Self-Guided Diffusion Models

Hu, V.T.; Zhang, D.W.; Asano, Y.M.; Burghouts, G.J.; Snoek, C.G.M.

DOI
10.48550/arXiv.2210.06462
10.1109/CVPR52729.2023.01766

Publication date
2023

Document Version
Author accepted manuscript

Published in
CVPR 2023

Citation for published version (APA):
Self-Guided Diffusion Models

Vincent Tao Hu† David W. Zhang†
Yuki M. Asano† Gertjan J. Burghouts‡
University of Amsterdam TNO

Abstract

Diffusion models have demonstrated remarkable progress in image generation quality, especially when guidance is used to control the generative process. However, guidance requires a large amount of image-annotation pairs for training and is thus dependent on their availability and correctness. In this paper, we eliminate the need for such annotation by instead exploiting the flexibility of self-supervision signals to design a framework for self-guided diffusion models. By leveraging a feature extraction function and a self-annotation function, our method provides guidance signals at various image granularities: from the level of holistic images to object boxes and even segmentation masks. Our experiments on single-label and multi-label image datasets demonstrate that self-labeled guidance always outperforms diffusion models without guidance and may even surpass guidance based on ground-truth labels. When equipped with self-supervised box or mask proposals, our method further generates visually diverse yet semantically consistent images, without the need for any class, box, or segment label annotation. Self-guided diffusion is simple, flexible and expected to profit from deployment at scale.

1. Introduction

Diffusion models have recently enabled tremendous advancements in many computer vision fields related to image synthesis, but counterintuitively this often comes with the cost of requiring large annotated datasets [49, 55]. For example, the image fidelity of samples from diffusion models can be spectacularly enhanced by conditioning on class labels [17]. Classifier guidance goes a step further and offers control over the alignment with the class label, by using the classifier gradient to guide the image generation [17]. Classifier-free guidance [28] replaces the dedicated classifier with a diffusion model trained by randomly dropping the condition during training. This has proven a fruitful line of research for several other condition modalities, such as text [50, 55], image layout [53], visual neighbors [3], and image features [20]. However, all these conditioning and guidance methods require ground-truth annotations. In many domains, this is an unrealistic and too costly assumption. For example, medical images require domain experts to annotate very high-resolution data, which is infeasible to do exhaustively [45]. In this paper, we propose to remove the necessity of ground-truth annotation for guided diffusion models.

We are inspired by progress in self-supervised learning [11, 13], which encodes images into semantically meaningful latent vectors without using any label information. It usually does so by solving a pretext task [2, 21, 24, 69] on image-level to remove the necessity of labels. This annotation-free paradigm enables the representation learning to upscale to

† Equal contribution
Source code at: https://taohu.me/sgdm/.
larger and more diverse image datasets [19]. The holistic image-level self-supervision has recently been extended to more expressive dense representations, including bounding boxes (e.g., [41, 57]) and pixel-precise segmentation masks (e.g., [22, 72]). Some self-supervised learning methods even outperform supervised alternatives [11, 24]. We hypothesize that for diffusion models, self-supervision may also provide a flexible and competitive, possibly even stronger guidance signal than ground-truth labeled guidance.

In this paper, we propose self-guided diffusion models, a framework for image generation using guided diffusion without the need for any annotated image-label pairs, the detailed structure is shown in Figure 1. The framework encompasses a feature extraction function and a self-annotation function, that are compatible with recent self-supervised learning advances. Furthermore, we leverage the flexibility of self-supervised learning to generalize the guidance signal from the holistic image level to (unsupervised) local bounding boxes and segmentation masks for more fine-grained guidance. We demonstrate the potential of our proposal on single-label and multi-label image datasets, where self-labeled guidance always outperforms diffusion models without guidance and may even surpass guidance based on ground-truth labels. When equipped with self-supervised box or mask proposals, our method further generates visually diverse yet semantically consistent images, without the need for any class, box, or segment label annotation.

2. Related Work

**Conditional generative models.** Earlier works on generative adversarial networks (GANs) have observed improvements in image quality by conditioning on ground-truth labels [8, 12, 42]. Recently, conditional diffusion models have reported similar improvements, while also offering a great amount of controllability via classifier-free guidance by training on images paired with textual descriptions [49, 50, 55], semantic segmentations [66], or other modalities [7, 60, 67]. Our work also aims to realize the benefits of conditioning and guidance, but instead of relying on additional human-generated supervision signals, we leverage the strength of pretrained self-supervised visual encoders.

