Increasing the Utility of Synthetic Images through Chamfer Guidance

Nicola Dall'Asen[1,2*] **Xiaofeng Zhang**[3,4,5†] **Reyhane Askari-Hemmat**[4†] **Melissa Hall**[4]

Jakob Verbeek[4] **Adriana Romero-Soriano**[3,4,6,7] **Michal Drozdzal**[4]

[1]University of Trento [2]University of Pisa [3]Mila - Québec AI Institute [4]FAIR at Meta
[5]Université de Montréal [6]McGill University [7]Canada CIFAR AI chair

Abstract

Conditional image generative models hold considerable promise to produce infinite amounts of synthetic training data. Yet, recent progress in generation quality has come at the expense of generation diversity, limiting the utility of these models as a source of synthetic training data. Although guidance-based approaches have been introduced to improve the utility of generated data by focusing on quality or diversity, the (implicit or explicit) utility functions oftentimes disregard the potential distribution shift between synthetic and real data. In this work, we introduce *Chamfer Guidance*: a training-free guidance approach which leverages a handful of real exemplar images to characterize the quality and diversity of synthetic data. We show that by leveraging the proposed Chamfer Guidance, we can boost the diversity of the generations *w.r.t.* a dataset of real images while maintaining or improving the generation quality on ImageNet-1k and standard geo-diversity benchmarks. Our approach achieves state-of-the-art few-shot performance with as little as 2 exemplar real images, obtaining 96.4% in terms of precision, and 86.4% in terms of distributional coverage, which increase to 97.5% and 92.7%, respectively, when using 32 real images. We showcase the benefits of the Chamfer Guidance generation by training downstream image classifiers on synthetic data, achieving accuracy boost of up to 15% for in-distribution over the baselines, and up to 16% in out-of-distribution. Furthermore, our approach does not require using the unconditional model, and thus obtains a 31% reduction in FLOPs *w.r.t.* classifier-free-guidance-based approaches at sampling time.

1 Introduction

In the last few years, conditional image generative models [31, 47, 55] have demonstrated extraordinary capabilities, producing highly realistic images from both textual descriptions and class labels. These models have rapidly evolved from experimental research tools to widely accessible applications, enabling creative expression for users across various domains. With this acceleration, researchers have also started exploring the use-cases of conditional image generative models as synthetic training data for downstream machine learning models [6]. However, recent research has revealed that as models grow in size and capability, they tend to produce images of higher quality, aligned with human preference, but with diminished diversity, hindering their utility as synthetic training data generators [5].

Learning-based approaches and guidance-based sampling techniques [4, 25, 26] have been introduced in the literature to mitigate the shortcomings of vanilla synthetic data. On the one hand,

*Work done during internship at FAIR, Meta. Currently at Fondazione Bruno Kessler. † equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 1: Our Chamfer Guidance addresses key limitations of existing image generation approaches, producing **high-quality** and **diverse** outputs. Base models (here $\text{LDM}_{3.5\text{M}}$) necessitate high CFG scales to achieve prompt adherence and quality, at the expense of diversity. Reference-free methods can introduce *ungrounded* diversity, failing to capture the underlying data distribution. While training-based solutions effectively narrow the fidelity gap with the reference distribution, they suffer from low subject diversity, particularly in backgrounds. Our Chamfer Guidance achieves superior image quality without using CFG, substantially improving grounded coverage (C) and aligning the generated images more precisely (P) with the reference distribution. Best viewed zoomed in.

learning-based approaches [21, 48] aim to improve the overall utility of synthetic data by reducing the distribution shift between generated samples and real samples. These techniques leverage real data samples to train or fine-tune parts of existing conditional image generative models. On the other hand, guidance-based sampling methods define reward functions to guide the process of generating synthetic data. These reward functions characterize the utility of the generated samples in terms of generation quality [64] and diversity [25]. However, in-depth analyses of conditional image generative models have highlighted a tension between these two desirable properties [5]. This tension becomes more apparent when defining generation diversity as the variation among generated samples without reference to any target distribution. Contextualizing the desired diversity through real exemplar data holds the promise of limiting potentially disadvantageous variations in synthetic data.

Using exemplar data at inference time to guide the generation process is reminiscent of few-shot in-context learning (ICL) approaches in large language models (LLMs), where a small number of examples are used to adapt to new tasks without requiring model fine-tuning. Yet, in conditional image generation, most few-shot approaches require partial model training or fine-tuning [21, 48], and the use of exemplar data *at inference time* remains underexplored. Following this line of research, Contextualized Vendi Score Guidance (c-VSG) [25] proposed to leverage a handful of contextualizing real images to ground the diversity of generations to that of the real data. However, contrary to ICL approaches [1], increasing the number of examples did not show to improve c-VSG's performance [25].

In this work, we introduce a new training-free guidance method to increase the utility of synthetic data. The method leverages exemplar real data to characterize the desired quality and diversity of synthetic data and scales gracefully with the number of exemplar real images. We define utility of synthetic data through the Chamfer distance [63] between generated and real data, and use this utility formulation as reward guidance to improve off-the-shelf state-of-the-art conditional image generative models. We demonstrate that our Chamfer Guidance approach increases the diversity of generated images compared to previous methods, while maintaining or improving generation quality. We validate these results on ImageNet-1k [13] and standard geo-diversity benchmarks [22, 44]. On ImageNet-1k, we reach a distributional coverage of 91.2% and 92.7% while reaching a precision of 95% and 97.5% for $\text{LDM}_{1.5}$ [47] and $\text{LDM}_{3.5\text{M}}$ [18], respectively. To further illustrate the benefits of our approach, we contrast Chamfer Guidance with alternative approaches in Figure 1. On the geodiversity task, we observe an improvement *w.r.t.* prior art of 6.5% and 5.7% in terms of average F_1 and worst-region F_1, respectively. To validate the utility of our Chamfer Guidance, we train a classifier for the ImageNet-1k [13] dataset on synthetic data, and we demonstrate that our guidance can boost downstream accuracy of up to 15% over the classifier-free guidance sampling. We further investigate out-of-distribution generalization, and our guidance outperforms the classifier-free guidance sampling on several ImageNet variants, and it can obtain a boost of up to 16% on ImageNet-Sketch [62] using $\text{LDM}_{3.5\text{M}}$. Finally, for $\text{LDM}_{3.5\text{M}}$, our approach does not require unconditional model to obtain state-of-the-art quality and diversity, obtaining a 31% reduction in FLOPs *w.r.t.* CFG-based approaches during sampling. With this work, we advance the synthetic training data field and show

that by designing appropriate guidance functions we can unlock the full potential of conditional image generative models and better exploit the knowledge encapsulated in them.

2 Related work

While most modern text-to-image diffusion models utilize classifier-free guidance [30], there has been increasing exploration of alternative guidance methods to further improve the sampling process. A subset of works focus on improving image quality with methods including modifications to classifier-free guidance to reduce over-saturation, such as Adaptive Projected Guidance (APG) [50], utilization of blurring [32, 33] or perturbations [2], and employment of a smaller, less-trained version of the model for guidance [35]. Other works explore guidance methods to improve the diversity of generated images, *e.g.*, Condition-Annealed Diffusion Sampler (CADS) [49] anneals the conditioning signal by incorporating scheduled, monotonically decreasing Gaussian noise into the conditioning vector during the diffusion chain to increase diversity of the generated samples, while Limited Interval guidance [37] removes the conditioning signal in a certain interval to improve ungrounded diversity. Particle Guidance [11] overcomes the common assumption of independent samples in a batched generation, and proposes an extension of sampling where a joint-particle time-evolving potential enforces diversity. c-VSG, which guides the denoising process to increase the diversity of a sample compared to images previously generated with the same prompt [25]. Still other works focus on safety-related guidance by *reducing* harmful representations, such as using classifier guidance to guide generations away from inappropriate content [9, 51] or "forgetting" concepts by zeroing out their cross-attention scores [65]. However, these methods tend to focus on generation quality/diversity/safety as ends in themselves, rather than as means to increasing a synthetic image corpus' *utility* in training downstream models.

Other recent works have begun to explore use of guidance methods to augment real image training datasets with synthetic samples for use in downstream tasks, such as using the loss and entropy of the downstream classifer to guide the synthetic image generator [26], promoting sample diversity to supplement small-scale datasets [66], and leveraging textual inversion [21] to generate images for the use of long-tail image classification [53]. There has also been increasing study into the use of guidance to help in model self-improvement, such as using generated samples as negative guidance to avoid mode collapse/degradation [3]. Our work complements these methods by utilizing features from a small set of real images in the guidance process to improve the utility of generated samples.

Finally, in lieu of guidance approaches to improve the generations of diffusion models, a standard practice is to perform fine-tuning. These methods include fine-tuning directly on additional data sources that better capture the desired qualities of the generations [12, 17, 18, 47] or utilizing reward models to incorporate preferences of model outputs [10, 15, 64]. For example, it is possible to directly fine-tune latent diffusion models by integrating human-preference feedback during the denoising process using Reward Feedback Learning (ReFL) [64]. Instead of the conventional generate-then-filter approach, ReFL leverages a reward model to provide gradient feedback. A key insight is that the quality of generated images becomes discernible only in the later denoising steps. At early timesteps, the reward scores are uniformly low, but once a sufficient number of steps have been taken (*e.g.*, , after 30 out of 40 steps), the reward model produces distinguishable scores. To balance stability and effective feedback, a random timestep from a designated (late) interval is chosen for fine-tuning. Although this approach has been proven successful for quality fine-tuning, it is not yet tested whether this approach would increase diversity in generation given the proper reward. In this work, we study how strong diversity and quality can be achieved by guidance *without* the need to perform fine-tuning.

3 Method

In this section, we introduce a training-free approach for diffusion and flow models, which leverages real data samples to characterize the desired quality and diversity of the synthetic data. Our method leverages the Chamfer distance as a mechanism to match sets of real and generated samples.

