Improving Geo-diversity of Generated Images with Contextualized Vendi Score Guidance

Reyhane Askari Hemmat[1 ⋆], Melissa Hall[1 ∗], Alicia Sun[1], Candace Ross[1], Michal Drozdzal[1], and Adriana Romero-Soriano[1,2,3,4]

[1] FAIR at Meta, [2] Mila, [3] McGill University, [4] Canada CIFAR AI Chair

Abstract. With the growing popularity of text-to-image generative models, there has been increasing focus on understanding their risks and biases. Recent work has found that state-of-the-art models struggle to depict everyday objects with the true diversity of the real world and have notable gaps between geographic regions. In this work, we aim to increase the diversity of generated images of common objects such that per-region variations are representative of the real world. We introduce an inference-time intervention, *contextualized Vendi Score Guidance (c-VSG)*, that guides the backwards steps of latent diffusion models to increase the diversity of a sample as compared to a "memory bank" of previously generated images while constraining the amount of variation within that of an exemplar set of real-world contextualizing images. We evaluate c-VSG with two geographically representative datasets and find that it substantially increases the diversity of generated images, both for the worst-performing regions and on average, while simultaneously maintaining or improving image quality and consistency. Additionally, qualitative analyses reveal that diversity of generated images is significantly improved, including along the lines of reductive region portrayals present in the original model. We hope that this work is a step towards text-to-image generative models that reflect the true geographic diversity of the world.[1]

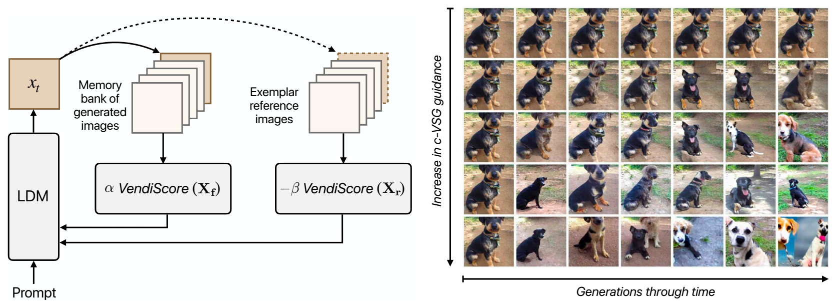
Keywords: geo-diversity · image generation · diffusion models

1 Introduction

The unprecedented results achieved by text-to-image systems [15,32,36,39] and their wide-spread use as plug-and-play solutions have propelled a body of research to understand their potential risks and biases [1,8,27]. Recent works have highlighted through qualitative and quantitative evaluations the disparities in demographic traits of people represented in the generated images [19,31,49], triggering the design of mitigation strategies based on textual interventions [16], attention-weight modification [17], and semantic guidance [14].

⋆ Equal Contribution.

[1] https://github.com/facebookresearch/Contextualized-Vendi-Score-Guidance



(a) (b)

Fig. 1: (a) We present **Contextualized Vendi Score Guidance (c-VSG)**, an inference-time intervention to increase the diversity of images generated by latent diffusion models (LDMs). c-VSG guides backwards steps of the diffusion process using the Vendi Score [13] to increase the diversity among a sample x_t and a memory bank of previous generations (with weight α) while constraining excessive variation using a small set of real, contextualizing exemplar images (with weight β). (b) Generations of `dog in Africa`, *all with the same seed.* First row has zero c-VSG guidance scale and as a result all samples are the same. As we increase the c-VSG guidance scale, we observe increased diversity in generations.

Beyond human-centric representations, research has shown that stereotypical biases also occur for representations of objects and their surroundings across the globe [2, 4, 18]. In particular, previous studies highlighted that progress in image quality or realism has come at the expense of representation diversity and text-image consistency [18]. This trade-off affects some regions more than others, hindering the progress towards conditional image generative systems that truly work well for all geographic regions. Notably, images generated with text prompts that include regional information contain limited object diversity and heightened geographic stereotypes. As shown in [18], many generations depict *limited diversity in object type*. Yet, to the best of our knowledge, no mitigation strategies have directly targeted geo-diversity in text-to-image systems.

In this paper, we aim to mitigate disparities in representation diversity of common objects across worldwide regions. More precisely, our goal is to increase the objects' representation diversity such that per-region variation in type of object, as observed in the real world, is better reflected in the generations. In doing so, we aim to maintain (or improve) the quality of images as well as their text-image consistency.

We devise a novel approach called Vendi Score Guidance (VSG) that increases the representation diversity of the generations by leveraging the Vendi Score (VS) [13] – a dataset diversity evaluation metric. In particular, VSG extends the guidance toolbox of diffusion models by driving the generation process towards samples that are substantially different from each other, intuitively maximizing the effective number of samples among the obtained generations. This requires the proposed guidance strategy to operate *auto-regressively* w.r.t. *the previous generations*. We implement this by keeping a *memory bank* of past generations that ensures that the current sample differs from those in the memory.do Yet, increasing diversity in an unconstrained way results in generations that lack either image realism or text-generation consistency. Therefore, we propose to contextualize VSG by providing a small set of randomly selected real images of objects, referred to as *exemplar images*, and guiding the generation process towards diverse samples that are grounded on those exemplar images. We evaluate the proposed approach by measuring precision [26,40], recall [26,40], and CLIPScore [24] to assess image realism/quality, representation diversity, and text-image consistency, respectively. We report results on two geographically diverse datasets that contain images of the common objects collected around the world, GeoDE [35] and DollarStreet [38].

