Leveraging style-based relations for text-conditioned style transfer

Surgan Jandial*[1] Silky Singh*[1] Simra Shahid*[1] Abhinav Java[1]
Shripad Deshmukh†[2]

[1]MDSR Labs, Adobe [2]UMass Amherst

Abstract

Text-conditioned style transfer is a promising area of research that enables the generation of stylistic images from natural language descriptions of desired styles. A common theme underlying existing methods that accomplish this is, using CLIP-based embeddings to direct image editing through directional losses. Such losses enforce alignment by ensuring the direction from a source text to the target text should match the direction from the source image to the stylized image. However, people often intuitively connect and relate a style with other stylistic attributes, such as associating "Starry Night by Van Gogh" with "blue hues" or "swirly lines" and not with "pop art" or "monochrome tones". In this work, we introduce StyleRel*, a framework that exploits such natural style relationships that people establish and propose a relational loss between style attributes, target style text, and the generated stylized image. Through comprehensive evaluations on global and localized style transfer tasks,* StyleRel *demonstrates a nuanced understanding of target style text and consistently outperforms existing state-of-the-art approaches, as evidenced by both qualitative and quantitative metrics.*

1. Introduction

Style Transfer [6, 10, 16, 23, 45] is a technique of applying the style from one image to another image while retaining the original content of the latter. The recent developments in Large Language Models (LLMs) [4, 27, 40, 46] and Visual Language Models (VLMs) [24, 35, 36, 39, 41] have paved exciting new directions in the field of style transfer allowing the expression of complex styles in the form of human interpretable textual descriptions [8, 22, 34]. This takes away the burden of obtaining specific stylized images and aims to transfer the semantic style from only a textual description to a content image by leveraging the representation capacity of large-scale foundation models like CLIP [35].

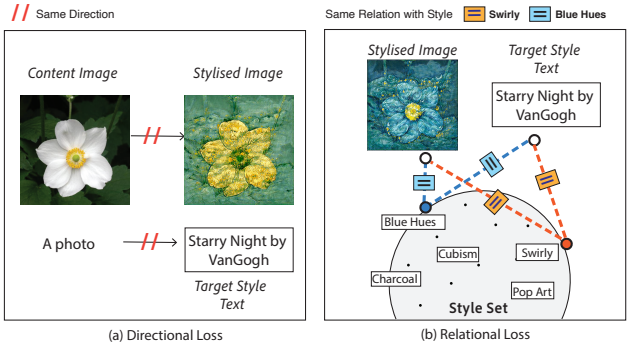

Figure 1. For the target style text 'Starry Night by Van Gogh', **(a) directional loss** aligns the direction of content image to stylized image with the direction of the text 'a photo' to target style text, preserving `directional consistency`. For a reference style in the set of styles, for e.g. 'Swirly', **(b) relational loss** ensures the relationship of stylised image with 'Swirly' is similar to the relationship of target style text with 'Swirly', as depicted by orange lines. In our work, we maintain this `relational consistency` for the entire set of reference styles using our relational loss. This is motivated by the intuition that if styles represented by target style text are similar to styles represented by stylised image, then their relationships will be consistent.

Broadly, the recent text-based style transfer approaches are optimized for a given prompt based on aligning the *directions* of the source text and target text to the direction of the source image and stylized image respectively, in the embedding space (e.g. CLIP) [9, 22, 34]. These approaches show promise by generating semantically consistent and diverse images from target textual descriptions.

However, in this work, we discuss an interesting yet crucial nature of the style descriptions by drawing upon the cognitive principles outlined by Gentner [11], Hofstadter [12], Hofstadter and Sander [13] that highlights comparative learning in understanding and differentiating concepts. More specifically, as humans we understand a style by intuitively connecting or relating it to the other styles from our knowledge. To elaborate upon our intuition, consider

*Equal contribution.
†Work done while at Adobe

an example style description: "photo in Starry Night by Van Gogh style" as illustrated in Fig 1, and we can experience how we implicitly interpret by associating "Starry Night by Van Gogh style" it with the other styles like "blue hues," "stars," and,"swirly lines," and not with "pop art" and "monochrome tones". We thus posit that in addition to matching directions, we can leverage these natural relationships between styles to improve text-conditioned style transfer.

To that end, we propose **StylRel**, a versatile and efficient framework that enhances the existing text-based style transfer approaches by introducing a relational loss term. Concretely, *first*, we sample a set of well-known style templates and encode them to form a "style tensor". *Next*, we compute similarities of "style tensor" with both the target style instruction text and generated stylized image and create text-style and image-style "relation vectors" respectively as illustrated in the Fig 2b. *Finally*, we compute the relational loss as the mean squared error between the image relation vector and the text relation vector averaged over multiple image patches. We summarize our contributions as follows:

- In this work, we show that leveraging the relationship between styles improves the effectiveness of text-based style transfer. Hence, we propose a framework StyleRel, that incorporates a flexible relational loss (Sec 4.2). The relation loss term consists of two key parts - a) grounding the style descriptions in a well-defined style vocabulary using a "style tensor", and b) computing a "relation vector" to describe the relationship of the stylized image and text with the style tensor.
- We compare our technique against the state-of-the-art text-based style transfer approaches (CLIPStyler [22], Gen-Art [47]). Our analysis in Sec 5 shows that StyleRel outperforms all baselines both qualitatively and quantitatively.
- We extend our application beyond global image stylization to the task of local style transfer and compare against a state-of-the-art baseline (Text2LIVE [3]). For a deeper analysis, we contribute a simple yet stronger baseline to account for limitations in the existing prior art.

