Gatha: Relational Loss for enhancing text-based style transfer

Surgan Jandial, Shripad Deshmukh, Abhinav Java, Simra Shahid, Balaji Krishnamurthy
Adobe MDSR
{jandial,sdeshmu,ajava,sshahid,kbalaji}@adobe.com

Abstract

Text-based style transfer is a promising area of research that enables the generation of stylistic images from plain text descriptions. However, the existing text-based style transfer techniques do not account for the subjective nature of prompt descriptions or the nuances of style-specific vocabulary during the optimization process. This severely limits the stylistic expression of the predominant models. In this paper, we address this gap by proposing Gatha, which incorporates subjectivity by introducing an additional loss function that enforces the relationship between stylized images and a proxy style set to be similar to the relationship between the text description and the proxy style set. We substantiate the effectiveness of Gatha through both qualitative and quantitative analysis against the existing state-of-the-art models and show that our approach allows for consistently improved stylized images.

1. Introduction

The recent developments in Large Language Models (LLMs [2, 18, 27, 30]) and Visual Language Models (VLMs [16, 22, 23, 25, 28]) has paved the way for many existing and new applications. One such application – Style transfer – has traditionally been image-based [4, 7, 11, 15, 29], which offers limited expressiveness in terms of complex styles. For instance, consider the prompt: *'Create an image that blends the classical style of a Renaissance painting with the futuristic aesthetic of a cityscape'*. This prompt involves merging two distinct styles, which is challenging to express through the visual modality. On the contrary, with natural language, one can describe an endless array of styles - with the power of LLMs, text-based style transfer [5, 31] has begun addressing this issue.

The recent methods in the nascent field of text-conditioned style transfer, *e.g*. CLVA [5], StyleCLIP [21] and CLIPstyler [14], show tremendous promise in terms of the quality and diversity of the generated outputs. These approaches base their solution on aligning the directions of text embeddings and image embeddings [21], in that they direct the stylized image to be faithful to the direction of the text. Nevertheless, these models do not consider the subjective nature of style descriptions or the nuances of style-specific vocabulary in their optimization process. Therefore, moving the image in the direction of the text vector may not necessarily ensure that the relationship b/w the image and the style is the same as the relationship b/w text and style.

Thus, we introduce **Gatha**, a framework to incorporate a relational loss in style transfer systems. More specifically, we sample a proxy set of well-known style templates (see Fig. 1(a)), and hypothesize that the relation/similarity of these templates with the stylized image should match their relation/similarity with the target text description. We show the efficacy of Gatha, by employing it with the current state-of-the-art approach, CLIPstyler [14] as the baseline. We demonstrate how Gatha, with simply an additional loss, produces consistently better outputs than the baseline (CLIPstyler), both qualitatively and quantitatively. Also, this modification of loss requires changing only a few lines of code. Owing to its simplistic nature, Gatha can be transferred to other text-guided computer vision approaches via the introduction of an arbitrary style basis. Our contributions can be summarized below:

1. We propose a simple framework, Gatha, that incorporates the subjective nature of style descriptions.
2. Our framework achieves this by leveraging a proxy set of well-known styles (in text) and then enforcing the relationship of the target style description and style set to be similar to that of the stylized image and style set.
3. We empirically show the efficacy of Gatha over the existing state-of-the-art techniques in text-guided style transfer.

2. Related Works

Our work belongs to the Text-guided Style Transfer research area [5, 14, 21], which modifies the content image using a natural language description of the desired style, leveraging the multimodal capabilities of VLMs such as CLIP [22], BLIP [16], and Kosmos [9].

While **neural style transfer** approaches [3, 4, 7, 10–13, 15,17,20,26,29], which rely on a target-style image as input were popular in the past, they have limited capability and expressiveness for transferring complex styles that require a textual description.

In the past, **text-guided style transfer** methods have utilized contrastive learning techniques to create representations of both text and image [5, 21]. The current state-of-the-art, CLIPstyler [14], has introduced a patch-level image-text matching loss that employs CLIP [22] embeddings along with several image augmentations to enhance the quality of text-guided transfer. While some extensions have been suggested [1, 8, 19], they are trained on general-purpose in-the-wild image captioning datasets and don't include style-specific knowledge, which presents an opportunity for improvement.

3. Preliminaries of Text-guided Style Transfer

Recent state-of-the-art style transfer methods like StyleCLIP [21], StyleGAN-NADA [6] and CLIPstyler [14] leverage the representation capability of CLIP [22] and formulate style transfer as,

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

where f is the stylized image generator (U-Net [24]), C is the frozen CLIP model, L_{dir} is the directional loss, and $L_{content}$ is the content loss.