Our results show that contextualized Vendi Score Guidance improves worst-region and average F1 for both datasets, including a relative improvement of 40% in worst-region F1 over the Latent Diffusion Model (LDM) baseline without any intervention when measured on GeoDE and 11% over the next closest baseline, and reduces performance disparities across regions. Even without contextualization, VSG shows relative improvements of up to 10% over the LDM baseline in F1. Upon visual inspection we find the object diversity in images to be significantly improved, with greater variation in object color, type, shape, and size. In addition, VSG improvements often coincide with improved regional representations beyond reductive portrayals present in the original model. Finally, we find in ablations that our combined use of the Vendi Score to encourage image diversity relative to previous generations while constraining variation to that of the real world with contextualizing exemplar images meaningfully navigates the diversity *vs*. quality tradeoff.

The contributions of this work can be summarized as follows:

- We introduce a new inference-time intervention, contextualized Vendi Score Guidance (c-VSG), to increase the diversity of images generated by LDMs using a memory bank of previously generated images and exemplar images.
- When evaluated with two geographically representative real world datasets, c-VSG shows significant improvement in diversity and quality over the vanilla LDM, while maintaining text-image consistency.
- c-VSG shows improvements in average and worst-region F1 over state-of-the-art baselines, while exhibiting reduced disparities in quality, diversity and text-generation consistency across regions.
- Our ablations demonstrate the efficacy of the memory and contextualizing components of the c-VSG criterion.

We hope this work contributes to image generations that better reflect the geographic diversity of the real world.

2 Method

In this work, we focus on increasing the diversity of latent diffusion models (LDMs). We first provide preliminary details about the components of the LDM generation process in which we apply our interventions, as well as the Vendi Score metric. Then, we introduce our proposed methodology, Vendi Score Guidance.

2.1 Preliminaries

Sampling in LDMs. In this work, we focus on a class of LDMs called denoising diffusion implicit models (DDIMs) [45]. At inference or generation time, DDIMs use a reverse process defined as:

$$x_{t-1} = \sqrt{\xi_{t-1}}\hat{x}_{0,t} + \underbrace{\sqrt{1-\xi_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(x_t)}_{\text{direction pointing to } x_t} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}, \tag{1}$$

where x_t is the sample at time step t, $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and ξ_t and σ_t are time-dependent coefficients, and $\hat{x}_{0,t}$ is the DDIM approximation of the denoised sample given current x_t,

$$\hat{x}_{0,t} = \mathrm{DDIMApprox}(x_t) := \frac{x_t - \sqrt{1-\xi_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\xi_t}}, \tag{2}$$

and $\epsilon_\theta^{(t)}$ is the pretrained denoising network parameterized by θ that approximates the score function conditioned on a class or prompt y,

$$\epsilon_\theta^{(t)} \approx \nabla_x \log p(x_t|y). \tag{3}$$

In this work, we modify this score function to guide the sampling process towards generating more diverse samples.

Guidance in LDMs. LDMs often condition generated samples on labels or text prompts in the score function. If one has access to a pretrained denoising network that only approximates the *unconditional* score function ($\nabla_x \log p(x)$), one can use classifier-guidance [12] to generate a *conditional* sample. Classifier-guidance modifies the score function using a pretrained classifier that provides $\log p(y|x_t)$,

$$\epsilon_\theta^{(t)} \approx \nabla_x \log p_\gamma(x_t|y) = \nabla_x \log p(x_t) + \gamma\nabla_x \log p(y|x_t), \tag{4}$$

where γ is a scaling factor controlling the strength of class-conditional generation. In this work, we adapt this concept of guidance by utilizing the Vendi Score [13], rather than a classifier, as our guidance function.

Vendi Score (VS). The VS is a metric for evaluating diversity in machine learning [13]. It applies a user-defined similarity function to a set of samples, without the need for a reference dataset as opposed to metrics such as recall or coverage [26,30,40]. VS measures the effective number of examples by computing the soft-rank of a similarity or kernel matrix. The similarity matrix is calculated using a similarity function that takes two samples and returns a positive value indicating how similar they are. Higher rank of the similarity matrix corresponds to higher Vendi Score and diversity.

Formally, given a set of samples $x_1, \ldots, x_n$, we can define a positive semi-definite similarity function k with $k(x, x) = 1$ for all x. The similarity function is applied on every pair of samples in our dataset to create a similarity matrix, $\mathcal{K}$ where $\mathcal{K}_{i,j} = k(x_i, x_j)$. VS is then defined as the soft-rank of $\mathcal{K}/n$, *i.e.*, the exponential of the entropy of the eigenvalues of $\mathcal{K}/n$ [13]:

$$\mathrm{VS}(x_1, \ldots, x_n) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right), \tag{5}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathcal{K}/n$.

A common choice for the similarity function is the cosine-similarity of every pair of samples in the dataset in a pre-defined feature space.

