StyleT2I: Toward Compositional and High-Fidelity Text-to-Image Synthesis

Zhiheng Li[1,2], Martin Renqiang Min[1], Kai Li[1], and Chenliang Xu[2]
[1]NEC Laboratories America, [2]University of Rochester
{renqiang,kaili}@nec-labs.com, {zhiheng.li,chenliang.xu}@rochester.edu

Abstract

Although progress has been made for text-to-image synthesis, previous methods fall short of generalizing to unseen or underrepresented attribute compositions in the input text. Lacking compositionality could have severe implications for robustness and fairness, e.g., inability to synthesize the face images of underrepresented demographic groups. In this paper, we introduce a new framework, StyleT2I, to improve the compositionality of text-to-image synthesis. Specifically, we propose a CLIP-guided Contrastive Loss to better distinguish different compositions among different sentences. To further improve the compositionality, we design a novel Semantic Matching Loss and a Spatial Constraint to identify attributes' latent directions for intended spatial region manipulations, leading to better disentangled latent representations of attributes. Based on the identified latent directions of attributes, we propose Compositional Attribute Adjustment to adjust the latent code, resulting in better compositionality of image synthesis. In addition, we leverage the ℓ_2-norm regularization of identified latent directions (norm penalty) to strike a nice balance between image-text alignment and image fidelity. In the experiments, we devise a new dataset split and an evaluation metric to evaluate the compositionality of text-to-image synthesis models. The results show that StyleT2I outperforms previous approaches in terms of the consistency between the input text and synthesized images and achieves higher fidelity.

1. Introduction

Text-to-image synthesis is a task to synthesize an image conditioned on given input text, which enables many downstream applications, such as art creation, computer-aided design, and training data generation for augmentation. Although progress has been made for this task, the compositionality aspect is overlooked by many previous methods [39]. As shown in Fig. 1, the input text "*he*[1] *is wearing lipstick*" describes an intersectional group [3] between two attributes—

		ControlGAN	DAE-GAN	TediGAN	**StyleT2I (Ours)**
Text Input: "*He* is *wearing lipstick*."					
Attribute Composition	*he*	✓	✓	✗	✓
	wearing lipstick	✗	✗	✓	✓
High Fidelity		✗	✗	✓	✓

Figure 1. When the text input contains underrepresented compositions of attributes, *e.g.*, (*he*, *wearing lipstick*), in the dataset, previous methods [30,51,64] incorrectly generate the attributes with poor image quality. In contrast, StyleT2I achieves better compositionality and high-fidelity text-to-image synthesis results.

"*he*" and "*wearing lipstick*," which is underrepresented in a face dataset [18]. The previous approaches [30,51,64] fail to correctly synthesize the image, which could be caused by overfitting to the overrepresented compositions, *e.g.*, ("*she*", "*wearing lipstick*") and ("*he*", not "*wearing lipstick*"), in the dataset. This leads to severe robustness and fairness issues by inheriting biases and stereotypes from the dataset. Therefore, it is imperative to improve the text-to-image synthesis results in the aspect of compositionality.

The crux of the compositionality problem is to prevent models from simply memorizing the compositions in the training data. First, in terms of the objective function, some previous methods [64,65] simply minimize the feature distance between pairwise matched image and text, leading to poor generalizability. In contrast, we propose a *CLIP-guided Contrastive Loss* to let the network better distinguish different compositions among different sentences, in which CLIP (Contrastive Language–Image Pre-training) [47] is pre-trained on large-scale matched image-text pairs as a foundation model [2]. Second, the compositional text-to-image model needs to be sensitive to each independent attribute described in the text. Most previous methods [30,68,71,75] mainly resort to attention mechanism [60], which focuses more on the correspondence between words and image features but falls short of separating individual attributes from a composition. Unlike previous approaches, our key idea is to identify disentangled representations [6,14] in the la-

[1]In this work, the gender and gender pronouns denote the visually perceived gender, which does not indicate one's actual gender identity.

tent space of a generative model, where each disentangled representation exclusively corresponds to one attribute in the dataset. By leveraging the disentangled representations of different attributes, we can improve the compositionality by ensuring that each attribute described in the sentence is correctly synthesized.

Motivated by these ideas, we present StyleT2I, a novel framework to improve the compositionality of text-to-image synthesis employing StyleGAN [19]. In specific, we propose a *CLIP-guided Contrastive Loss* to train a network to find the StyleGAN's latent code semantically aligned with the input text and better distinguish different compositions in different sentences. To further improve the compositionality, we propose a *Semantic Matching Loss* and a *Spatial Constraint* for identifying attributes' latent directions that induce intended spatial region manipulations. This leads to a better disentanglement of latent representations for different attributes. Then we propose *Compositional Attribute Adjustment* to correct the wrong attribute synthesis by adjusting the latent code based on identified attribute directions during the inference stage. However, we empirically found that optimizing the proposed losses above can sometimes lead to degraded image quality. To address this issue, we employ *norm penalty* to strike a nice balance between image-text alignment and image fidelity.

To better evaluate the compositionality of text-to-image synthesis, we devise a test split for the CelebA-HQ [18] dataset, where the test text only contains unseen compositions of attributes. We design a new evaluation metric for the CUB [61] dataset to evaluate if the synthesized image is in the correct bird species. Extensive quantitative results, qualitative results, and user studies manifest the advantages of our method on both image-text alignment and fidelity for compositional text-to-image synthesis.

We summarize our contributions as follows: (1) We propose StyleT2I, a compositional text-to-image synthesis framework with a novel *CLIP-guided Contrastive Loss* and *Compositional Attribute Adjustment*. To the best of our knowledge, this is the first text-to-image synthesis work that focuses on improving the compositionality of different attributes. (2) We propose a novel *Semantic Matching Loss* and a *Spatial Constraint* for identifying attributes' latent directions that induce intended variations in the image space, leading to a better disentanglement among different attributes. (3) We devise a new test split and an evaluation metric to better evaluate the compositionality of text-to-image synthesis.

