PARASOL: Parametric Style Control for Diffusion Image Synthesis

Gemma Canet Tarrés[1], Dan Ruta[1], Tu Bui[1], John Collomosse[1,2]
[1]University of Surrey, [2]Adobe Research
{g.canettarres, d.ruta, t.v.bui, j.collomosse}@surrey.ac.uk

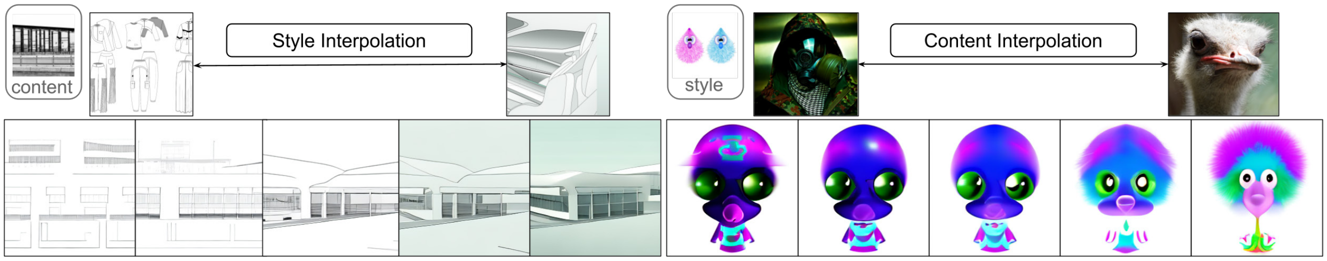

Figure 1. Images generated using PARASOL. Leveraging parametric style control and multi-modal synthesis, we demonstrate the ability of our model to synthesize creative images by interpolating different styles (left) or contents (right).

Abstract

We propose PARASOL, a multi-modal synthesis model that enables disentangled, parametric control of the visual style of the image by jointly conditioning synthesis on both content and a fine-grained visual style embedding. We train a latent diffusion model (LDM) using specific losses for each modality and adapt the classifer-free guidance for encouraging disentangled control over independent content and style modalities at inference time. We leverage auxiliary semantic and style-based search to create training triplets for supervision of the LDM, ensuring complementarity of content and style cues. PARASOL shows promise for enabling nuanced control over visual style in diffusion models for image creation and stylization, as well as generative search where text-based search results may be adapted to more closely match user intent by interpolating both content and style descriptors.

1. Introduction

Deep generative models have immense potential for creative expression, yet their controllability remains limited. While diffusion models excel in synthesizing high-quality and diverse outputs, their fine-grained attribute control, especially in visual style, is often limited by coarse-grained inputs such as textual descriptions [35], structural visual cues [53] or style transfer [5, 55]. As shown in Fig. 2, these inputs present significant limitations: (i) they restrict the nuances that can be inherited from style inputs, (ii) without specifically disentangling both attributes, they hinder the model's ability to distinguish between content and style information. In contrast, visual search models often use *parametric style embeddings* to achieve this more nuanced control. Leveraging such embeddings for guiding image synthesis, we propose Parametric Style Control (PARASOL) to bridge this gap. PARASOL is a novel synthesis model that enables disentangled parametric control over the fine-grained visual style and content of an image, conditioning synthesis on both a semantic cue and a fine-grained visual style embedding [37]. We show how the use of parametric style embeddings also enable various applications, including (i) interpolation of multiple contents and/or styles (Fig. 1), (ii) refining generative search. Additionally, for enhanced user control, we introduce test-time features in our pipeline that enable more control over the influence of each attribute on the output. Our approach holds relevance in real-world contexts such as fashion design, architectural rendering, and personalized content creation, where precise control over image style and content is essential for creative expression and practical utility. Thus, our technical contributions are:

Fine-grained style-conditioned diffusion. We synthesize images using a latent diffusion model (LDM) conditioned on multi-modal input describing independent content and style descriptors. A joint loss is introduced for encouraging disentangled control between the two modalities of interdependent nature. At inference time we invert the content image back to its noised latent, and re-run the denoising process incorporating content and style conditioning as well as modality-specific classifier-free guidance, enabling fine-grained control over the influence of each modality.

Cross-modal disentangled training. We use auxiliary semantic and style based search models to form triplets (content input, style input, image output) for supervision of the LDM training, ensuring complementarity of content and style cues to encourage disentangled control at inference.

Extensive evaluation We evaluate our model thoroughly by comparing to conditional generation models and style transfer methods and show that PARASOL outperforms the state-of-the-art in various metrics and user studies.

2. Related Work

Style transfer and representation. Neural style transfer (NST) has classically relied upon aligning statistical features [19, 28] extracted from pretrained models (e.g. VGG [12]) in the output image with those of an exemplar style image. More recently, style feature representations are learned via self-attention [31, 41]. PAMA [29] further improve upon this through progressive multi-stage alignment of features, improving consistency of style across the stylized image. ContraAST [5] introduce contrastive losses and domain-level adversarial losses to improve similarity of stylized images to real style images. CAST [55] refine this by including ground truth style images into the contrastive objective. InST [56] leverage inversion in diffusion models for transferring a new style. NST methods aim to alter texture while retaining exact content of an image, hindering creativity and controllability in image generation. In this work, we aim at bridging this gap.

Cross-modal/Style search. Early cross-modal retrieval work focused on canonical correlation analysis [13, 24]. Deep metric learning approaches typically explore dual encoders unified via recurrent or convolutional layers combined with metric learning over joint embeddings. Recently transformers are popular in such frameworks, including BERT [22] and derivatives for language and ViT [9] based image representations. Vision-language models (e.g. CLIP [32]) have been shown effective for search and conditional synthesis.