2.2 Vendi Score Guidance

In this section, we introduce our proposed approach, called Vendi Score Guidance (**VSG**). Our goal is to increase the diversity of the generated samples by incorporating the Vendi Score as an auto-regressive guidance mechanism during diffusion model inference. In other words, every new generation will take into consideration all previously generated samples and will aim to generate a sample which is substantially different from the previous ones.

Let us define a *memory bank* which contains all the previously generated samples as $\mathbf{X_f} : \{x^1, ..., x^n\}$. Given the $\mathbf{X_f}$, and a newly generated sample x, we can compute $\mathrm{VS}(x, \mathbf{X_f})$. Our goal is to modify the generation process of x by steering it towards part of the manifold that increases VS, and thus, augment the sample diversity. We can achieve this via the following score function,

$$\epsilon_\theta^{(t)} \approx \nabla_x \log p_{\gamma,\alpha}(x_t|y, \mathbf{X_f}) = \nabla_x \log p_\gamma(x_t|y) + \alpha \nabla_x \mathrm{VS}(\hat{x}_{0,t}, \mathbf{X_f}), \tag{6}$$

where α is a scaling factor that controls the strength of the Vendi Score guidance and $\hat{x}_{0,t}$ is the predicted denoised sample.

A naive way of finding $\hat{x}_{0,t}$ is to apply all the T steps of the backward diffusion process and use the denoised image to compute the Vendi Score. However, this is computationally expensive. Thus, we use the efficient DDIM approximation of a denoised sample, given in Eq. 2.

We note that the original Vendi Score formulation in Eq. 5 is not differentiable, since it requires finding the singular values of the similarity matrix, and singular value decomposition is not differentiable. To overcome this limitation, we developed a simple algorithm to compute the Vendi Score in a differentiable way and present it in Appendix 6.2.

Contextualized Vendi Score Guidance (c-VSG). In experimentation, we found that unconstrained VSG can have limited efficacy, as generated images may drift from real world representations. Thus, in the second part of our method, we use Vendi Score a second time with a small set of randomly selected exemplar images, $\mathbf{X_r}$. This set of exemplar images is used to contextualize the Vendi Score computation and thus, the generation process. Similar to Eq. 6, we compute $\mathrm{VS}(x, \mathbf{X_r})$. However, because we want to *reduce* excessive drift from the exemplar representation of images, we calculate the gradients with respect to the *negative* of this score. Overall, we use the following score function which combines VSG with contextualized samples:

$$\nabla_x \log p(x_t|y, \mathbf{X_f}, \mathbf{X_r}) = \nabla_x \log p_{\gamma,\alpha}(x_t|y, \mathbf{X_f}) - \beta \nabla_x \mathrm{VS}(\hat{x}_{0,t}, \mathbf{X_r}), \tag{7}$$

where β is a scaling factor that controls the strength of VSG contextualization. This formulation allows the model to generate samples that remain close to the exemplar images without the need for additional training. Intuitively, the modified score function (Eq. 7) steers the generation process towards parts of the space which would result in a sample that increases the rank of the memory bank while staying close to the exemplar images.

Algorithm 1 presents the c-VSG computation step by step. Note that we use $Gfreq$ to control the rate at which c-VSG is applied in the diffusion process, allowing for a balance between efficiency and diversity. To generate image x^n, we apply the the proposed guidance in the backward diffusion process. In the intermediate steps and when the sample is not yet generated, we use DDIM's approximation of $\hat{x}_0$ and feed it to the VS computation. After the final diffusion step and once the image is completely denoised, x^n is added to the memory bank of generated images and used for future image generations.

3 Experiments

In this section, we describe evaluation datasets and metrics, baselines, and set-up. We then present experimental results and ablations.

3.1 Datasets and metrics

In our experiments, we report worst-region and average F1, recall, precision, and CLIPScore. We follow [18] and evaluate performance *w.r.t.* two geographically diverse datasets containing images of objects in multiple regions: GeoDE [35] and DollarStreet [38]. Both datasets contain images of objects in their everyday settings and are intended to be geographically representative. However, images in GeoDE were submitted by people living in the regions of interest and were collected with specific guidelines, *e.g.*, that the object of focus fills at least 25% of the image, while images in DollarStreet were collected by photographers who travelled to various regions, with a special focus on “disadvantaged and isolated areas” [38]. Thus, DollarStreet likely has more income variation than GeoDE.

Algorithm 1 Contextualized Vendi Score Guidance (c-VSG)

Require: Class label y, a single fake image x_f, a tensor of M real images $\mathbf{X_r}$, pre-trained diffusion model ϵ_θ, N number of generations to obtain, T number of diffusion steps, $Gfreq$ guidance frequency, α, β and γ guidance weights