2. Related Work

Image Style Transfer. [10] introduced style transfer as a pixel-level optimization problem that uses an image of the reference style, and then jointly optimizes the style and content losses. [17, 42] extended the formulation in [10] by training a style-specific generative model using similar losses or adding additional perceptual losses. Adaptive Instance Normalization (AdaIn) [15] on the other hand, proposed to apply the mean and standard deviation of the style images to the normalized statistics of the content images. More recently and owing to the success of attention mechanisms, several works have leveraged it to compute image-style correlations. Style Attention Network [32] utilized cross-attention, while Adaptive Attention Normalization [25] explored an improved version of attention-based style transfer by learning a spatial attention score from both the shallow and deep features. In a different line of work, [1, 28] benefitted with the use of neural flows and VAEs to model style transfer respectively. We note that most of the discussed works employ convolutional networks (CNNs), and subsequently found the recent methods to extend the use of transformers for the aforementioned problem. StyleFormer [44] proposes a transformer-based style composition module whereas [5] discusses the limitations of CNN-localization for style transfer, and uses vision transformers to incorporate the long-range dependencies in images.

Given the existing and the recent image style transfer methods, we emphasize their limitation to require a reference style image and thus focus on the problem of text-based style transfer in subsequent discussions.

Text-conditioned Style Transfer. This overlaps with the broader areas of text-conditioned generation, editing, and VLM that allow us to relate both image and text in a shared embedding space. CLVA [7] introduced the new task of language-driven artistic style transfer. However, they utilise annotated styled images with text descriptions during training. The majority of the recent methods [3, 18, 22, 29, 47] rely on the representational capability of CLIP [35] for guiding the pre-trained generators to use text conditions. StyleCLIP [34] proposed text-guided attribute manipulation by exploring novel editing directions in the pre-trained StyleGAN [19] latent space using CLIP. Moving forward, StyleGAN-NADA [9] adapts pre-trained generators for novel domains without requiring additional training images and with the text descriptions using CLIP supervision. However, the pre-trained knowledge of generative models poses a bottleneck in both these methods, and the edits are either restricted within the trained domain of pre-trained StyleGAN [19] or only work for attributes seen during the training. CLIPStyler [22] alleviates this and employs global as well as patch CLIP losses to transfer the source image using target style description/text, regardless of their source domains. Text2LIVE [3] extended this to perform edits in a localized region, hence, localized text conditioned style transfer. Further, CLIPStyler suffers from over-stylization and content-mismatch problems, which Sem-CS [18] and Generative Artisan (Gen-Art) [47] address. However, both input pre-computed segmentation masks, which restricts their application to human portraits/faces and other standard classes. Gen-Art requires portrait segmentation masks from FCN-ResNet101[30], while Sem-CS from Deep Spectral Segmentation[31]. By definition, each of the above approaches works with a CLIP-based directional loss (see Fig 1a), and we contrast this by augmenting their existing loss functions with our intuitive yet simple relational loss

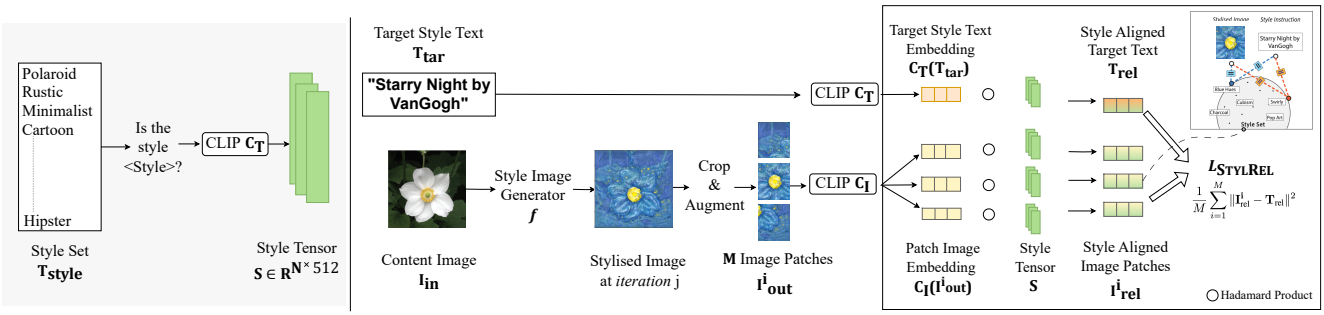

Figure 2. (a) demonstrates the creation of the Style Tensor **S**. First, each style in the set of styles T_{style} is framed as a question like 'Is the style <STYLE>?', which is then embedded in the CLIP space with the CLIP text encoder C_T to create the Style Tensor **S**. (b) shows our proposed **StylRel** framework, which aligns the relationship of target text style (T_{tar}) and **S**, represented by text-relation vector ($T_{\mathbf{rel}}$) with generated stylized image patches ($I_{\mathbf{out}}$) and **S** represented by image relation vector ($I^i_{\mathbf{rel}}$). The framework aligns the target style text with the stylized image with the proposed relational loss L_{StylRel}.

(see Fig 1b). In a different line of work, recent methods DiffStyler [14], Stable Diffusion [37] use diffusion to perform text-conditioned image editing and stylization. However, we note that our current formulation of relational loss assumes style understanding of the *CLIP embedding space*, and an adaptation of this phenomenon to diffusion's latent representation is part of our future work.