In terms of style, earlier style transfer work [10] explore representation learning of artistic style in the context of fine art, learning 10 styles to generalize a style transfer model to more than a single style. [6] leverage a labelled triplet loss to learn a metric style representation over a subset of the BAM dataset [46], but are limited by the 7 available style labels in BAM. ALADIN [37] first explored a fine-grained style representation learning through multi-layer AdaIN [19] feature extraction, trained over their newly introduced BAM-FG dataset. This was evolved in StyleBabel [38] by replacing the architecture with a vision transformer [9], which we use in our work.

Multi-modal conditional image generation. Due to the outstanding quality of recent image generation methods, conditional image generation is currently becoming a focus of attention in research. These aim to achieve a more controllable synthesis. Most conditional methods consider a single input modality (*e.g.* text [14, 30, 34, 35, 39], bounding box layouts [44, 45, 57], scene graphs [1, 21]) and a few accept multi-modal conditions. VAEs [23] have often been used for this task, since they allow learning a joint distribution over several modalities while enabling joint inference given a subset [25, 43, 48].

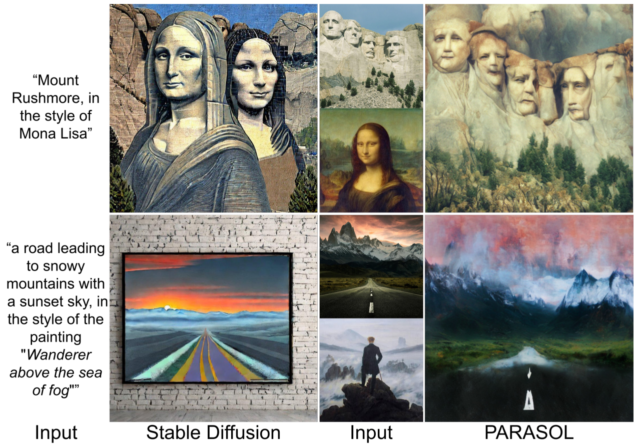
“Mount Rushmore, in the style of Mona Lisa”
“a road leading to snowy mountains with a sunset sky, in the style of the painting "*Wanderer above the sea of fog*""
Input
Stable Diffusion
Input
PARASOL

Figure 2. Comparison to Stable Diffusion [35]. Stable Diffusion encounters difficulty in disentangling content and style as well as transferring the particular requested style, while PARASOL adeptly combines fine-grained details of both into its output.

GAN-based methods follow different conditioning strategies. TediGAN [50] relies on a pretrained unconditional generator, IC-GAN [4] is based on a conditional GAN that leverages search for synthesizing new images. PoE-GAN [20] uses a product of experts GAN to combine several modalities (sketch, semantic map, text) into an image.

Make-A-Scene [11] and CoGS [15] leverage the benefits of transformers and learned codebooks for incorporating multimodal conditioning. They both encode each modality independently into discretized embeddings that are modelled jointly by the transformer.

More recently, several multi-modal diffusion-based image generation methods have been proposed. DiffuseIT [26] and CDCD [58] combine the different modalities by tuning specific losses. Others make use of multimodal embeddings, such as CLIP [32] for combining text and image-based inputs [18] or even retrieve auxiliary similar images to further condition the network [3, 36, 42]. eDiffi [2] uses specific single-modality encoders and incorporates all embeddings through cross-attention at multiple resolutions. ControlNet [53] and MCM [16] explore the integration of new input modalities as extra conditioning signals by leveraging the frozen Stable Diffusion model [35]. In contrast, we train our latent diffusion with explicit governing loss for each input modality (i.e. content and style).

3. Methodology

We propose PARASOL; a method to creatively synthesize new images with disentangled parametric control over the fine-grained visual style and content. The main design choices making that possible are as follows:

- By incorporating a **parametric style encoder** in our pipeline, our model is able to incorporate a more *fine-grained style* to the generated image.
- An **inverse diffusion step** is incorporated at sampling

time for enabling control over the amount of *content details and structure* to be preserved.

- A specific **classifier-free guidance** format is introduced for each modality to *independently influence* the output.
- The **metric properties** of both modality-specific encoders allow multiple styles and/or semantics to be *combined and interpolated* for generating creative content.

The core of PARASOL is a pre-trained LDM [35] fine-tuned for accommodating multimodal conditions. The pipeline (Fig. 3) consists of six components: 1) An Autoencoder ($\mathcal{E}$, $\mathcal{D}$) for encoding/decoding images into the diffusion latent space; 2) A U-Net based denoising network; 3) A parametric style encoder A enabling fine-grained control over visual appearance; 4) A semantic encoder C to express content control; 5) A projector network $\mathcal{M}()$ to bridge both modality embeddings into a same feature space; 6) An optional post-processing step for colour correction.

3.1. Obtaining Training Supervision Data via Cross-Modal Search

A dataset of triplets (x, y, s) is required to train PARASOL. In each set, the output image x is a stylistic image matching the artistic style of the style image s, and semantic content of the content image y.

This dataset is built via cross-modal search for the input modalities: content and style. Given an image x, its semantics descriptor $c_x = C(x)$ and style descriptor $a_x = A(x)$ are computed using parametric modality specific encoders (C and A). Leveraging their parametric properties, the most similar images for each modality can be retrieved by finding the nearest neighbours in the respective feature spaces.

Certain restrictions are applied to ensure disentanglement between both input modalities. First, different data is indexed for each modality's search. A set of images with stylistic and aesthetic properties is defined as the "Style Database" ($\mathcal{S}$). These are indexed using a parametric style encoder and used to find the style image s. In parallel, a set of photorealistic images with varied content is defined as the "Semantics Database" ($\mathcal{C}$) and is indexed using a parametric semantics encoder for finding the content image y.

Using a_x as query, the top k most style similar images in $\mathcal{S}$ are retrieved as candidates for the style image s. Their similarity to x in the semantics feature space is computed and they are discarded as candidates if this similarity is over a certain threshold. Finally, the style image s is picked as the closest image in $\mathcal{S}$ fullfilling all restrictions. Using the semantics description c_y as query, same procedure is conducted for finding the content image y in $\mathcal{C}$. Visualization of such triplets can be found in SuppMat.

