# More Control for Free! Image Synthesis with Semantic Diffusion Guidance

Xihui Liu[1,4*]     Dong Huk Park[1]     Samaneh Azadi[1]     Gong Zhang[2,3]
Arman Chopikyan[2]     Yuxiao Hu[2]     Humphrey Shi[2,3]     Anna Rohrbach[1]     Trevor Darrell[1]
[1]UC Berkeley    [2]Picsart AI Research (PAIR)    [3]University of Oregon    [4]The University of Hong Kong

**(a) Image synthesis with language guidance**

A smiling woman with straight, blonde hair wearing sunglasses.

**(b) Image synthesis with image guidance**

Image Guidance                    Generated images

**(c) Image synthesis with both language and image guidance**

Language Guidance + Image Guidance                              Generated images

A photo of a woman with curly hair. **+**

Figure 1: We incorporate flexible semantic guidance into diffusion models for image synthesis. Our method allows fine-grained control via language, image, or multimodal guidance. It can be applied to datasets without paired image-text data.

## Abstract

*Controllable image synthesis models allow creation of diverse images based on text instructions or guidance from a reference image. Recently, denoising diffusion probabilistic models have been shown to generate more realistic imagery than prior methods, and have been successfully demonstrated in unconditional and class-conditional settings. We investigate fine-grained, continuous control of this model class, and introduce a novel unified framework for semantic diffusion guidance, which allows either language or image guidance, or both. Guidance is injected into a pretrained unconditional diffusion model using the gradient of image-text or image matching scores, without re-training the diffusion model. We explore CLIP-based language guidance as well as both content and style-based image guidance in a unified framework. Our text-guided synthesis approach can be applied to datasets without associated text annotations. We conduct experiments on FFHQ and LSUN datasets, and show results on fine-grained text-guided image synthesis, synthesis of images related to a style or content reference image, and examples with both textual and image guidance.[1]*

## 1. Introduction

Image synthesis has made great progress in recent years [24, 5, 37, 42, 10]. In addition to the goal of generating high-quality photo-realistic images, fine-grained control over the generated images is also an important desideratum when assisting users with art creation and design.

Previous works have explored controllable image synthesis by adding different conditions, including language [57, 54, 42], attributes [59, 46, 53], scene graphs [9], and user sketches or scribbles [4]. Specifically, text-to-image synthesis, as shown in Figure 1-(a), aims to generate images based on text instructions, by adding text embeddings as conditional information to the image generation network. However, most previous text-to-image synthesis methods require image-caption pairs for training, and can-

---

*This work was done when Xihui Liu was a postdoc in UC Berkeley.

[1]Project page `xh-liu.github.io/sdg/`

not generalize to datasets without text annotations.

Besides text instructions, users may want to guide the image generation model with a reference image. E.g., a user might want to generate cat images which look similar to a given photo of a cat in terms of its appearance. This information cannot be easily described by language, but can be provided via a reference image, as shown in Figure 1-(b). Furthermore, a user may want to provide both language and image guidance. For example, a user might seek to generate "a woman with curly hair" that looks similar to a reference image of a woman with red hair, as shown in Figure 1-(c).

Current image-conditioned synthesis techniques either only transfer the "style" of a reference image to a target image [21, 3] or are restricted to the domains with well-defined structure, e.g. human or animal faces [21, 53]. They cannot generate diverse images with various poses, structures and layouts based on a single reference image.

We propose *Semantic Diffusion Guidance (SDG)*, a unified framework for text-guided and image-guided synthesis that overcomes these limitations. Our model is based on denoising diffusion probabilistic models (DDPM) [19] which generate an image from a noise map and iteratively removes noise to approach the data distribution of natural images.

We inject the semantic input by using a guidance function to guide the sampling process of an unconditional diffusion model. This enables more controllable generation in diffusion models and gives us a unified formulation for both language and image guidance. Specifically, our language guidance is based on the image-text matching score predicted by CLIP [41] finetuned on noised images. As for the image guidance, depending on what information we seek in the image, we define two options: content and style guidance. The flexibility of the guidance module allows us to inject either language or image guidance alone or both at once into any unconditional diffusion model without the need for re-training. We propose a self-supervised scheme to finetune the CLIP image encoder without text annotations, from which we obtain the guidance model with minimal cost.