3.1 Preliminaries

Diffusion Models. Diffusion models [31, 55] are generative models that transform a unit Gaussian prior into a data distribution via iterative denoising. They consist of a forward and a backward

process. The forward process progressively adds noise to data x_0 over T timesteps following a Markovian process $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$, where $\beta_t \in (0,1)$ is a variance schedule. The backward process learns to reverse this by denoising a sample x_t to x_{t-1}, modeled as $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$. Ho et al. [31] show how we can equivalently train a denoising neural network $\epsilon_\theta(x_t,t)$ to predict the noise ϵ using the following objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right]. \tag{1}$$

Sampling starts with $x_T \sim \mathcal{N}(0,\mathbf{I})$, and the learned reverse process is iteratively applied to obtain x_0, *i.e.*, $x_{t-1} = \text{denoise}(x_t, \epsilon_\theta) := \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right) + \sigma_t\varepsilon$, with $\varepsilon \sim \mathcal{N}(0,\mathbf{I})$, and σ a weighting parameter [31].

Flow matching. Flow Matching (FM) is a generative modeling approach that offers a deterministic alternative to diffusion models by learning continuous transformations between probability distributions. While diffusion models rely on stochastic processes that gradually add and then remove noise through stochastic differential equation (SDE), FM employs first-order ordinary differential equations (ODEs) to define a vector field guiding the transformation from a simple base distribution to the target data distribution. This deterministic nature results in faster inference and reduced computational costs. Moreover, FM can work with optimal transport interpolations, which can lead to more direct and stable mappings compared to the often curved trajectories in diffusion models [39].

CFG. Classifier-free guidance (CFG) [30] enhances controllability in diffusion models by combining conditional and unconditional outputs during sampling as follows:

$$\epsilon_{\text{guidance}}(x_t,t,c) = (1+\omega)\epsilon_\theta(x_t,t,c) - \omega\epsilon_\theta(x_t,t), \tag{2}$$

where ω adjusts guidance strength, $\epsilon_\theta(x_t,t,c)$ is the noise prediction conditioned on c (commonly a class label or text prompt), and $\epsilon_\theta(x_t,t)$ is the unconditional counterpart. This extrapolation trades diversity for fidelity without requiring auxiliary classifiers.

Classifier-guidance. Another approach to control the generation is classifier-guidance [14], which uses a pretrained classifier $p_\phi(y|x_t)$ to modify the unconditional score function $\nabla_{x_t}\log p_\theta(x_t)$, yielding the following modified score function for conditional generation:

$$\nabla_{x_t}\log p_\theta(x_t|y) = \nabla_{x_t}\log p_\theta(x_t) + \gamma\nabla_{x_t}\log p_\phi(y|x_t), \tag{3}$$

where γ is a scaling factor that controls the strength of the guidance signal. Classifier guidance can also be used on top of the CFG score prediction from Equation (2).

DDIM. At sampling time, using the insights of DDIM [56], we can obtain an approximate denoised image for each timestep t,

$$\hat{x}_{0,t} = \text{DDIMApprox}(x_t) := \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

Access to the approximated denoised sample is useful and necessary to apply guidance on the denoising process at inference time without modifying the diffusion model network ϵ_θ [25].

3.2 Chamfer distance for image distribution matching

The Chamfer distance [7] is a metric used to quantify the similarity between two sets of points by measuring the average closest-point distance between each point in one set to the other and vice versa. It has been widely used in 3D computer vision and computer graphics to register point clouds [58].

In this work, we represent image distributions as sets of points and leverage the Chamfer distance for distribution matching. Given a set of real images $\mathcal{X}$ and a set of generated images $\mathcal{Y}$, our goal is to encourage the generated set of images to be as close as possible to a set of real target images. We compute the Chamfer distance between these sets of images as:

$$\mathcal{L}_{\text{Chamfer}}(\mathcal{X},\mathcal{Y}) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}}||x-y||^2 + \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}||x-y||^2. \tag{5}$$

The distance consists of two terms: The first term matches each real sample with its closest generated counterpart. This term is reminiscent to *diversity* metrics used to evaluate synthetic data [36, 41],

and may be viewed as a particular instance of Implicit Maximum Likelihood Estimation [38], which inherently resists mode collapse by encouraging coverage of the entire data distribution. The second term matches each generated sample with its closest real counterpart. This term is reminiscent to *fidelity* metrics used to evaluate synthetic data [36, 41].

Representation space projection. Before computing the Chamfer distance, we project both real and synthetic images into a representation space that induces more semantically meaningful distances than a naive ℓ_2 metric in RGB pixel-space. We use DINOv2 [42], which captures semantic information through self-supervised learning. Unlike CLIP's contrastive learning with text-image pairs and Inception's supervised ImageNet training, DINOv2's self-supervised learning approach provides a feature representation that balances between focusing on important objects and holistic image structure [57]. Moreover, DINOv2's features have been shown to capture human-perceived similarity better [24] than representations such as CLIP [43] and Inceptionv3 [59].

3.3 Inference-time guidance

To improve the alignment between the generated and real data distributions at inference time, without requiring any fine-tuning, we introduce a guidance mechanism based on the Chamfer distance. This builds on the general framework of guidance in diffusion models, where external signals are used to steer the generative process toward a desired objective [4, 14, 25, 26, 52].

Chamfer Guidance. We introduce Chamfer Guidance, where the external signal is derived from the Chamfer distance between a batch of generated samples and a batch of real samples. Let $\mathcal{X}$ denote a set of features extracted from real images, and let $\hat{x}_{0,t}$ be the denoised approximation of a batch of generated samples x_t at time step t (see Equation (4)). The guidance score becomes:

$$\nabla_{x_t} \log p_\theta(x_t|c, \mathcal{X}) = \nabla_{x_t} \log p_\theta(x_t|c) - \gamma \nabla_{x_t} \mathcal{L}_{\text{Chamfer}}(\mathcal{X}, \hat{x}_{0,t}), \tag{6}$$

where $\mathcal{L}_{\text{Chamfer}}$ is the Chamfer distance between the real sample set $\mathcal{X}$ and the current batch of denoised samples, as defined in Equation (5). The negative sign in Equation (6) reflects the goal of minimizing the Chamfer distance, thereby encouraging the generated samples to be close to the real data (fidelity) and to cover its modes (diversity).

Efficient approximation. Evaluating the Chamfer distance across the full reverse diffusion trajectory is computationally expensive. To mitigate this, similar to [4, 26], we adopt the DDIM approximation in Equation (4), which provides a first-order estimate of the final denoised sample at each time step t. This allows us to compute a differentiable reward signal and its gradient during the intermediate steps of the diffusion process without completing all T reverse steps.

Chamfer as a reward signal. Chamfer Guidance can also be interpreted in the context of reward-based generation methods, such as ReFL. In our experiments we find, however, that applying Chamfer-based guidance at inference time leads to stronger distributional alignment than using the same reward function for fine-tuning.

4 Experiments

4.1 Experimental setup

We consider three primary experimental settings. The first is an *object-centric* setup, where the focus lies on guiding or evaluating the generation of specific object categories. This setting allows us to assess the capability of conditional diffusion models in producing accurate and semantically consistent visual representations of well-defined object classes. The second setting addresses *geographic representation* and targets the well-documented limitations of T2I models in handling geographically grounded content. In this setting, we aim to evaluate how well the models capture regional diversity and mitigate the visually biased depiction of locations [23, 25]. Finally, we evaluate the downstream utility of the data generated by our Chamfer Guidance by training an image classifier purely on synthetic data.

Datasets. We utilize three publicly available datasets. For the object-centric setting, we use ImageNet-1k [13], a large-scale image classification dataset containing over one million images across 1,000 object categories. For the geodiversity representation, we use GeoDE [44] and DollarStreet [22]. GeoDE contains a curated set of images annotated with geographic provenance and object labels.

DollarStreet includes photographs of everyday household items from around the world, labeled by country and income level, thus enabling fine-grained analysis of regional visual representations, in line with prior works [23–25]. For downstream utility, we employ ImageNet-1k [13]. We further employ ImageNet-V2 [46], ImageNet-Sketch [62], ImageNet-R [27], and ImageNet-A [28] to measure out-of-distribution generalization.

Models. We consider two state-of-the-art T2I models, in particular latent diffusion models (LDMs), $\text{LDM}_{1.5}$ [47], and $\text{LDM}_{3.5\text{M}}$ [18], which we use in a class-conditional and text-conditional way.

Implementation Details. All experiments are implemented using the `diffusers` library [61], using the default samplers with 40 denoising steps. For the latent projection of our Chamfer Guidance, we primarily use the DINOv2 [42] (ViT-L) feature space, which offers strong semantic representations suited for fine-grained perceptual alignment and whose features have been shown to correlate better with human judgement of similarity [24]. To enable comparison with prior studies [25], we also report results using CLIP embeddings [43] for the geographic diversity scenario. Evaluation in the object-centric setting is conducted using the `dgm-eval` library [57], while the geographic representation scenario utilizes the publicly available evaluation code from DIG-In [23]. The Chamfer distance implementation is from `PyTorch3D` library [45]. As in c-VSG [25], we set the inference-time guidance frequency to $G_{\text{freq}} = 5$, *i.e.*, we apply Chamfer Guidance once every five denoising steps.

4.2 Object-centric scenario: ImageNet-1k

Baselines. We employ reference-free, and reference-based baselines, where reference-based baselines use a few real data samples as reference. As reference-free baselines, we test the default capabilities of LDMs with different CFG [30] values, as well as advanced guidance techniques such as APG [50], CADS [49], Limited Interval guidance [37], and Particle Guidance [11]. Regarding reference-based approaches, for $\text{LDM}_{1.5}$ we test c-VSG [25], which employs a memory bank of prior outputs and real-world exemplar images to guide the generation process, balancing ungrounded and grounded diversity through the use of two Vendi Scores [19]. Additionaly, we employ two training-based solutions, vanilla fine-tuning and reward-based fine-tuning [64]. For vanilla fine-tuning, we fine-tune the model through the standard denoising loss in Equation (1), using the set of real reference images $\mathcal{X}$ as fine-tuning data. Drawing inspiration from ImageReward [64] and their ReFL algorithm, we implement a reward-based fine-tuning approach which leverages a reward function derived from the negative Chamfer distance, *i.e.*, $r = -\mathcal{L}_{\text{Chamfer}}$. As such, this approach requires a few real data samples to encourage the model to generate samples that minimize the distributional distance to the real data. Unlike traditional fine-tuning which focuses on individual samples, this approach explicitly optimizes for distribution-level properties.