3.2. Encoding the Style and Semantics Inputs

Given a triplet (content image y, style image s, output image x), the diffusion model is trained to reconstruct x by conditioning on y and s, after encoding the fine-grained style and content inputs through the respective encoders. Opting for y over x as the content image enables better style matching, acknowledging that perfect content alignment isn't always desirable (e.g., transferring a childish sketch style or a hand-drawn diagram requires changes in overall content shapes). Therefore, training the network using y, which has similar content structure and semantics to x, enables flexibility to accommodate a wider range of styles.

Arguably, the same encoder could be used on s and y to describe both style and content modalities, as in [3]. However, the use of modality-specific encoders pre-trained on a task-specific curated data has been shown to be beneficial for conditioning the diffusion process [39]. Furthermore, using a condition-specific encoder (i.e. style or semantics specific) strongly contributes to disentangling the content and style features. Nonetheless, using a different encoder for each modality poses a new challenge. When each modality is encoded into a distinct feature space, the LDM must be conditioned on both. Fine-tuning an LDM that was pre-trained on one conditional modality to understand and incorporate the information of two separate ones can be very data and compute demanding. Hence, we train an MLP-based *projector network* $\mathcal{M}()$ to obtain a joint space for both descriptors. $\mathcal{M}()$ takes the style descriptor a_s of s as input and returns m_s, a new embedding that lies in the same feature space as c_y, the semantics descriptor of y. This allows for the LDM to be easily conditioned on both modalities while maintaining the disentanglement and details encoded in their respective embeddings.

3.3. Incorporating Latent Diffusion Models

PARASOL is built with a diffusion backbone. This LDM has two components: An Autoencoder (consisting of an Encoder $\mathcal{E}$ and a Decoder $\mathcal{D}$) and a U-Net denoising network. For each image x, the encoder $\mathcal{E}$ embeds it into $z = \mathcal{E}(x)$, while the decoder $\mathcal{D}$ can reconstruct $x' = \mathcal{D}(z)$ from z. The diffusion process takes place in the Autoencoder's latent space. This process can be interpreted as a sequence of denoising autoencoders $\epsilon_\theta(z_t, t)$ that estimate z_{t-1} from its noisier version z_t.

Conditioning through Cross-Attention During our training, we freeze ($\mathcal{E}, \mathcal{D}$) and fine-tune the U-Net to incorporate multimodal conditions. In particular, the diffusion process needs to be conditioned on two independent signals: m_s and c_y. We adapt the denoising autoencoder ϵ_θ to condition on both signals as well as on the noisy sample z_t and the timestamp t: $\epsilon_\theta(z_t, t, m_s, c_y)$. Inspired by [3, 35], we incorporate these new signals into the U-Net backbone through the use of cross-attention by stacking the two signals m_s and c_y together and mapping them into every intermediate layer of the U-Net via cross-attention layers. As a result, at each timestep t, the output z_{t-1} of the model ϵ_θ is computed taking both conditions m_s and c_y into account.

Conditioning using Classifier-free Guidance As presented in [8, 30], samples from conditional diffusion models can be improved by the use of a classifier(-free) guidance. The mean and variance of the diffusion model are then additively perturbed by the gradient of the log-probability of

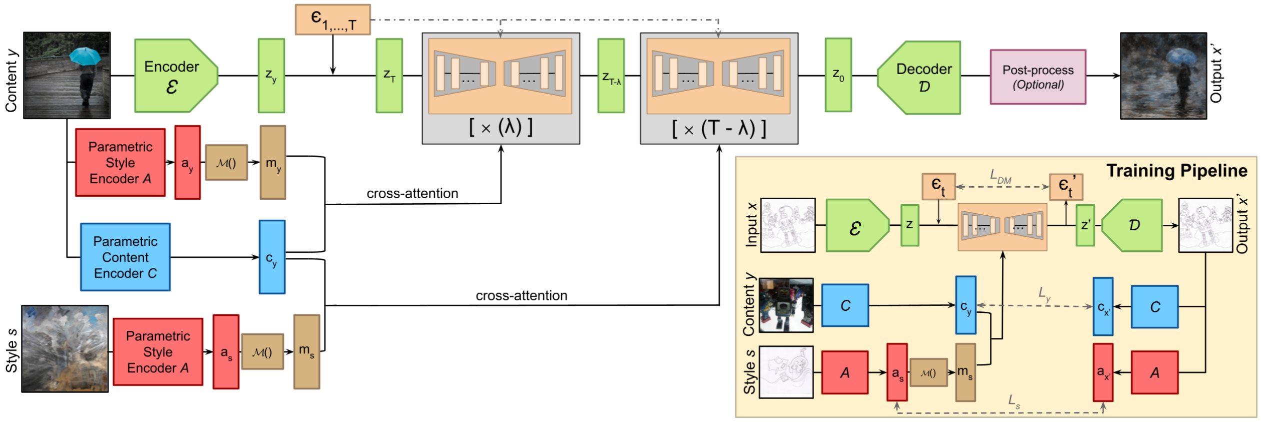

Figure 3. Illustration of the full PARASOL pipeline. It consists of six components: A parametric style encoder A (red); A projector network $\mathcal{M}()$ (brown); A semantics encoder C (blue); An Autoencoder $(\mathcal{E}, \mathcal{D})$ (green); A denoising U-Net (orange); An optional post-processing step (purple). At training time (bottom-right corner), two modality-specific losses (L_s and L_y) are used to encourage disentanglement. They are combined with L_{DM} and minimized in the training. At inference time (big pipeline), a parameter $\lambda \in [0, T]$ is introduced. After λ denoising steps, the style condition is changed for transferring a new style.

the conditioning modality leading to the generation of a particular image.