Our unified framework is flexible and allows fine-grained semantic control in image synthesis as shown in Figure 1. We show that our model can handle: (1) Text-guided image synthesis with fine-grained text queries on any dataset *without language annotations*; (2) Image-guided image synthesis with content or style control from the input image, which generates diverse images with different pose, structure, and layout; (3) Multi-modal guidance for image synthesis with both language and image input. Our flexible guidance network can be injected into off-the-shelf unconditional diffusion models, *without the need for re-training* the diffusion model. We further present a self-supervised efficient finetuning scheme for the CLIP guidance model which does not require textual annotations. We conduct experiments on FFHQ [12] and LSUN [56] datasets to validate the quality, diversity, and controllability of our generated images, and show various applications of our proposed Semantic Diffusion Guidance.

## 2. Related Work

**Text-guided Synthesis** Pioneered by GAN-INT-CLS [44] and GAWWN [45], conditional generative adversarial networks (GANs) [15] have been the dominant framework for text-based image synthesis. Various methods have been studied leading to significant improvements in editing quality and correctness [55, 60, 51, 27, 11, 61, 33, 35, 17, 28, 7, 23, 40]. Recent work DALL-E [42] shows promising results with transformers [52] and discrete VAE [43] by leveraging web-scale data. A concurrent work GLIDE [36] adapts classifier-free guidance for large diffusion models and large-scale training for text-guided image synthesis. Despite great advancements, prior methods require paired image-text annotations which limits the application to certain datasets or requires large amount of data and computational resources for training. Our proposed framework is able to generate images on multiple domains given detailed text prompts, requiring neither image-text paired data from those domains nor large amount of compute to train the text-guided image synthesis model.

**Image-guided Synthesis** Image-guided synthesis aims to generate diverse images with the constraint that they all should resemble a given reference image in terms of content or style. Many style transfer works fall under this category where the content of the input image must be preserved while the style of the reference image is transferred [13, 29, 16, 14, 26, 20, 31, 30, 38, 1], yet they struggle to generate diverse images. Some works study image synthesis guided by the content of the reference images. ILVR [8] proposes a way to iteratively inject image guidance to a diffusion model, yet it exhibits limited structural diversity of the generated images. Instance-Conditioned GAN [6] uses nearest neighbor images of a given reference for adversarial training to generate structurally diverse yet semantically relevant images. Yet, it requires training the GAN model with instance-conditioned techniques. Our approach demonstrates better controllability as different types of image guidance are proposed where users can decide how much semantic, structural, or style information to preserve by using different types and scales of guidance, while not requiring to re-train the unconditional diffusion model.

**Diffusion Models** Diffusion models are a new type of generative models consisting of a forward process (signal to noise) and a reverse process (noise to signal). The denoising diffusion probabilistic model (DDPM) [47, 19] is a latent variable model where a denoising autoencoder gradually transforms Gaussian noise into signal. Score-based generative model [48, 49, 50] trains a neural network to

predict the score function which is used to draw samples via Langevin Dynamics. Diffusion models have demonstrated comparable or superior image quality compared to GANs while exhibiting better mode coverage and training stability. They have also been explored for conditional generation such as class-conditional generation, image-guided synthesis, and super-resolution [50, 10, 8, 34]. Concurrent work [2] explored text-guided image editing with diffusion models. Dhariwal *et al.* [10] proposed classifier guidance for class-conditional image synthesis with diffusion models. Based on the guidance algorithm proposed in [10], we further explore whether diffusion models can be semantically guided by text or image, or both to synthesize images.

**CLIP-guided Generation and Manipulation** CLIP [41] is a powerful vision-language joint embedding model trained on large-scale images and texts. Its representations have been shown to be robust and general enough to perform zero-shot classification and various vision-language tasks on diverse datasets. StyleCLIP [39] and StyleGAN-NADA [12] have demonstrated that CLIP enables text-guided *image manipulation* and *domain adaptation of image generation* without domain-specific image-text pairs. DiffusionCLIP [25] uses CLIP for language-based image editing, and Blended Diffusion [2] explores mask-guided image editing with CLIP. However, the application to *image synthesis* has not been explored. Our work investigates text and/or image guided synthesis using CLIP and unconditional DDPM.

# 3. Semantic Diffusion Guidance

We propose *Semantic Diffusion Guidance (SDG)*, a unified framework that incorporates different forms of guidance into a pretrained unconditional diffusion model. SDG can leverage language guidance, image guidance, and both, enabling controllable image synthesis. The guidance module can be injected into any off-the-shelf unconditional diffusion model without re-training or finetuning it. We only need to finetune the guidance network, which is a CLIP [41] model in our implementation, on the images with different levels of noise. We propose a self-supervised finetuning scheme, which is efficient and does not require paired language data to finetune the CLIP image encoder.