Metrics. To comprehensively assess both image quality and diversity aspects, we report Precision and Recall [36], Density and Coverage [41], F_1 measured as the harmonic mean of Precision to Coverage, as well as Fréchet Distance (FD) [29]. Note that these metrics are grounded on a reference dataset. We use DINOv2 features to compute all the metrics as recent literature [24, 57] found that DINOv2 space provides more accurate estimations of perceptual similarity. In addition, we report the standard FD using InceptionV3 features.

Number of real images. We investigate the effect of increasing the number of real images used in both fine-tuning and inference-time guidance approaches. This allows us to establish the data efficiency and scaling properties of each method. For the number of real images we consider $k \in [1, 2, 4, 8, 16, 32]$.

Implementation details. For all the experiments with $k \leq 8$, we use a single H100 GPU to perform training and inference. We use multiple GPUs for $k = [16, 32]$. Vanilla fine-tuning and Chamfer fine-tuning solutions are trained for at most 5,000 steps with checkpoints taken every 1,000 steps. We use a constant learning rate of 10^{-6} across all experiments. For $\text{LDM}_{1.5}$, we fine-tune the entire U-Net backbone, while for $\text{LDM}_{3.5\text{M}}$ we employ LoRA [34] fine-tuning with a rank $r = 4$ applied on the key, query, value, and output layers of attention modules. For ReFL hyperparameters, we use the official implementation and set $\lambda = 10^{-3}, T = 40, T_1 = 30, T_2 = 39$.

4.2.1 Quantitative Results

We show the quantitative evaluation on ImageNet-1k dataset for $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$ in Tables 1a and 1b, respectively. For each k, we report the metrics corresponding to the best F_1 score *w.r.t.* the validation set of each dataset. Our analysis reveals that existing reference-free solutions are unable to

Table 1: Quantitative results on ImageNet-1k using $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$. Our Chamfer guidance consistently achieves state-of-the-art fidelity, diversity, and image quality compared to reference-free and training-based approaches across both models.

(a) Results using $\text{LDM}_{1.5}$.

Method	k	ω	F_1 (P, C) ↑	Precision ↑	Coverage ↑	Density ↑	Recall ↑	FDD ↓	FID ↓
$\text{LDM}_{1.5}$	–	1.0	0.507	0.723	0.391	0.551	0.656	431.241	31.289
$\text{LDM}_{1.5}$	–	7.5	0.709	0.862	0.603	0.775	0.415	248.731	16.116
APG [50]	–	4.0	0.723	0.855	0.626	0.752	0.533	217.937	13.391
CADS [49]	–	4.0	0.718	0.850	0.621	0.743	0.546	217.959	13.434
Limited Interval [37]	–	4.0	0.708	0.837	0.613	0.686	0.631	219.168	11.405
Particle Guidance [11]	–	4.0	0.719	0.846	0.625	0.744	0.544	222.264	14.516
c-VSG [25]	2	2.0	0.660	0.788	0.568	0.632	**0.738**	236.337	10.742
Vanilla fine-tuning	2	7.5	0.733	0.867	0.635	0.797	0.495	220.089	14.869
Vanilla fine-tuning	32	7.5	0.741	0.869	0.646	0.796	0.503	214.045	14.723
Chamfer fine-tuning	2	2.0	0.759	0.886	0.664	0.868	0.420	211.714	15.464
Chamfer fine-tuning	32	2.0	0.766	0.898	0.668	0.875	0.404	209.999	15.492
Chamfer Guidance (Ours)	2	2.0	0.886	0.947	0.833	1.108	0.480	156.179	13.670
Chamfer Guidance (Ours)	32	2.0	**0.931**	**0.950**	**0.912**	**1.213**	0.649	**113.301**	**8.935**

(b) Results using $\text{LDM}_{3.5\text{M}}$.

Method	k	ω	F_1 (P, C) ↑	Precision ↑	Coverage ↑	Density ↑	Recall ↑	FDD ↓	FID ↓
$\text{LDM}_{3.5\text{M}}$	–	1.0	0.599	0.752	0.498	0.560	**0.667**	314.732	17.268
$\text{LDM}_{3.5\text{M}}$	–	2.0	0.727	0.872	0.623	0.797	0.502	231.890	15.673
APG [50]	–	2.0	0.723	0.856	0.625	0.783	0.485	237.623	15.433
CADS [49]	–	2.0	0.717	0.851	0.620	0.749	0.518	238.125	15.139
Vanilla fine-tuning	2	4.0	0.727	0.861	0.628	0.799	0.433	238.856	18.539
Vanilla fine-tuning	32	4.0	0.728	0.855	0.635	0.789	0.472	232.911	18.098
Chamfer fine-tuning	2	2.0	0.741	0.906	0.626	0.906	0.255	284.654	21.538
Chamfer fine-tuning	32	2.0	0.731	0.899	0.616	0.890	0.276	286.708	22.061
Chamfer Guidance (Ours)	2	1.0	0.912	0.964	0.864	1.245	0.469	134.305	**8.878**
Chamfer Guidance (Ours)	32	1.0	**0.950**	**0.975**	**0.927**	**1.366**	0.550	**121.403**	9.606

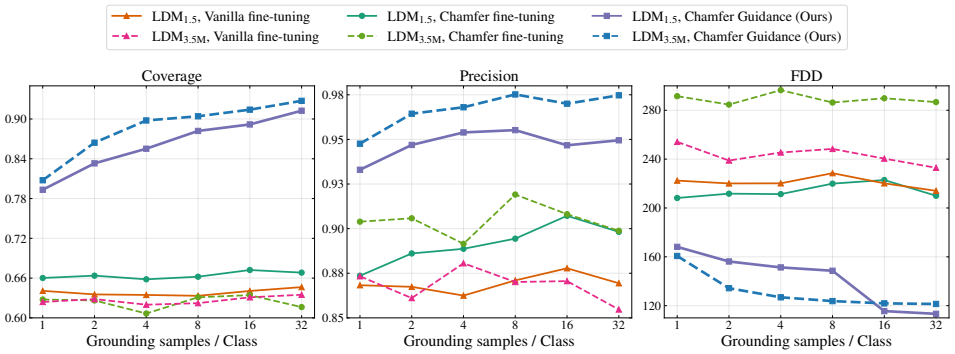


Figure 2: Effect of the number of real reference samples k on $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$ for ImageNet-1k. We can see that only our Chamfer Guidance can effectively leverage the the increased number of reference images, consistently obtaining favorable trends across Coverage, Precision, and FDD.

deliver satisfactory outcomes, offering only marginal improvements over the base LDM performance in terms of diversity. c-VSG achieves the best recall, at the expense of coverage, indicating the generation of outlier images. Furthermore, we observe that neither vanilla fine-tuning nor Chamfer fine-tuning improve their performance when trained with an increased number of available training samples. Interestingly, on $\text{LDM}_{3.5\text{M}}$ our Chamfer Guidance obtains state-of-the-art results in fidelity and diversity without using CFG, *i.e.*, $\omega = 1.0$. This brings a significant computational reduction *w.r.t.* CFG-based approaches of $\approx 31\%$ for the case $k = 32$. The detailed computation of efficiency is presented in the Appendix.

Figure 2 shows how our Chamfer Guidance effectively increases the diversity when a higher number of real samples become available. Meanwhile, training-based approaches only marginally increase precision when learning on higher number of real samples k, while not improving the image quality as indicated by the FDD metric. Our results on Chamfer Guidance are in line with recent trends in LLMs, where test-time compute can be more effective than training or fine-tuning-time compute [54].

Table 2: Comparison on the GeoDE dataset under the geographical representation benchmark of c-VSG, with model selection on $F_1(P, C)$. Metrics are computed in Inception space. † indicates re-implemented results. AF: Africa, WAS: West Asia, AM: Americas, EU: Europe. "label" is region label, "desc" is text description, and "img" is exemplar images. Our Chamfer Guidance achieves state-of-the-art in terms of F_1, and coverage scales with an increased amount of available samples.

Method	Ref. Info	k	Worst-Reg.	F_1 (P, C) ↑		Precision ↑		Coverage ↑		CLIPScore ↑	
				Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.
$LDM_{1.5}$ †	✗	–	AF	0.412	0.346	0.459	0.378	0.374	0.319	0.251	0.239
Synonyms†	✗	–	AF	0.339	0.297	0.350	0.298	0.328	0.297	0.215	0.203
VSG† [25]	✗	–	AF	0.353	0.312	0.349	0.307	0.357	0.317	0.180	0.191
Paraphrasing†	desc	–	WAS	0.329	0.301	0.338	0.309	0.320	0.293	0.231	0.228
Semantic Guidance†	label	–	AF	0.412	0.344	0.458	0.376	0.375	0.317	0.251	0.239
FG CLIP (Loss)†	label	–	AF	0.418	0.391	0.441	0.422	0.397	0.365	0.246	0.244
FG CLIP (Entropy)†	label	–	AF	0.414	0.357	0.437	0.393	0.393	0.327	0.238	0.236
Textual Inversion [21]†	img	4	EU	0.300	0.267	0.356	0.308	0.260	0.236	0.212	0.214
c-VSG† [25] (CLIP)	img	2	AM	0.435	0.412	0.424	0.408	0.446	0.416	0.254	0.254
c-VSG† [25] (CLIP)	img	4	AF	0.412	0.357	0.428	0.382	0.398	0.335	0.253	0.253
Chamfer Guidance (Ours, CLIP)	img	2	AF	0.495	0.461	0.525	0.487	0.468	0.437	0.265	0.263
Chamfer Guidance (Ours, CLIP)	img	4	AM	0.492	0.457	0.513	0.449	**0.473**	**0.465**	**0.267**	**0.266**
Chamfer Guidance (Ours, DINOv2)	img	2	AF	0.488	0.456	0.539	**0.513**	0.446	0.411	0.257	0.248
Chamfer Guidance (Ours, DINOv2)	img	4	AF	**0.500**	**0.469**	**0.549**	0.512	0.459	0.432	0.257	0.249

4.3 Geographic Diversity

We conduct a thorough comparative analysis of our approach against existing geographic diversity enhancement techniques, particularly focusing on the contextualized Vendi Score Guidance (c-VSG) method [25], which is the closest in spirit to our approach. This comparison reveals several key advantages of our Chamfer Guidance.