Zhou et al. [71] train a GAN for text-to-image generation without any image-text pairs, by leveraging the CLIP [48] model that was pretrained on a large collection of paired data. In this work, we do not assume any paired data for the generative models and rely purely on images. Additionally, image layouts are difficult to be expressed by text, thus our self-boxed and self-segmented methods are complementary to text conditioning. Instance-Conditioned GAN [12], Retrieval-augmented Diffusion [6] and KNN-diffusion [3] are three recent methods that utilize nearest neighbors as guidance signals in generative models. Similar to our work, these methods rely on conditional guidance from an unsupervised source, we differ from them by further attempting to provide more diverse spatial guidance, including (self-supervised) bounding boxes and segmentation masks.

**Self-supervised learning in generative models.** Self-supervised learning [2, 10, 11, 13] has shown great potential for representation learning in many downstream tasks. As a consequence, it is also commonly explored in GAN for evaluation and analysis [43], conditioning [12, 40], stabilizing training [14], reducing labeling costs [39] and avoiding mode collapse [1]. Our work focuses on translating the benefits of self-supervised methods to the generative domain and providing flexible guidance signals to diffusion models at various image granularities. In order to analyze the feature representation from self-supervised models, Bordes et al. [7] condition on self-supervised features in their diffusion model for better visualization in data space. We instead condition on the compact clustering after the self-supervised feature, and further introduce the elasticity of self-supervised learning into diffusion models for multi-granular image generation.

3. Approach

Before detailing our self-guided diffusion framework, we provide a brief background on diffusion models and the classifier-free guidance technique.

3.1. Background

**Diffusion models.** Diffusion models [27, 58] gradually add noise to an image $x_0$ until the original signal is fully diminished. By learning to reverse this process one can turn random noise $x_T$ into images. This diffusion process is modeled as a Gaussian process with Markovian structure:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$$

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\alpha_t} x_0, (1 - \alpha_t) I),$$

(1)

where $\beta_1, \ldots, \beta_T$ is a fixed variance schedule on which we define $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. All latent variables have the same dimensionality as the image $x_0$ and differ by the proportion of the retained signal and added noise.

Learning the reverse process reduces to learning a denoiser $x_t \sim q(x_t|x_0)$ that recovers the original image as $(x_t - (1 - \bar{\alpha}_t) e_{\theta}(x_t, t)) / \sqrt{\bar{\alpha}_t} \approx x_0$. Ho et al. [27] optimize the parameters $\theta$ of noise prediction network by minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{e,x,t} \left[ \|e_{\theta}(x_t, t) - e\|^2_2 \right],$$

(2)

in which $e \sim \mathcal{N}(0, I)$, $x \in D$ is a sample from the training dataset $D$ and the noise prediction function $e_{\theta}(\cdot)$ are encouraged to be as close as possible to $e$.

The standard sampling [27] requires many neural function evaluations to get good quality samples. Instead, the faster
We propose to replace the supervised labeling process. We denote the human annotation process as the function $g \colon x \rightarrow g_\phi(x), \forall x \in D$, and $f$ is a self-annotation function parameterized by $\psi$ to map the raw feature representation to the ultimate guidance signal $k, f_\psi : g_\phi(\cdot; D) \rightarrow k$. The guidance signal $k$ can be any form of annotation, e.g., label, box, pixel, that can be paired with an image, which we derive by $k= f_\psi(g_\phi(x; D); D)$. The choice of the self-annotation function $f$ can be non-parametric by heuristically searching over dataset $D$ based on the extracted feature $g_\phi(\cdot; D)$, or parametric by fine-tuning on the feature map $g_\phi(\cdot; D)$.

For the noise prediction function $\epsilon_\theta(\cdot)$, we adopt the traditional UNet network architecture [54] due to its superior image generation performance, following [27, 49, 55, 61].

Stemming from this general framework, we present three methods working at different spatial granularities, all without relying on any ground-truth labels. Specifically, we cover image-level, box-level, and pixel-level guidance by setting the feature extraction function $g_\phi(\cdot)$, self-annotation function $f_\psi(\cdot)$, and guidance signal $k$ to an approximate form.