We extend the idea of classifier-free guidance for *independently* accommodating multiple modalities. At training time, both input conditions are replaced by a null condition $\epsilon_\theta(z_t, t, \emptyset, \emptyset)$ with a fixed probability so that the network can learn how to produce unconditional outputs. Then, at sampling time, this output is guided towards $\epsilon_\theta(z_t, t, m_s, \emptyset)$ and $\epsilon_\theta(z_t, t, \emptyset, c_y)$ and away from $\epsilon_\theta(z_t, t, \emptyset, \emptyset)$ as:

$$\begin{aligned} \epsilon_\theta(z_t, t, m_s, c_y) &= \epsilon_\theta(z_t, t, \emptyset, \emptyset) \\ +g_s &\left[\epsilon_\theta(z_t, t, m_s, \emptyset) - \epsilon_\theta(z_t, t, \emptyset, \emptyset)\right] \\ +g_y &\left[\epsilon_\theta(z_t, t, \emptyset, c_y) - \epsilon_\theta(z_t, t, \emptyset, \emptyset)\right] \end{aligned} \quad (1)$$

The parameters g_s and g_y are introduced to determine how much weight the style or semantics inputs have in the image generation process. Thus, by tuning the ratio of the two parameters g_s and g_y, the user can approximate the degree of influence the style and semantics inputs have over the image synthesis process. However, it should be noted that high values of either parameter will lead to higher quality and lower diversity.

3.4. Training Pipeline

At each training step, a random timestep $t \in [1, T]$ is selected. As shown in Fig. 3 (bottom-right), each training image x is encoded into z using the pre-trained encoder $\mathcal{E}$ and noised with Gaussian noise ϵ_t. Similar images in terms of style s and content y are encoded into m_s and c_y. The embedding z_t is then fed into the U-Net denoising autoencoder, which is also conditioned on m_s and c_y through cross-attention. We simultaneously train the U-Net and the projector network $\mathcal{M}()$ while freezing all other modules.

Training Objectives Training is done by minimizing a combination of 3 losses:

Diffusion Loss: Through a reparametrization trick [17, 35], minimizing the distance between the predicted noise $\epsilon_t' := \epsilon_\theta(z_t, t)$ and the true noise ϵ_t is equivalent to minimizing the distance between z_t and z_{t-1}. Thus, the loss to minimize is:

$$L_{DM} = \mathbb{E}_{z_t, \epsilon_t \sim \mathcal{N}(0,1), t}\left[\|\epsilon_t - \epsilon_t'\|^2\right]. \quad (2)$$

Modality-Specific Losses. For encouraging the output image to have the same style as s and the same semantics as y, the style and semantics of the reconstructed image x' are encoded through the use of a style and a semantics encoder, respectively. The two losses are thus computed as:

$$L_s = \text{MSE}(a_s, a_{x'}), L_y = \text{MSE}(c_y, c_{x'}). \quad (3)$$

Total Loss. All three losses are combined in a weighted sum and simultaneously optimized:

$$L = L_{DM} + \omega_s \cdot L_s + \omega_y \cdot L_y, \quad (4)$$

with ω_s and ω_y being two weight parameters.

3.5. Sampling Pipeline

When sampling (Fig. 3), an inversion process [49] takes place to ensure the fine-grained content details in y can be preserved in the final image. First, the semantics image y is encoded via $\mathcal{E}$ and noised through a complete forward diffusion process. During this process, the noise ϵ_t introduced at each $t = 1, ..., T$ is being saved. These ϵ_t values are then used for sequentially denoising the image in the reverse diffusion process. If the input conditions are unchanged throughout the whole denoising process, the image y is faithfully reconstructed. For offering the possibility of transferring a new style while preserving all fine-grained content details in the image, $\lambda \in [1, T]$ is introduced. In the first λ denoising steps, the U-Net is conditioned through cross-attention on the style and semantics descriptors of y, while in the last $T - \lambda$ steps, the style condition is switched to m_s, the encoded style from s (*i.e.*, for $T = 50$, $\lambda = 20$, the style and semantics of y are used in 20 steps and the target style in the remaining 30). Therefore, setting λ close to T leads to more structurally similar images to y while lower λ values generate images stylistically closer to s.

Colour Distribution Post-Processing A challenge with perceptually matching the style of a reference image s through diffusion, is mis-matched colour distribution, despite image fidelity to y. We offer the possibility of addressing this issue via additional post-processing steps inspired by ARF [52]. In this optional step, the generated image is modified to match the mean and covariance of the style image, shifting the colour distribution.

4. Experiments

We discuss our experimental setup and evaluation metrics in Section 4.1 and compare to baselines (Section 4.2) and different ablations (Section 4.3). We showcase user control of PARASOL in experiments in Section 4.4, and its use for interpolation and as a generative search model for improving fine-grained control is investigated in Section 4.5. See SuppMat for more experiments and visualizations.

4.1. Experimental Setup

Network and training parameters. We use a pre-trained ALADIN [37] as parametric style encoder and the pre-trained "ViT-L/14" CLIP [32] for encoding the content input. The Autoencoder and U-Net are used from [3]. The U-Net is fine-tuned using a multimodal loss (Eq. 4) with weights $\omega_s = 10^5$ and $\omega_y = 10^2$. At sampling time, unless stated otherwise, we use $\lambda = 20$, $g_s = 5.0$, and $g_y = 5.0$ and no post-processing in our experiments and figures. The training of PARASOL takes $\sim$ 10 days on an 80GB A100 GPU, while the sampling of an image takes $\sim 5 - 90$s depending on T and λ.

Dataset. The final model and all the ablations are trained using our own set of 500k triplets obtained as described in Section 3.1. Images x are obtained from BAM-FG [37], while style images s are extracted from BAM (Behance Artistic Media Dataset) [47] and content images y are parsed from *Flickr*. BAM-FG is a dataset that contains 2.62M images grouped into 310K style-consistent groupings. Only a subset of stylized non-photorealistic images from BAM-FG with semantically rich content are considered for building our training triplets.