Ensure: $\mathbf{X_f}$ a tensor with N diverse model samples

```
X_f ← x_f                                          ▷ Tensor with diverse model samples
for n = 1, ..., N − 1 do
    x^n ∼ N(0, I)                                  ▷ Initialize sample with random noise
    for t = T, ..., 1 do                           ▷ Apply backward diffusion steps
        if t % Gfreq == 0 then
            x̂^n_{0,t} = DDIMApprox(x^n_t)                                      ▷ Eq. 2
            ε_θ^(t) ← ∇_x log p_γ(x_t|y) + α∇_x VS(x̂^n_{0,t}, X_f) − β∇_x VS(x̂^n_{0,t}, X_r)   ▷ Eq. 7
            x^n_{t−1} ← one step of DDIM                                       ▷ Eq. 1
        else
            ε_θ^(t) ← ∇_x log p_γ(x_t|y)
            x^n_{t−1} ← one step of DDIM                                       ▷ Eq. 1
        end if
    end for
    X_f ← cat(X_f, x^n_0)                          ▷ Add final denoised sample to the memory
end for
```

Furthermore, DollarStreet contains large imbalances between regions. We filter GeoDE to obtain 180 images from each region for a total of 27 objects. The remaining images are used as the pool for randomly selected exemplar real images. We subsample DollarStreet so that images contain only one concrete class, *e.g.*, removing classes corresponding to favorite objects and images with multiple classes, and use the original distribution exclusive of the images used for contextualizing. For CLIPScore, we compute the average of the $10th$-percentile for each object class, as prior work suggests that the lower tail of the distribution tends to be most informative in capturing generation inconsistencies [18,20].

3.2 Baselines

We split our discussion between baselines that do not utilize any information specific to the task in order to improve generation diversity, and those that use small amounts of information from the task. To see examples of generations for different methods, see Appendix 6.4.

Without any additional information:

- `LDM`: This is the baseline setup where the unaltered LDM is used with the prompt `{object} in {region}` to condition the generation process.
- `Synonyms`: This strategy maps each object class to its corresponding ImageNet [11] class and WordNet [29] synset. Each synset contains lemmas representing a specific "sense" of the given class. We generate images stratified across all possible lemmas (including the original object word), for each class, *i.e.*, `{synonym} in {region}` is used to condition the generation process. The list of synonyms can be found in Appendix 6.1.

With additional information:

- `Paraphrasing`: We use the LLaMA-2-70B-chat [48] large language model to generate paraphrases of the original prompt template, `{object} in {region}`. We include the specifications and descriptions used in the collection of GeoDE and DollarStreet. The metaprompts and paraphrases, as well as our method of tuning prompts and model specifications, are included in Appendix 6.1.
- `Semantic Guidance`: As observed in [18] and highlighted in Fig. 3, generated images tend to contain diversity issues related to the magnification of region-specific representations of objects beyond those in the evaluation task. To reduce this amplification of region information in the generated images, we employ Semantic Guidance [6,14] by adding negative guidance corresponding to the region term for each generation. Note that this baseline requires knowing the biases embedded in the text-to-image system upfront.
- `Feedback Guidance (FG)`: Following [22,43], during generation we use feedback from an external CLIP-based [34] classifier to predict region labels of the generated images[2]. We experiment with two versions of feedback guidance to encourage more diverse generations: one that maximizes the `loss` of the classifier and one that maximizes the `entropy` of predicted class distributions.
- `Textual Inversion`: Textual inversion injects a learned token embedding using small subset of images from the evaluation task to better represent the reference images [16]. We apply textual inversion by learning an embedding for each object in the evaluation dataset using four images per object.

3.3 Experimental set-up

We perform all experiments using an open sourced version of an LDM trained on a large scale dataset [39]. For c-VSG, we use $\alpha = 1$ and $\beta = 2$ unless otherwise specified, with two exemplar images as real world context and set $Gfreq$ to 5. We calculate the Vendi Score in the CLIP ViT-B32 [34] feature space. For each method, we perform a hyperparameter search across guidance scales, as well as the weighting of contextual information, where applicable. We perform one-region-out hyper-parameter selection, selecting the best hyper-parameters for a given region based on the F1 for all the other regions. For computational complexity and efficiency see Appendix Sec. 6.5.

3.4 Results

Overall results. On GeoDE (reported in Tab. 1), VSG without contextualization achieves a **relative** improvement in average F1 of 9.6% over generations obtained with the LDM (VSG achieves F1 of 0.399 and LDM achieves F1 of 0.364), an improvement over baselines that do not leverage any additional information. Similarly, VSG yields relative improvement in worst-region F1 by 10.6%.

[2] In this baseline, we do not implement the additional real/fake discriminator as suggested in [43], as we focus only on methods that utilize existing classifiers.

Table 1: Comparison to baselines on GeoDE. Contextualized Vendi Score Guidance contributes to greater diversity improvements (recall) with increases, or little cost to, quality (precision) and consistency (CLIPScore), both on average and for the worst performing region (determined by F1). AF: Africa, WAS: West Asia, "label" refers to region label, "desc" to text description, and "img" to exemplar images.