**Self-labeled guidance.** To achieve self-labeled guidance, we need a self-annotation function $f$ that produces a representative guidance signal $k \in \mathbb{R}^K$. Firstly, we need an embedding function $g_\phi(x), x \in D$ which provides semantically meaningful image-level guidance for the model. We obtain $g_\phi(\cdot)$ in a self-supervised manner by mapping from image space, $g_\phi(\cdot) : \mathbb{R}^{W \times H \times 3} \rightarrow \mathbb{C}$, where $W$ and $H$ are image width and height and $C$ is the feature dimension. We may use any type of feature for the feature embedding function $g$, which we will vary and validate in the experiments. As the image-level feature $g_\phi(\cdot; D)$ is not compact enough for guidance, we further conduct a non-parametric clustering algorithm, e.g., $k$-means, as our self-annotation function $f$. For all features $g_\phi(\cdot)$, we obtain the self-labeled guidance via self-annotation function $f_\psi(\cdot) : \mathbb{C} \rightarrow \mathbb{R}^K$. Motivated by [52], we use a one-hot embedding $k \in \mathbb{R}^K$ for each image to achieve a compact guidance.

We inject the guidance information into the noise prediction function $\epsilon_\theta$ by concatenating it with timestep embedding $t$ and feed the concatenated information $\text{concat}[t, k]$ into every block of the UNet. Thus, the noise prediction function $\epsilon_\theta$ is rewritten as:

$$\epsilon_\theta(x_t, t; k) = \epsilon_\theta(x_t, \text{concat}[t, k]),$$

where $k=f_\psi(g_\phi(x; D); D)$ is the self-annotated image-level guidance signal. For simplicity, we ignore the self-annotation function $f_\psi(\cdot)$ here and in the later text. Self-labeled guidance focuses on image-level global guidance. Next, we consider a more fine-grained spatial guidance.

**Self-boxed guidance.** Bounding boxes specify the location of an object in an image [9, 51] and complement the content information provided by class labels. Our self-boxed guidance approach aims to attain this signal via self-supervised...
models. We represent the bounding box as a binary mask \( k_s \in \mathbb{R}^{W \times H} \) rather than coordinates, where 1 indicates that the pixel is inside the box and 0 outside. This design directly aligns the image and mask along the spatial dimensions. We propose the self-annotation function \( f \) that obtains bounding box \( k_s \) by mapping from feature space \( H \) to the bounding box space via \( f_s(x_t; \cdot, \mathcal{D}) : \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{W \times H} \), and injects the guidance signal along the channel dimension: \( x_t := \text{concat}[x_t, k_s] \). Usually in self-supervised learning, the derived bounding box is class-agnostic \([64, 65]\). To inject a self-supervised pseudo label to further enhance the guidance signal, we again resort to clustering to obtain \( k \) and concatenate it with the timestep embedding \( t := \text{concat}[t, k] \). To incorporate such guidance, we reformulate the noise prediction function \( \epsilon_\theta \) as:

\[
\epsilon_\theta(x_t, t; k_s, k) = \epsilon_\theta(\text{concat}[x_t, k_s], \text{concat}[t, k]),
\]

in which \( k_s \) is the self-supervised box guidance obtained by self-annotation functions \( f_s \), \( k \) is the self-supervised image-level guidance from clustering. \( k_s \) and \( k \) denotes the location and class information, respectively. The design of \( f_s \) is flexible as long as it obtains self-supervised bounding boxes by \( f_s(x_t; \cdot, \mathcal{D}) : \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{W \times H} \). Self-boxed guidance guides the diffusion model by boxes, which specifies the box area in which the object will be generated. Sometimes, we may need an even finer granularity, e.g., pixels, which we detail next.

**Self-segmented guidance.** Compared to a bounding box, a segmentation mask is a more fine-grained signal. Additionally, a multichannel mask is more expressive than a binary foreground-background mask. Therefore, we propose a self-annotation function \( f \) that acts as a plug-in built on feature mapping \( g_\psi(x_t; \cdot, \mathcal{D}) : \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{W \times H \times K} \), where \( K \) is the number of segmentation clusters.