For the Generative Search experiment, a subset of 1M images from BAM is indexed and used for search while ensuring no intersection with the images used for training.

Evaluation metrics. We measure several properties of the style transfer quality in our experiments. LPIPS [54] as a perceptual metric for the semantic similarity between the content and stylized image, SIFID [40] to measure style distribution similarity, and Chamfer distance to calculate colour similarity. We normalize Chamfer distance by the number of pixels in the image to maintain comparable values for any image resolution. We compute the MSE based on ALADIN embeddings to measure style similarity between the synthesized image and the input style and the MSE based on CLIP embeddings for quantifying its semantic similarity to the content input.

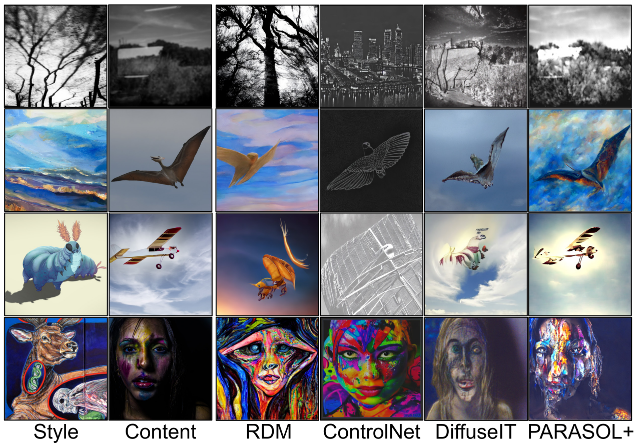

Figure 4. Comparison to Generative Multimodal Models (RDM [3], ControlNet [53], DiffuseIT [26]) and PARASOL+.

4.2. Comparison to State-Of-The-Art Methods

We separately compare to (i) generative multimodal diffusion-based models, (ii) NST models. For fairness, we fine-tune each of them on our training data.

A) Comparison to Generative Multimodal Models We compare to several SoTA models that can be conditioned on an example-based style. *RDM [3]* share its backbone with PARASOL, the main difference being using CLIP for encoding all conditions. We fine-tune it on our training triplets. *ControlNet [53]* propose the training of a neural network for incorporating task-specific conditions into the generation process. Since it only accepts textual prompts as input, we extract captions from our content images using BLIP [27] and train the network to be conditioned by style images via our training triplets. *DiffuseIT [26]* propose a diffusion-based image translation method guided by a semantic cue. Several losses ensure consistency in style and structure. Since no training code is provided, we resource to their publicly available model pretrained on ImageNet.

Visual examples of all generative baselines are shown in Fig. 4 and quantitative evaluations using the metrics described in Section 4.1 are provided in Tab. 1 (top).

Our complete method produces the best SIFID and ALADIN-MSE scores proving to accurately transfer the specific input style. The measure of semantics similarity to the content input through CLIP-MSE and LPIPS is comparable to the other methods. Although our Chamfer score is improvable, we highlight the ability of our method for tuning the influence of style and structure. If style and colour similarity are a priority for the user, the different parameters in the model can be adapted for providing more stylistically similar images to the style input.

The optional post-processing step drastically improves Chamfer distances, and SIFID metrics. However, the LPIPS

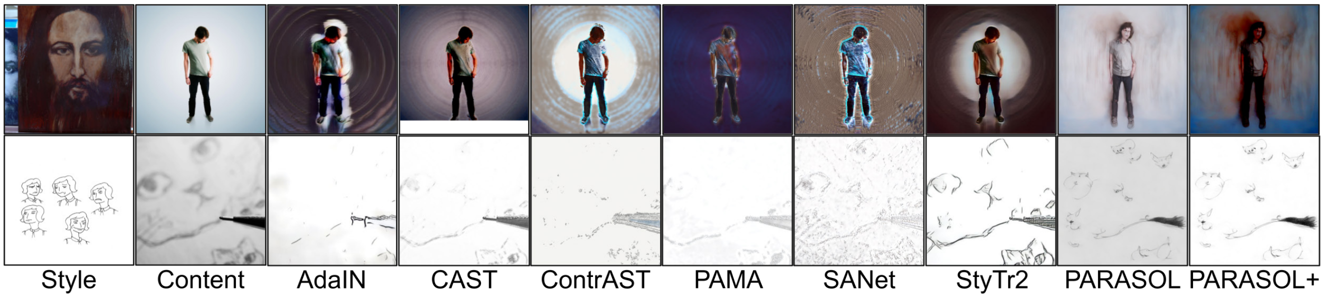

Figure 5. Comparison to Style Transfer Models (AdaIN [19], CAST [55], ContrAST [5], PAMA [29], SANet [31], StyTr2 [7]), PARASOL and PARASOL+.

Method	SIFID↓	LPIPS↓	ALADIN-MSE↓	CLIP-MSE↓	Chamfer↓
RDM [3]	3.937	0.736	3.945	**12.792**	1.431
ControlNet [53]	4.265	0.757	5.148	16.944	3.612
DiffuseIT [26]	2.572	0.677	4.460	15.159	**0.184**
PARASOL	2.994	**0.525**	4.054	15.12	1.847
PARASOL+	**2.293**	0.679	**3.891**	14.999	0.340
AdaIN [19]	2.290	0.631	4.107	16.129	0.456
CAST [55]	6.106	0.608	4.633	16.732	0.188
ContrAST [5]	2.791	0.651	4.171	16.111	**0.126**
PAMA [29]	2.509	0.659	4.411	16.194	0.287
SANet [31]	2.608	0.671	4.200	16.187	0.301
StyTr2 [7]	**1.966**	0.588	4.057	16.083	0.535
PARASOL	2.994	**0.525**	4.054	15.12	1.847
PARASOL+	2.293	0.679	**3.891**	**14.999**	0.340

Table 1. Quantitative comparison of style (SIFID, ALADIN-MSE, Chamfer) and content (LPIPS, CLIP-MSE) metrics. (*Top:* Generative Multimodal Models; *Bottom:* Style Transfer Models) We include our optional post-processing as PARASOL+. Chamfer scores are scaled down $\times 10^{-3}$.