3. Preliminaries of Text-conditioned Style Transfer

Recent style transfer methods like StyleCLIP [34], StyleGAN-NADA [9], CLIPstyler [22] primarily divide their loss function into a CLIP-based directional loss, and a content loss:

$$\min_{\theta_f}[L_{\text{dir}}(\theta_f, \theta_C) + L_{\text{content}}(\theta_f, \theta_C)] \tag{1}$$

where f is an image generator (e.g. U-Net [38]), θ_{C} are the parameters of a frozen CLIP model, L_{dir} is the directional loss, and L_{content} is the content loss. Next, we discuss each loss term given in Eqn 1.

Directional Loss (L_{dir}). We first compute the unit vector joining CLIP text embeddings of the placeholder textual description T_{in} of content image I_{in} (e.g. *"a photo"*), and the target style text T_{tar} (e.g. *"photo in Starry Night by Van Gogh style."*), respectively. Likewise, we compute unit vector joining CLIP image embeddings of I_{in} and its desired stylized output $I_{\text{out}} = f(I_{\text{in}})$, respectively. Let C_I and C_T denote the clip image and text encoders, the direction vectors are defined as:

$$\mathbf{T}_{\text{dir}} = \frac{C_T(T_{\text{tar}}) - C_T(T_{\text{in}})}{\|C_T(T_{\text{tar}}) - C_T(T_{\text{in}})\|_2}$$

$$\mathbf{I}_{\text{dir}} = \frac{C_I(I_{\text{out}}) - C_I(I_{\text{in}})}{\|C_I(I_{\text{out}}) - C_I(I_{\text{in}})\|_2}$$

where C_I and C_T are CLIP image and text encoders respectively.

The directional loss aligns the CLIP-space direction between the input-output text pairs $(T_{\text{in}}, T_{\text{tar}})$ with the input-output image pairs $(I_{\text{in}}, I_{\text{out}})$. The final loss is given by:

$$L_{\text{dir}} = 1 - \mathbf{T}_{\text{dir}} \cdot \mathbf{I}_{\text{dir}} \tag{2}$$

where $\mathbf{T}_{\text{dir}}, \mathbf{I}_{\text{dir}}$ are the text direction and image direction vectors respectively.

Content Loss (L_{content}). It is the mean-squared error between the features of content and output stylized images, both extracted from the pre-trained VGG-19 networks [10]. Further, methods like [22, 47] also employ a total variation regularization loss L_{tv} to handle the artifacts and ensure the spatial consistency of the final image.

4. Methodology

In this section, we describe our proposed framework for text-based style transfer, StylRel, that goes beyond the traditional directional losses (Eqn 2) and incorporates a more intuitive method of associating styles with other stylistic attributes and grounding these relationships in the generated stylized images. To achieve this we propose a *relational loss* by which aligns the target style text and the stylized image with the well-known style attributes. *First*, we compose a vocabulary of these well-known style attributes with colors, tones, style arts (Sec 4.1). *Next*, we compute the relationship vectors of these styles with the generated stylized image and the target style text, and, *finally*, we match these individual relationships to better align the output image with the target style text, not just directionally but also

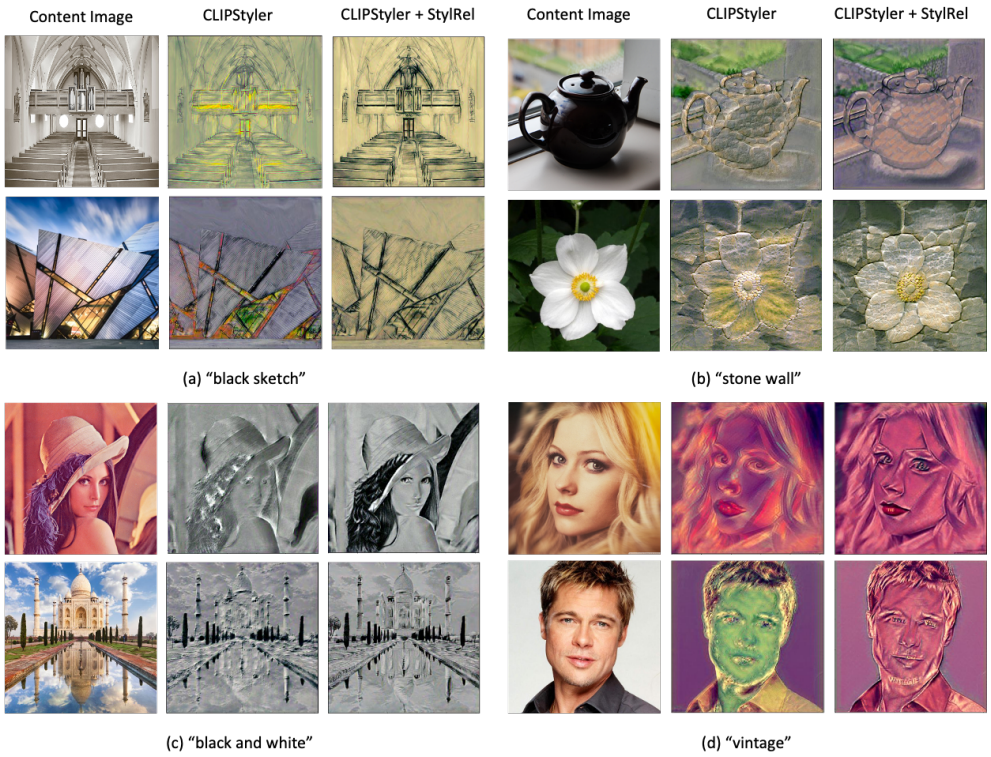

Figure 3. **Qualitative comparison of L_{StyleRel} v/s CLIPStyler**. CLIPStyler with our framework StyleRel (CLIPStyler + StylRel) is better able to produce realistic textures that reflect the target style texts, and also preserves the content of the original image. Additional results will be provided in the appendix.

relationally (Sec 4.2). In the following sections we discuss each of these key components in detail.