In Section 3.1, we review the preliminaries on diffusion models, and introduce our approach for injecting guidance for controllable image synthesis. In Section 3.2, we describe the language guidance which enables the unconditional diffusion model to perform text-to-image synthesis. In Section 3.3, we propose two types of image guidance, which take the content and style information from the reference image as the guidance signal, respectively. In Section 3.5, we explain how we finetune the CLIP network without requiring text annotations in the target domain.
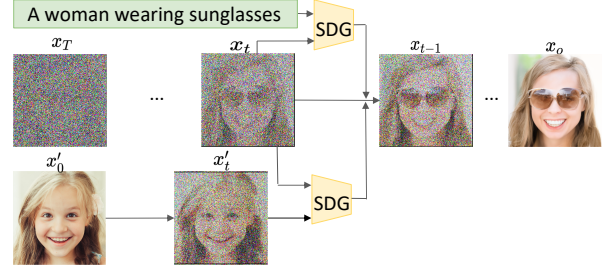


Figure 2: Our method is based on the DDPM model which generates an image from a noise map by iteratively removing noise at each timestep. We control the diffusion generation process by Semantic Diffusion Guidance (SDG) with language and/or a reference image. SDG is iteratively injected at each step of generation process. We only illustrate the guidance at one timestep $t$ in the figure.

## 3.1. Guiding Diffusion Models for Controllable Image Synthesis

Diffusion models define a Markov chain where random noise is gradually added to the data, known as the forward process. Formally, given a data point sampled from a real-data distribution $x_0 \sim q(x)$, the forward process sequentially adds Gaussian noise to the sample over $T$ timesteps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I})$$
$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

where $\{\beta\}_{t=1:T}$ denotes a constant or learned variance schedule that controls the noise step size. A property of the forward process is that we can sample $x_t$ from $x_0$ in a closed form:

$$q(x_t|x_0) = \sqrt{\overline{\alpha_t}}x_0 + \epsilon\sqrt{1-\overline{\alpha_t}}, \epsilon \sim \mathcal{N}(0,1) \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$.

Generative modeling is done by learning the backward process where the forward process is reversed via a parameterized diagonal Guassian transition:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \sigma_\theta^2(x_t)\mathbf{I}) \tag{3}$$

We choose the notation $p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta, \sigma_\theta^2\mathbf{I})$ for brevity. In order to learn the backward process, neural networks are trained to predict $\mu_\theta$ and $\sigma_\theta^2$.

The formulations above explain the unconditional backward process $p_\theta(x_{t-1}|x_t)$; with an extra guidance signal $y$, the sampling distribution becomes:

$$p_{\theta,\phi}(x_{t-1}|x_t, y) = Zp_\theta(x_{t-1}|x_t)p_\phi(y|x_{t-1}), \tag{4}$$

where $Z$ is a normalizing constant. It is proven [10] that the new distribution after incorporating the guidance can be

approximated by a Gaussian distribution with shifted mean:

$$p_\theta(x_{t-1}|x_t)p_\phi(y|x_{t-1}) = \mathcal{N}(\mu + \Sigma g, \Sigma), \quad (5)$$

where $\mu = \mu_\theta$, $\Sigma = \sigma_\theta^2 \mathbf{I}$, $g = \nabla_{x_{t-1}} \log p_\phi(y|x_{t-1})$.

Class-guided synthesis was explored in [10] where $y$ is a discrete class label, and $p_\phi(y|x_{t-1})$ is the probability of $x_{t-1}$ belonging to class $y$. Here, we generalize $y$ to a continuous embedding for language, image or multimodal guidance. In the following, we introduce the guidance function $F_\phi(x_t, y, t) = \log p_\phi(y|x_t)$ for different guidance types.

Figure 2 and Algorithm 1 summarize the proposed Semantic Diffusion Guidance. Note that there is an additional scaling factor $s$ for semantic guidance in Algorithm 1, a user-controlled hyperparameter that determines the strength of the guidance. We discuss its effect in Section 4.

---

**Algorithm 1:** Semantic Diffusion Guidance

**Input:** guidance $y$, scaling factor $s$
**Given:** diffusion model ($\mu_\theta$, $\sigma_\theta$), Guidance function
$\quad F_\phi(x_t, y, t)$
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
**for** $t = T, \cdots 1$ **do**
$\quad \mu, \Sigma \leftarrow \mu_\theta, \sigma_\theta^2 \mathbf{I}$
$\quad x_{t-1} \leftarrow$ sample from
$\quad \mathcal{N}(\mu + s\Sigma\nabla_{x_t}F_\phi(x_t, y, t), \Sigma)$
**end for**
**return** $x_0$

---

## 3.2. Language Guidance

Language is one of the most intuitive ways that a user can control the generation model. In order to incorporate language information to the image synthesis process, we use a visual-semantic embedding model for image-text alignment. Specifically, given an image $x$ and a text prompt $l$, the model embeds them into the joint embedding space using an image encoder $E_I$ and a text encoder $E_L$, respectively. The similarity between the embeddings $E_I(x)$ and $E_L(l)$ is calculated as the cosine distance, and we utilize this to formulate the language guidance function.