Method	Pref. Overall	Pref. Style Fidelity	Pref. Content Fidelity
RDM [3]	17.60%	18.40%	9.20%
ControlNet [53]	1.20%	0.00%	1.60%
DiffuseIT [26]	29.20%	34.80%	27.20%
PARASOL	**52.00%**	**46.80%**	**62.00%**
CAST [55]	20.40%	11.20%	33.60%
ContrAST [5]	15.20%	12.00%	10.40%
PAMA [29]	17.60%	8.40%	8.40%
SANet [31]	16.40%	9.60%	7.20%
PARASOL	**30.40%**	**58.80%**	**40.40%**

Table 2. Evaluation of our method vs. different baselines based on AMT experiments. (*Top:* Generative Multimodal Models; *Bottom:* Style Transfer Models). Given a content image, a style image and a set of images generated using our method and different baselines, we conduct three separate experiments. In each experiment, workers are asked to choose their preferred image based on: 1) image quality, 2) style fidelity, and 3) content fidelity. For fairness, we compare to PARASOL without the post-processing step.

scores are worse. While this is not unexpected, it highlights a weakness with LPIPS, as a metric in measuring content similarity. The re-colouring steps change only the colours in the image, not modifying any content.

We undertake a user study using Amazon Mechanical Turk (AMT), to support our quantitative metrics. As show in Tab. 2 (top), PARASOL was chosen as the preferred method in all categories, measured via majority consensus voting (3 out of 5 workers).

B) Comparison to Style Transfer Models Since most diffusion-based models are text-based, we mostly resource to state of the art non-diffusion based models. We evaluate our method against five recent NST models (CAST [55], PAMA [29], SANet [31], ContraAST [5] and StyTr2 [7]), as well as a traditional one (AdaIN [19]). Each of these methods is prompt-free and directly comparable to our method. Note that PARASOL is *not* designed to be a style transfer model. However, we display in Tab. 1 (bottom) how our model is able to preserve both style and content comparably to the SoTA style transfer models, which were specifically trained for texture-based stylization instead of generation. Additionally, the user studies in Tab. 2 (bottom) show how users prefer PARASOL over the most popular NST SoTA models in all experiments. One benefit of PARASOL when compared to most NST models is that, by relaxing the constraint of preserving exact content, it can adapt to more complex styles while maintaining unchanged semantics (See Fig. 5).

4.3. Ablation Study

Considering the pre-trained RDM conditioned on CLIP embeddings as our baseline, we justify the addition of each of our components through an ablation study (Tab. 3).

Switching the style encoder from CLIP to ALADIN offers a lot of benefits in terms of disentanglement and fine-grained style information. However, even when fine-tuning the network on the new descriptors, it is not able to understand the nature of the new representations. Tab. 3 shows the substantial improvement of all metrics when the projector $M()$ is introduced, proving its crucial role in the pipeline. As illustrated in Fig. 6, the multimodal loss assists the network in further disentangling both modalities.

Despite some lower metrics when using inversion, we consider it beneficial overall, due to the added controllability, and improved style transfer metrics (SIFID, LPIPS), making PARASOL+ our best model overall.

4.4. Controllability Experiments

We demonstrate several ways PARASOL can be used to exert control over the image synthesis process.

Method	SIFID↓	LPIPS↓	ALADIN-MSE↓	CLIP-MSE↓	Chamfer↓
PARASOL -(I, L, M, Ft, A)	3.077	0.749	4.312	15.428	**0.127**
PARASOL -(I, L, M, Ft)	7.759	0.813	5.540	17.457	1.184
PARASOL -(I, L, M)	6.883	0.777	4.356	15.887	0.788
PARASOL -(I, L)	4.269	0.748	3.573	14.740	0.174
PARASOL -(I)	4.329	0.747	**3.564**	**14.714**	0.155
PARASOL	2.994	**0.525**	4.054	15.12	1.847
PARASOL+	**2.293**	0.679	3.891	14.999	0.340

Table 3. Quantitative evaluation metrics for different ablations of our final model. Considering pre-trained RDM as the baseline, this study justifies: 1) The use of a parametric style encoder (A); 2) The need to fine-tune the model (Ft); 3) The addition of a projector network for the style embedding (M); 4) The effect of modality-specific losses (L); 5) The design choice of inverting the diffusion process (I); 6) The use of the optional post-processing step (PARASOL+).

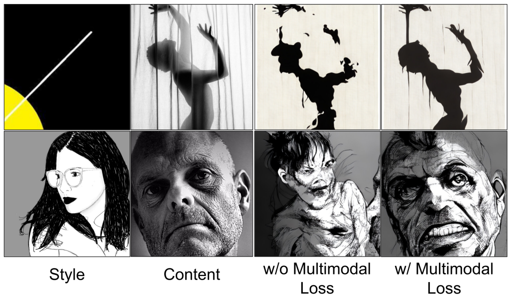

Figure 6. Effect of the multimodal loss. Adding the multimodal loss encourages the model to better combine the information from each modality when their descriptors are not fully disentangled.