Datasets and evaluation metrics. We follow the same evaluation protocol of c-VSG, but we employ the F_1 score between precision and coverage metrics instead of precision and recall. This is to avoid inflated results due to generated outliers. We use Inception-based metrics and prompt the models with `{object} in {region}` for a fair comparison with c-VSG. We report results computed in DINOv2 space in the Appendix. When reporting results for "worst region", we follow the original protocol to first identify the worst region in terms of the F_1 score, and then report the other metrics corresponding to the same region. This may lead to cases where the worst region has a higher value for the metric *w.r.t.* the average.

Implementation details. We use $LDM_{1.5}$ as in c-VSG and apply the same filtering on the GeoDE dataset as described in prior work [23, 25]. We choose the DINOv2 [42] latent space to compute the Chamfer distance, but we also report results using CLIP [43] for a fair comparison with c-VSG.

4.3.1 Quantitative Results

To enable a direct comparison, our experiments replicate the settings of c-VSG [25]. Table 2 presents a comparative analysis of our Chamfer Guidance against c-VSG on the GeoDE dataset, using recomputed baselines to use the revised F_1 score. Details about the baselines are presented in the Appendix. Notably, our optimal approach (DINOv2 as the feature extractor, and 4 grounding samples) achieves a nearly 7% improvement in F_1 score over c-VSG. We also observe a consistent scaling in coverage, exhibiting the same trend as in the object-centric scenario. This indicates that the observed behavior persists even in the more challenging context of geographical diversity. Furthermore, our method substantially enhances worst-region coverage by up to 4.9%, demonstrating its efficacy in mitigating regional misrepresentation. As expected, the highest image-text alignment, measured by CLIPScore, is obtained when employing the CLIP latent space for projecting images during Chamfer distance computation. This methodology effectively guides synthetic images towards the CLIP subspace corresponding to the target object. Then, we evaluate our approach on DollarStreet, with results presented in the Appendix, where we demonstrate a 3.5% improvement in the F_1 score compared to c-VSG, due to the enhanced quality and diversity of the generated data. In summary, we can draw the following conclusions from these results: Chamfer Guidance enhances the quality (precision) and diversity (coverage) of generated data, and demonstrates superior image-text alignment. Unlike c-VSG, it exhibits favorable scaling with increased real data volume. We show qualitative results in the Appendix, where our Chamfer Guidance showcases increased quality, less saturation and more diversity.

Our analysis reveals two additional significant advantages of our approach: (1) Unlike c-VSG, which necessitates storing intermediate results to continuously update the Vendi Score computation, our approach operates with significantly lower memory overhead. (2) Our method eliminates the need to balance between a diversity term and a contextualization term, removing one hyperparameter

from the optimization process. This simplification makes our approach more accessible and easier to tune, reducing the complexity of the diversity enhancement pipeline.

Table 3: Validation accuracy on real images of classifiers trained with real data, synthetic data, and a mix of synthetic and training data for ImagetNet-1k. All models are trained for 200k iterations.

(a) Real data only.

Real images	IN1k	IN-v2	IN-Sk	IN-R	IN-A
2k	5.01	3.94	0.63	0.89	0.21
32k	34.05	25.38	4.17	5.04	0.53
1.3M	82.60	70.90	32.50	44.60	29.40

(b) Synthetic only and mixed data. Real and synthetic images refer to the number of samples used to train the classifier.

Real images	Syn. images	Guidance	$LDM_{1.5}$					$LDM_{3.5M}$				
			IN1k	IN-v2	IN-Sk	IN-R	IN-A	IN1k	IN-v2	IN-Sk	IN-R	IN-A
0	1.3M	$\omega = 2$	47.67	40.33	20.49	17.49	1.45	37.83	34.07	17.60	11.53	0.88
		Chamfer k=2	52.88	45.37	28.07	19.60	1.71	52.14	44.27	33.47	20.26	1.93
		Chamfer k=32	**54.91**	**46.43**	**28.08**	**19.78**	**5.11**	**53.66**	**45.46**	**34.44**	**20.67**	**5.28**
2k	1.3M	$\omega = 2$	48.47	41.07	21.21	16.96	1.57	40.89	31.72	20.03	12.31	1.32
		Chamfer k=2	**53.57**	**46.42**	**29.48**	**21.25**	**1.65**	**52.95**	**45.26**	**33.52**	**20.51**	**1.99**
32k	1.3M	$\omega = 2$	59.07	49.77	25.04	20.10	2.44	55.65	45.65	21.64	14.97	1.54
		Chamfer k=32	**63.81**	**53.84**	**32.34**	**22.40**	**2.72**	**62.61**	**52.58**	**34.49**	**21.85**	**2.36**

4.4 Training downstream image classifiers on synthetic data

We complement our analysis by evaluating the downstream utility of our Chamfer guidance in the "static" ImageNet-1k setup introduced in [4], and generate a dataset of $1,300,000$ synthetic images using our approach, with different k real exemplar images from the training set. Each synthetic image is generated with a simple prompt `(class name)`. We report the accuracy of a ViT-B [16] classifier trained on this synthetic data and tested on real validation data.

We compare the performance of $LDM_{1.5}$ and $LDM_{3.5M}$ against a classifier trained on limited real images (k times the number of classes), with results shown in Table 3. We initially train the classifier on only limited real images and present the results in Table 3a to serve as reference. Our findings demonstrate that this limited real data alone fails to achieve satisfactory performance. Next, we leverage the available real data per class to generate synthetic images using our Chamfer distance approach, presenting these results in Table 3b. We study the effect of images generated without and with our Chamfer guidance, and we show that our synthetic-only approach using Chamfer guidance achieves substantial improvements over the base LDM generations, with gains of $+12.01$ and $+15.83$ accuracy, respectively. Finally, we find that combining real and synthetic data yields optimal results, achieving up to 63.81% and 62.61% accuracy when using only 32 real images per class alongside our generated synthetic data using $LDM_{1.5}$ and $LDM_{3.5M}$, respectively. These results also show how our Chamfer guidance can effectively close the gap in grounded diversity between the two models. The difference is more than 5% when using only the synthetic data, and we reduce it to 1.25% with $k = 32$. An analogous behavior emerges when leveraging real data, reducing the gap from 3.42 to 1.20 with 32k real images.

We also test on ImageNet-V2 [46], ImageNet-Sketch [62], ImageNet-R [27], and ImageNet-A [28] to measure out-of-distribution (OOD) generalization. Models trained with data generated by our Chamfer guidance exhibit strong performance on these out-of-distribution datasets, surpassing models trained on only real data on all variants, given the same number of real images. For ImageNet-Sketch, $LDM_{3.5M}$ with our Chamfer guidance surpasses the performance of a classifier trained on the full (1.3M samples) ImageNet-1k. Our Chamfer guidance always obtains substantial gains over the default sampling of $LDM_{1.5}$ and $LDM_{3.5M}$ across all ImageNet variants, up to 16% when using synthetic data only for ImageNet-Sketch on $LDM_{3.5M}$. Interestingly, on ImageNet-R and ImageNet-A, adding real data to our synthetic ones harms the model's performance. In these domains, which differ significantly from ImageNet-1k, the gain resulting from synthetic data is less pronounced *w.r.t.* ImageNet-V2 and ImageNet-Sketch. R(endition) and A(dversarial) represent more challenging domains where more natural-looking images bring reduced benefit. Nevertheless, our Chamfer guidance still surpasses their real-only counterparts. These results confirms the downstream utility of our generated samples for OOD tasks as well. Similarly to the in-domain results, we observe how our Chamfer guidance can effectively close the performance gap between $LDM_{1.5}$ and $LDM_{3.5M}$, and for ImageNet-Sketch, the performance of $LDM_{3.5M}$ is superior to $LDM_{1.5}$.

Finally, we would like to observe how the improvement in quality between $LDM_{1.5}$ and $LDM_{3.5M}$ does not correspond to a greater utility of the generated data. What we generally observe is a decrease

in performance between $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$, and that our Chamfer guidance can restore the utility of the most recent models.

4.5 Ablations

We conduct our ablation studies on $\text{LDM}_{1.5}$, using the ImageNet-1k dataset. The goal is to understand the robustness of our Chamfer Guidance to the relevant hyperparameters, *i.e.*, ω from Equation (2) (in Table 4) and the strength of the Chamfer Guidance γ from Equation (6) (in the Appendix).

CFG ablation. Table 4 presents the impact of varying the CFG scale ω on our method. The results clearly demonstrate that a moderate ω value of 2.0 with $k = 32$ guiding images achieves optimal performance, yielding the highest F_1 score (0.931) and competitive precision (0.950) while achieving high coverage (0.912). This balance is crucial for generating both accurate and diverse images. Notably, the FID score of 8.935 at $\omega = 2.0$ indicates better image quality compared to higher guidance settings, which deteriorates to 14.388. This confirms that excessive guidance strength introduces oversaturation artifacts, compromising perceptual quality. The low FDD score (113.301) at $\omega = 2.0$ further validates the effectiveness of this configuration in producing faithful, high-quality, and diverse image generations. Interestingly, using the conditional model only ($\omega = 1.0$) produces competitive results, confirming that our Chamfer Guidance approach enables state-of-the-art results while reducing computational complexity. We present a more detailed study on efficiency in the Appendix.

Table 4: ω ablation on ImageNet-1k using $\text{LDM}_{1.5}$. Our Chamfer Guidance can achieve near state-of-the-art results employing only the conditional model, reducing the needed inference computation.