3.1. Directional Loss(L_{dir})

This represents the cosine distance between the *direction vectors* of text and image modalities. Concretely, we first compute unit vector joining clip text embeddings of T_{in} – the placeholder textual description of content image I_{in} (*"a photo"*), and T_{tar} – the user-specified target style description (*"A vintage photo of Brad Pitt."*), respectively. Likewise, we compute unit vector joining clip image embeddings of I_{in} and its desired stylised output $I_{out} = f(I_{in})$, respectively. Mathematically, if C_I and C_T denote the clip image and text encoders, the direction vectors are given as,

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

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

The directional loss L_{dir} is then given by

$$L_{dir} = 1 - T_{dir} \cdot I_{dir} \tag{2}$$

CLIPstyler applies L_{dir} at two levels: global image level (${L_{dir}}^{glob}$), and image patch level (${L_{dir}}^{patch}$).

3.2. Content Loss($L_{content}$)

This represents the mean-squared error between the features of content and output stylized images both extracted from the pre-trained VGG-19 networks [7]. Additionally, [14] and other methods also deploy a total variation regularization loss L_{tv} to handle the artifacts from irregular pixels.

4. Proposed Methodology

We propose a novel approach to style transfer that goes beyond directional losses (Eqn 2) commonly used in the field. Our method introduces a relational loss that captures the nuanced relationships between image and text descriptions and improves the discriminative ability of the underlying style transfer model. We achieve this by comparing the relationship that a generated image forms with a set of style templates to the relationship that its target style text description forms with the same style templates.

In the following sections, we describe the building blocks of our model: Style Templates, Style Tensor, and Relational Loss. Then we show the step-by-step architecture of our model in figure 1.

4.1. Creating the Style Tensor

We demonstrate the creation of style tensor in Figure (a) 1. We first collect a list of 261 commonly used style templates from popular design platforms. Each style template (or basis) is preprocessed into a question format, and the resulting list of questions is encoded into a style tensor using an embedding size of 512. The question prompt is of the following format – *"Is the style <Style>?"*.

The style tensor, denoted as S, has dimensions of $N \times 512$, where N is the number of style templates in the list. We encode the basis in a style tensor S using C_T as follows:

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

4.2. Relational Loss (L_{gatha})

Our proposed relational loss aims to ensure that the relationship between the generated stylized image and the style tensor is similar to the relationship between the target style text description and the style tensor. This captures the nuanced relationships between the image and text descriptions more effectively, resulting in better performance in the style transfer task. We discuss the comparison of the relational loss with directional loss in Figure 2.

To compute the relationship between the target text description T_{tar} and the style tensor S, we use a similarity

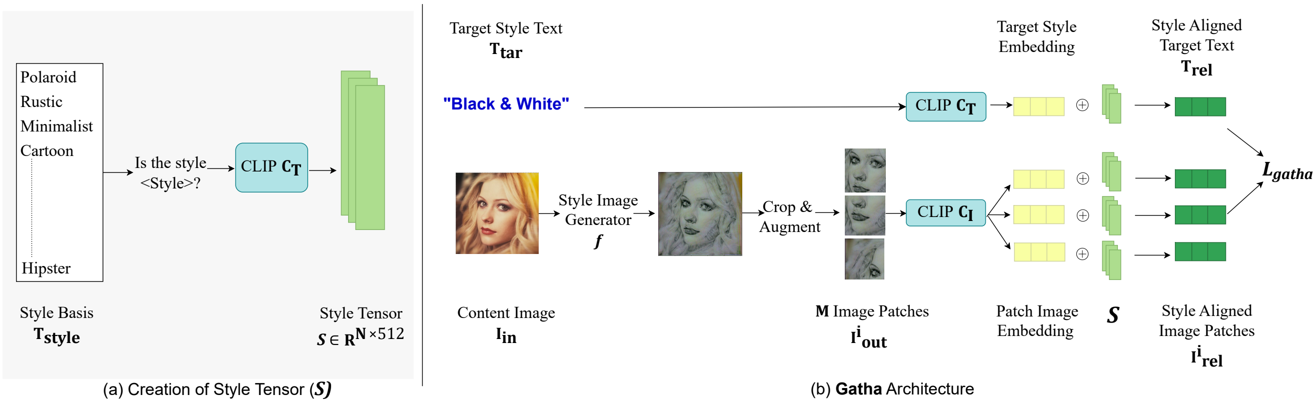

Figure 1. (a) (Left) demonstrates the creation of the Style Tensor S from what we call as our Style Basis. First, each style is prompted using the template **'Is the style ⟨Style⟩?'**, and further embedded through the CLIP C_T encoder into S . (b) (Right) illustrates the proposed flow of **Gatha**, which uses the Style Tensor S to establish a relationship between Patched Images and Target Style Text. Further, L_{gatha} enforces a relational constraint in the proxy manifold of the style tensor which improves the style transfer model's discriminative ability. . Note that we currently use only one template for prompting, and that we can extend this for an improved L_{gatha}.

score T_{rel}, obtained by projecting the target text embedding $C_T(T_{tar})$ over the style tensor S.

$$T_{rel} = S_{N\times 512} \times (C_T(T_{tar}))^T{}_{512\times 1} \tag{4}$$

where $T_{rel} \in \mathcal{R}^{N\times 1}$ is the relation vector independent of the embedding dimension of the VLM, where N is the number of styles in the style tensor.