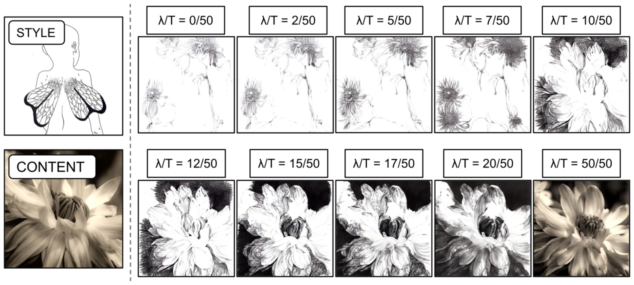

Figure 7. Effect of λ. Higher λ values lead to more content preservation, while lower values further encourage style transfer.

A) Disentangled Style and Content We propose different ways of controlling the influence of each input modality.

Via inversion: λ enables choosing at which step of the inversion process the style condition is changed (Fig. 7).

Via classifier-free guidance parameters: The values g_s and g_y (Eq. 1) determine how much weight each condition has in the image generation process. (Fig. 8).

B) Textual Captions as Conditioning Inputs Leveraging the multimodal properties of CLIP, content cues can be provided as either captions or images at sampling time. When a textual prompt is provided, our diffusion model is used for generating an image conditioned on its CLIP descriptor and an empty style descriptor. Considering this generated image as y, the sampling procedure continues as in Section 3.5.

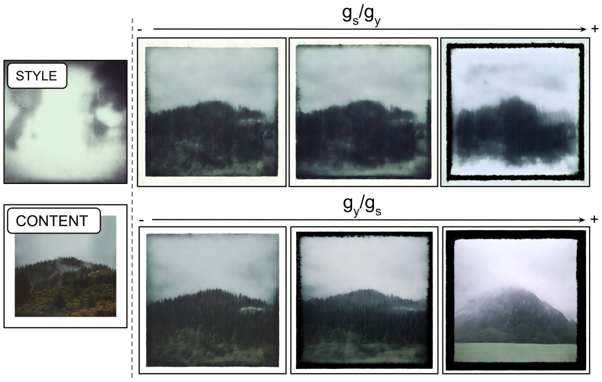

Figure 8. Effect of g_s and g_y. *Top:* as the value of g_s increases, so does the influence of the style. *bottom:* g_y is increased, offering a more notable semantic influence from the content cue.

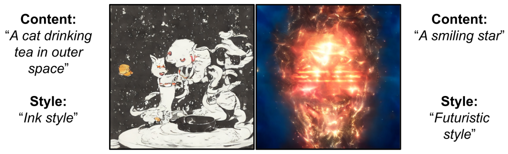

Figure 9. Images generated from text prompts for both modalities. PARASOL accepts both visual and textual inputs for better transferring the user's intent.

The style cue can also be provided in textual format. The prompt can be encoded using CLIP and projected to a joint space through a pre-trained projector network [38]. By indexing all images in the "Style Database" $\mathcal{S}$ through ALADIN and projecting them to the joint feature space, a similar style image can be retrieved and used as input s to the sampling pipeline (Fig. 9).

C) Content Diversity with Consistent Semantics Although PARASOL offers the option of preserving the fine-grained content details in y, it can also be used for synthesizing images with consistent semantics and style yet diverse details and image layouts (Fig. 10). Following a similar procedure to that in Section 4.4 (B), an image can first be generated from y and used to guide the inversion process for transferring the semantics of y with new fine-grained content details.

4.5. Applications

A) Content and Style Interpolation PARASOL can synthesize images by interpolating different styles and/or contents, unlocking the potential of generating a wider range of creative images (Fig. 11, Fig. 1). For demonstrating this capacity, a crowd-source AMT evaluation is performed (Tab. 4), positioning PARASOL as the preferred method by users in terms of both content and style interpolation.

B) Generative Visual Search PARASOL can be applied for the task of Generative Search, where search results are not constrained to a static corpus but fused with generation to yield images that more closely match a user's search in-

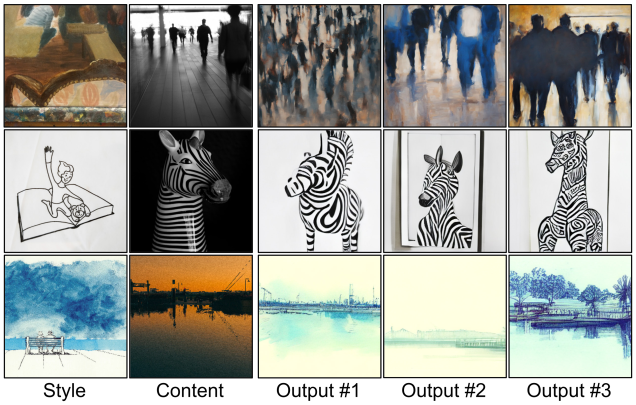

Figure 10. Diversity of fine-grained content. PARASOL can be used for synthesizing images with same semantics and fine-grained style yet diverse fine-grained content.

User Study	RDM	DiffuseIT	PARASOL
Pref. Content Interpolation	12.037%	43.518%	**44.444%**
Pref. Style Interpolation	16.808%	31.932%	**51.260%**

Table 4. AMT user evaluation of content and style interpolations, comparing PARASOL with RDM [3] and DiffuseIT [26]. Model preference is measured via majority consensus voting (3 out of 5).

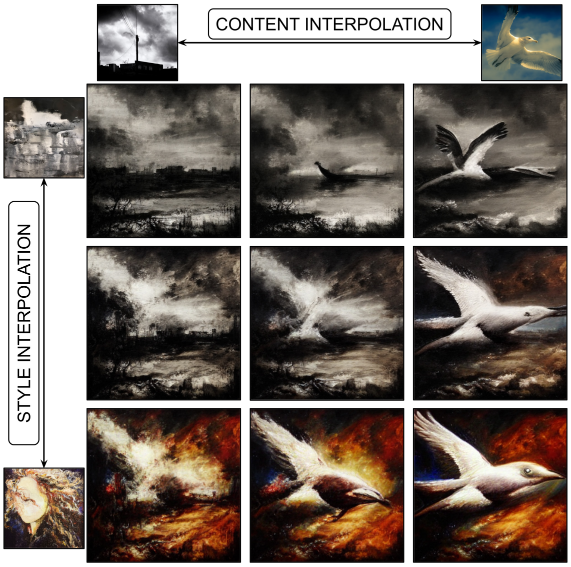

Figure 11. Style and content interpolation. Images generated by interpolating two styles and two content images.