4.1. Creating the Style Tensor

We begin by creating a set of well-known style attributes consisting of art styles, colors, line strokes, and textures (see Fig 2a). Each style is framed as a question like *"Is the style* <Style>*?"* where each <Style> is one of the style in the vocabulary. These questions are then encoded into the CLIP embedding space to create the style tensor $\mathbf{S} \in \mathbb{R}^{N\times 512}$ where N is the number of styles and 512 is the embedding size of CLIP. The style tensor is created as:

$$\mathbf{S} = C_T(T_{\text{style}}) \in \mathbb{R}^{N\times 512} \tag{3}$$

where C_T is the CLIP text encoder and T_{style} is the list of N preprocessed questions. The choice of our question is substantiated in the ablation studies in Sec 5.4. We illustrate this visually in Fig 2a. More details about the set of styles are included in the Appendix.

This formulation of Style Tensor framed as a question helps us to understand (i) how the target text description (T_{tar}) is associated with each of these styles, and, (ii) how the stylized image (I_{out}) is associated with each of these styles. Next, we show how we leverage this formulation.

4.2. Relational Loss (L_{StyleRel})

Text-Style Relation: We represent the relationship between target style text with $\mathbf{S}$ as $\mathbf{T}_{\text{rel}}$ and define it as the similarity score of the target text with each style in $\mathbf{S}$:

$$\mathbf{T}_{\text{rel}} = \mathbf{S} \circ (C_T(T_{\text{tar}}))^\intercal \tag{4}$$

where $\circ$ is Hadamard product, $\mathbf{T}_{\text{rel}} \in \mathbb{R}^{N\times 1}$ is the relation vector, N is the number of styles in the style tensor.

Image-Style Relation: We represent the relationship between stylized image with $\mathbf{S}$ as $\mathbf{I}^{\mathbf{i}}_{\text{rel}}$. We compute this relationship for every patch of the stylized image I_{out}. This is inspired from CLIPStyler [22] which introduces a patch level loss for a fine-grained style transfer which aligns individual patches with the target style text. Similarly, for each image patch $\text{aug}(\cdot)$ we encode it with clip image encoder as $C_I(\text{aug}(I^i_{\text{out}}))$, and then compute the similarities of each patch with style relations in the form of the image relation vector ($\mathbf{I}^{\mathbf{i}}_{\text{rel}}$) as follows:

$$\mathbf{I}^{\mathbf{i}}_{\text{rel}} = \mathbf{S} \circ (C_I(\text{aug}(I^i_{\text{out}})))^\intercal \tag{5}$$

Relational Loss: Given the relation vectors for image and text domains, we optimize f with L_{StyleRel} that aligns the

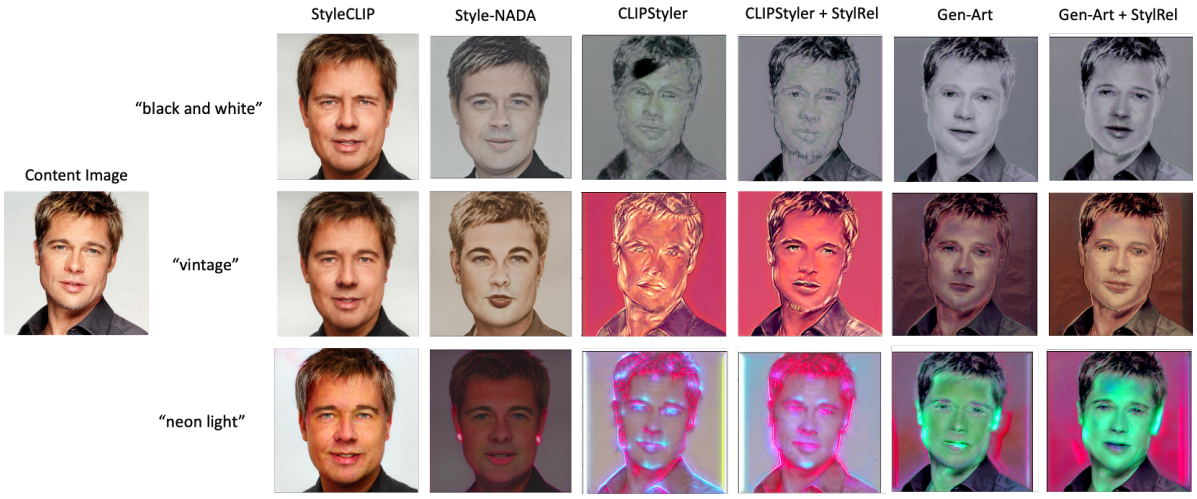

Figure 4. Comparison with baselines: StyleCLIP [34], Style-NADA [9], CLIPStyler[22], Gen-Art [47]. First column is the content image, and the adjoining texts are style text description prompts. Best viewed in zoom and color.

relation vector of the image with the relation vector of the text domain using the mean squared error averaged over all image patches, and is given by:

$$L_{\text{StylRel}} = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{I}^{\mathbf{i}}_{\text{rel}} - \mathbf{T}_{\text{rel}}\|^2 \tag{6}$$

where M denotes the total number of patches. We illustrate the proposed loss visually in Fig 2b.