However, the models for backward process and guidance in Equation 5 are time-dependent, and take noisy images as input. This means that the image encoder $E_I$ needs to incorporate the timestep $t$ as input and be further trained on noisy images at different timesteps as well. We denote such time-dependent image encoder for noisy images as $E_I'$. Finally, the language guidance function can be defined as:

$$F(x_t, l, t) = E_I'(x_t, t) \cdot E_L(l), \quad (6)$$

where $E'$ denotes the image encoder trained on noisy images with additional timestep input. In Section 3.5, we give details on adapting a CLIP model [41] to become time-dependent with minimal architecture changes, and present a self-supervised finetuning strategy for noisy images.

## 3.3. Image Guidance

Sometimes, an image can convey information that is difficult to express in language. For example, users may want to generate a photo of a cat that looks similar to another cat, or want to generate a photo of a bedroom in the style of Van Gogh's painting "The Starry Night". They may also want to generate realistic images given an emoji or a painting. We thus propose an approach for image-guided diffusion that effectively controls the content or style information according to an image. We present two types of image guidance, namely *image content guidance* and *image style guidance*.

**Image Content Guidance** aims to control the content of the generated image, with or without structural constraints, based on a reference, and is formulated as the cosine similarity of the image feature embeddings. Let $x_0'$ denote the noise-free reference image. We perturb $x_0'$ per Equation 2 to get $x_t'$. Then, the guidance signal at timestep $t$ is,

$$F(x_t, x_t', t) = E_I'(x_t, t) \cdot E_I'(x_t', t). \quad (7)$$

Similar to language guidance, we use an image encoder finetuned with noised images to define the image guidance function and extract embeddings that mostly capture the high-level semantics. An interesting property of using image encoders for guidance is that one can control how much structural information such as pose and viewpoint is maintained from the reference image. For instance, the embeddings used in Equation 7 do not have spatial dimensions, resulting in samples with great variations in pose and layout. However, by utilizing spatial feature maps and forcing alignment between features in corresponding spatial locations, we can guide the generated image to additionally share similar structure with the reference image as follows.

$$F(x_t, x_t', t) = -\sum_j \frac{1}{C_j H_j W_j}||E_I'(x_t, t)_j - E_I'(x_t', t)_j||_2^2$$

$$(8)$$

where $E_I'()_j \in \mathcal{R}^{C_j \times H_j \times W_j}$ denotes the spatial feature maps of the $j$-th layer of the image encoder $E_I'$.

**Image Style Guidance** allows style transfer from the reference image. It is formulated similarly, except the alignment between the Gram matrices of the intermediate feature maps is enforced:

$$F(x_t, x_t', t) = -\sum_j ||G_I'(x_t, t)_j - G_I'(x_t', t)_j||_F^2, \quad (9)$$

where $G_I'()_j$ is the Gram matrix [22] of the $j$-th layer feature map of the image encoder $E_I'$.

## 3.4. Multimodal Guidance

In some application scenarios, image and language may contain complementary information, and allowing both image and language guidance at the same time provides further

Table 1: Quantitative evaluation of our proposed SDG and comparison to prior work on FFHQ dataset with image guidance and text guidance. For FID, the lower, the better. For other scores, the higher, the better.

|  |  | Quality | Diversity | Correctness (retrieval evaluation) | | | |
|---|---|---|---|---|---|---|---|
|  |  | FID | LPIPS | Top 1 | Top 5 | Top 10 | Top 20 |
| image guidance | ILVR (N=32) [8] | 17.15 | 0.439 | 0.205 | 0.416 | 0.556 | 0.727 |
|  | SDG | **14.37** | **0.583** | **0.520** | **0.742** | **0.816** | **0.906** |
| text guidance | StyleGAN+CLIP | 57.45 | 0.578 | **0.749** | **0.934** | **0.974** | **0.996** |
|  | SDG | **28.38** | **0.610** | 0.553 | 0.795 | 0.878 | 0.947 |

flexibility for user control. Our pipeline can easily incorporate both by a weighted sum of the two guidance functions, with their scaling factors as weights.

$$F_{\phi_0}(x_t, y, t) = s_1 F_{\phi_1}(x_t, y, t) + s_2 F_{\phi_2}(x_t, y, t). \quad (10)$$

By adjusting the weighting factors of each modality, users can balance between the language and image guidance.