ω	k	F_1 (P, C) ↑	Precision ↑	Coverage ↑	Density ↑	Recall ↑	FDD ↓	FID ↓
1.0	2	0.849	0.890	0.811	0.904	**0.736**	150.748	13.217
1.0	32	0.899	0.923	0.876	1.086	0.735	117.834	9.759
2.0	2	0.881	0.932	0.835	1.051	0.637	124.191	8.840
2.0	32	**0.931**	0.950	**0.912**	1.213	0.649	**113.301**	**8.935**
7.5	2	0.886	0.947	0.833	1.108	0.480	156.179	13.670
7.5	32	0.925	**0.957**	0.894	**1.238**	0.498	153.111	14.388

5 Conclusion

We introduced Chamfer Guidance, a novel training-free approach to improve the utility of synthetic data from conditional image models. By leveraging a small set of exemplar real images to guide the generation process, our method balances quality and diversity while addressing distribution shifts between synthetic and real data. Experimental results on ImageNet-1k and geo-diversity benchmarks show Chamfer Guidance achieves state-of-the-art performance and scales its effectiveness as exemplars increase. Additionally, our synthetic data can be used in downstream applications. Furthermore, our approach eliminates the computational overhead of CFG, reducing computational requirements while maintaining superior performance in quality and diversity. These contributions advance synthetic data generation and demonstrate that thoughtfully designed guidance can unlock the full potential of conditional image generative models for downstream applications.

Limitations. Our evaluation of reference-based diversity and quality is built on automated metrics, which inherit intrinsic biases, *e.g.*, , the distribution of the dataset we compare against, or the reliance on pre-existing feature extractors that might not capture subtle differences. While we rely on previous works to provide the most accurate metrics possible, these are statistical aggregators and do not account for individual preferences. Currently, our approach is designed for class-conditional models and does not directly support text-to-image generation. This limitation opens up several exciting avenues for future work. To extend our method to text-to-image models, we envision a retrieval-based pipeline. First, a large text-image dataset would be embedded into semantic vectors to create an offline retrieval database. Then, at inference time, for a user's text query, the top-k relevant images would be retrieved to serve as exemplars for our guidance. Another interesting extension would be a zero-shot, data-free pipeline. We could envision a self-bootstrapping technique that first generates initial candidate images for a class, then automatically selects a diverse subset maximizing coverage (or "diameter") in a robust feature space, and finally uses this synthetically-generated set as the guidance exemplars.

Societal Impact. Our work builds on image generative models, and it inherits some of the societal challenges of image content creation. However, our Chamfer Guidance provides a new way to guide the generations towards a set of exemplary images. As such, it equips the user with a broader inference-time tool set to better control the sampling process.

Acknowledgments. This work was sponsored by the project FAIR Future AI Research (PE00000013), funded by NextGeneration EU.

References

[1] Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *Proceedings of IEEE/CVF European Conference on Computer Vision (ECCV)*, 2024.

[3] Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data. *arXiv preprint arXiv:2408.16333*, 2024.

[4] Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. *arXiv preprint arXiv:2502.15588*, 2025.

[5] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024.

[6] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research (TMLR)*, 2023.

[7] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*, 1977.

[8] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[9] Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Mitigating inappropriateness in image generation: Can there be value in reflecting the world's ugliness? *arXiv preprint arXiv:2305.18398*, 2023.

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

[11] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *International Conference on Learning Representations*, 2024.

[12] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[15] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research (TMLR)*, 2023.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[17] Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. Mitigating stereotypical biases in text to image generative systems. *arXiv preprint arXiv:2310.06904*, 2023.

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[19] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research (TMLR)*, 2022.

[20] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint at arXiv:2302.10893*, 2023.

[21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[22] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[23] Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. Dig in: Evaluating disparities in image generations with indicators for geographic diversity. *Transactions on Machine Learning Research (TMLR)*, 2023.

[24] Melissa Hall, Samuel J Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2024.

[25] Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *Proceedings of IEEE/CVF European Conference on Computer Vision (ECCV)*, 2024.

[26] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *Transactions on Machine Learning Research (TMLR)*, 2024.

[27] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[29] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[30] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[32] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[33] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[34] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[35] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[36] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[37] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[38] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018.

[39] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[40] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[41] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2023.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[44] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[45] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[46] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arxiv 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[49] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[50] Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[51] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[52] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[53] Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023.

[54] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[57] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[58] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[61] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[62] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[63] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Balanced chamfer distance as a comprehensive metric for point cloud completion. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[64] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[65] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[66] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This work studies the use of Chamfer distance as a guidance technique for generative models to increase grounded diversity in their generation. Description of the approach is presented in Section 3, quantitative and qualitative evaluations *w.r.t.* previous works are presented in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The work includes a "Limitations" section, where we describe the limitations of automated metrics when dealing with distributional representation. We rely on previous benchmarks but recognize that a more complete evaluations should include human judges.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include all the relevant implementation details in the Experimental section (Section 4) of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We put an effort to make the paper self contained to facilitate reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We report used hyperparameters and data splits in the Experiment section when novel. For the geographical diversity setup we refer to hyperparameters and data splits of previous works.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Although our approach enables the possibility of conducting multiple evaluations and statistical significance analyses, we follow previous works [25] and use a single seed to report results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the type of resources needed to run the experiments in the Implementation details section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully inspected the code of ethics, and believe our work conforms with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a "Societal Impact" section, where we discuss the potential harm a generative model *in the wild* could bring. Effective filters need to be put in place before these models are released to a broader audience, to avoid the generation of harmful content.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We employ pre-trained diffusion models and we use the safeguards put in place by the original authors. Additionally, we perform our analyses on non-controversial datasets, such as ImageNet and geo-diversity datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original creators of all models, datasets and algorithms used in this work are properly credited, with citations in the manuscript. We used their material for only non-commercial purpose of developing this research paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release any new dataset with our submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The study does not involve human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs have been used only for editing and formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Efficiency analysis

We analyze the computational overhead and efficiency gains of our proposed Chamfer guidance method compared to traditional Classifier-Free Guidance (CFG) when applied to $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$. This analysis quantifies the significant computational savings achieved by our approach. We use `pytorch's FlopCounterMode` to compute the FLOPs of each module.

A.1 Baseline Diffusion Model Computation

$\text{LDM}_{1.5}$. The standard diffusion process using $\text{LDM}_{1.5}$ requires ≈ 800GFLOPs per forward pass. With the default implementation using 40 denoising steps, the total computation for generating a single sample amounts to 40×800 GFLOPs $= 32$ TFLOPs.

When applying CFG, which requires a doubled batch size per step, the computational requirements double, 40×1.6 TFLOPs $= 64$ TFLOPs.

$\text{LDM}_{3.5\text{M}}$. The standard diffusion process using $\text{LDM}_{3.5\text{M}}$ requires approximately 6 TFLOPs per forward pass. With the implementation using 40 denoising steps, the total computation for generating a single sample amounts to 40×6 TFLOPs $= 240$ TFLOPs.

When applying CFG, which requires a doubled batch size per step, the computational requirements double, 40×12 TFLOPs $= 480$ TFLOPs.

We also have to take into consideration the decoding of the final samples, which account for $\approx$ 2TFlops for $\text{LDM}_{1.5}$, and ≈ 10TFlops for $\text{LDM}_{3.5\text{M}}$.

Therefore we estimate 34 and 66 TFLOPs to generate an image without and with CFG for $\text{LDM}_{1.5}$, and 250 and 490 TFLOPs to generate an image without and with CFG for $\text{LDM}_{3.5\text{M}}$.

A.2 Our Chamfer Guidance Approach

The Chamfer guidance method introduces two additional computational components:

Reference Image Processing We use DINOv2 (ViT/L) for feature extraction, which requires $\approx$ 160 GFLOPs per image. Given our largest reference set of $k = 32$ real images, the total computation for reference embedding is: 32×160 GFLOPs ≈ 5.1 TFLOPs. This is a one-time cost incurred at the beginning of the sampling process.

During generation, we apply our guidance at a frequency of $G_{\text{freq}} = 5$, meaning we compute the guidance every 5 steps. Each guidance computation using DINOv2 requires first decoding the samples, and then encoding in DINO space, therefore the cost for $\text{LDM}_{1.5}$ is $(2\text{TFLOPs} + 160\text{ GFLOPs}) \times 8 \approx 17.3$ TFLOPs, while for $\text{LDM}_{3.5\text{M}}$ is: $(10\text{TFLOPs} + 160\text{ GFLOPs}) \times 8 \approx 81$ TFLOPs.

For the entire sampling process with 40 steps, the reference encoding and the Chamfer guidance adds approximately ≈ 22.4 TFLOPs for $\text{LDM}_{1.5}$, and ≈ 86 TFLOPs for $\text{LDM}_{3.5\text{M}}$.

The total computation required by our Chamfer guidance method is therefore about 56.4 TFLOPs for $\text{LDM}_{1.5}$, and 336 TFLOPs for $\text{LDM}_{3.5\text{M}}$.

A.3 Efficiency Comparison

Comparing our method with the standard CFG approach:

$$\text{Efficiency gain LDM}_{1.5} = 1 - \frac{\text{Computation}_{\text{Chamfer}}}{\text{Computation}_{\text{CFG}}} \tag{7}$$

$$= 1 - \frac{56.4 \text{ TFLOPs}}{66 \text{ TFLOPs}} \tag{8}$$

$$\approx 0.15 \text{ or } 15\% \tag{9}$$

$$\text{Efficiency gain LDM}_{3.5\text{M}} = 1 - \frac{\text{Computation}_{\text{Chamfer}}}{\text{Computation}_{\text{CFG}}} \tag{10}$$

$$= 1 - \frac{336 \text{ TFLOPs}}{490 \text{ TFLOPs}} \tag{11}$$

$$\approx 0.31 \text{ or } 31\% \tag{12}$$

This analysis demonstrates that our Chamfer guidance method achieves superior generation quality while reducing the computational requirements by approximately 31% compared to traditional CFG on $\text{LDM}_{3.5\text{M}}$. This translates to ≈ 4s. to generate a sample with $\text{LDM}_{3.5\text{M}}$ on a RTX A6000.

B On the evaluation metrics

Through our experimentation, we identified several limitations in commonly used metrics for evaluating generative diversity, leading us to choose the make the following choices to the evaluation framework:

From Recall to Coverage: We examine the use of recall as a grounded diversity metric and find it to be highly sensitive to outliers in the generated distribution [41]. A single generated sample that happens to be far from the other generated points can disproportionately inflate the recall value. As an alternative, we advocate for the use of coverage, which provides a more reliable assessment of how well the generated distribution matches the target diverse distribution. Coverage is more robust to outliers, as the manifold is computed *w.r.t.* real points [41].