Method	**Ref. Info.**	**Worst-Reg.**	**F1**		**Precision**		**Recall**		**CLIPScore**	
			Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.
LDM	✗	AF	0.364	0.322	0.413	0.273	0.337	0.395	0.242	0.218
Synonyms	✗	AF	0.357	0.306	0.350	0.298	0.366	0.315	0.215	0.203
VSG (ours)	✗	AF	0.399	0.356	0.349	0.307	0.470	0.424	0.180	0.191
Paraphrasing	desc	WAS	0.384	0.354	0.338	0.309	0.449	0.415	0.231	0.228
Semantic Guidance	label	WAS	0.420	0.401	**0.459**	**0.519**	0.391	0.326	0.245	**0.253**
FG CLIP (Loss)	label	WAS	0.409	0.378	0.387	0.383	0.436	0.373	0.228	0.223
FG CLIP (Ent.)	label	AF	0.380	0.337	0.344	0.329	0.429	0.345	0.224	0.227
Textual Inversion	img	AF	0.369	0.363	0.409	0.444	0.338	0.308	0.234	0.232
c-VSG (Ours)	img	AF	**0.455**	**0.444**	0.424	0.417	**0.493**	**0.476**	**0.254**	0.**253**

Table 2: Comparison to baselines on DollarStreet. Contextualized Vendi Score Guidance contributes to greater diversity improvements (recall) with little cost to quality (precision) and consistency (CLIPScore), both on average and for the worst performing region (determined by F1). c-VSG outperforms all the methods by at least 8% relative F1 improvement. AF: Africa, AS: Asia, "label" refers to region label, "desc" to text description, and "img" to exemplar images.

Method	**Ref. Info.**	**Worst-Reg.**	**F1**		**Precision**		**Recall**		**CLIPScore**	
			Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.
LDM	✗	AS	0.448	0.442	0.428	0.434	0.472	0.450	0.231	0.235
Synonyms	✗	AS	0.464	0.457	0.451	0.448	0.479	0.467	0.216	0.220
VSG (ours)	✗	AS	0.457	0.444	0.413	0.388	0.516	0.518	0.191	0.198
Paraphrasing	desc	AF	0.454	0.445	0.445	0.454	0.465	0.437	0.226	0.215
Semantic Guidance	label	AS	0.470	0.458	0.447	0.449	0.498	0.467	0.230	0.233
FG CLIP (Loss)	label	AS	0.437	0.394	0.401	0.321	0.488	0.510	0.223	0.206
FG CLIP (Entropy)	label	AS	0.465	0.462	0.412	0.404	0.**535**	**0.540**	0.222	0.219
Textual Inversion	img	AS	0.425	0.398	0.478	**0.491**	0.386	0.335	0.217	0.219
c-VSG (Ours)	img	AS	**0.497**	**0.483**	**0.486**	0.486	0.511	0.479	**0.234**	**0.238**

Interestingly, using Synonyms tends to decrease performance, perhaps due to its enforcement of a strict, yet limited, distribution of object names.

When considering methods that include additional information, we observe that leveraging textual descriptions of the evaluation dataset (Paraphrasing) is useful to improve average and worst-region F1. However, leveraging region information (Semantic Guidance, FG CLIP) yields better results. Notably, when the additional region information corresponds to *known biases* embedded in text-to-image systems, as is the case of Semantic Guidance, we observe substantial relative performance improvements of 15.4% and 24.5% in average and worst-region F1 over the LDM baseline.

Leveraging exemplar images to contextualize VSG results in further performance gains. In particular, c-VSG shows a 25% and 37.9% relative improvement in average and worst F1 compared to the LDM baseline. Comparing c-VSG with

(a) No VSG, *i.e.*, `{object} in {region}` baseline

(b) Vendi Score Guidance

(c) Contextualized Vendi Score Guidance

Fig. 2: Generated images of cooking pots (**Left**) and cars (**Right**). The same six seeds are shared among the examples, and the box colors indicate images pertaining to Africa, Europe, and Southeast Asia. Vendi Score Guidance increases the diversity of generated images, including object type, positioning, and quality. Contextualization with exemplar images increases similarity to real world diversity. (More examples are shown in Appendix Figures 7 and 8.)

the closest competitor, Semantic Guidance, we observe a relative improvement of 8.3% and 10.7% in average and worst-region F1.

Tab. 2 presents results on DollarStreet. In this case, uncontextualized VSG only moderately improves over the LDM baseline in average and worst-region F1. However, when including exemplar images per c-VSG, the patterns of improvement are consistent with GeoDE. c-VSG achieves 5.7% and 4.5% improvements over the closest baselines in average and worst-region F1, respectively.

In the reminder of this section, we analyze the findings described above in the context of improvements to diversity, quality, and consistency.

VSG improves diversity. The improvements in F1 relate to consistent boosting of both worst-region and average recall: for GeoDE (Tab. 1, recall), VSG both with and without contextualization show improvements for all baselines, and by at least 7.3% (and up to 46.2%) over the LDM baseline. Similarly, VSG exhibits among the highest recall values in DollarStreet (Tab. 2, recall). FG CLIP (Ent.) achieves the highest recall values, but those come with significant drops in precision. Notably, for DollarStreet, VSG has better recall when exemplar images are *not* used. Again, this comes at a cost to precision.

Diversity improvements are also reflected in qualitative examples. Fig. 2 reveals that VSG leads to greater variation in object color, type, and shape, such as more colorful cars for Africa, sports-cars for Europe, and lengthened hoods for Southeast Asia. With VSG, image backgrounds also show greater variety in textures and diversity in indoor *vs*. outdoor settings. In addition, VSG enables more *inter-region diversity*. For example, in Fig. 2, the same seed can look much more similar across regions in the baseline than with c-VSG.