2. Related Work

Text-to-Image Synthesis Many previous works [8,15,25,29,30,32,46,48,50,58,68,70–73,75] have studied text-to-image synthesis. DALL·E [48] trains dVAE [58] that autoregressively predicts the image tokens on a large-scale dataset. Zhang *et al.* [71] use cross-modal contrastive loss on real image-text and fake image-real image pairs to adversarially train the conditional GAN. In contrast, StyleT2I's *CLIP-guided Contrastive Loss* enjoys a simpler training scheme by using the pretrained CLIP as a conditional discriminator to contrast fake image-text pairs. While DAE-GAN [51] extracts aspects from the language with the attention mechanism to improve image-text alignment, StyleT2I identifies attribute's latent directions and explicitly manipulates the latent code with proposed *Compositional Attribute Adjustment*, which is more interpretable. TediGAN [64,65] uses pretrained StyleGAN [19] as the generator and trains a text encoder by deterministically minimizing the feature distances between paired image and text in either StyleGAN's latent space [64] or CLIP's feature space [65], which suffers from memorizing the dataset's compositions. TediGAN also needs to conduct a manual analysis to find the layer-wise control for each attribute. In comparison, StyleT2I automatically finds disentangled latent directions for different attributes with a novel *Semantic Matching Loss* and a *Spatial Constraint*. Wang *et al.* [62] perform text-to-face synthesis based on attribute's latent direction identified by using additional attribute labels as supervision, whereas StyleT2I does not need additional attribute labels. Tan *et al.* [57] focus on the compositionality problem for multi-object scene image synthesis. Very recently, Park *et al.* [39] propose a new benchmark, revealing that many previous methods suffer from the compositionality problem, which motivates us to propose StyleT2I to address this issue.

Disentangled Representation Unsupervised disentangled representation learning focuses on training generative models [11,24] with different latent dimensions interpreting independent factors of data variations, and most of such models are based on VAE [5,14,21,23,26] and GAN [43,63], enabling many downstream applications [27,31,55]. However, Locatello *et al.* [35] show that unsupervised disentanglement is impossible without inductive bias or supervision. Zhu *et al.* [76] modify the generative model's architecture with an additional loss to improve spatial constriction and variation simplicity. Some supervised disentanglement methods use a pre-trained classifier [53], regressor [77], or multi-attribute annotation [1] as the full supervision to identify latent attribute directions. In contrast, StyleT2I finds disentangled attribute directions in the unmodified StyleGAN's latent space based on the supervision from text, which has a much lower annotation cost than multi-attribute labels.

3. Overview of StyleT2I

Figure 2 gives an overview of our StyleT2I framework. Unlike most previous end-to-end approaches [51,68,71,75], we leverage a pre-trained unconditional generator, StyleGAN [19], and focus on finding a text-conditioned latent code in the generator's latent space that can be decoded into a high-fidelity image aligned with the input text.

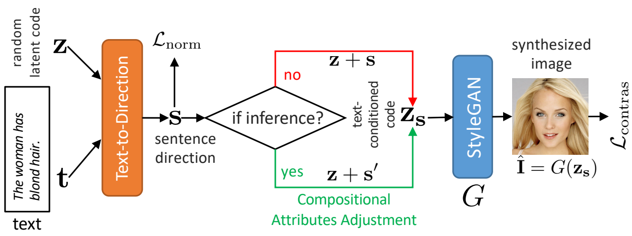

Figure 2. An overview of StyleT2I. The *Text-to-Direction* module takes a text $\mathbf{t}$ and a random latent code $\mathbf{z}$ as inputs and outputs a sentence direction $\mathbf{s}$ to edit $\mathbf{z}$, resulting in a text-conditioned latent code $\mathbf{z_s} = \mathbf{z} + \mathbf{s}$ in StyleGAN's latent space for image synthesis. The *Text-to-Direction* module is trained with novel *CLIP-guided Contrastive Loss* (Sec. 4.1) with *norm penalty* employed (Sec. 4.2). During the inference stage (lower branch), *Compositional Attribute Adjustment* (Sec. 5.3) is performed by adjusting $\mathbf{s}$ to $\mathbf{s}'$, leading to better compositionality.

To achieve this, in Sec. 4, we present a *Text-to-Direction* module (see Fig. 2) trained with a novel *CLIP-guided Contrastive Loss* for better distinguishing different compositions (Sec. 4.1) and a *norm penalty* (Sec. 4.2) to preserve the high fidelity of the synthesized image.

To further improve the compositionality of the text-to-image synthesis results, in Sec. 5, we propose a novel *Semantic Matching Loss* (Sec. 5.1) and a *Spatial Constraint* (Sec. 5.2) for identifying disentangled attribute latent directions, which will be used to adjust the text-conditioned latent code during the inference stage (Sec. 5.3) with our novel *Compositional Attribute Adjustment* (*CAA*). The pseudocode of the complete algorithm is in Appendix A.1.

4. Text-conditioned Latent Code Prediction

As many previous works [42,53,54,69,77] show that the latent direction in StyleGAN's latent space can represent an attribute—traversing a latent code along the attribute's latent direction can edit the attribute in the synthesized image, we hypothesize that there exists a latent direction that corresponds to the composition of multiple attributes described in the input text, *e.g.*, "*woman*" and "*blond hair*" attributes in text "*the woman has blond hair*." Therefore, to find a text-conditioned latent code in a pre-trained StyleGAN's latent space, we propose a *Text-to-Direction* module that takes the text $\mathbf{t}$ and a randomly sampled latent code $\mathbf{z}$ from the latent space of the pre-trained StyleGAN as input. The output is a latent direction $\mathbf{s}$, dubbed sentence direction, to edit the latent code $\mathbf{z}$, resulting in the text-conditioned code $\mathbf{z_s} = \mathbf{z} + \mathbf{s}$. As a result, $\mathbf{z_s}$ is fed into the StyleGAN generator G to synthesize the image $\hat{\mathbf{I}} = G(\mathbf{z_s})$.