To inject the self-segmented guidance into the noise prediction function \( \epsilon_\theta \), we consider two pathways for injection of such guidance. We first concatenate the segmentation mask to \( x_t \) in the channel dimension, \( x_t := \text{concat}[x_t, k_s] \), to retain the spatial inductive bias of the guidance signal. Secondly, we also incorporate the image-level guidance to further amplify the guidance signal along the channel dimension. As the segmentation mask from the self-annotation function \( f_s \) already contains image-level information, we do not apply the image-level clustering as before in our self-labeled guidance. Instead, we directly derive the image-level guidance from the self-annotation result \( f_\psi \) via spatial maximum pooling: \( \mathbb{R}^{W \times H \times K} \rightarrow \mathbb{R}^K \), and feed the image-level guidance \( k \) into the noise prediction function via concatenating it with the timestep embedding \( t := \text{concat}[t, k] \).

The concatenated results will be sent to every block of the UNet. In the end, the overall noise prediction function for self-segmented guidance is formulated as:

\[
\epsilon_\theta(x_t, t; k_s, k) = \epsilon_\theta(\text{concat}[x_t, k_s], \text{concat}[t, k]),
\]

in which \( k_s \) is the spatial mask guidance obtained from self-annotation function \( f_s \), \( k \) is a multi-hot image-level guidance derived from the self-supervised learning mask \( k_s \).

We have described three variants of self-guidances by setting the feature extraction function \( g_\psi(\cdot) \), self-annotation function \( f_s(\cdot) \), guidance signal \( k \) to an approximate form. In the end, we arrive at three noise prediction functions \( \epsilon_\theta \), which we utilize for diffusion model training and sampling, following the standard guided \([28]\) diffusion approach as detailed in Section 3.1.

### 4. Experiments

In this section, we aim to answer the overarching question: Can we substitute ground-truth annotations with self-annotations? First, we consider the image-label setting, in which we examine what kind of self-labeling is required to improve image fidelity. In addition, we explore what semantic concepts are induced by self-labeling approaches that broaden the control over the content beyond the standard ground-truth labels. Next, we look at image-bounding box pairs. Finally, we examine whether it is possible to gain fine-grained control with self-labeled image-segmentation pairs. We first present the general settings relevant for all experiments.

**Evaluation metric.** We evaluate both diversity and fidelity of the generated images by the Fréchet Inception Distance (FID) \([26]\), as it is the de facto metric for the evaluation of generative methods, e.g., \([8, 17, 30, 55]\). It provides a symmetric measure of the distance between two distributions in the feature space of Inception-V3 \([62]\). We use FID as our main metric for the sampling quality.

**Baselines & implementation details.** As baselines, we compare against both the unconditional diffusion model and a diffusion model trained with classifier-free guidance using ground-truth annotations \([28]\). We use the same neural network and hyperparameters for the baselines and our method. Note that applying more training steps generally tends to further improve the performance \([32, 55]\), thus to facilitate a fair comparison we use the same computational budget in every experiment when comparing the baselines to our proposed method. We use DDIM \([59]\) samplers with 250 steps, \( \sigma_r = 0 \) to efficiently generate samples. For details on the hyperparameters, we refer to Appendix B.
We consider 1 to 10,000 clusters on the extracted CLS token. We also measure the Inception Score (IS) \[56\], following to put our sampling results in perspective, we also provide Table 1.

<table>
<thead>
<tr>
<th></th>
<th>FID↓</th>
<th>IS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Label-supervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet50</td>
<td>22.00</td>
<td>8.23</td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>22.30</td>
<td>7.81</td>
</tr>
<tr>
<td><strong>Self-supervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE ViTBase</td>
<td>32.58</td>
<td>8.20</td>
</tr>
<tr>
<td>SimCLR-v2</td>
<td>23.16</td>
<td>9.35</td>
</tr>
<tr>
<td>MSN ViT-B/16</td>
<td>21.16</td>
<td>10.59</td>
</tr>
<tr>
<td>DINO ViT-B/16</td>
<td>19.35</td>
<td>10.41</td>
</tr>
</tbody>
</table>

Table 1. Choice of feature extraction function on ImageNet32. DINO and MSN ViT-B/16 obtain good trade-offs between FID and IS.

4.1. Self-Labeled Guidance

We use ImageNet32/64 \[16\] and CIFAR100 \[33\] to validate the efficacy of self-labeled guidance. On ImageNet, we also measure the Inception Score (IS) \[56\], following common practice \[8, 17, 30\]. IS measures how well a model fits into the full ImageNet class distribution.