Contextualized VSG improves quality. In GeoDE, c-VSG also allows for improvement in both average and worst-region precision over the LDM baseline, the latter by up to 52.7% (Tab. 1, precision). Similar precision gains are observed in DollarStreet (Tab. 2, precision): leveraging exemplar images results in relative improvements of 12% in worst-region over the LDM baseline. Comparing c-VSG with VSG shows relative gains of 17.7% and 25.3% in average and worst-region precision, respectively. Interestingly, Semantic Guidance allows for higher precision than c-VSG on GeoDE, whose images were submitted by people living in the regions of interest. The lower precision achieved by Semantic Guidance on DollarStreet might be related to the datasets' focus on disadvantaged and isolated areas. Thus, Semantic Guidance requires an understanding of whether applying region information helps or hurts the diversity of the generations.

Contextualized VSG improves consistency. In both GeoDE and DollarStreet, we observe that VSG without contextualization slightly reduces consistency (Tab. 1 and 2, CLIPScore). This highlights the importance of contextualization when using VSG: leveraging exemplar images in c-VSG results in the highest CLIPscores across the board, on average and for the worst-region.

Contextualized VSG reduces disparities across regions. When comparing c-VSG to VSG, we observe that disparities across regions appear substantially reduced on both datasets. On GeoDE, contextualization reduces the performance gap between average and worst-region in F1, recall, precision, and CLIPScore. Similar trends may be observed for DollarStreet, where c-VSG exhibits remarkably lower disparities across all metrics when compared to the LDM baseline. It is worth noting that Semantic Guidance also has low disparities.

Contextualized VSG shows early improvements in region representation. While this work does not focus particularly on mitigation of reductive regional representations, we find in initial qualitative inspections that representations of regions can change positively after applying contextualized VSG. We find that for Africa, objects like cars and cooking pots are less dilapidated with VSG, as compared to the original generations. In addition, some objects tend to be larger and more central in the image, reducing the prevalence of unrepresentative background fixtures, as shown in Appendix 6.2. The improvement of consistency with contextualization similarly reduces the dominance of non-representative regional background information. Remaining images without objects tend to have less repetition in reductive backgrounds associated with regional tropes. For example, Appendix 6.2 shows that generated images of bags in Europe can have more variety in their backgrounds when using c-VSG.

3.5 Ablations

We next discuss ablations on the variations of VSG criteria and the strength, quantity, and type of exemplar images used in c-VSG.

Variations of c-VSG Criteria. We compare the use of the full contextualized Vendi Score Guidance criterion from our approach ($\alpha > 0$, $\beta > 0$, c-VSG) to Vendi Score guidance without contextualization ($\alpha > 0$, $\beta = 0$, VSG), a criterion using the exemplar contextualizing images exclusively ($\alpha = 0$, $\beta > 0$), and the LDM without intervention ($\alpha = 0$, $\beta = 0$). Results are shown in Tab. 3 and visual examples in Appendix 6.2.

When evaluating with GeoDE, we find that Vendi Score guidance without contextualization can lead to images with improved diversity but unrealistic object shapes and colors, as well as extreme variation in styles, such as black-and-white or film-style photos. While using exclusively contextualizing images helps in almost all aspects as opposed to only the Vendi Score guidance on the bank of generated images, the methods combined tend to have the best diversity. For DollarStreet, there is a slight trade-off in precision and recall between the inclusion of contextualizing images versus previously generated images in the criterion: while the combination of the two has moderately better measures over only contextualizing images, the use of only previously generated images shows the highest recall.

Strength of exemplar images. Tab. 4 shows ablations of the strength of exemplar images used in the contextualized VSG criterion. Unsurprisingly, lower weight β of real exemplar images yields greater recall while a higher weight leads to a larger precision, emphasizing the quality *vs*. diversity trade-off. For example, increasing β can correlate with changes in camera angle of cars to more closely match those in the reference images, as shown in Appendix 6.2. In addition, consistency is improved with additional weight to exemplar images.

Table 3: Ablation study on variations of VSG criteria.

Dataset	α	β	**F1**		**Precision**		**Recall**		**CLIPScore**	
			Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.
GeoDE	0	0	0.364	0.322	0.413	0.273	0.337	0.395	0.242	0.218
GeoDE	tuned	0	0.399	0.356	0.349	0.307	0.470	0.424	0.180	0.191
GeoDE	0	tuned	0.446	0.427	**0.431**	0.409	0.464	0.447	**0.261**	**0.254**
GeoDE	tuned	tuned	**0.455**	**0.444**	0.424	**0.417**	**0.493**	**0.476**	0.254	0.253
DollarStreet	0	0	0.448	0.442	0.428	0.434	0.472	0.450	0.231	0.235
DollarStreet	tuned	0	0.457	0.444	0.413	0.388	**0.516**	**0.518**	0.191	0.198
DollarStreet	0	tuned	0.492	0.480	**0.486**	0.483	0.500	0.476	**0.240**	**0.242**
DollarStreet	tuned	tuned	**0.497**	**0.483**	**0.486**	**0.486**	0.511	0.479	0.234	0.238

Table 4: Ablation study on exemplar images strength in contextualized VSG.