4.1. CLIP-guided Contrastive Loss

The *Text-to-Direction* module should predict the sentence direction that is aligned with the input text and avoid simply memorizing the compositions in the training data. To achieve this, we leverage a foundational model CLIP [56] pre-trained on a large-scale dataset with matched image-caption pairs to learn a joint embedding space of text and image, as a conditional discriminator. We propose a novel *CLIP-guided Contrastive Loss* based on CLIP and contrastive loss [4] to train the *Text-to-Direction* module. Formally, given a batch of B text $\{\mathbf{t}_i\}_{i=1}^B$ sampled from the training data and the corresponding fake images $\hat{\mathbf{I}}_i$, we compute the *CLIP-guided Contrastive Loss* of the i-th fake image as:

$$\mathcal{L}_{\text{contras}}(\mathbf{I}_i) = -\log \frac{\exp(\cos(E_{\text{CLIP}}^{\text{img}}(\hat{\mathbf{I}}_i), E_{\text{CLIP}}^{\text{text}}(\mathbf{t}_i)))}{\sum_{j\neq i}^{B}\exp(\cos(E_{\text{CLIP}}^{\text{img}}(\hat{\mathbf{I}}_i), E_{\text{CLIP}}^{\text{text}}(\mathbf{t}_j)))}, \tag{1}$$

where $E_{\text{CLIP}}^{\text{img}}$ and $E_{\text{CLIP}}^{\text{text}}$ denote the image encoder and text encoder of CLIP, respectively. $\cos(\cdot,\cdot)$ denotes the cosine similarity. *CLIP-guided Contrastive Loss* attracts paired text embedding and fake image embedding in CLIP's joint feature space and repels the embedding of unmatched pairs. In this way, the *Text-to-Direction* module is trained to better align the sentence direction $\mathbf{s}$ with the input text $\mathbf{t}$. At the same time, *CLIP-guided Contrastive Loss* forces the *Text-to-Direction* module to contrast the different compositions in different texts, *e.g.*, "*he is wearing lipstick*" and "*she is wearing lipstick*," which prevents the network from overfitting to compositions that predominate in the training data.

4.2. Norm Penalty for High-Fidelity Synthesis

However, the experimental results (Fig. 7) show that minimizing the contrastive loss alone fails to guarantee the fidelity of the synthesized image. We observe that it makes the *Text-to-Direction* module predict $\mathbf{s}$ with a large ℓ_2 norm, resulting in $\mathbf{z_s}$ shifted to the low-density region in the latent distribution, leading to degraded image quality. Therefore, we penalize the ℓ_2 norm of sentence direction $\mathbf{s}$ when it exceeds a threshold hyperparameter θ:

$$\mathcal{L}_{\text{norm}} = \max(||\mathbf{s}||_2 - \theta, 0). \tag{2}$$

Our ablation study (Fig. 7) shows that adding the *norm penalty* strikes a nice balance between the text-image alignment and quality.

To summarize, the **full objective function** for training the *Text-to-Direction* module is:

$$\mathcal{L}_s = \mathcal{L}_{\text{contras}} + \mathcal{L}_{\text{norm}}. \tag{3}$$

5. Compositionality with Attribute Directions

To further improve the compositionality, we first identify the latent directions representing the attributes with a novel *Semantic Matching Loss* (Sec. 5.1) and a *Spatial Constraint* (Sec. 5.2). Then, we propose *Compositional Attribute Adjustment* (Sec. 5.3) to adjust the sentence direction by the identified attribute directions to improve the compositionality of text-to-image synthesis results.

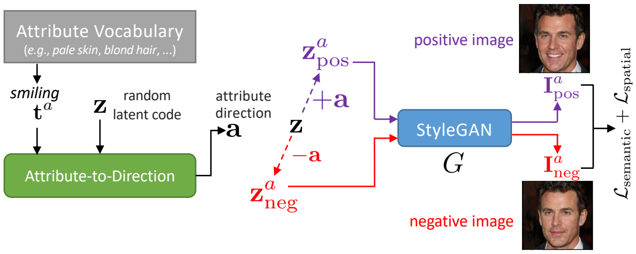

Figure 3. Identifying disentangled attribute directions by training an *Attribute-to-Direction* module with a *Semantic Matching Loss* ($\mathcal{L}_{\text{semantic}}$) and a *Spatial Constraint* ($\mathcal{L}_{\text{spatial}}$).

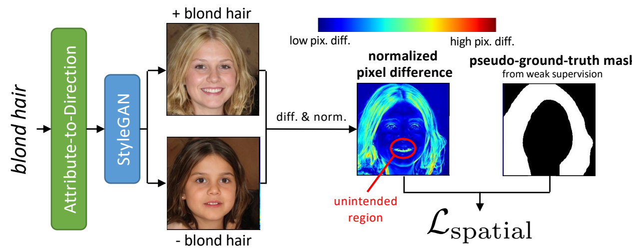

Figure 4. *Spatial Constraint* ($\mathcal{L}_{\text{spatial}}$) to train *Attribute-to-Direction* module. We compute the pixel-level difference between the positive and negative image to measure the changing region on the image space (red: high pixel differences; blue: low pixel variations). $\mathcal{L}_{\text{spatial}}$ supervises the pixel-level differences by the mask (obtained from a weak-supervised segmentation method) of the intended region (*e.g.*, hair) for the given attribute (*e.g.*, "*blond hair*") to suppress changes on other unintended areas (*e.g.*, mouth), leading to better disentanglement among different attributes.

5.1. Identify Attribute Directions via a Semantic Matching Loss

To identify the latent directions of attributes existing in the dataset, we first build a vocabulary of attributes, *e.g.*, "*smiling*," "*blond hair*," attributes in a face dataset, where each attribute is represented by a word or a short phrase. Then, we extract the attributes from each sentence in the dataset based on string matching or dependency parsing. For example, "*woman*" and "*blond hair*" attributes are extracted from the sentence "*the woman has blond hair*."