$\mathbf{F_1}$ with precision and coverage: To evaluate both fidelity and diversity in a single metric, we propose to modify the F_1 score in previous works [25] to harmonically combine precision and coverage. We specifically chose precision over density [41] despite their similar purposes, as precision is naturally bounded between 0 and 1, making it more intuitive to interpret and combine with other metrics. In contrast, density is potentially unbounded, which complicates its use in composite metrics.

From Inception to DINO space for evaluation: We assess the choice of feature space for computing similarity-based metrics. While the Inception feature space has been the *de facto* standard in generative model evaluation, we advocate for the use of the DINO feature space instead, as supported by previous works [24, 57]. Self-supervised latent spaces better align with human perception of image realism, while Inception focuses on specific objects rather than holistic image features, often ignoring important aspects of images.

C User Study

While the primary focus of our study is the utility of the generated data for representation and as a training source for downstream tasks, which is confirmed by our extensive experimental evaluation with downstream classifier training, we also evaluate human perception of our generated images.

To this end, we conduct a small-scale user study to complement our quantitative findings. We collected 965 data points from more than twenty anonymous annotators. In this study, users were presented with samples generated from prompts based on the ImageNet dataset. Their task was to choose their preferred generation in a side-by-side comparison between images from the base $\text{LDM}_{3.5\text{M}}$ model and images generated with our Chamfer Guidance applied to the same model. Users were also presented with real images from the dataset to ground their evaluation in real-world quality and coherence, as we show in Figure 3.

The results showed a strong preference for our method: images generated with our Chamfer Guidance were preferred in $\mathbf{92}\% \pm \mathbf{2}\%$ of the cases. This suggests that the automatic evaluation of quantitative improvements in downstream utility also correlates with enhanced human-perceived quality and fidelity to the target concept distribution.

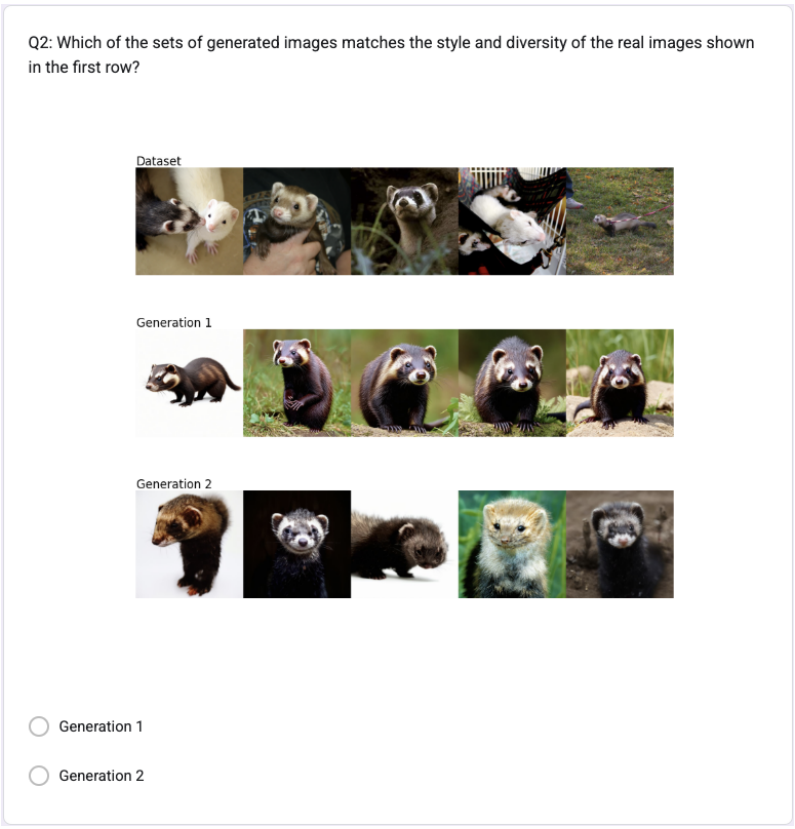


Figure 3: Example of user study question.

D Geographic diversity baselines

We employ the same baselines defined in c-VSG [25], which uses reference-free and reference-based solutions. We report descriptions for them as follows.

Without any additional information:

- $\text{LDM}_{1.5}$: This is the baseline setup where the base $\text{LDM}_{1.5}$ is used with the prompt `{object} in {region}` to condition the generation process.
- `Synonyms`: This strategy maps each object class to its corresponding ImageNet [13] class and WordNet [40] synset. For each class, we generate images that cover all possible meanings (lemmas), including the original object word. This means that each synset, which groups lemmas for a specific sense of the class, guides the image generation process through `{synonym} in {region}` .

With additional information:

- `Paraphrasing`: LLaMA-2-70B-chat [60] large language model is used to generate paraphrases of the original prompt template, `{object} in {region}` . The authors include the specifications and descriptions used in the collection of GeoDE and DollarStreet. The metaprompts and paraphrases, as well as the method of tuning prompts and model specifications, are included in the original paper Supplementary material [25].
- `Semantic Guidance`: Generated images often exhibit diversity problems, as highlighted by previous works [23], due to an amplification of region-specific object features that go beyond what is present in the evaluation task. To mitigate this overemphasis on regional data, the authors utilized Semantic Guidance [8, 20], which involves applying negative guidance corresponding to the regional term for each generated image.
- `Feedback Guidance (FG)`: Inspired by [26, 52], the authors employ an external CLIP-based classifier [43] to provide feedback during image generation by predicting region labels of these

images. To promote greater diversity in the generated outputs, they investigate two feedback guidance strategies: the first maximizes the `loss` of the classifier, while the second maximizes the `entropy` of its predicted class distributions.

- `Textual Inversion`: Textual Inversion is a technique used to teach text-to-image models new, specific visual concepts from just a few example images. It achieves this by creating a new "pseudo-word" or token in the model's vocabulary that becomes associated with the visual characteristics of the provided images. We apply textual inversion by learning an embedding for each object in the dataset using four images per object.

E Additional geographic diversity results

In Table 5 we report the same experiments on GeoDE as in Table 2 of the main body, but with metrics in DINOv2 [42] space, which is known to align better with human evaluation [24, 57]. We observe state-of-the-art results *w.r.t.* relevant baselines, and in particular in terms of F_1 by a significant margin ($+12\%$) over c-VSG [25], due to an increased precision ($+26.7\%$) and coverage ($+7.7\%$). The results reported here present a different scale *w.r.t.* the object-centric scenario due to the k of the kNN computation of metrics, which we maintain equal to c-VSG benchmark, and set it to 3, differently from the default value of 5 when using `dgm-eval` [57].

Table 5: Comparison on the GeoDE dataset under the geographical representation benchmark of c-VSG setting, with model selection on $F_1(P, C)$. Metrics are computed in DINOv2 space. All the results are re-implemented. AF: Africa, AS: Asia, EU: Europe. "label" refers to region label, "desc" to text description, and "img" to exemplar images.

Method	Ref. Info	k	Worst-Reg.	F_1 (P, C) ↑		Precision ↑		Coverage ↑		CLIPScore ↑	
				Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.	Avg.	Worst-Reg.
$LDM_{1.5}$	✗	–	AF	0.166	0.126	0.345	0.284	0.109	0.081	0.244	0.233
Semantic Guidance	label	–	AF	0.160	0.117	0.345	0.289	0.104	0.073	0.248	0.234
FG CLIP (Loss)	label	–	AF	0.186	0.154	0.339	0.308	0.129	0.103	0.246	0.244
FG CLIP (Entropy)	label	–	AF	0.178	0.133	0.328	0.272	0.122	0.088	0.238	0.236
c-VSG [25] (CLIP)	img	2	AF	0.183	0.139	0.337	0.288	0.126	0.091	0.254	0.252
c-VSG [25] (CLIP)	img	4	AF	0.184	0.133	0.338	0.289	0.127	0.086	**0.254**	**0.253**
Chamfer Guidance (Ours, DINOv2)	img	2	AF	0.246	0.221	0.551	0.507	0.159	0.141	0.248	0.242
Chamfer Guidance (Ours, DINOv2)	img	4	AF	**0.304**	**0.292**	**0.605**	**0.564**	**0.204**	**0.197**	0.250	0.244

In Table 6 we present the results computed in InceptionV3 space for DollarStreet [22], using the F_1 selection on precision and coverage. Our Chamfer guidance outperforms c-VSG [25] by 3.5% in terms of F_1 due to both an increased fidelity (precision) and diversity (coverage) of the generated samples. As previously reported for GeoDE, our method exhibits an increased image-text alignment when using CLIP as the feature extractor for the guidance.

Table 6: Comparison on the DollarStreet dataset under the geographical representation benchmark of c-VSG setting, with model selection on $F_1(P, C)$. Metrics are computed in Inception space. † indicates re-implemented results. AF: Africa, AS: Asia, EU: Europe. "label" refers to region label, "desc" to text description, and "img" to exemplar images. Our Chamfer guidance achieves state-of-the-art in terms of F_1, obtaining the best grounded diversity.