### 3.5. Self-supervised Finetuning of CLIP without Text Annotations

CLIP [41] is a powerful vision and language model trained on large-scale image-text data. We leverage its semantic knowledge to achieve controllable synthesis for diffusion models. To act as a guidance function, CLIP is expected to handle noisy images $x_t$ at any timestep $t$. We make a minor architectural change to CLIP image encoder $E_I$ to accept an additional input $t$ by converting batch normalization layers to adaptive batch normalization layers, where the prediction of scale and bias terms is conditioned on $t$. We denote this modified CLIP image encoder as $\widetilde{E_I}$. The parameters of $\widetilde{E_I}$ are initialized by the parameters of the pretrained CLIP model $E_I$, except for the parameters for the adaptive batch normalization layers.

To finetune $\widetilde{E_I}$, we propose a self-supervised approach in which we force an alignment between features extracted from clean and noised images. Formally, given a batch of $N$ pairs of clean and noised images $\{x_0^i, x_{t_i}^i\}_{i=1}^N$ where $t_i$ is the timestep sampled for the $i$-th image that governs the amount of noise, we encode $x_0^i$ and $x_{t_i}^i$ with $E_I$ and $\widetilde{E_I}$, respectively. We rely on CLIP's contrastive objective to maximize the cosine similarity of the $N$ positive pairs while minimizing the similarity of the remaining negative pairs. We fix the parameters of $E_I$ and use the contrastive objective to finetune the parameters of $\widetilde{E_I}$. With our finetuned CLIP model, the diffusion model can be guided by image or language information that users provide. Moreover, the CLIP model is finetuned in a self-supervised manner without requiring any language data for the target dataset.

## 4. Experiments
### 4.1. Dataset and Implementation Details

We conduct experiments on FFHQ [24] and LSUN [56] cat, horse, and bedroom subsets. FFHQ dataset contains 70,000 images of human faces. LSUN contains 3 million bedroom images, 2 million horse images, and 1.7 million cat images. We use unconditional DDPMs from [10, 8], and finetune CLIP [41] RestNet $50 \times 16$ models on noised images on each dataset with initial learning rate $10^{-4}$ and weight decay $10^{-3}$, with a batch size of 256. When synthesizing images with our SDG, the scaling factor is a hyperparameter that we manually adjust for each guidance, which will be discussed in Sec. 4.3. The default scaling factor is 100 for image guidance and 120 for language guidance.

### 4.2. Quantitative Evaluation

**Evaluation Setup** Since our SDG is the first method that unifies text guidance and image guidance for image synthesis, there is no previous work on image synthesis with both image and language guidance. So we evaluate the language-guided image synthesis and image-guided image synthesis separately in order to compare with previous work. We evaluate the **language-guided** generation on FFHQ dataset. For that we define 400 text instructions based on combinations of gender and face attributes from CelebA-Attributes [32]. For example, "A photo of a smiling man with glasses". We generate 25 images for each text query, which results in 10,000 images in total. We compare our language-guided generation with StyleGAN+CLIP[2], which uses CLIP [41] loss to optimize the randomly initialized latent codes of StyleGAN [24] for text-guided image synthesis. StyleGAN+CLIP removes the GAN inversion module of StyleCLIP [24] so that it can be applied for language-based image synthesis. Since our model does not require text annotation for training, our text-guided image synthesis experiments are conducted on image-only datasets without paired text annotations. So our method cannot be directly compared with other text-based image synthesis methods which have to be trained on text-image paired datasets. To evaluate **image-guided** image synthesis, we randomly

---

[2]https://colab.research.google.com/drive/
1br7GP_D6XCgulxPTAFhwGaV-ijFe084X

Table 2: Ablation study of our proposed SDG with image guidance. The numbers in the brackets after "SDG" indicates the scaling factor. For FID, the lower, the better. For other scores, the higher, the better.

| | | Quality | Diversity | Correctness (retrieval evaluation) | | | |
|---|---|---|---|---|---|---|---|
| | | FID | LPIPS | Top 1 | Top 5 | Top 10 | Top 20 |
| LSUN | SDG (100) | **16.02** | **0.617** | 0.178 | 0.443 | 0.592 | 0.766 |
| Cat | SDG (200) | 16.23 | 0.565 | **0.278** | **0.533** | **0.738** | **0.880** |
| LSUN | SDG (100) | **10.30** | **0.597** | 0.165 | 0.418 | 0.568 | 0.704 |
| Horse | SDG (200) | 11.22 | 0.585 | **0.298** | **0.609** | **0.738** | **0.863** |
| LSUN | SDG (100) | **5.18** | **0.633** | 0.364 | 0.745 | 0.866 | 0.942 |
| Bedroom | SDG (200) | 5.19 | 0.550 | **0.445** | **0.805** | **0.900** | **0.951** |