Then, we present an *Attribute-to-Direction* module (see Fig. 3) that takes the random latent code $\mathbf{z}$ and word embedding of attributes $\mathbf{t}^a$ sampled from the attribute vocabulary as the inputs, outputting the attribute direction $\mathbf{a}$. To ensure that $\mathbf{a}$ is semantically matched with the input attribute, we propose a novel *Semantic Matching Loss* to train the *Attribute-to-Direction* module. Concretely, $\mathbf{a}$ is used to edit $\mathbf{z}$ to obtain the positive latent code $\mathbf{z}^a_{\text{pos}} = \mathbf{z} + \mathbf{a}$ and negative latent code $\mathbf{z}^a_{\text{neg}} = \mathbf{z} - \mathbf{a}$. $\mathbf{z}^a_{\text{pos}}$ is used to synthesize the positive image $\mathbf{I}^a_{\text{pos}} = G(\mathbf{z}^a_{\text{pos}})$ that can reflect the semantic meaning of the attribute, *e.g.*, the smiling face in Fig. 3. While $\mathbf{z}^a_{\text{neg}} = G(\mathbf{z}^a_{\text{neg}})$ is used to synthesize the negative image $\mathbf{I}^a_{\text{neg}} = G(\mathbf{z}^a_{\text{neg}})$ that does *not* contain the information of the given attribute, *e.g.*, the *not* smiling face in Fig. 3. Based on the triplet [52] of ($\mathbf{t}^a$, $\mathbf{I}^a_{\text{pos}}$, $\mathbf{I}^a_{\text{neg}}$), the *Semantic Matching Loss* is computed as:

$$\begin{aligned}\mathcal{L}_{\text{semantic}} = \max(&\cos(E^{\text{img}}_{\text{CLIP}}(\mathbf{I}^a_{\text{neg}}), E^{\text{text}}_{\text{CLIP}}(\mathbf{t}^a)) \\ &- \cos(E^{\text{img}}_{\text{CLIP}}(\mathbf{I}^a_{\text{pos}}), E^{\text{text}}_{\text{CLIP}}(\mathbf{t}^a)) + \alpha, 0),\end{aligned} \tag{4}$$

where α is a hyperparameter as the margin. $\mathcal{L}_{\text{semantic}}$ attracts attribute text embedding and positive image embedding and repels the attribute text embedding against negative image embedding in CLIP's feature space, rendering the attribute direction $\mathbf{a}$ semantically matched with the attribute.

5.2. Attribute Disentanglement with a Spatial Constraint

However, the *Semantic Matching Loss* cannot ensure that the given attribute is disentangled with other attributes. For example, in Fig. 4, while the *Attribute-to-Direction* module is expected to predict an attribute direction of "*blond hair*," the mouth region is also changing. To mitigate this issue, we propose a novel *Spatial Constraint* as an additional loss to train the *Attribute-to-Direction* module. Our motivation is to restrict the spatial variation between the positive and negative images to an intended region, *e.g.*, the hair region for the "*blond hair*" attribute. To achieve this, we capture the spatial variation by computing the pixel-level difference $\mathbf{I}^a_{\text{diff}} = \sum_c |\mathbf{I}^a_{\text{pos}} - \mathbf{I}^a_{\text{neg}}|$, where c denotes image channel dimension. Then, min-max normalization is applied to rescale its range to 0 to 1, denoted as $\tilde{\mathbf{I}}^a_{\text{diff}}$. We send the positive image to a weakly-supervised (*i.e.*, supervised by attributes extracted from text) part segmentation method [17] to acquire the pseudo-ground-truth mask $\mathbf{M}^a$ (Sec. 6.2), *e.g.*, hair region mask in Fig. 4. Finally, *Spatial Constraint* is computed as:

$$\mathcal{L}_{\text{spatial}} = \text{BCE}(\tilde{\mathbf{I}}^a_{\text{diff}}, \mathbf{M}^a), \tag{5}$$

where BCE denotes binary cross-entropy loss. Minimizing $\mathcal{L}_{\text{spatial}}$ will penalize the spatial variations out of the pseudo-ground-truth mask. In this way, the *Attribute-to-Direction* module is forced to predict the attribute direction that can edit the image in the intended region.

In addition, similar to the *norm penalty* used for *Text-to-Direction* module, we also add it here to ensure the image quality. As a summary, the **full objective function** for training the *Attribute-to-Direction* module is:

$$\mathcal{L}_a = \mathcal{L}_{\text{semantic}} + \mathcal{L}_{\text{spatial}} + \mathcal{L}_{\text{norm}}. \tag{6}$$

5.3. Compositional Attribute Adjustment

After training the *Attribute-to-Direction* module, we propose novel *Compositional Attribute Adjustment* (*CAA*) to ensure the compositionality of the text-to-image synthesis results. The key idea of *Compositional Attribute Adjustment* is two-fold. First, we identify the attributes that the sentence direction $\mathbf{s}$ incorrectly predicts based on its agreement with attribute directions. Second, once we identify the wrongly predicted attributes, we add these attribute directions as the correction to adjust the sentence direction.

Concretely, during the inference stage, as described in Sec. 4, we first sample a random latent code $\mathbf{z}$ and send it to *Text-to-Direction* module along with the input text $\mathbf{t}$ to obtain the sentence direction $\mathbf{s}$. At the same time, we also extract K attributes $\{\mathbf{t}_i^a\}_{i=1}^K$ from the sentence $\mathbf{t}$ and then feed it into the *Attribute-to-Direction* module along with the random latent code $\mathbf{z}$ to obtain the attribute directions $\{\mathbf{a}_i\}_{i=1}^K$. Here K is not a hyperparameter but is decided by the number of attributes described in the sentence, and the same $\mathbf{z}$ is used as the input for both the *Text-to-Direction* module and the *Attribute-to-Direction* module. Based on the attribute directions, we adjust the sentence direction $\mathbf{s}$ to $\mathbf{s}'$:

$$\mathbf{A} = \{\mathbf{a}_i \mid \cos(\mathbf{a}_i, \mathbf{s}) \leq 0\}, \quad \mathbf{s}' = \mathbf{s} + \sum_{\mathbf{a}_i \in \mathbf{A}} \frac{\mathbf{a}_i}{||\mathbf{a}_i||_2}, \quad (7)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and $\mathbf{s}'$ stands for the attribute-adjusted sentence direction. $\mathbf{A}$ is a set of attribute directions that have a less or equal to zero cosine similarity with the sentence direction. When $\cos(\mathbf{a}_i, \mathbf{s}) \leq 0$, the sentence direction $\mathbf{s}$ is not agreed with the i-th attribute direction $\mathbf{a}_i$, indicating that $\mathbf{s}$ fails to reflect the i-th attribute in the input text. By adding the i-th attribute direction $\frac{\mathbf{a}_i}{||\mathbf{a}_i||_2}$, the adjusted sentence direction $\mathbf{s}'$ is corrected to reflect the i-th attribute. Then, it replaces $\mathbf{s}$ to edit the latent code $\mathbf{z}$ to obtain the new text-conditioned code $\mathbf{z}_\mathbf{s} = \mathbf{z} + \mathbf{s}'$ (lower branch in Fig. 2), which is used to synthesize the final image, enhancing compositionality of the text-to-image synthesis.

6. Experiments

6.1. Experiment Setup

Dataset We use two datasets to conduct the experiments. The first dataset is CelebA-HQ [18], which contains 30,000 celebrity face images. We use the text annotations provided by Xia *et al.* [64], where each text description is based on the facial attributes, *e.g.*, "*She is wearing lipstick.*" We remove the texts that mention the "*attractiveness*" attribute due to the ethical concern [45]. The second dataset is CUB [61], which contains 11,788 bird images in 200 bird species. We use the text annotations collected by Reed *et al.* [49], where each sentence describes the fine-grained attributes of the bird.

Test Split for Compositionality Evaluation To better evaluate the compositionality of the text-to-image synthesis results, we carefully choose the test split on each dataset. We observe that about half of the texts in the standard test split [28] of CelebA-HQ dataset contain compositions of attributes seen in the training split. Therefore, we exclude these texts with seen compositions from the test split. As a result, the texts in the new test split only contain the unseen compositions of attributes, which can better evaluate the compositionality results. Proposed Split (PS) [66,67] is a CUB dataset split to benchmark the compositional zero-shot learning by splitting the dataset based on bird species. We choose the "unseen test" in PS as the test split, which can evaluate the model's capability of synthesizing images in 50 unseen bird categories.

Evaluation Metrics

FID. We use FID [13] to evaluate image quality results. Lower values indicate better image quality.

R-Precision. We use R-Precision [68] that evaluates the top-1 retrieval accuracy as the major evaluation metric in image-text alignment. We follow [39] to use the CLIP finetuned on the whole dataset (including the test split) to compute the R-Precision results, which has been shown to be more aligned with human evaluation results. Higher R-Precision values indicate better alignment between text and image.

Bird Species Classification Accuracy. As the models are expected to synthesize birds in unseen species on CUB dataset, we regard that a model that can more accurately synthesize birds in unseen bird species has better compositionality for disentangling different attributes from seen bird species. To this end, we propose a new evaluation metric—bird species classification accuracy for evaluating compositionality. Concretely, we finetune a ResNet-18 [12] on the test split of CUB dataset with real images and bird species labels to classify 50 bird species. In evaluation, the test split contains (text, bird species label) pairs, where text is used to synthesize images. We use the finetuned classifier to predict bird species of the synthesized image. We report the top-1 accuracy based on the prediction and bird species labels (Tab. 2). However, a text may not contain enough discriminative information for classifying the bird species. Therefore, we train a text classifier, implemented as a GRU followed by an MLP, (last row in Tab. 2) that directly takes the text as input to predict the bird species. We train this text classifier on 80% of texts in the test split, and we evaluate its classification accuracy on the rest 20%, which can serve as the upper bound for the text-conditioned bird species classification results.

User Study. The quantitative evaluation metrics above cannot substitute human evaluation. Therefore, we invite 12 subjects to conduct the user study on the two datasets to evaluate image quality and text alignment. Following [71], each question contains synthesized images from different methods conditioned on the same text input. Participants are invited to rank the synthesized images from different methods based on the image quality and image-text alignment. More details of the user study, *e.g.*, user interface and use of human subjects, are in Appendix E.

Comparison Methods We compare with four recent text-to-image synthesis methods—ControlGAN [30], DAE-

	CelebA-HQ		CUB	
	R-Precision ↑	FID ↓	R-Precision ↑	FID ↓
ControlGAN	0.435	31.38	0.137	29.03
DAE-GAN	0.484	30.74	0.145	26.99
TediGAN-A	0.044	16.45	0.071	**16.38**
TediGAN-B	0.306	**15.46**	0.121	16.79
StyleT2I (**Ours**)	**0.625**	17.46	**0.264**	20.53
StyleT2I-XD (**Ours**)	**0.698**	18.02	**0.350**	19.19

Table 1. Text-to-Image synthesis results on CelebA-HQ [64] and CUB [61] datasets. ↑: high values mean better results. ↓: lower values indicate better results.

GAN [51], TediGAN-A [64] TediGAN-B [65]. ControlGAN focuses on controllable generation based on attention mechanism. DAE-GAN extracts "aspects" information from text, which is related to the attributes studied in this paper. TediGAN-A trains a text encoder to minimize the distance between encoded text and encoded image in StyleGAN's latent space. TediGAN-B uses CLIP to optimize the StyleGAN's latent code iteratively for each input text. For a fair comparison, we use the official code of each comparison method to conduct the experiments.