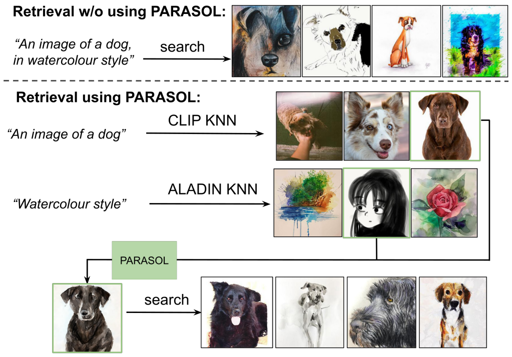

Figure 12. Generative Search. Our model can be used for aiding in image retrieval process (See Section 4.5 (B)).

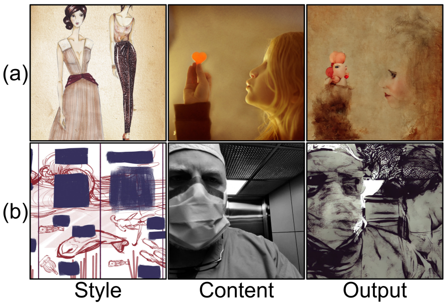

Figure 13. Failure cases - more details in Section 5.

tent. Assuming an initial query in textual form, we use pre-existing text encoders (per Section 4.4 (B)) for CLIP and ALADIN to retrieve two sets of images that match their intent semantically and stylistically (Fig. 12, middle). The user picks images from each set, each of which may be interactively weighted to reflect their relevance to the users' intent. Using the fine-grained interpolation method of Section 4.5 (A), PARASOL is then used to generate an image which may either be accepted by the user or used as a further basis for semantic and style search. PARASOL thus provides a fine-grained way to disambiguate a single text prompt by expressing a pair of modalities and can help surface images beyond those that would be available in a static text corpus.

5. Conclusion

We introduced PARASOL, a method that leverages parametric style embeddings for multimodal image synthesis with fine-grained parametric control over style. We show PARASOL can be applied to text, image or embedding conditioned generation enabling disentangled control over style and content at inference time. We also propose a novel method for obtaining paired training data leveraging cross-modal search. We show a use case in generative search, providing an image that can be used as query for a more fine-grained search.

Limitations While the use of modality-specific encoders offers numerous advantages in attribute disentanglement and parametric control, facilitating interpolation and search, certain styles can still present challenges due to ambiguity regarding which features should be considered as part of the content or style (Fig. 13a). Additionally, addressing challenging contents such as faces (Fig. 13 b) often necessitates specific additional training. Future research could explore alternative modalities (i.e. sketches or segmentation maps) to target regions locally with fine-grained cues [51].

6. Acknowledgements

This work was funded in part by DECaDE – the UKRI Centre for Decentralized Creative Economy – under the EPSRC grant EP/T022485/1.

References

[1] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 1

[3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 2, 3, 5, 6, 8, 1

[4] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021. 2

[5] 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, 2021. 1, 2, 6

[6] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *Proceedings of the IEEE international conference on computer vision*, pages 2660–2668, 2017. 2

[7] 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. 6

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 2

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 2

[12] 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. 2

[13] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 529–545. Springer, 2014. 2

[14] 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*, pages 10696–10706, 2022. 2

[15] Cusuh Ham, Gemma Canet Tarrés, Tu Bui, James Hays, Zhe Lin, and John Collomosse. Cogs: Controllable generation and search from sketch and style. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 632–650. Springer, 2022. 2

[16] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[18] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1085–1094, 2022. 2

[19] 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. 2, 6

[20] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 91–109. Springer, 2022. 2

[21] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image synthesis from reconfigurable layout and style. In *Proc. CVPR*, 2018. 2

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186, 2019. 2

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[24] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2

[25] Svetlana Kutuzova, Oswin Krause, Douglas McCloskey, Mads Nielsen, and Christian Igel. Multimodal variational autoencoders for semi-supervised learning: In defense of product-of-experts. *arXiv preprint arXiv:2101.07240*, 2021. 2

[26] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2, 5, 6, 8, 1

[27] 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. 5, 1

[28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 2

[29] Xuan Luo, Zhen Han, Lingkang Yang, and Lingling Zhang. Consistent style transfer. *arXiv preprint arXiv:2201.02233*, 2022. 2, 6

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021. 2, 3

[31] 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. 2, 6

[32] 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. 2, 5, 1, 3

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1

[34] 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. 2

[35] 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*, pages 10684–10695, 2022. 1, 2, 3, 4

[36] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022. 2

[37] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. Aladin: all layer adaptive instance normalization for fine-grained style similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11926–11935, 2021. 1, 2, 5, 3

[38] Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, Jo Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, et al. Stylebabel: Artistic style tagging and captioning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 219–236. Springer, 2022. 2, 7

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 3

[40] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 5

[41] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8242–8250, 2018. 2

[42] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. kNN-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[43] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[44] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019. 2

[45] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2647–2655, 2021. 2

[46] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, pages 1202–1211, 2017. 2

[47] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John P. Collomosse, and Serge J. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. *CoRR*, abs/1704.08614, 2017. 5

[48] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018. 2

[49] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2023. 4

[50] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 2

[51] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and Vishal M Patel. Scenecomposer: Any-level semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22468–22478, 2023. 8

[52] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 717–733. Springer, 2022. 5

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 5, 6

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[55] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 1, 2, 6

[56] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 2

[57] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019. 2

[58] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. In *The Eleventh International Conference on Learning Representations*. 2