Table 3: Ablation study of our proposed SDG with language guidance on FFHQ dataset. The numbers in the brackets after "SDG" is the scaling factor. For FID, the lower, the better. For other metrics, the higher, the better.

| | | Quality | Diversity | Correctness (retrieval accuracy) | | | |
|---|---|---|---|---|---|---|---|
| | | FID | LPIPS | Top 1 | Top 5 | Top 10 | Top 20 |
| | SDG (120) | **19.60** | **0.650** | 0.248 | 0.526 | 0.654 | 0.795 |
| FFHQ | SDG (160) | 22.63 | 0.644 | 0.263 | 0.548 | 0.679 | 0.801 |
| | SDG (320) | 28.38 | 0.610 | **0.553** | **0.795** | **0.878** | **0.947** |

choose 10,000 images from each dataset as guidance and synthesize new images based on the guidance images. We compare our image-guided results to ILVR [8].

We present quantitative results and comparison with previous work in Table 1 with the following evaluation metrics.

**FID for image quality evaluation.** We report FID score [18] calculated on 10,000 images for each dataset to evaluate the quality of generated images. Lower FID indicates better generation quality. Our SDG outperforms comapred methods for both image-guided synthesis and language-guided synthesis.

**LPIPS for diversity evaluation.** We calculate the LPIPS score [58] between paired images generated from the same image guidance or the same text guidance, as shown in Table 1. Higher LPIPS indicates more diversity. Our model generates more diverse images compared to previous work ILVR [8] and StyleGAN+CLIP. The images generated by ILVR follows the same structure and layout, with variations in details. While our method is able to generate diverse images with different pose, structure, and layout, as shown in Figure 6(a). The images generated by StyleGAN+CLIP also suffers from low diversity, as shown in Figure 6(b). The high FID score of StyleGAN+CLIP is also because of the low diversity of the generated images.

**Retrieval accuracy to evaluate consistency with guidance.** We use text-to-image retrieval or image retrieval by an original CLIP ResNet $50\times16$ model without finetuning to evaluate how well the generated images matches the guidance. For an image generated with text guidance, we randomly select 99 real images from the training set as negative images, and evaluate the text-to-image retrieval performance. Similarly, for an image synthesized with a reference image, we use the reference image to retrieve the generated

image from the 99 randomly selected real images[3]. Style-GAN+CLIP has a very high retrieval performance because the latent codes of the StyleGAN model are directly optimized to minimize the CLIP score calculated by the CLIP model used for retrieval. So the high retrieval performance of StyleGAN+CLIP comes at the cost of low generation diversity, as indicated by the high FID and low LPIPS scores.

### 4.3. Ablation Study

As demonstrated in Section 3.1 and Algorithm 1, the scaling factor $s$ is a user-controllable hyper-parameter that controls the strength of the guidance. We explore the effect of the scaling factor in Table 2 and Table 3. A visual example of the effect of different scaling factors is shown in the Appendix. We observe the trade-off between semantic correctness and diversity of generated images. As the scaling factor gets larger, the guidance signal has more control on the generation results, as indicated by the increased semantic consistency with the guidance. While larger scaling factor also leads to lower diversity of generated images. Users can adjust the scaling factor to control how diverse they expect the generated images to be.

### 4.4. Qualitative Results

**Text-guided and image-guided synthesis results** Our model combines the language and image guidance in a unified framework, and is easy to adapt to various applications. In Figure 3 we show the synthesis results with image content guidance (Equation 7). With the image guided diffusion, the model is able to synthesize new images with diverse structures that match the semantics of the guidance image. Figure 4 shows the language-guided diffusion re-

---

[3]The selected negative images are disjoint with the guidance images we used for synthesizing images.

Figure 3: Image synthesis results with image content guidance on LSUN and FFHQ datasets. Given a guidance image, the model is able to generate semantically similar images with different pose, layout, and structure.



Figure 4: Image synthesis results with language guidance on LSUN and FFHQ datasets. Our model is able to generate images based on fine-grained language instructions.

sults, where our model is able to handle complex and fine-grained descriptions, such as "A smiling woman with curly brown hair and lipstick.", or "A bedroom with a wooden closet and a painting on the wall." We can also incorporate language and image guidance jointly, as shown in Figure 5. The image and language guidance provide complementary information, and our semantic diffusion guidance is able to generate images that align with both. For example, we can generate a bedroom similar to the guidance bedroom image but with windows, or generate a woman according to a guidance image but with a new attribute defined the language guidance (e.g., "smiling" or "short hair" or "sunglasses").