For the stylized image I_{out}, we consider its patches and encode their relationships independently. We first augment each patch, compute its clip image embedding $C_I(aug(I^i_{out}))$, and then compute its similarity score I^i_{rel} as follows:

$$I^i_{rel} = S_{N\times 512} \times (C_I(aug(I^i_{out})))^T{}_{512\times 1} \tag{5}$$

To incorporate the relational loss, L_{gatha}, we measure the mean squared error between the image and text style relation vector averaged over all patches -

$$L_{gatha} = \frac{1}{M}\sum_{i=1}^{M} \|I^i_{rel} - T_{rel}\|_2^2 \tag{6}$$

5. Experiments and Results

In this section, we present qualitative and quantitative results for our approach. To ensure consistency, the dataset along with the resolution of content images and the other configuration settings exactly follow CLIPstyler (CS) [14].

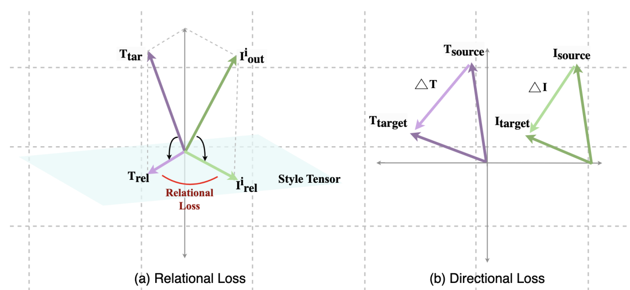

Figure 2. This figure demonstrates the contrast between the directional loss and our proposed relation loss. Figure (a) (Left) illustrates how a relational loss, such as L_{gatha}, is computed in vector terms. The image and text vectors are projected to a Style Space, and the relational loss computes the distance between them in this manifold. Figure (b) (Right) shows the directional loss, where the vector connecting the generated and source images ΔI is required to be parallel to the vector connecting the source and target text, ΔT.

5.1. Result Discussion

Table 1 corroborates our quantitative findings. We report CLIP score [14] and SSIM as our quantitative metric. CLIP score evaluates the correlation of the stylized image with the style text, and SSIM evaluates its structural similarity with the original content image. Table 1(a) demonstrates the importance of individual losses (L_{dir}, L_P) in CS, while Table 1 (b) demonstrates quantitative measures of our method over multiple weight (β) choices.
We observe the original CS configuration achieves overall higher metrics than its counterparts. Though CS without (w/o) L_P achieves higher SSIM, it's significantly low CLIP score suggests that it was not able to stylize the