Choice of feature extraction function \(g\). We first measure the influence of the feature extraction function \(g\) used before clustering. We consider two supervised feature backbones: ResNet50 \[25\] and ViT-B/16 \[18\], and four self-supervised backbones: SimCLR \[13\], MAE \[23\], MSN \[4\] and DINO \[11\]. To assure a fair comparison we use 10k clusters for all architectures. From the results in Table 1, we make the following observations. First, features from the supervised ResNet50, and ViT-B/16 lead to a satisfactory FID performance, at the expense of relatively limited diversity (low IS). However, they still require label annotation, which we strive to avoid in our work. Second, among the self-supervised feature extraction functions, the MSN- and DINO-pretrained ViT backbones have the best trade-off in terms of both FID and IS. They even improve over the label-supervised backbones. This implies that the benefits of guidance is not unique to human annotated labels and self-supervised learning can provide a much more scalable alternative. Since DINO ViT-B/16 achieves the best FID performance, from now on we pick it as our self-supervised feature extraction function \(g\).

Effect of number of clusters. Next, we ablate the influence of the number of clusters on the overall sampling quality. We consider 1 to 10,000 clusters on the extracted CLS token from the DINO ViT-B/16 feature. For efficient comparison, we train each version for 20 epochs on ImageNet32. To put our sampling results in perspective, we also provide results for the no guidance and ground-truth guidance.

In Figure 2 we see that our model’s performance improves monotonically as the cluster number increases from 1 to 5,000, consistently outperforming the no guidance baseline. At 1,000 clusters, self-labeled guidance is competitive with the baseline trained using ground-truth labels. For 5,000 clusters, we find a sweet spot where our method outperforms the model using ground-truth labels, with an FID of 16.4 versus 17.9 and an IS of 10.35 versus 9.94. We can understand this result by considering (pseudo-)label conditioning as a method of transforming a single diffusion model into multiple specialized models, with each one focused on a distinct set of semantically coherent images. Increasing the granularity of the groups, such as by increasing the number of clusters to 5,000, improves the semantic coherence of each group and simplifies the distribution. However, if the cluster number becomes too large, the self-supervised clusters may pick up on dataset-specific details that no longer correspond to general semantic concepts, leading to a deterioration in cluster quality and FID performance. Nevertheless, we observe that samples generated from the same cluster ID exhibit high semantic coherence, indicating that the self-supervised clusters represent meaningful concepts that can be used to control the generation process. We discuss assigning semantic descriptions to clusters further in Appendix C.

Importance of self-supervised clusters. In the previous paragraph, we observed that training a diffusion model with 5,000 clusters can outperform the 1,000 ground-truth labels. Here, we check whether we can reproduce this result on another dataset and examine how the performance varies when we inject different degrees of noise into the cluster assignment. On CIFAR100 \[33\] we compare the ground-truth 100 labels with 400 self-supervised clusters. We corrupt the cluster assignments at different levels by randomly shuffling the cluster id for 25% to 100% of the images before training. The results in Figure 3 highlight the importance of using self-supervised features for assigning clusters and that assigning cluster ids for a subset of the dataset is already sufficient to see improvements.

Self-labeled comparisons on ImageNet32/64. We compare our self-labeled guidance method against the baseline trained with ground-truth labels. For a fair comparison, we use the same compute budget for all runs. In particular, each model is trained for 100 epochs taking around 6 days on four RTX A5000 GPUs. Results on ImageNet32 and ImageNet64 are in Table 2. Similar to \[17\], we observe that any guidance setting improves considerably over the unconditional & no-guidance model. Surprisingly, our self-labeled model even outperforms the ground-truth labels by a large gap in terms of FID of 1.9 and 4.7 points respectively. We hypothesize that the ground-truth taxonomy might be suboptimal for learning generative models and the self-supervised clusters offer a better guidance signal due to better alignment with the visual similarity of the images. In Figure 4 we report...
Figure 2. Effect of number of clusters. Self-labeled guidance outperforms DDPM without any guidance beyond a single cluster, is competitive with classifier-free guidance beyond 1,000 clusters and is even able to outperform guidance by ground-truth (GT) labels for 5,000 clusters. We visualize generated samples from ImageNet64 (middle) and ImageNet32 (right) for ground-truth labels guidance (top) and self-labeled guidance (bottom). More qualitative results in Appendix C.

Figure 3. Corruption of cluster assignments on CIFAR100. Self-labeled guidance with 400 clusters outperforms the baseline trained with ground-truth labels. The FID performance deteriorates monotonically with the percentage of corrupted cluster assignments, underscoring the importance of assigning cluster ids via the self-supervised features.