Method	Ref. Info	k	Worst-Reg.	F_1 (P, C) ↑		Precision ↑		Coverage ↑		CLIPScore ↑	
				Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.
$LDM_{1.5}$ †	✗	–	AS	0.473	0.445	0.504	0.504	0.447	0.398	0.249	**0.255**
Synonyms†	✗	–	AS	0.445	0.435	0.448	0.439	0.443	0.432	0.216	0.219
VSG† [25]	✗	–	AS	0.424	0.404	0.421	0.419	0.428	0.390	0.195	0.194
Paraphrasing†	desc	–	AF	0.440	0.436	0.445	0.451	0.436	0.422	0.226	0.215
Semantic Guidance†	label	–	AS	0.473	0.446	0.504	0.507	0.447	0.398	0.249	**0.255**
FG CLIP (Loss)†	label	–	AS	0.468	0.429	0.481	0.461	0.456	0.401	0.243	0.244
FG CLIP (Entropy)†	label	–	AS	0.468	0.440	0.487	0.500	0.453	0.393	**0.245**	0.253
Textual Inversion†	img	4	AS	0.076	0.037	0.505	0.491	0.042	0.019	0.213	0.216
c-VSG† [25] (CLIP)	img	4	EU	0.517	0.504	0.510	0.481	0.526	**0.529**	0.241	0.241
Chamfer Guidance (Ours, CLIP)	img	4	AS	0.528	0.512	0.515	0.514	0.542	0.510	0.244	0.245
Chamfer Guidance (Ours, DINOv2)	img	4	AS	**0.552**	**0.532**	**0.545**	**0.542**	**0.560**	0.520	0.240	0.244

In Table 7 and Table 8 we report the result with selection of the F_1 score between precision and recall, as in the original c-VSG benchmark. Although recall does not represent the best metric for evaluating

grounded diversity [41], our Chamfer guidance surpasses c-VSG by 3.9% in terms of recall in GeoDE, and by 2.7% on DollarStreet, leading to state-of-the-art F_1 in both datasets. Interestingly, when using recall, grounding the generation on more samples does not lead to increased diversity with CLIP, and to a marginal one with DINOv2. This was the case when using coverage, and we consider these results to be a result of the inflated recall manifold, which promotes outlier generated images.

Table 7: Comparison on the GeoDE dataset under the geographical representation benchmark of c-VSG setting, with model selection on $F_1(P, R)$. Metrics are computed in Inception space. AF: Africa, AS: Asia, EU: Europe. "label" refers to region label, "desc" to text description, and "img" to exemplar images.

Method	Ref. Info	k	Worst-Reg.	F_1 (P, R) ↑		Precision ↑		Recall ↑		CLIPScore ↑	
				Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.
$LDM_{1.5}$	✗	–	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 [25]	✗	–	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 (Entropy)	label	–	AF	0.380	0.337	0.340	0.329	0.429	0.345	0.224	0.227
Textual Inversion	img	4	AF	0.369	0.363	0.409	0.444	0.338	0.308	0.234	0.232
c-VSG [25] (CLIP)	img	2	AF	0.455	0.444	0.424	0.417	0.493	0.476	0.254	0.253
Chamfer Guidance (Ours, CLIP)	img	2	AF	0.454	0.440	0.398	0.401	**0.532**	0.489	0.245	0.245
Chamfer Guidance (Ours, CLIP)	img	4	AF	**0.463**	**0.449**	0.427	0.413	0.509	**0.492**	0.251	0.246
Chamfer Guidance (Ours, DINOv2)	img	2	AF	0.451	0.428	0.435	0.402	0.470	0.457	0.235	0.225
Chamfer Guidance (Ours, DINOv2)	img	4	AF	0.460	0.437	0.453	0.401	0.472	0.481	0.236	0.225

Table 8: Comparison on the DollarStreet dataset under the geographical representation benchmark of c-VSG setting, with model selection on $F_1(P, R)$. Metrics are computed in Inception space. † indicates re-implemented results. AF: Africa, AS: Asia, EU: Europe. "label" refers to region label, "desc" to text description, and "img" to exemplar images.

Method	Ref. Info	k	Worst-Reg.	F_1 (P, R) ↑		Precision ↑		Recall ↑		CLIPScore ↑	
				Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.	Avg.	Worst Reg.
$LDM_{1.5}$	✗	–	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	0.467	0.467	0.216	0.220
VSG [25]	✗	–	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	0.467	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	4	AS	0.425	0.398	0.478	0.491	0.386	0.335	0.217	0.219
c-VSG† (CLIP) [25]	img	4	AS	0.497	0.483	0.486	0.486	0.511	0.479	0.234	0.238
Chamfer Guidance (Ours, CLIP)	img	4	AF	0.492	0.478	0.464	0.458	0.524	0.498	0.231	0.225
Chamfer Guidance (Ours, DINOv2)	img	4	AF	**0.508**	**0.484**	**0.482**	**0.483**	**0.538**	0.486	0.225	0.215

F Additional ablation

γ ablation. This ablation examines the performance using different Chamfer guidance strengths (γ). The results in Table 9 show that this hyperparameter significantly impacts model performance, with $\gamma = 0.07$ and $k = 32$ achieving the best overall results. This configuration yields the highest F_1 score (0.931) and precision (0.950), indicating superior fidelity and diversity. Interestingly, while stronger guidance generally improves these metrics, there is a trade-off with image quality, the FID score increases from 8.840 at $\gamma = 0.05$ to 13.670 at $\gamma = 0.07$. This suggests that while stronger Chamfer guidance ($\gamma = 0.07$) produces more accurate samples, it somewhat compromises the distribution statistics compared to the moderate guidance setting ($\gamma = 0.05$), which maintains a better balance between precision/coverage performance and image quality. This might also be due to the use of the Inception network for the computation of FID instead of DINO.

Feature extractor ablation. To further validate the choice of DINOv2, we conducted a preliminary empirical study comparing its performance as a feature extractor for performing Chamfer guidance against CLIP on the GeoDE dataset in an "object-centric" setting, *e.g.*, , with prompts like `"a photo of a car."`. These experiments were run using $LDM_{1.5}$.

Our findings were as follows:

Table 9: Chamfer guidance strength γ ablation on ImageNet-1k using $\text{LDM}_{1.5}$. Our Chamfer guidance obtains the best coverage results when using a high ($\gamma = 0.07$) strength, but obtains the most balanced image quality (FID) when using a milder amount ($\gamma = 0.05$).

γ	k	$\mathbf{F_1}$ **(P, C)** ↑	**Precision** ↑	**Coverage** ↑	**Density** ↑	**Recall** ↑	**FDD** ↓	**FID** ↓
0.02	2	0.837	0.896	0.786	0.910	**0.702**	145.390	9.268
0.02	32	0.872	0.912	0.835	0.994	0.699	138.917	9.301
0.05	2	0.881	0.932	0.835	1.051	0.637	124.191	**8.840**
0.05	32	0.914	0.946	0.884	1.162	0.650	114.847	8.906
0.07	2	0.886	0.947	0.833	1.108	0.480	156.179	13.670
0.07	32	**0.931**	**0.950**	**0.912**	**1.213**	0.649	**113.301**	8.935

- DINOv2 consistently yielded the best performance, showing substantial gains in both diversity and fidelity. Similarity in DINOv2 space also correlates to human-perceived similarity slightly more than CLIP, as reported in [24].
- CLIP improved upon the baselines, and in particular, coverage scales with the number of guiding samples. With higher k we observe reduced marginal improvements, and we deem this to the fact that the model tended to converge towards generating an average representation of the object. We hypothesize this is because CLIP's pre-training is "concept-centric" (aligning images to general text concepts), whereas DINOv2's is "instance-centric" due to its self-supervised training, making it better at preserving the unique features of a specific reference image.

We report these results in Table 10. This analysis confirms that DINOv2 is the most effective choice for our method, with CLIP being an alternative.

Table 10: Comparison of feature extractors across different k values. DINOv2 shows strong improvements in F_1, precision, and coverage compared to $\text{LDM}_{1.5}$ and CLIP, while also achieving lower FID.

Feature Extractor	k	$\mathbf{F_1}(\mathbf{P}, \mathbf{C})$ ↑	**Precision** ↑	**Coverage** ↑	**Density** ↑	**Recall** ↑	**FDD** ↓	**FID** ↓
$\text{LDM}_{1.5}$ $\omega = 1.0$	–	0.2334	0.4363	0.1593	0.1731	**0.6135**	684.81	35.27
$\text{LDM}_{1.5}$ $\omega = 2.0$	–	0.3277	0.5433	0.2346	0.2647	0.4859	524.59	24.04
$\text{LDM}_{1.5}$ $\omega = 7.5$	–	0.2960	0.6222	0.1942	0.3629	0.2025	693.31	42.60
DINOv2	2	0.4242	0.6354	0.3184	0.4027	0.4614	410.84	19.73
DINOv2	4	0.5313	0.8296	0.3907	0.8933	0.1501	368.27	19.08
DINOv2	8	0.6251	0.9010	0.4785	1.3575	0.0708	353.80	18.81
DINOv2	16	**0.7527**	**0.9461**	**0.6250**	**2.0309**	0.0547	**323.11**	**18.44**
CLIP	2	0.3888	0.6367	0.2798	0.3987	0.3604	446.77	19.66
CLIP	4	0.4063	0.5951	0.3084	0.3501	0.4827	421.20	19.01
CLIP	8	0.4163	0.5974	0.3195	0.3567	0.5030	416.75	18.25
CLIP	16	0.4088	0.5757	0.3169	0.3180	0.5735	415.02	18.22

G Additional object-centric results

In Table 11 we show more results of base $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$ sampling, and additional set of parameters for CADS [49] and APG [50]. These results extend those presented in Table 1, and show how these reference-free approaches cannot significantly increase the fidelity and diversity of the samples, and that applying APG on a small ω brings moderate improvement over the best base sampling, for both $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$.

H Additional classification results

We complement our downstream utility analysis by evaluating the downstream utility of images generated using our Chamfer Guidance. We follow the "static" ImageNet-100 setup introduced in [4], and generate a dataset of $130,000$ synthetic images using our Chamfer Guidance, using $k = 32$ real exemplar images from the training set. Each synthetic image is generated with a simple prompt `(class name)`. We report the accuracy of a ViT-B [16] classifier trained on this synthetic data and tested on real validation data. In Table 12 we compare to the base performance of several LDMs, and

Table 11: Comparison of different LDM configurations with APG and CADS guidance on ImageNet-1k.