5. Experiments and Results

5.1. Experimental Setup

Our proposed framework is simple and can be easily integrated with existing text-conditioned style transfer methods, requiring only a few additional lines of code. Note that our StylRel loss function is added to the existing set of losses used by the baselines, and hence makes no changes to their inputs or architecture.

Dataset and Metrics. We utilize the test dataset provided in [22] and in total obtain 100 stylized outputs for comparison. We employ SSIM (Structural Similarity Index Measure) [43], CLIP score [35], and PatchCLIP score for quantitative evaluation. SSIM measures image similarity between the stylized output and the content image, while CLIP score estimates the association of the stylized output images with the target style description. To capture more intricacies, we also evaluate the CLIP score over randomly sampled image patches and specify it as the PatchCLIP score. For each of these metrics, a higher score signifies superior performance.

Hyperparameters and Compute Details. We set the input image resolution to 512×512, and use Adam optimizer [20]. In the case of StylRel, the number of patches M is always kept as 64. Besides, the hyperparameters specific to baselines are set using their official implementations and further elaborated in appendix. We run all of our experiments using Pytorch v1.12 [33] on a single NVIDIA GeForce RTX 3090 (24GB) GPU.

5.2. Comparison with baselines

Comparison with CLIPStyler (CS) [22]. In this experiment, we add L_{StylRel} to CLIPStyler which uses a patch-wise directional loss (L_{patch}), a global directional loss (L_{dir}), a content loss ($L_{content}$) and a total variation loss (L_{tv}):

$$L_{\text{CS}} = \lambda_d L_{\text{dir}} + \lambda_p L_{\text{patch}} + \lambda_c L_{\text{content}} + \lambda_{tv} L_{tv} \tag{7}$$

The overall loss function is thus given as:

$$L_{\text{CS+StylRel}} = L_{\text{CS}} + \beta L_{\text{StylRel}} \tag{8}$$

where β is the loss balancing parameter (here $1e$-4). The quantitative results are shown in Tab 1. We begin by reporting the effect of each of global L_{dir} and patch-wise L_{patch} on CLIPStyler. Across the experiments, we observe that L_{patch} has more influence on performance than L_{dir}, and that adding our loss function consistently improves CS on all the specified metrics.

The qualitative comparison of our approach on CLIPStyler (CS) is shown in Fig 3. We examine over multitude of style descriptions and observe our approach is able to preserve the original content of the images while faithfully transferring the intended styles. On the other hand, artefacts in CLIPStyler's outputs are clearly evident, for example, in (c) – first row, "Black and White", the woman in the CS's stylized image is missing the original structure of the input

image. Moreover, the object boundaries get distorted in CS outputs (second rows in (b), (c), first row in (d)). Also, we report for some examples in (a) and (d) that CS's outputs do not reflect the style described by the target style texts.

Setting	CLIP Score (↑)	PatchCLIP Score (↑)	SSIM (↑)
CS	23.82	22.95	0.3895
CS w/o L_{patch}	19.73	19.99	**0.4005**
CS w/o L_{dir}	23.33	22.57	0.3886
CS + StylRel	**24.07**	**23.85**	0.3913

Table 1. Quantitative comparison of StylRel with CLIPStyler(CS) and its variants. We notice the addition of StylRel consistently improves CS on the standard metrics

Comparison with other baselines. In this experiment, we compare against the existing text-guided manipulation methods – (i). StyleCLIP [34], (ii). Style-NADA [34], and an extension of CLIPStyler – (iii). Generative Artisan CLIPStyler (Gen-Art) [47]. Baselines (i, ii) are based on CLIP and StyleGAN pre-trained on human faces, while (iii) improves over CLIPStyler's overstylization problem which distorts human faces, by leveraging human segmentation masks from FCN(ResNet-101)[30]. More specifically, Gen-Art operates over human selfies and portraits and is a simple extension that uses the same loss function as Eqn 7. Hence, we adopt a consistent setting and a human example for the results in Fig 4.
Similar to [22], we found that StyleCLIP and Style-NADA are restricted and inherently different in behavior to CLIPStyler for the current problem. They either fail to transfer the style satisfactorily and/or change the face identity of the source image. However, CLIPStyler and Gen-Art both perform decently while handling face styles, with Gen-Art being better (see Fig 4). Further, it can be observed that adding our L_{StylRel} to both CLIPStyler and Gen-Art considerably improves their outputs. Note that we extend Gen-Art with our approach in the same way as Eqn 8.

5.3. Text-conditioned Local Style Transfer

In this experiment, the text descriptions are used to transfer the styles to a specific localized region of the image rather than the entire image, e.g."*building* on fire" instead of "fire". Thus, it deems a more challenging problem that necessitates an accurate delineation of the region of interest. Here, we first compare against the baseline Text2LIVE [2], and then extend over its limitations to propose a simple yet effective baseline, and further demonstrate the efficacy of StylRel on the aforementioned task.

