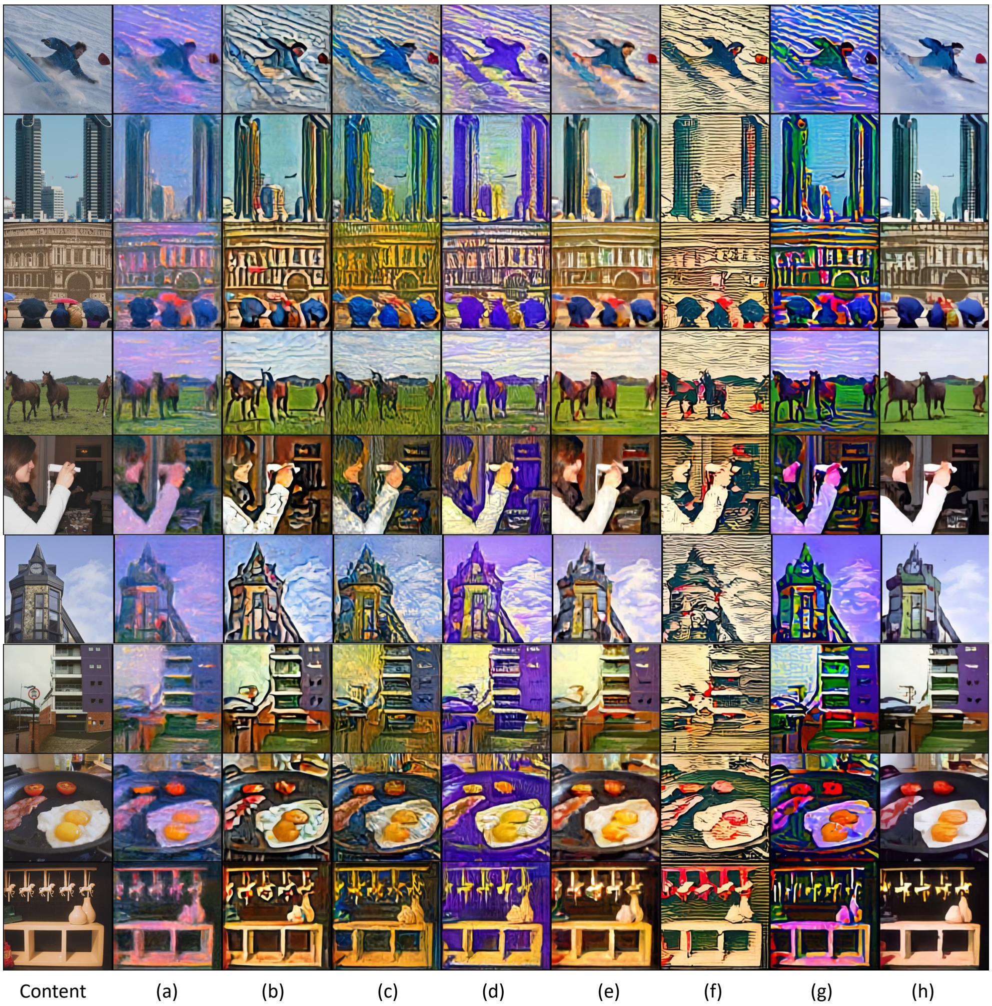

Figure 3: Language-guided stylized results of StylerDALL-E on: (a) Monet Impression Sunrise, (b) Picasso cubism, (c) Van Gogh blue color, (d) Van Gogh purple color, (e) warm and relaxing, (f) ukiyo-e print, (g) fauvism, (h) pixel art illustration.

5.1 Results

We show the language-guided style transfer results of StylerDALL-E in Fig. 1 and Fig. 3, where we conduct multiple experiments using language instructions vary from: a) artistic painting types, e.g., "oil painting" and "watercolor painting"; b) abstract artistic styles, e.g., "fauvism", "cartoon" and "pop art"; c) artist-specific styles, e.g., "Monet" and "Van Gogh"; d) artist-specific styles with additional descriptions, e.g., "Monet Impression Sunrise" and "Van Gogh blue color"; and e) emotional effects, e.g., "warm and relaxing". For evaluation, we use the COCO val-set.

The stylized images in Fig. 1 and Fig. 3, jointly with the images included in the Appendix A.2, show that: 1) StylerDALL-E can transfer abstract style concepts which go beyond the texture and color features and are similar to the typical trait of the artist/artistic target style; 2) each style corresponds to generated images which are different from those of other styles; 3) the image content is well preserved; and 4) StylerDALL-E can be applied to open-domain content images (i.e., the image content can contain animals, human beings, daily objects, buildings, etc.).