6.2. Implementation Details

Architecture and Hyperparameters We choose StyleGAN2 [20] as the generator for synthesizing images in 256^2 resolution. We use $\mathcal{W}+$ space as the latent space, where latent directions are more disentangled than the input noise space [19]. GloVe [44] is used to obtain the word embeddings of text, which will be used as the input to *Text-to-Direction* and *Attribute-to-Direction* modules. The two modules have the same architecture—a GRU [9] to extract the text feature, which is concatenated with the random latent code to send to a multi-layer perceptron with two fully-connected layers and one ReLU activation function [37]. We set the value $\theta = 8$ in Eq. (2) and $\alpha = 1$ in Eq. (4). More details are in Appendix A.2. The code is written in PyTorch [41] and is available at `https://github.com/zhihengli-UR/StyleT2I`.

Attributes Vocabulary and Attributes Extraction For the vocabulary of attributes (Sec. 5.1), we use the attributes defined in [34] (*e.g.*, "*wearing lipstick*") as the attributes of CelebA-HQ dataset, and the attributes defined in [61] (*e.g.*, "*red belly*") as the attributes of CUB dataset. Note that we do not use any attribute annotations. To extract attributes from sentences, we use string matching (*i.e.*, the word "lipstick" in the sentence indicates "*wearing lipstick*" attribute) on CelebA-HQ dataset. We use part-of-speech tag and dependency parsing implemented in spaCy [16] to extract attributes from the text on CUB dataset. More details are shown in Appendix A.3.

Pseudo-Ground-Truth Mask For the *Spatial Constraint* (Sec. 5.2), we obtain the pseudo-ground-truth mask based on a weakly-supervised part segmentation method [17], where

Method	Accuracy ↑
ControlGAN	0.071
DAE-GAN	0.056
TediGAN-A	0.063
TediGAN-B	0.036
StyleT2I w/o (*CAA*) (**Ours**)	0.115
StyleT2I (**Ours**)	**0.125**
StyleT2I-XD (**Ours**)	**0.142**
Text Classifier (upper bound)	0.204

Table 2. Unseen bird species classification results. Our method outperforms other methods, and the results are closer to the upper bound, which demonstrates that StyleT2I can better synthesize unseen bird species based on the input text description, indicating better compositionality of our method.

we train image classier supervised by attributes extracted from text. More details are presented in Appendix A.4.

Finetune CLIP We empirically find that directly using the CLIP trained on the original large-scale dataset [47] performs poorly for the proposed losses (Eqs. (1) and (4)) on two datasets. We suspect the reason is the domain gap between in-the-wild images in the large-scale dataset [47] and face or birds images with fine-grained attributes. Therefore, we finetune the last few layers of CLIP on the training splits of CelebA-HQ and CUB datasets, respectively. Note that the CLIP used for training differs from the one used for evaluating R-Precision, where the latter is trained on the whole dataset. More details are in Appendix A.5.

Cross-dataset Synthesis (StyleT2I-XD) Since StyleT2I is based on a pretrained StyleGAN generator, we can train the StyleGAN generator on a different image dataset with more image samples and diversity to further improve the results. We denote this method as **StyleT2I-XD**. Concretely, we pretrain StyleGAN on FFHQ [19] dataset, a face dataset with more variation on various attributes (*e.g.*, age), to synthesize images conditioned on the text from CelebA-HQ dataset. Similarly, we pretrain StyleGAN on NABirds [59] dataset with more bird species (the unseen bird species in the test split are still excluded) and image samples to synthesize images conditioned on the text from CUB dataset.

6.3. Results on Text-to-Image Synthesis

Quantitative Results The quantitative results of text-to-image synthesis on CelebA-HQ and CUB datasets are shown in Tab. 1. In terms of R-Precision, our StyleT2I outperforms other comparison methods by a large margin, showing that our method has a better compositionality to synthesize faces in novel compositions and birds in novel bird species. Although TediGAN-A is also based on StyleGAN, it performs poorly on both datasets, which suggests that deterministically minimizing the distance between the latent codes of text and image in StyleGAN's latent space leads to poor generalizability to the unseen compositions. The bird species

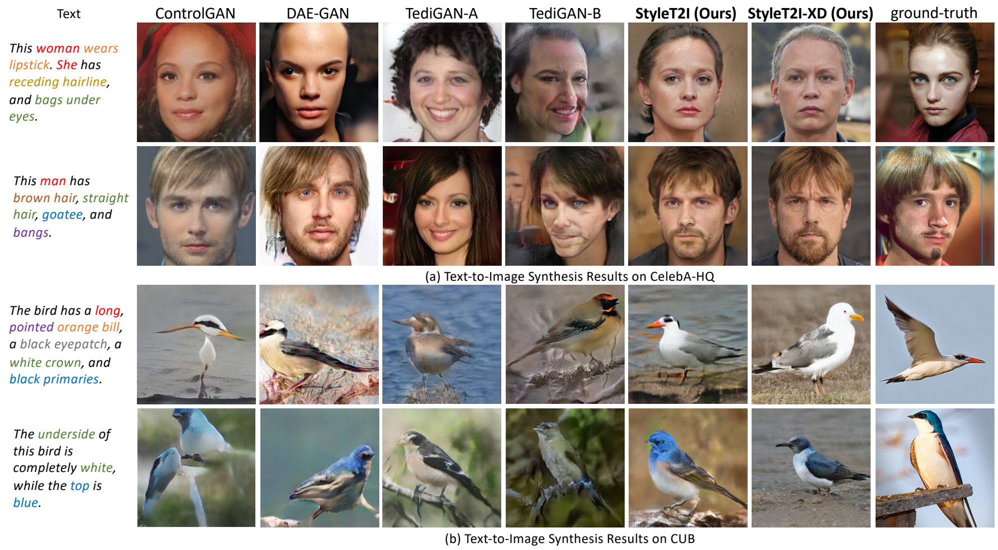

Figure 5. Qualitative comparison of text-to-image synthesis on CelebA-HQ and CUB datasets. Different attributes in the text are highlighted in different colors. More examples are in Appendix D.