Comparison with Text2LIVE (T2L). Given the source content image I_{in} and target style text T_{tar}, Text2LIVE (T2L) [2] consists of a generator that synthesizes the edit layer = $\{\alpha, C\}$ where α is an opacity map which acts as a segmentation mask to localize the region, and C is an image which defines edits in that localized region.:

$$I_o = \alpha C + (1 - \alpha) I_{\text{in}} \tag{9}$$

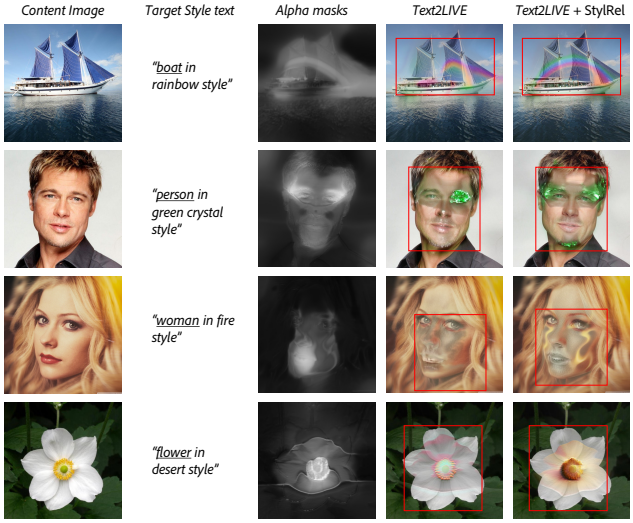

Figure 5. **Qualitative comparison with Text2LIVE [2] on local style transfer**. Clearly, StylRel allows to better preserve the nuances of the target text style. Text2LIVE also produces alpha masks (shown in 3rd column) which highlight the regions where edits take place. The regions of interest corresponding to underlined text in target style texts are highlighted in red boxes.

For their supervision, they also compose the edit layer with a green background(I_{green}) to obtain I_{screen} as:

$$I_{screen} = \alpha C + (1 - \alpha) I_{green} \tag{10}$$

Their objective function comprises of L_{comp} – a combination of CLIP based cosine distance and directional loss to ensure I_o conforms to T_{tar}, L_{screen} – a CLIP based supervision to learn the edit layer, $L_{structure}$ which preserves content of I_o w.r.t I_{in}, and L_{reg} which is a regularization term:

$$L_{\text{T2L}} = L_{\text{comp}} + \lambda_g L_{\text{screen}} + \lambda_s L_{\text{structure}} + \lambda_r L_{\text{reg}} \tag{11}$$

where $\lambda_g, \lambda_s, \lambda_r$ control the relative weights of the individual loss functions. In this case, we apply StylRel on both $I_{\text{edit}=\{\alpha, C\}}, I_{\text{screen}}$ with $\beta = 4e\text{-}7$:

$$L_{\text{T2L+StylRel}} = L_{\text{T2L}} + \beta(L^{\text{edit}}_{\text{StylRel}} + L^{\text{screen}}_{\text{StylRel}}) \tag{12}$$

We compare with Text2LIVE for local style transfer task, quantitatively in Tab 2, and qualitatively in Fig 5 where we observe that StylRel exhibits better fine-grained understanding in interpreting target style instructions (compare for prompts *flower in desert style* and *woman in fire style*

in Fig 5) and focusing on regions of interest (compare for prompts *boat in rainbow style* and *person in green crystal style* in Fig 5) than Text2Live.

Proposing a new baseline – Mask-CLIPStyler. Due to the complexity of prompts and difficulty to learn localizations under constraints, Text2LIVE is bottlenecked by its inaccurate opacity maps. In their current form, they employ a U-Net architecture to obtain these maps, and from Fig 5 (Col. 2) we clearly observe they are not satisfactory which affects the composition results in Eqn 9.

To address this and propose a stronger baseline for evaluating our approach, we extend CLIPStyler with an additional pipeline which utilizes the state-of-the art method Grounding DINO [26] followed by SAM [21] (referred to as g-SAM) to extract segmentation masks for the target regions mentioned in the text prompt. We call this baseline – Mask-CLIPStyler. In particular, this pipeline is divided into three steps (i) Extract segmentation masks M from g-SAM, (ii) input the segmentation mask and I_{in} to a U-Net [38] based generator f (iii) optimize the stylized output $I_{\text{mask-CS}} = f(I_{\text{in}}, M)$ using modified loss functions. Specifically, we modify the CLIP-based directional loss as follows:

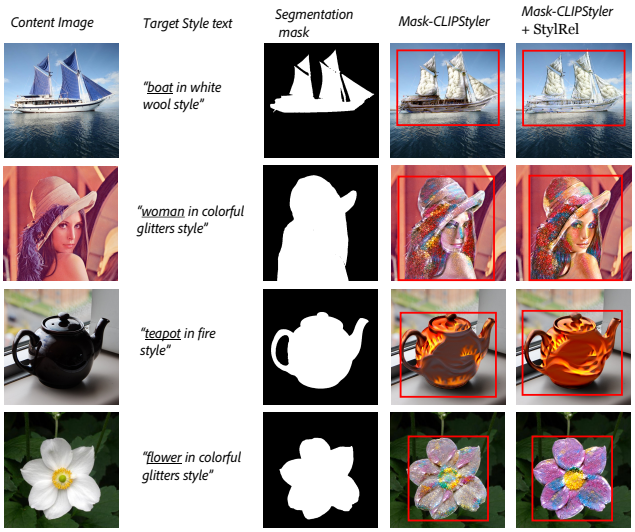

Figure 6. **Qualitative results of Mask-CLIPStyler and Mask-CLIPStyler+StylRel on local style transfer**. With StylRel, the output stylized images reflect the style text descriptions faithfully in the local regions. Segmentation masks in Col 3 are obtained using g-SAM. Best viewed in color and zoom.