Dataset	α	β	F1		Precision		Recall		CLIPScore	
			Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.
GeoDE	tuned	α	0.451	0.434	0.384	0.380	**0.547**	**0.506**	0.216	0.221
GeoDE	tuned	2α	**0.455**	**0.444**	0.424	0.417	0.493	0.476	0.254	0.253
GeoDE	tuned	4α	0.448	0.437	**0.428**	**0.435**	0.472	0.438	**0.261**	**0.261**

Table 5: Ablation study on the number of exemplar images per object used in the contextualized Vendi Score computation, or M in Algorithm 1. Results are reported based on GeoDE dataset.

Method	# imgs	F1		Precision		Recall		CLIPScore	
		Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.
c-VSG	2 per obj.	0.455	**0.444**	0.424	0.417	**0.493**	**0.476**	0.254	0.253
c-VSG	4 per obj.	0.448	0.437	0.428	**0.435**	0.472	0.438	0.261	**0.257**
c-VSG	8 per obj.	0.455	0.442	**0.445**	0.436	0.466	0.449	**0.262**	**0.257**
c-VSG	20 per obj.	**0.460**	0.433	0.440	0.434	0.485	0.433	0.251	0.246
c-VSG	4 per obj.	0.448	0.437	0.428	**0.435**	0.472	0.438	0.261	**0.257**
c-VSG	4 per obj.-reg.	**0.454**	**0.445**	**0.436**	0.431	**0.475**	**0.459**	0.259	0.253

Quantity and type of exemplar images. In Tab. 5, we first study how the number of images selected as exemplar images in c-VSG affects image generation diversity. Generally, using fewer exemplar images tends to help with diversity while increasing this number helps with quality, although these trends appear minor in some cases. In Appendix 6.2, we visually inspect images and find that trends in quality and consistency for a single seed can also reverse as the number of images increases, first improving, then worsening (or vice versa). In addition, we study whether exemplar images randomly selected across all possible regions or the same number images selected specifically for a given region yields better diversity. We find that region-stratified exemplar images yield results within 1-2% of those when using randomly selected images for all measures except for worst-region recall, for which there is a 5% improvement.

4 Related Work

Prior work often focuses on mitigating *general* diversity issues in text-to-image models. For example, [37, 50] focus on preventing mode collapse while [44] generates a balanced distribution, focusing on the long-tail, by using synthetic images. In addition, there is growing research focused on issues related to diversity through a lens of reducing *person-related bias*. One line of work reduces people-related biases, *e.g.*, related to gender, through concept forgetting or erasure [23, 33, 52], rather than diversification. The idea is that, by modifying a model's understanding of a concept, the harmful or biased content is not generated. [10] augments the original prompt with real-world distributions of gender and ethnicity using a finetuned language model, and [25, 51] learn token embeddings for a diverse set of sensitive attributes and concatenate them to the original prompt. Building on work in NLP [5], other approaches remove the components of sensitive attributes like race and gender from the text embeddings altogether [9, 47] or modify the cross attention layer [17, 33] to remove implicit assumptions about the world. However, there is little work investigating improvements to *geographic diversity*. To the best of our knowledge, prior work primarily discusses the (minimal) improvement in diversity when including more granular geographic information in text prompts [2, 18].

Our proposed method is most relevant to works that use classifier guidance during inference time. [7, 42] use classifier guidance to guide generations away from inappropriate content. In the context of augmenting a dataset with synthetic samples, [22] has extended the idea of classifier guidance to feedback guidance where loss and entropy are used as the guidance mechanism. While the Vendi Score has been used at training time to increase the diversity of generative adversarial networks (GANs) across modes in the training dataset [3], to the best of our knowledge, we are the first to apply it to diffusion models and as guidance at inference time.

5 Conclusion

In this work, we introduced an inference time intervention that extends the guidance toolbox of diffusion models by driving the generation process towards samples that are substantially different from each other while still representative of the real world. Through extensive experiments, we showed that our approach produces images with higher intra- and inter-region diversity, while exhibiting increased image quality and text-image consistency; overall resulting in reduced disparities in generation diversity, quality and consistency.

Limitations. In this work, we focus on improving worst-region F1 while also analyzing worst and average region precision, recall, and consistency. Automatic metrics are susceptible to several challenges, including region representations in the reference dataset, reliance on pre-existing feature extractors, and the composition of geographic groupings. The metrics are aggregates and do not account for individual preferences. A human evaluation study would be necessary for capturing subjective and personal perspectives about the effect of c-VSG. Finally, our approach is only a single effort at improving the diversity of text-to-image generative models, and there remains future work for their further improvement.

Societal Impacts. This work involves representations of geographic regions. While our qualitative analyses show that c-VSG tends to help mitigate reductive representations of regions, our method may not always remove harmful representations in the underlying generative models. Furthermore, increasing the diversity of text-to-image models may yield unexpected generations. Thorough study of the full range of potential outcomes is necessary before deployment of these methods in real world systems.

Acknowledgments

We would like to thank Oscar Mañas and Pietro Astolfi for helpful discussion and pointers regarding in-context learning and paraphrasing approaches. We also thank Mohammad Pezeshki for useful discussions, feedback and support. In addition, we thank the participants of the qualitative study of c-VSG, discussed in Appendix 6.3.