**Comparison to prior work** Since there is no prior work that incorporates text and image guidance in the same unified framework, we compare our approach to previous text-guided and image-guided synthesis work. In image-guided synthesis, the most related to our work is ILVR [8].As shown in Fig. 6(a), our model can generate images in different poses and structures, while ILVR can only generate images of the same pose and structure. We compare our language-guided image synthesis with StyleGAN+CLIP in

Fig. 6(b). Although StyleGAN+CLIP is able to generate high-quality images, diversity is lacking in their results, while our model is able to generate high-quality and diverse results based on the language instructions.

**Other applications** In Fig. 7(a,b), we demonstrate the results of style (Equation 9) and structure-preserving (Equation 8) image guidance. With the style guidance, the model trained on LSUN bedroom is able to synthesize bedrooms in the unseen style. With the structure-preserving content guidance, the synthesized images preserve the structure, pose, and layout from the reference image. Fig. 7(c) shows that the model is able to take an out-of-domain image as guidance, and synthesize photo-realistic images which are semantically similar to the guidance cartoon image.

## 5. Conclusion and Discussions

We propose Semantic Diffusion Guidance (SDG), a unified framework for diffusion-based image synthesis with language, image, or multi-modal guidance. The flexible guidance module allows us to inject various types of guid-
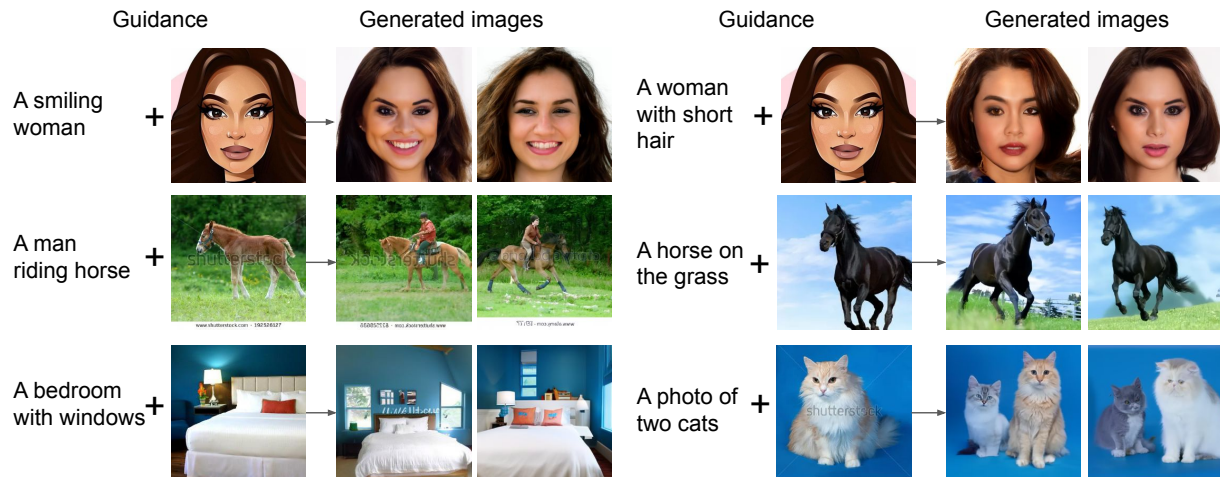
Figure 5: Image synthesis results with both image and language guidance. The image and language guidance provides complementary information, and our model generates images that matches both sources of guidance.
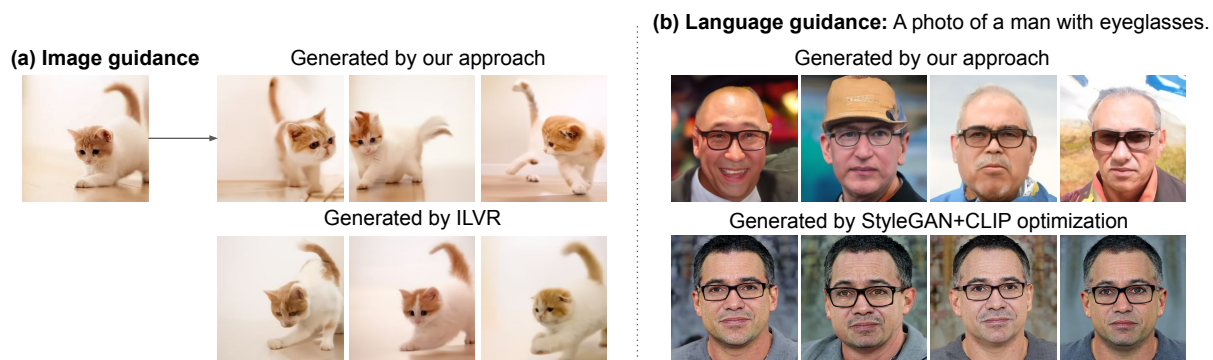