5.2 Comparisons with Other Methods

Comparison with other Language-Guided Methods. We compare StylerDALL-E with CLIPStyler (Kwon & Ye, 2022), a language-guided style transfer model which is the most directly comparable to ours. Both StylerDALL-E and CLIPStyler can be applied to open-domain content images and use only language instructions to describe the target style. Kwon & Ye (2022) propose two methods: CLIPStyler-Optimization and CLIPStyler-Train. The former optimizes a style transfer network *for each content image*, thus it is very time-consuming. The latter trains a network *for each style*, and then it can be used with different content images (similarly to StylerDALL-E, which fine-tunes a network for each style, Sec. 4.2).

As shown in Fig. 4, CLIPStyler-Optimization generates diverse stylized results but it suffers from inharmonious artifacts. For instance, in the column "Monet Impression Sunrise", there are multiple suns in the background and on the umbrella held by the woman. Other artifacts are generated with the other styles, e.g., "fauvism", "Monet" and "Van Gogh". On the other hand, the images generated by CLIPStyler-Train do not contain artifacts but there is less variation among different styles. Importantly, it is hard to recognize the typical trait of each artistic style, and the main differences among the styles are the colors. In contrast, the results of StylerDALL-E are much closer to the artworks of the specific artistic style, they show distinct differences among different styles, and they have no artifact issues. Moreover, StylerDALL-E is much less time-consuming as compared to CLIPStyler-Optimization. A magnified comparison is illustrated in the appendix A.2.

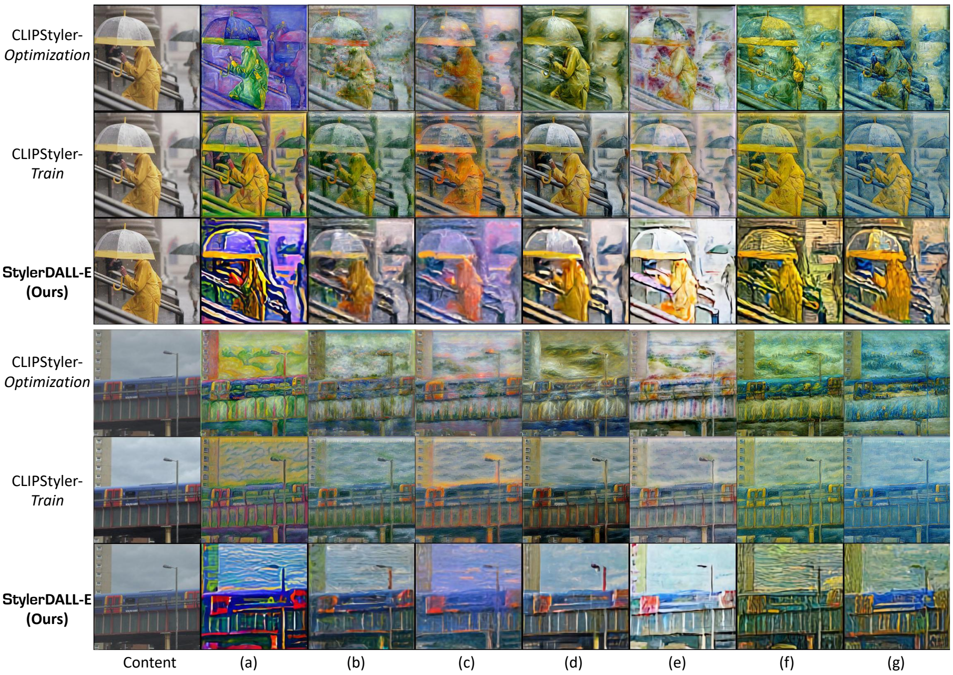

Figure 4: A comparison with language-guided methods (clearer using magnification): (a) fauvism, (b) Monet, (c) Monet Impression Sunrise, (d) oil painting, (e) watercolor painting, (f) Van Gogh, (g) Van Gogh blue color.

Comparisons with Reference Image-Based Methods. We compare StylerDALL-E with different state-of-the-art reference image-based methods: 1) AesUST (Wang et al., 2022), an arbitrary style transfer method which enhances the aesthetic reality using a GAN with an aesthetic discriminator trained with a collection of artworks; 2) StyTr2 (Deng et al., 2022b), an arbitrary style transfer method which uses a transformer to eliminate the biased content representation issues of CNN-based methods; and 3) AST (Sanakoyeu et al., 2018), which can transfer an artist style by training a GAN on a set of artworks of that artist.