Table 2. Self-labeled comparisons on ImageNet32/64. Self-labeled guidance surpasses the no-guidance baseline by a large margin on both datasets and even outperforms the guided diffusion model trained using ground-truth class labels.

Table 3 shows that self-labeled guidance significantly outperforms the baselines, indicating the effectiveness of our method for high-resolution images. Note that the lack of ground-truth labels for LSUN-Churches reflects an advantage of our method since most real-world images are unlabeled. We show generated samples for two different clusters in Figure 5. The samples are diverse and reflect shared characteristics for samples guided by the same cluster. For more qualitative and quantitative results we refer to Appendix A.

Figure 4. Performance at different compute budgets. After training for ~100 GPU hours self-labeled guidance achieves persistent FID reduction over training with ground-truth labels.

Higher resolution image generation. Finally, we verify the effectiveness of self-labeled guidance on larger images. We report on the ImageNet-100 [63] (a subset of ImageNet-1k with 100 classes) and the LSUN-Churches dataset [68], both with images of size 256 × 256. Notably, the latter does not come with any annotations rendering a ground-truth guided baseline infeasible. Too limit computation, we use the Latent Diffusion Model [53] which is much more efficient at training with large image sizes than directly learning a diffusion model in the pixel space.

Table 3 shows that self-labeled guidance significantly outperforms the baselines, indicating the effectiveness of our method for high-resolution images. Note that the lack of ground-truth labels for LSUN-Churches reflects an advantage of our method since most real-world images are unlabeled. We show generated samples for two different clusters in Figure 5. The samples are diverse and reflect shared characteristics for samples guided by the same cluster. For more qualitative and quantitative results we refer to Appendix A.
Higher resolution image generation. ImageNet-100 and LSUN-Churches results for images of size 256×256.

Self-boxed guidance. For the ground-truth labels guidance baseline, we condition on a class embedding. Since there are now multiple objects per image, we represent the ground-truth class with a multi-hot embedding. Aside from the class embedding which is multi-hot in our method, all other settings remain the same for a fair comparison. The results in Table 4, confirm that the multi-hot class embedding is indeed effective for multi-label datasets, improving over the no-guidance model by a large margin. This improvement comes at the cost of manually annotating multiple classes per image. Self-boxed guidance further improves upon this result, by reducing the FID by an additional 5.1 and 3.3 points respectively without using any ground-truth annotation. In Figure 6, we show our method generates diverse and semantically well-aligned images.

Self-segmented guidance. For COCO-Stuff, any form of guidance results in a considerable FID reduction over the no-guidance model. Self-segmented guidance improves over ground-truth multi-hot labels guidance and narrows the gap with guidance by annotation-intensive ground-truth masks. The results in Table 5 demonstrate that our self-segmented guidance, we apply STEGO as our self-annotation function in Equation (6). We set the cluster number to 27 for COCO-Stuff, and 21 for Pascal VOC, following STEGO. We train all models on images of size 64×64, for 800 epochs on Pascal VOC, and for 400 epochs on COCO-Stuff. We report the train FID for Pascal VOC and both train and validation FID for COCO-Stuff. More details on the dataset and experimental setup are provided in Appendix B.

Self-segmented guidance still outperforms the ground-truth labels guidance baseline on both datasets. The comparison between ground-truth labels and segmentation masks reveals an improvement in

![Generated samples at 256 × 256 resolution using self-labeled guidance on LSUN-Churches.](image)

Table 3. Higher resolution image generation. ImageNet-100 and LSUN-Churches results for images of size 256×256.

<table>
<thead>
<tr>
<th>Diffusion Method</th>
<th>Annotation free?</th>
<th>ImageNet-100 FID</th>
<th>LSUN-Churches FID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-truth labels guidance</td>
<td>✓</td>
<td>21.2</td>
<td>64.1</td>
</tr>
<tr>
<td>No guidance</td>
<td>✓</td>
<td>42.1</td>
<td>41.1</td>
</tr>
<tr>
<td>Self-labeled guidance</td>
<td>✓</td>
<td>16.1</td>
<td>78.3</td>
</tr>
</tbody>
</table>

Table 4. Self-boxed comparisons on Pascal VOC and COCO_20K. Self-boxed guidance outperforms the no-guidance baseline FID considerably for multi-label datasets and is even better than a label-supervised alternative.