Figure 6: Comparison to previous work. (a) Image-guided image synthesis is compared with ILVR, (b) text-guided image synthesis is compared with StyleGAN+CLIP.
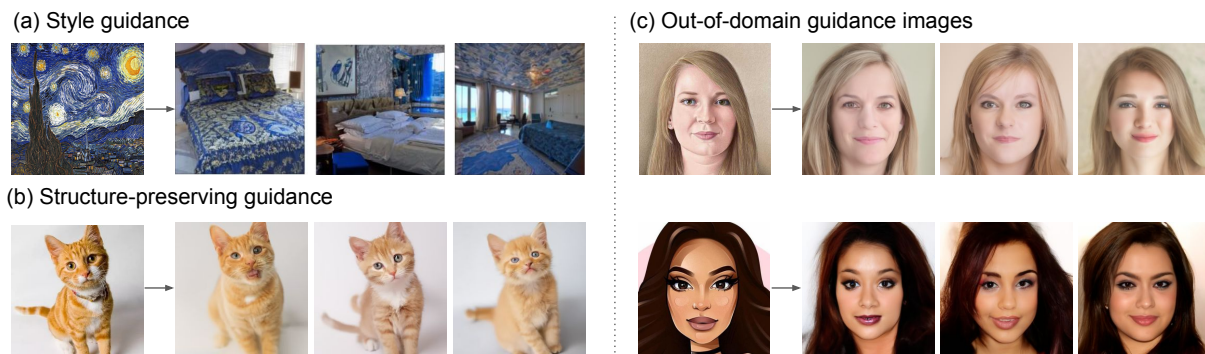


Figure 7: Different applications of SDG. (a) Style-guided synthesis. (b) Structure-preserving synthesis when the user does not want to generate diverse structures. (c) Synthesizing photo-realistic images with out-of-domain image guidance.

ance into any off-the-shelf unconditional diffusion model without re-training or finetuning the diffusion model. We further present a self-supervised efficient finetuning scheme for the CLIP guidance model which does not require textual annotations. However, image generation has as much potential for misuse as it has for beneficial applications. We should be aware of the potential negative social impact if image synthesis is used for generating fake images to mislead people.

## Acknowledgements

# References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *arXiv preprint arXiv:2111.14818*, 2021.

[3] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018.

[4] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (TOG)*, 38(4):59, 2019.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Instance-conditioned gan. *arXiv preprint arXiv:2109.05070*, 2021.

[7] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018.

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[9] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5213–5222, 2020.

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[11] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

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

[13] 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.

[14] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3985–3993, 2017.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8222–8231, 2018.

[17] Mehmet Günel, Erkut Erdem, and Aykut Erdem. Language guided fashion image manipulation with feature-wise transformations. *arXiv preprint arXiv:1808.04000*, 2018.

[18] 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*, pages 6626–6637, 2017.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[20] 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.

[21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[23] KJ Joseph, Arghya Pal, Sailaja Rajanala, and Vineeth N Balasubramanian. C4synth: Cross-caption cycle-consistent text-to-image synthesis. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 358–366. IEEE, 2019.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[25] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021.

[26] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.

[27] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019.

[28] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[29] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2479–2486, 2016.

[30] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019.

[31] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[33] Xiaofeng Mao, Yuefeng Chen, Yuhong Li, Tao Xiong, Yuan He, and Hui Xue. Bilinear representation for language-based image editing using conditional generative adversarial networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2047–2051. IEEE, 2019.

[34] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[35] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

[36] 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. *arXiv preprint arXiv:2112.10741*, 2021.

[37] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CVPR*, 2019.

[38] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*, 2020.

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

[40] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

[41] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[43] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019.

[44] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[45] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. *Advances in neural information processing systems*, 29:217–225, 2016.

[46] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. *arXiv preprint arXiv:2101.02477*, 2021.

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[48] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

[49] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[51] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[52] 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*, pages 5998–6008, 2017.

[53] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 168–184, 2018.

[54] 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. *CVPR*, 2018.

[55] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[56] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *ICCV*, 2017.

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[59] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.

[60] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019.

[61] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision*, pages 1680–1688, 2017.