References

1. Bansal, H., Yin, D., Monajatipoor, M., Chang, K.W.: How well can text-to-image generative models understand ethical natural language interventions? In: EMNLP (Short) (2022)
2. Basu, A., Babu, R.V., Pruthi, D.: Inspecting the geographical representativeness of images from text-to-image models. arXiv preprint arXiv:2305.11080 (2023)
3. Berns, S., Colton, S., Guckelsberger, C.: Towards mode balancing of generative models via diversity weights (2023)
4. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale (2022)
5. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016)
6. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing text-to-image models using semantic guidance (2023)
7. Brack, M., Friedrich, F., Schramowski, P., Kersting, K.: Mitigating inappropriateness in image generation: Can there be value in reflecting the world's ugliness? arXiv preprint arXiv:2305.18398 (2023)
8. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers (2022)
9. Chuang, C.Y., Jampani, V., Li, Y., Torralba, A., Jegelka, S.: Debiasing vision-language models via biased prompts. arXiv preprint arXiv:2302.00070 (2023)
10. Clemmer, C., Ding, J., Feng, Y.: Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8596–8605 (2024)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410 (2022)
14. Friedrich, F., Schramowski, P., Brack, M., Struppek, L., Hintersdorf, D., Luccioni, S., Kersting, K.: Fair diffusion: Instructing text-to-image generation models on fairness (2023)
15. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV. pp. 89–106. Springer (2022)
16. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
17. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023)
18. Hall, M., Ross, C., Williams, A., Carion, N., Drozdzal, M., Soriano, A.R.: Dig in: Evaluating disparities in image generations with indicators for geographic diversity (2023)

19. Hao, S., Shelby, R., Liu, Y., Srinivasan, H., Bhutani, M., Ayan, B.K., Poddar, S., Laszlo, S.: Harm amplification in text-to-image models (2024)
20. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574 (2022)
21. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? (2023), https://arxiv.org/abs/2210.07574
22. Hemmat, R.A., Pezeshki, M., Bordes, F., Drozdzal, M., Romero-Soriano, A.: Feedback-guided data synthesis for imbalanced classification (2023)
23. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems **36** (2024)
24. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
25. Kim, E., Kim, S., Shin, C., Yoon, S.: De-stereotyping text-to-image models through prompt tuning (2023)
26. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models (2019)
27. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Analyzing societal representations in diffusion models (2023)
28. Mañas, O., Astolfi, P., Hall, M., Ross, C., Urbanek, J., Williams, A., Agrawal, A., Romero-Soriano, A., Drozdzal, M.: Improving text-to-image consistency via automatic prompt optimization (2024)
29. Miller, G.A.: WordNet: A lexical database for English. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), https://aclanthology.org/H94-1111
30. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning. pp. 7176–7185. PMLR (2020)
31. Naik, R., Nushi, B.: Social biases through the text-to-image generation lens (2023)
32. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. CoRR **abs/2112.10741** (2021), https://arxiv.org/abs/2112.10741
33. Orgad, H., Kawar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. arXiv preprint arXiv:2303.08084 (2023)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
35. Ramaswamy, V.V., Lin, S.Y., Zhao, D., Adcock, A.B., van der Maaten, L., Ghadiyaram, D., Russakovsky, O.: Beyond web-scraping: Crowd-sourcing a geodiverse dataset. In: arXiv preprint (2023)
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
37. Rangwani, H., Jaswani, N., Karmali, T., Jampani, V., Babu, R.V.: Improving gans for long-tailed data through group spectral regularization. In: European Conference on Computer Vision. pp. 426–442. Springer (2022)

38. Rojas, W.A.G., Diamos, S., Kini, K.R., Kanter, D., Reddi, V.J., Coleman, C.: The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CoRR **abs/2112.10752** (2021), https://arxiv.org/abs/2112.10752
40. Sajjadi, M.S.M., Bachem, O., Lučić, M., Bousquet, O., Gelly, S.: Assessing Generative Models via Precision and Recall. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
41. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
42. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
43. Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., Ferrer, C.C.: Generating high fidelity data from low-density regions using diffusion models (2022)
44. Shin, J., Kang, M., Park, J.: Fill-up: Balancing long-tailed data with generative models. arXiv preprint arXiv:2306.07200 (2023)
45. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
46. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
47. Tanjim, M.M., Singh, K.K., Kafle, K., Sinha, R., Cottrell, G.W.: Discovering and mitigating biases in clip-based image editing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2984–2993 (2024)
48. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
49. Wan, Y., Chang, K.W.: The male ceo and the female assistant: Probing gender biases in text-to-image models through paired stereotype test (2024)
50. Yu, N., Li, K., Zhou, P., Malik, J., Davis, L., Fritz, M.: Inclusive gan: Improving data and minority coverage in generative models. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 377–393. Springer (2020)
51. Zhang, C., Chen, X., Chai, S., Wu, C.H., Lagun, D., Beeler, T., De la Torre, F.: Iti-gen: Inclusive text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3969–3980 (2023)
52. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2303.17591 (2023)