<table>
<thead>
<tr>
<th>Diffusion Method</th>
<th>Annotation free?</th>
<th>Pascal VOC Train</th>
<th>Pascal VOC Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-truth labels guidance</td>
<td>✓</td>
<td>21.5</td>
<td>19.3</td>
</tr>
<tr>
<td>No guidance</td>
<td>✓</td>
<td>58.6</td>
<td>42.5</td>
</tr>
<tr>
<td>Self-boxed guidance</td>
<td>✓</td>
<td>18.4</td>
<td>16.0</td>
</tr>
<tr>
<td>Ground-truth boxes guidance</td>
<td>✓</td>
<td>13.2</td>
<td>9.6</td>
</tr>
</tbody>
</table>

Table 5. Self-segmented comparisons on Pascal VOC and COCO-Stuff. Any form of guidance results in a considerable FID reduction over the no-guidance model. Self-segmented guidance improves over ground-truth multi-hot labels guidance and narrows the gap with guidance by annotation-intensive ground-truth masks.

4.2. Self-Boxed Guidance

We run experiments on Pascal VOC and COCO_20K to validate the efficacy of self-boxed guidance. The self-boxed guidance model takes a bounding box in addition to the cluster-ID as guidance signal. To obtain the class-agnostic object bounding boxes, we use LOST as our self-annotation function \( f \) in Equation (6). For the clustering, we empirically found \( k=100 \) to work well for both datasets, as both are relatively small in scale when compared to ImageNet. We train our diffusion model for 800 epochs with images of size 64x64. We report train FID for Pascal VOC and train/validation FID for COCO_20K. We evaluate the performance on the validation split by extracting the guidance signal from the training dataset to ensure that there is no information leakage. See Appendix B for more details.

4.3. Self-Segmented Guidance

Finally, we validate the efficacy of self-segmented guidance on Pascal VOC and COCO-Stuff. For COCO-Stuff we follow the split from [15, 22, 29, 70], with a train set of 49,629 images and a validation set of 2,175 images. Classes are merged into 27 (15 stuff and 12 things) categories. For self-segmented guidance, we apply STEGO as our self-annotation function \( f \) in Equation (6). We set the cluster number to 27 for COCO-Stuff, and 21 for Pascal VOC, following STEGO. We train all models on images of size 64×64, for 800 epochs on Pascal VOC, and for 400 epochs on COCO-Stuff. We report the train FID for Pascal VOC and both train and validation FID for COCO-Stuff. More details on the dataset and experimental setup are provided in Appendix B.
image quality when using the more fine-grained segmentation mask as the condition signal. But these segmentation masks are one of the most costly types of image annotations that require every pixel to be labeled. Our self-segmented approach avoids the necessity for annotations while narrowing the performance gap, and more importantly offering fine-grained control over the image layout. We demonstrate this controllability with examples in Figure 7 and explain how to assign semantic descriptions to the clusters in Appendix C. These examples further highlight a robustness against noise in the segmentation masks, which our method acquires naturally due to training with noisy segmentations.

5. Conclusion

We have explored the potential of self-supervision signals for diffusion models and propose a framework for self-guided diffusion models. By leveraging a feature extraction function and a self-annotation function, our framework provides guidance signals at various image granularities: from the level of holistic images to object boxes and even segmentation masks. Our experiments indicate that self-supervision signals are an adequate replacement for existing guidance methods that generate images by relying on annotated image-label pairs during training. Furthermore, both self-boxed and self-segmented approaches demonstrate that we can acquire fine-grained control over the image content, without any ground-truth bounding boxes or segmentation masks. Though in certain cases, clusters can capture visual concepts that are challenging to articulate in everyday language, such as in the case of LSUN-Churches. For future research, it would be interesting to investigate the efficacy of our self-guidance approach on feature extractors trained on larger datasets or with image-text pairs [48]. Ultimately, our goal is to enable the benefits of self-guided diffusion for unlabeled and more diverse datasets at scale, wherein we believe this work is a promising first step.

Acknowledgements. The work of DWZ is part of the research programme Perspectief EDL with project number P16-25 project 3, which is financed by the Dutch Research Council (NWO) domain Applied and Engineering Sciences (TTW). We thank EscherCloud AI for the European compute resources.
References


[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In CVPR, 2022. 5