$$\begin{aligned} \Delta T &= C_T(T_{\text{tar}}) - C_T(T_{in}) \\ \overline{\Delta I} &= C_I(I_{\text{mask-CS}} \odot M) - C_I(I_{\text{in}} \odot M) \\ \overline{L_{dir}} &= 1 - \frac{\Delta T \cdot \overline{\Delta I}}{|\Delta T||\overline{\Delta I}|} \end{aligned} \tag{13}$$

Setting	CLIP Score (↑)	SSIM (↑)
Text2Live	22.64	0.8363
Text2Live + StylRel	**22.65**	**0.8397**
Mask-CLIPStyler	22.60	0.6470
Mask-CLIPStyler + StylRel	**22.62**	**0.6536**

Table 2. Quantitative comparison with baselines on the task of local style transfer. We notice the addition of StylRel consistently improves both Text2Live [3], and proposed baseline Mask-CLIPStyler on the standard metrics.

where C_T and C_I denote the CLIP text and image encoders respectively. The overall objective function for Mask-CLIPStyler is then given as:

$$L_{\text{MCS}} = \lambda_d \overline{L_{\text{dir}}} + \lambda_p \overline{L_{\text{patch}}} + \lambda_c L_{\text{content}} + \lambda_{tv} L_{tv} \tag{14}$$

Finally, we add StylRel to Mask-CLIPStyler with $\beta = 5e\text{-}5$:

$$L_{\text{MCS+StylRel}} = L_{\text{MCS}} + \beta' \overline{L_{\text{StylRel}}} \tag{15}$$

We document the quantitative comparisons in Tab 2. In Fig 6, a qualitative comparison with Mask-CLIPStyler reveals that StylRel outperforms in generating realistic interpretations of both target text and content image. Specifically, for prompts such as "*flower in colorful glitters style*" and "*woman in colorful glitters style*", StylRel adeptly captures the nuanced attributes of the target style "colorful glitter," rendering them onto the content image realistically. This can be seen by the distinct colors of flower elements (flower petals are glittery pink and pistil is yellow glitter) and the nuanced application of glitter on the woman's face, as opposed to a less refined glitter application observed in the other baseline.

5.4. Ablation and Analysis

In this section, we study the influence of components used by our approach. We select CLIPStyler for these experiments and divide this section into questions regarding our choice of β, list of styles in T_{style}, question prompt in T_{style} and deciding where to apply StylRel– global v/s local (patch-wise). Finally, we discuss some limitations of our approach in the later part of this section.

On varying β in Tab 3b, we observe our performance first increases on increasing β, and then subsequently decreases for its higher values. We expect this to happen as assigning considerably high weights to L_{StylRel} will create an overall imbalance for other loss functions used in Eqn 7.

To vary the choice of styles in T_{style}, we experiment with three different style vocabularies of varied lengths and report the results in Tab 3a. We observe that StylRel is able to improve over baseline CLIPStyler in all three cases, with a slight variation of performance in each case. We attribute this variation to the phenomenon of quantity v/s quality of styles in each vocabulary, and look forward to it as an inter-

Setting	CLIP Score (↑)	SSIM (↑)
CS	23.82	0.3895
I	24.07	**0.3913**
II	**24.18**	0.3887
III	24.02	0.3909

(a)

Setting	CLIP Score (↑)	SSIM (↑)
β_{5e-5}	24.01	**0.3919**
β_{1e-4}	**24.07**	0.3913
β_{4e-4}	23.45	0.3808
β_{1e-3}	22.59	0.3687

(b)

Table 3. (a). Performance of our framework across a variation of style sets {I, II, III}. (b). Performance of L_{STYLREL} across multiple β parameters.

esting future direction. More details for these vocabularies are attached in the appendix.

To vary the choice of question prompt in T_{style}, we choose three different prompts each of which are –"is the style <STYLE>?", "is the image in style <STYLE>?", and "is it looking like <STYLE>?". To effectively quantify this comparison, we combine the SSIM score, CLIP score, and PatchCLIP score into an overall score O_s = 100 * SSIM * (CLIP Score + PatchCLIP Score) / 2. We then report percent improvement in overall score (O_s) over the base CLIPStyler as $\Delta\%$ in Fig 7, and observe that STYLREL consistently improved over CLIPStyler across different prompt choices. We acknowledge that different prompts improve differently, however, developing methodologies to predict the right prompt for STYLREL is left as part of our future work.

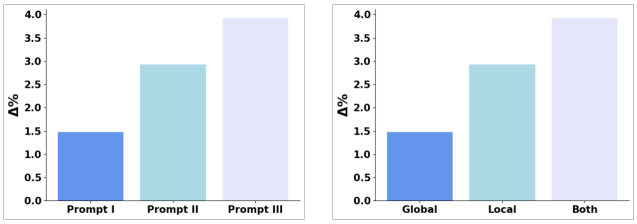

Figure 7. **Ablations**. (a) Ablation study on different sets of question prompts (I, II, III) used to create our style tensor S, (b) Ablation on three variants of our proposed STYLREL loss- Local, Global, and both. We observe a superior performance compared to CS across different sets of prompts, as well as variations of the STYLREL loss applied.