To make the comparison feasible, we show the results of AesUST and StyleTr2 using two Van Gogh paintings as the input reference images, the results of AST trained with a collection of Van Gogh paintings, and the results of StylerDALL-E trained using language guidance. As shown in Fig. 5, the results of both AesUST and StyleTr2 are highly affected by the colors and the textures of the reference images and often in an unnatural way. For example, in "AesUST w/ ref2", the generated images contain a lot of eyes spread over different objects, which have presumably been transferred as texture details from the Van Gogh self-portrait. In "StyTr2 w/ ref1", the objects of all the images have an unrealistic blue color, which is the same as the source Starry Night image, and the orange bottle at the bottom loses its original color while showing a starry texture. As compared to AesUST and StyleTr2, AST generates more realistic results, showing harmonious styles, colors and textures across the images. However, the edges of the objects are sharp as those in the content images, making the result less like the Van Gogh trait, which describes the real-world scene in a more abstract way. Moreover, all the generated images have colors close to light green, while Van Gogh artworks are usually in bright and expressive colors. Among all the compared methods, the results of StylerDALL-E are much more distinctive and they are closer to the Van Gogh style with respect to the brushstroke, the bright colors, and the general feeling.

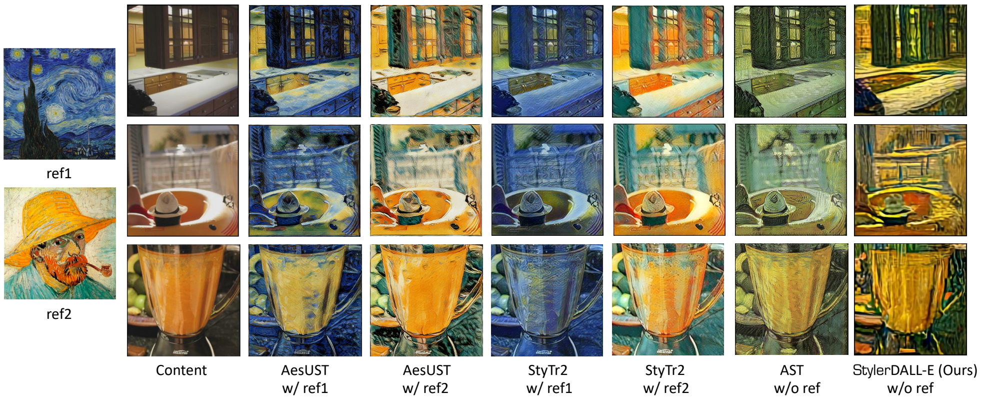

Figure 5: A comparison with reference image-based methods using Van Gogh style transfer (clearer using magnification). "w/" and "w/o" refer to whether the method uses a reference image as input.

6 Conclusion

We present StylerDALL-E, a language-guided style transfer method which uses the latent spaces of DALL-E and CLIP to incorporate visual-language semantics in the style transfer task. Specifically, inspired by natural language translation, we propose a non-autoregressive token sequence translation approach to manipulate the discrete latent space of DALL-E and a RL approach to include a CLIP-based reward in the fine-tuning stage. Differently from previous work, StylerDALL-E can transfer abstract style concepts which are implicitly represented in DALL-E and CLIP and which cannot be easily obtained using reference images. Moreover, using the DALL-E latent space as the basic image representation, makes it possible to reduce the artifacts and the semantic incoherence better than the previous work that operates at the pixel level.

REFERENCES

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*, 2015.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021a.

Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 872–881, June 2021b.

Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 872–881, 2021c.

Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 134–143, 2021.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11326–11336, June 2022a.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11326–11336, 2022b.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 341–346, 2001.

Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision (ECCV)*, 2022.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *ICLR*, 2018.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 327–340, 2001.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4422–4431, 2019.

Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18062–18071, 2022.

José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. *arXiv preprint arXiv:2209.04439*, 2022.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3920–3928, 2017.

Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5141–5150, June 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6649–6658, 2021a.

Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5880–5888, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 698–714, 10 2018.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP*, 2019.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6924–6932, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, and Dongming Lu. Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7789–7798, 2020.

Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Aesust: Towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, may 1992. ISSN 0885-6125. doi: 10.1007/BF00992696.

Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1467–1475, 2019b.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5943–5951, 2019.

Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. UFC-BERT: unifying multi-modal controls for conditional image synthesis. *Advances in Neural Information Processing Systems*, 34:27196–27208, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.