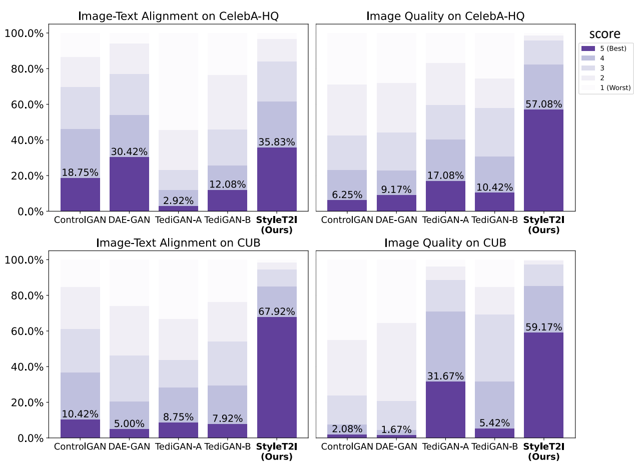

Figure 6. User study results on CelebA-HQ and CUB datasets.

classification results are shown in Tab. 2. Our StyleT2I outperforms other methods in the accuracy results by a large margin, which is also closer to the text classifier accuracy upper bound. This indicates that StyleT2I can more accurately synthesize the unseen bird species based on the text description, demonstrating better compositionality of StyleT2I. Concerning FID, our method achieves strong image quality results, which are also comparable with TediGAN. The FID results also show the advantage of StyleGAN-based methods (TediGAN and our StyleT2I) over methods with customized generator architectures (*i.e.*, ControlGAN and DAE-GAN) for achieving high-fidelity synthesis results.

Qualitative Results We also show qualitative results in Fig. 5. ControlGAN and DAE-GAN, although they reflect most attributes in the text, achieve poor images quality results. For example, in the first row of Fig. 5, they both exaggerate the "*receding hairline*" as bald. Although TediGAN can synthesize high-quality images, the images are barely aligned with the text, *e.g.*, wrong gender in the second row of Fig. 5. In contrast, the synthesized images by StyleT2I are in high fidelity and aligned with the attributes in text, *e.g.*, "*orange bill*" in Fig. 5 (b).

	R-Precision ↑	FID ↓
w/o *CLIP-guided Contrastive Loss*	0.205	**18.64**
w/o *norm penalty*	**0.333**	<u>23.86</u>
w/o *Spatial Constraint*	0.246	**19.17**
w/o *Compositional Attribute Adjustment*	0.238	**19.17**
w/o finetune CLIP	<u>0.145</u>	19.91
Full Model	**0.264**	19.19

Table 3. Ablation Study of StyleT2I on CUB dataset. Top-2 results are bolded and the worst results are underlined.

User Study The user study results are shown in Fig. 6. Compared with other methods, StyleT2I receives higher ranking scores from the human participants in terms of both image-text alignment and image quality, which further manifests the advantages of our method.

Cross-dataset Synthesis Our cross-dataset text-to-image synthesis (StyleT2I-XD) can further improve the results. The quantitative results are shown in Tabs. 1 and 2. StyleT2I-XD achieves even stronger R-Precision and bird species classification accuracy results, demonstrating the effectiveness of cross-dataset training. Although StyleT2I-XD does not improve FID values, our qualitative results in Fig. 5 show that StyleT2I-XD achieves photo-realistic image quality.

6.4. Ablation Studies

We conduct ablation studies to verify the effectiveness of each component of our method. More ablation study results

Figure 7. Ablation study of *norm penalty* for improving image quality. More examples are shown in Appendix D.

are included in Appendices B and C.

CLIP-guided Contrastive Loss An alternative loss to Eq. (1) is minimizing the cosine distance between the paired fake image feature and text feature in CLIP's feature space, which is initially proposed in StyleCLIP [42] and used in TediGAN-B [65] for text-to-image synthesis. The result of this alternative loss is shown on the first row of Tab. 3. Although it slightly improves the FID result, the R-Precision result significantly decreases, demonstrating the necessity of contrasting unmatched (image, text) pairs to distinguish the difference of compositions better.

Norm Penalty As shown in Tab. 3 and Fig. 7, Although it lowers the performance in terms of R-Precision, using the proposed *norm penalty* can effectively improve the FID results and perceptual quality, striking a better balance between image-text alignment and fidelity.

Spatial Constraint The R-Precision results in Tab. 3 show that *Spatial Constraint* can improve the alignment between text and image. The qualitative results in Fig. 8 show that *Spatial Constraint* effectively constrains the spatial variation within the intended region, *e.g.*, hair region for "*blond hair*" attribute. These more disentangled attribute directions help StyleT2I achieve better R-Precision performance by adjusting the sentence direction during the inference stage.

Compositional Attribute Adjustment Tab. 3 shows that Compositional Attribute Adjustment (*CAA*) improves the R-Precision results and achieves a similar FID result. In Tab. 2, *CAA* can also improve the unseen bird species classification results, demonstrating its effectiveness for improving compositionality. In Fig. 9, we show that *(CAA)* can not only detects wrong attributes, *e.g.*, "*brown hair*", but also correct these wrong attributes by adjusting the sentence direction based on the identified attribute directions.

Finetune CLIP As introduced in Sec. 6.2, we finetune the CLIP on the training split of the dataset. The R-Precision results in Tab. 3 show that finetuning can greatly improve performance. Although trained on a large-scale dataset, the results suggest that CLIP will underperform for text-to-image synthesis with fine-grained attributes, proving the necessity to finetune on the dataset for better results.

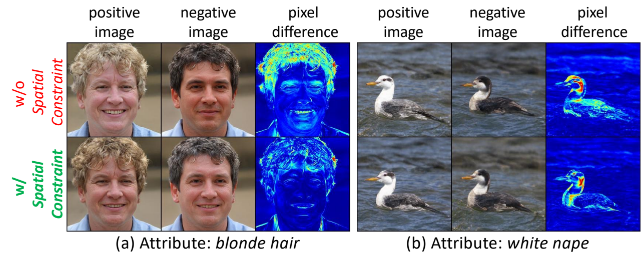

Figure 8. Ablation study of *Spatial Constraint* for identifying attribute directions. Without our *Spatial Constraint* (first row), there are also changes in the other regions (*e.g.*, brows and mouth regions for the *blond hair* attribute; the wings region for the *white nape* attribute). Our *Spatial Constraint* (second row) successfully suppresses the variations in other unintended regions, leading to better disentanglement among different attributes.