To analyze where to apply STYLREL, we consider three possible configurations namely global, local, and both (global + local). Global refers to applying STYLREL only on the entire image i.e. keeping patch size equal to the size of the image and number of patches M = 1. Local refers to the Eqn 6, and both refer to incorporating both the losses simultaneously. We present our results in Fig 7, and observe that the influence of local is significantly higher than the global, even though applying both works better overall. Thus, for simplicity, we use local in all our experiments.

Limitations. We found that our method depends on CLIP's capacity to encode question prompts (see Sec 4.1), and the performance of the underlying baseline such as Text2LIVE. In the case of the former, we experiment with two longer length and twisted question prompts — 1 : "Would it be easy for someone looking at this image to say it is in the style of <STYLE>?", 2 : "Is it easy for someone looking at this image to infer that it is in the style of <STYLE>?", and report our results in Fig. 8a. It can be observed that the performance of STYLREL gets affected and we attribute this to the loss of semantics in CLIP's text encodings.
For the latter, we sample images from Text2LIVE wherein it severely distorts the output image, and intuitively found that application of STYLREL leads to little or no improvement in such cases (see Fig 8b).

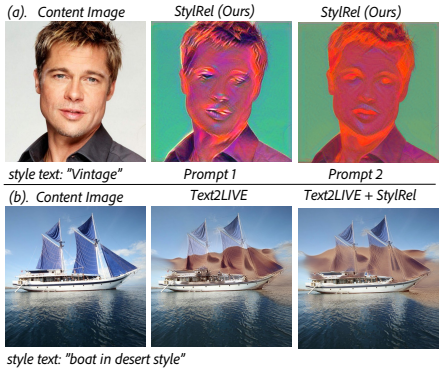

Figure 8. **Limitations**. (a) CLIP text embeddings for longer and twisted question prompts can lose semantics and affect the performance of our approach. We experiment with two such prompts (denoted Prompt 1 & 2), refer Sec 5.4-Limitations for their definition. (b) In cases where baseline (here Text2LIVE) fails considerably, STYLREL applied on top offers little to no discriminable correction.

6. Conclusion and Future Work

In this work, we present a novel strategy for enhancing text-based style transfer by leveraging relationships between various textual style descriptions with standard styles. We propose a versatile framework, STYLREL, which seamlessly integrates into existing text-based style transfer techniques, requiring only minimal alterations. Compared to strong baseline approaches, the use of our proposed STYLREL shows significant improvements in both qualitative outputs and quantitative performance metrics. We extend STYLREL to facilitate fine-grained localized style transfer with only text instructions. Going forward, our research will focus on examining the inherent graphical relationships within style spaces to offer an interpretable understanding of these interconnections. Additionally, we intend to broaden the applicability of the relational loss concept to both text domains and diffusion models.

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021. ii

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. vi

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. ii, vii

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. i

[5] 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*, pages 11326–11336, 2022. ii

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. i

[7] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pages 717–734. Springer, 2022. ii

[8] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 717–734. Springer, 2022. i

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. i, ii, iii, v

[10] 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*, pages 2414–2423, 2016. i, ii, iii

[11] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983. i

[12] Douglas R Hofstadter. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538, 2001. i

[13] Douglas R Hofstadter and Emmanuel Sander. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books, 2013. i

[14] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Yong Zhang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *arXiv preprint arXiv:2211.10682*, 2022. iii

[15] 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*, pages 1501–1510, 2017. ii

[16] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. i

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. ii

[18] Chanda Grover Kamra, Indra Deep Mastan, and Debayan Gupta. Sem-cs: Semantic clipstyler for text-based image style transfer. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 395–399. IEEE, 2023. ii

[19] 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*, pages 8110–8119, 2020. ii

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

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. vii

[22] 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*, pages 18062–18071, 2022. i, ii, iii, iv, v, vi

[23] Haochen Li. A literature review of neural style transfer. 2018. i

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. i

[25] 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*, pages 6649–6658, 2021. ii

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. vii

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. i

[28] Zhi-Song Liu, Vicky Kalogeiton, and Marie-Paule Cani. Multiple style transfer via variational autoencoder. In *2021*

IEEE International Conference on Image Processing (ICIP), pages 2413–2417. IEEE, 2021. ii

[29] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: Text-guided artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3529–3533, 2023. ii

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. ii, vi

[31] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, 2022. ii

[32] 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*, pages 5880–5888, 2019. ii

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. v

[34] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. i, ii, iii, v, vi

[35] 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*, pages 8748–8763. PMLR, 2021. i, ii, v

[36] 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*, pages 8821–8831. PMLR, 2021. i

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. iii

[38] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. iii, vii

[39] 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. i

[40] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. i

[41] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. i

[42] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*, 2016. ii

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. v

[44] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14618–14627, 2021. ii

[45] Keiji Yanai and Ryosuke Tanno. Conditional fast style transfer network. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 434–437, 2017. i

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. i

[47] Zhenling Yang, Huacheng Song, and Qiunan Wu. Generative artisan: A semantic-aware and controllable clipstyler, 2022. ii, iii, v, vi