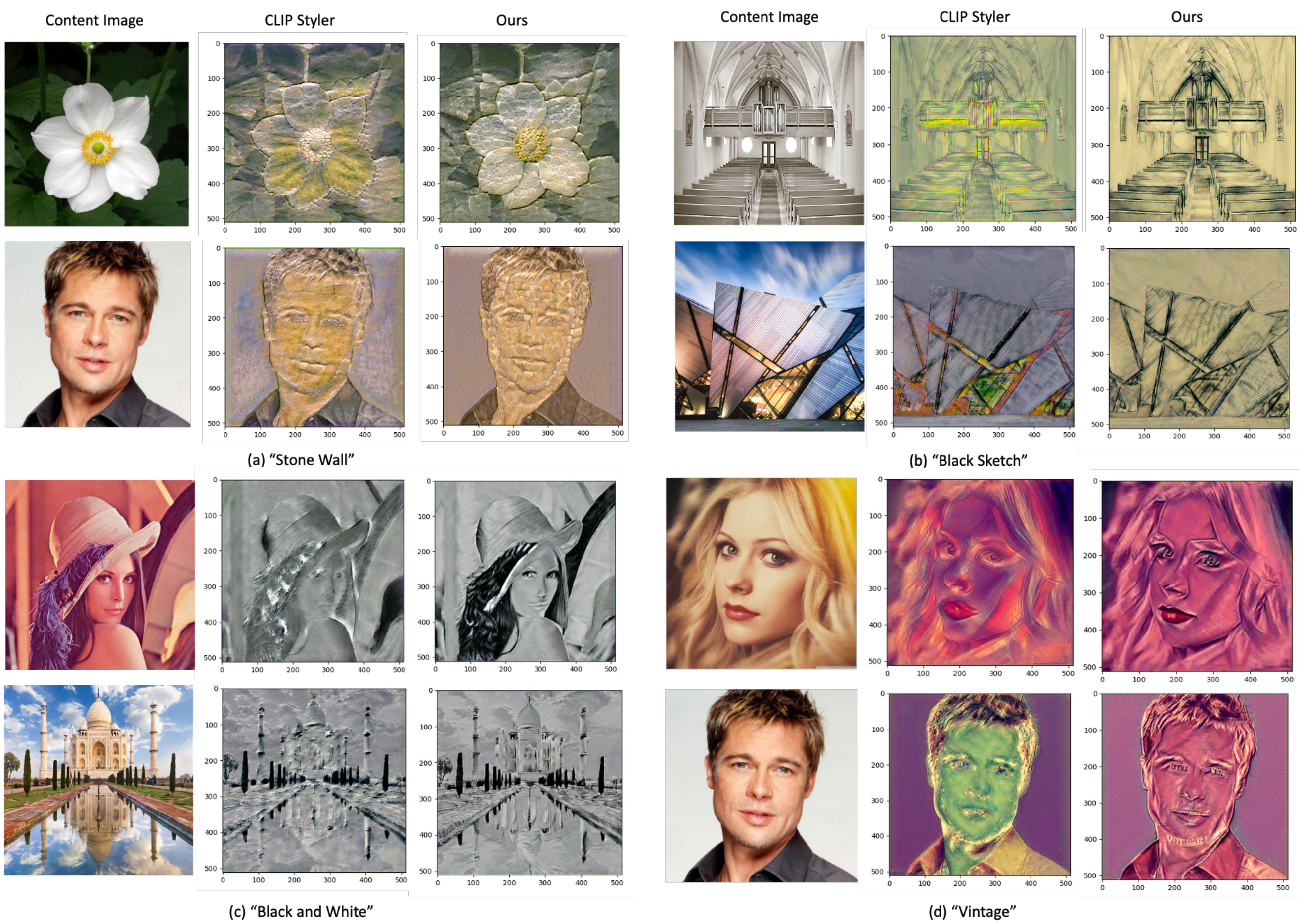

Figure 3. Qualitative comparison of L_{gatha} v/s CLIP Styler. L_{gatha} better preserves the style understanding and produces more realistic samples during the transfer.

image, and so the image almost looks similar to the original (unstylized) image. In contrast to this, in Table 1 (b), we observe our approach to outperform CS by considerable margins in both CLIP scores and SSIM. Furthermore, Fig. 3 qualitatively establishes the efficacy of Gatha over CS across multiple reference images, and text prompts.

Setting	CLIP Score	SSIM
CS	23.82	0.3895
CS w/o $L_{dir}{}^{patch}$	19.73	**0.4005**
CS w/o $L_{dir}{}^{glob}$	23.33	0.3886
CS w/o L_{TV}	**23.91**	0.3880

(a)

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

(b)

Table 1. (a). Performance of CLIPstyler across a variation of losses. (b). Performance of L_{gatha} across multiple β parameters.

Metrics	CS	I	II	III
CLIP Score	23.82	24.07	**24.18**	24.02
SSIM	0.3895	**0.3913**	0.3887	0.3909

Table 2. Experimenting L_{gatha} with different style basis {I, II, III}, where CS is CLIPstyler.

5.2. Ablations

We experiment L_{gatha} with different weight (β) parameters in Table. 1(b), and observe that increasing β beyond a point results in worse behaviors. Also, style labels used in our loss functions are currently sampled randomly across the web, and hence we probe the performance effects of their choice. In Table 2, we thus consider 3 different types of style basis and find that L_{gatha} by its design is able to show improvements in all three cases, irrespective of their variations.

References

[1] Yunpeng Bai, Jiayue Liu, Chao Dong, and Chun Yuan. Itstyler: Image-optimized text-based style transfer. *arXiv*

preprint arXiv:2301.10916, 2023. ii

[2] 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

[3] 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*, pages 872–881, 2021. ii

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

[5] 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, ii

[6] 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. ii

[7] 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

[8] 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. ii

[9] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. ii

[10] 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

[11] 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, ii

[12] 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

[13] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Bjorn Ommer. A content transformation block for image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10032–10041, 2019. ii

[14] 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

[15] Haochen Li. A literature review of neural style transfer. 2018. i, ii

[16] 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, ii

[17] 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

[18] 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

[19] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022. ii

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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. ii

[25] 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

[26] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd

style transfer. In *proceedings of the European conference on computer vision (ECCV)*, pages 698–714, 2018. ii

[27] 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

[28] 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

[29] 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, ii

[30] 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

[31] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. i