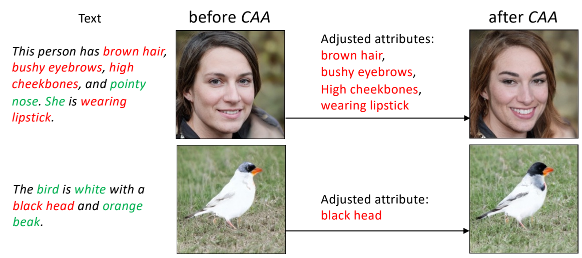

Figure 9. *Compositional Attribute Adjustment* (*CAA*) automatically detect the attributes that are failed to be synthesized (highlighted in red) and adjust the sentence direction with the attribute directions to improve the compositionality of the text-to-image synthesis results.

7. Conclusion

We propose StyleT2I, a new framework for achieving compositional and high-fidelity text-to-image synthesis. We propose a novel *CLIP-guided Contrastive Loss* to better distinguish different compositions, a *Semantic Matching Loss* and a *Spatial Constraint* to identify disentangled attribute directions, and *Compositional Attribute Adjustment* to correct wrong attributes in the synthesis results. StyleT2I outperforms previous approaches in terms of image-text alignment and achieves image fidelity. Admittedly, our work has some limitations. For example, our *Spatial Constraint* is not helpful to disentangle a few attributes that share the same spatial region, *e.g.*, "*bushy eyebrow*" and "*arched eyebrow*." One potential negative societal impact is that StyleT2I's high-fidelity synthesis may be maliciously used for deception. We will mitigate it by asking the users to follow ethical principles when releasing the model. A promising future direction for StyleT2I is complex scene images synthesis for disentangling different objects and backgrounds.

Acknowledgment This work has been partially supported by the National Science Foundation (NSF) under Grant 1909912 and 1934962 and by the Center of Excellence in Data Science, an Empire State Development-designated Center of Excellence. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms. In *The European Conference on Computer Vision (ECCV)*, 2020.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*, 2021.

[3] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *ACM Conference on Fairness, Accountability, and Transparency*, 2018.

[4] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative Pretraining from Pixels. In *International Conference on Machine Learning*, 2020.

[5] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, 2018.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2016.

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv:1504.00325 [cs]*, 2015.

[8] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *Advances in Neural Information Processing Systems Workshop*, 2014.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

[15] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic Object Accuracy for Generative Text-to-Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[17] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring Image Segmentation via Cross-Modal Progressive Comprehension. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*, 2018.

[19] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[21] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Inter-

pretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*, 2018.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

[23] Durk P Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*, 2018.

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[25] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-Image Generation Grounded by Fine-Grained User Attention. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[26] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. In *International Conference on Learning Representations*, 2018.

[27] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[28] Hyodong Lee, Joonseok Lee, Joe Yue-Hei Ng, and Paul Natsev. Large Scale Video Representation Learning via Relational Graph Clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[29] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-Scale Few-Shot Learning: Knowledge Transfer With Class Hierarchy. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable Text-to-Image Generation. In *Advances in Neural Information Processing Systems*, 2019.

[31] Zhiheng Li and Chenliang Xu. Discover the Unknown Biased Attribute of an Image Classifier. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[32] Jiadong Liang, Wenjie Pei, and Feng Lu. CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis. In *The European Conference on Computer Vision (ECCV)*, 2020.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *The European Conference on Computer Vision (ECCV)*, 2014.

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[35] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning*, 2019.

[36] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[37] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.

[38] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 1979.

[39] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for Compositional Text-to-Image Synthesis. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[40] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, 2019.

[42] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[43] William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement. In *The European Conference on Computer Vision (ECCV)*, 2020.

[44] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing*, 2014.

[45] Vinay Uday Prabhu, Dian Ang Yap, Alexander Wang, and John Whaley. Covering up bias in CelebA-like datasets with Markov blankets: A post-hoc cure for attribute prior avoidance. In *International Conference on Machine Learning Workshop*, 2019.

[46] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. MirrorGAN: Learning Text-To-Image Generation by Redescription. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021.

[48] 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*, 2021.

[49] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning Deep Representations of Fine-Grained Visual Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[50] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adver-

sarial Text to Image Synthesis. In *International Conference on Machine Learning*, 2016.
[51] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
[52] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[53] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[54] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[55] Jing Shi, Ning Xu, Haitian Zheng, Alex Smith, Jiebo Luo, and Chenliang Xu. SpaceEdit: Learning a Unified Editing Space for Open-domain Image Color Editing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[56] Kihyuk Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems*, 2016.
[57] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2Scene: Generating Compositional Scenes From Textual Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[58] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, 2017.
[59] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a Bird Recognition App and Large Scale Dataset With Citizen Scientists: The Fine Print in Fine-Grained Dataset Collection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017.
[61] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011.
[62] Tianren Wang, Teng Zhang, and Brian Lovell. Faces a la Carte: Text-to-Face Generation via Attribute Disentanglement. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.
[63] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal Jacobian Regularization for Unsupervised Disentanglement in Image Generation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
[64] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[65] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards Open-World Text-Guided Face Image Generation and Manipulation. *arXiv:2104.08910 [cs]*, 2021.
[66] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[67] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning — The Good, the Bad and the Ugly. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[68] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[69] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A Latent Transformer for Disentangled Face Editing in Images and Videos. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
[70] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics Disentangling for Text-to-Image Generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[71] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[72] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.
[73] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[74] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[75] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[76] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and What? Examining Interpretable Disentangled Representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[77] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G. Schwing. Enjoy Your Editing: Controllable GANs for Image Editing via Latent Space Navigation. In *International Conference on Learning Representations*, 2021.