Method	APG	CADS	ω	$\mathbf{F_1}(\mathbf{P}, \mathbf{C})\uparrow$	Precision↑	Coverage↑	Density↑	FDD↓	FID↓
LDM 1.5									
LDM 1.5	—	—	1.0	0.507	0.723	0.391	0.551	431.2	31.3
LDM 1.5	—	—	2.0	0.673	0.802	0.580	0.648	226.2	**10.8**
LDM 1.5	—	—	7.5	0.709	**0.862**	0.603	**0.775**	248.7	16.1
LDM 1.5	✓	✗	2.0	0.677	0.809	0.582	0.644	226.0	**10.8**
LDM 1.5	✓	✗	4.0	**0.723**	0.855	**0.626**	0.752	**217.9**	13.4
LDM 1.5	✓	✗	7.5	0.713	0.853	0.612	0.768	247.9	16.2
LDM 1.5	✓	✗	10.0	0.707	0.858	0.601	0.766	261.2	17.0
LDM 1.5	✗	✓	1.0	0.487	0.713	0.370	0.526	431.4	31.2
LDM 1.5	✗	✓	2.0	0.676	0.806	0.582	0.641	226.5	10.9
LDM 1.5	✗	✓	4.0	0.718	0.850	0.621	0.743	218.0	13.4
LDM 1.5	✗	✓	7.5	0.715	0.856	0.615	0.771	247.0	16.1
LDM 1.5	✗	✓	10.0	0.707	0.857	0.601	0.770	260.8	16.9
LDM 3.5									
LDM 3.5	—	—	1.0	0.599	0.752	0.498	0.560	314.7	17.3
LDM 3.5	—	—	2.0	**0.727**	0.872	0.623	0.797	**231.9**	15.7
LDM 3.5	—	—	7.5	0.690	**0.878**	0.568	0.829	329.6	24.4
LDM 3.5	✓	✗	2.0	0.723	0.856	**0.625**	0.783	237.6	15.4
LDM 3.5	✓	✗	4.0	0.722	0.877	0.613	**0.832**	287.4	21.4
LDM 3.5	✓	✗	7.5	0.695	0.874	0.576	0.820	334.7	23.7
LDM 3.5	✓	✗	10.0	0.681	0.870	0.559	0.798	349.8	23.5
LDM 3.5	✗	✓	1.0	0.582	0.751	0.475	0.529	346.6	19.9
LDM 3.5	✗	✓	2.0	0.717	0.851	0.620	0.749	238.1	**15.1**
LDM 3.5	✗	✓	4.0	0.712	0.870	0.603	0.807	282.3	20.9
LDM 3.5	✗	✓	7.5	0.688	0.868	0.570	0.799	328.7	23.2
LDM 3.5	✗	✓	10.0	0.679	0.866	0.559	0.778	345.6	23.1

observe substantial gains when using our Chamfer Guidance. Interestingly, the initial performance gap of 6.5 points between $\text{LDM}_{1.5}$ and $\text{LDM}_{3.5\text{M}}$ (presumably because reduced sample diversity in $\text{LDM}_{3.5\text{M}}$) is reduced to 2.4 points when using Chamfer Guidance. These results confirm the ability of Chamfer Guidance to improve downstream utility of synthetic data.

Table 12: Validation accuracy on real images of classifiers trained with synthetic data for ImagetNet-100 classes. All models are trained for 50k iterations.

	$\text{LDM}_{1.4}$	$\text{LDM}_{2.1}$	LDM_{XL}	$\text{LDM}_{1.5}$	$\text{LDM}_{1.5}$ + Chamfer	$\text{LDM}_{3.5\text{M}}$	$\text{LDM}_{3.5\text{M}}$ + Chamfer
Real Val. Acc.	59.06	55.92	52.8	59.24	**67.82**	52.72	**65.42**

I Licenses

We report the licenses for datasets and models used in Table 13.

J Additional qualitative results

Figure 4 shows $\text{LDM}_{1.5}$ with $\omega = 1.0$ without and with our Chamfer guidance, and $\omega = 2.0$ with our Chamfer guidance. $\text{LDM}_{1.5}$ does not have a "strong" conditional only model ($\omega = 1.0$), as quantitatively highlighted in Table 11, which exhibits poor coverage and density, and a high FID. Qualitatively, we observe bad-looking images that often lack the correct subject, or present a distorted one. Our Chamfer guidance, even when applied only the conditional model, can effectively recover the correct texture, shape, and proportions of the subject. For $\text{LDM}_{1.5}$, the best results are obtained by applying our Chamfer guidance on top of a low ω value, *e.g.*, 2.0.

Figure 5 shows that $\text{LDM}_{3.5\text{M}}$ without CFG, *i.e.*, $\omega = 1.0$ produces poor quality images, often with distorted or wrong subjects. We can appreciate that Chamfer distance, applied with moderate strength γ, can effectively steer the generation towards the right subject, increasing the quality and diversity

Table 13: Links and licenses for the datasets and pre-trained models we use in our study.

Name	Link	License
ImageNet	https://www.image-net.org	Apache License 2.0
GeoDE	https://geodiverse-data-collection.cs.princeton.edu/	CC BY 4.0
DollarStreet	https://mlcommons.org/datasets/dollar-street/	CC-BY-SA 4.0
DINOv2	https://github.com/facebookresearch/dinov2	Apache License 2.0
CLIP	https://github.com/openai/CLIP	MIT License
Llama2-70b	https://huggingface.co/meta-llama/Llama-2-70b-chat	Llama 2 Community License Agreement
$\text{LDM}_{1.5}$	https://huggingface.co/ruwnayml/stable-diffusion-v1-5	CreativeML Open RAIL-M
$\text{LDM}_{3.5\text{M}}$	https://huggingface.co/stabilityai/stable-diffusion-3.5-medium	Stability AI Community License
dgm-eval	https://github.com/layer6ai-labs/dgm-eval	MIT License

of the generation. When using a high CFG strength ω *i.e.*, 7.5, steering the generation with our Chamfer guidance requires a higher strength γ before becoming effective. Our Chamfer guidance helps to recover naturalness of the image by reducing oversaturation, and to increase the variety in the backgrounds.

Figure 6 qualitatively shows the effect of increasing the number of exemplar images k when using our Chamfer guidance. We can see a greater diversity in both subjects and background when using a higher number for k, in agreement with the quantitative results presented in Figure 2.

In Figure 7 we show the comparison of the geographic diversity benchmark, in particular the GeoDE dataset. We show how base generations using $\text{LDM}_{1.5}$ introduce stereotypical elements when using the geographical indicator, *e.g.*, dilapidated cars, or rusty pans. c-VSG [25] partially mitigates the issue, but it comes at the expense of quality, generating oversaturated images that do not resemble the dataset. Our Chamfer guidance effectively mitigates both issues, by generating natural-looking images without stereotypical elements, due to the guidance of real exemplars.

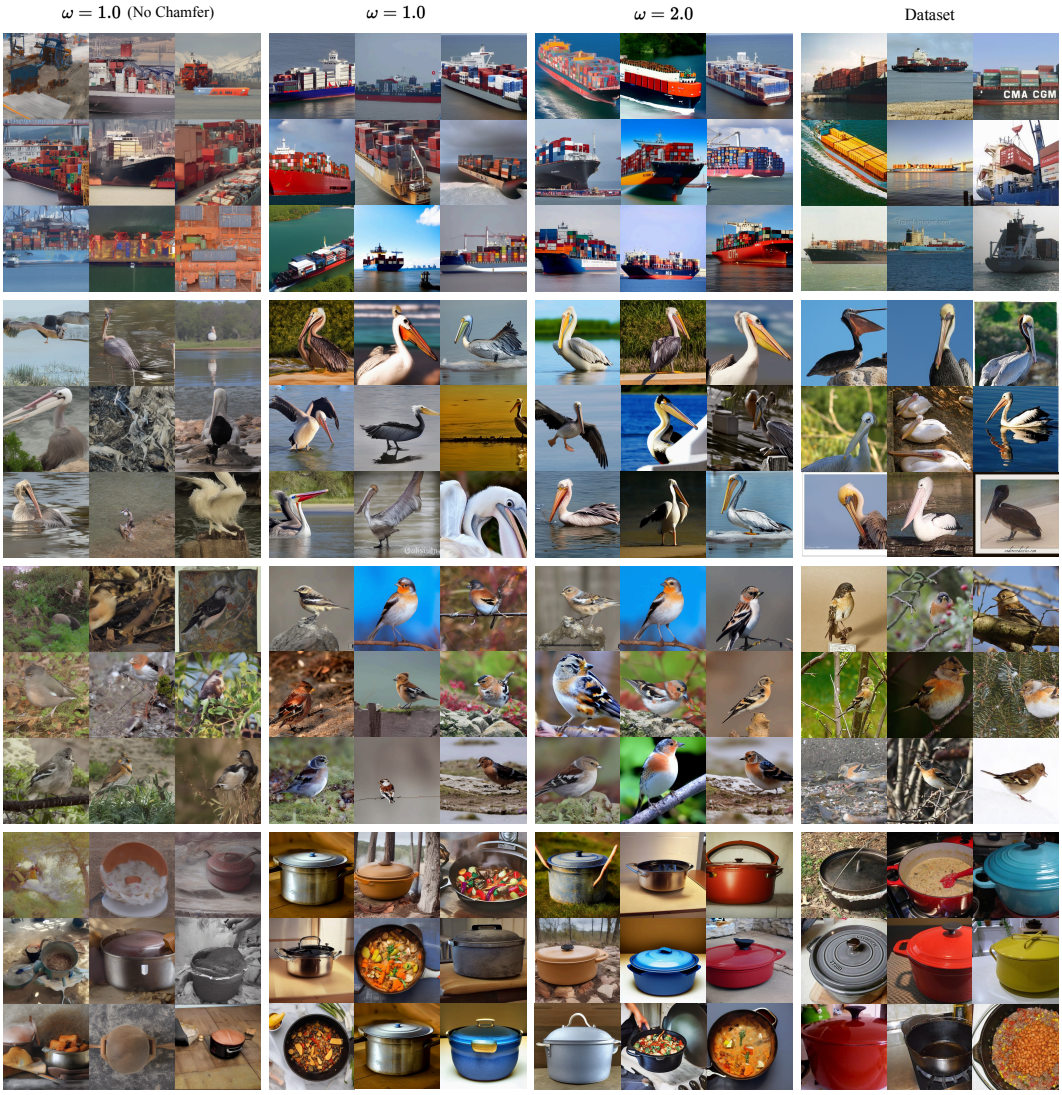


Figure 4: $\text{LDM}_{1.5}$ generations on ImageNet-1k with different ω values. $k = 32$, $\gamma = 0.07$ for our Chamfer guidance. The classes are from top to bottom: `container ship`, `pelican`, `brambling`, and `dutch oven`.

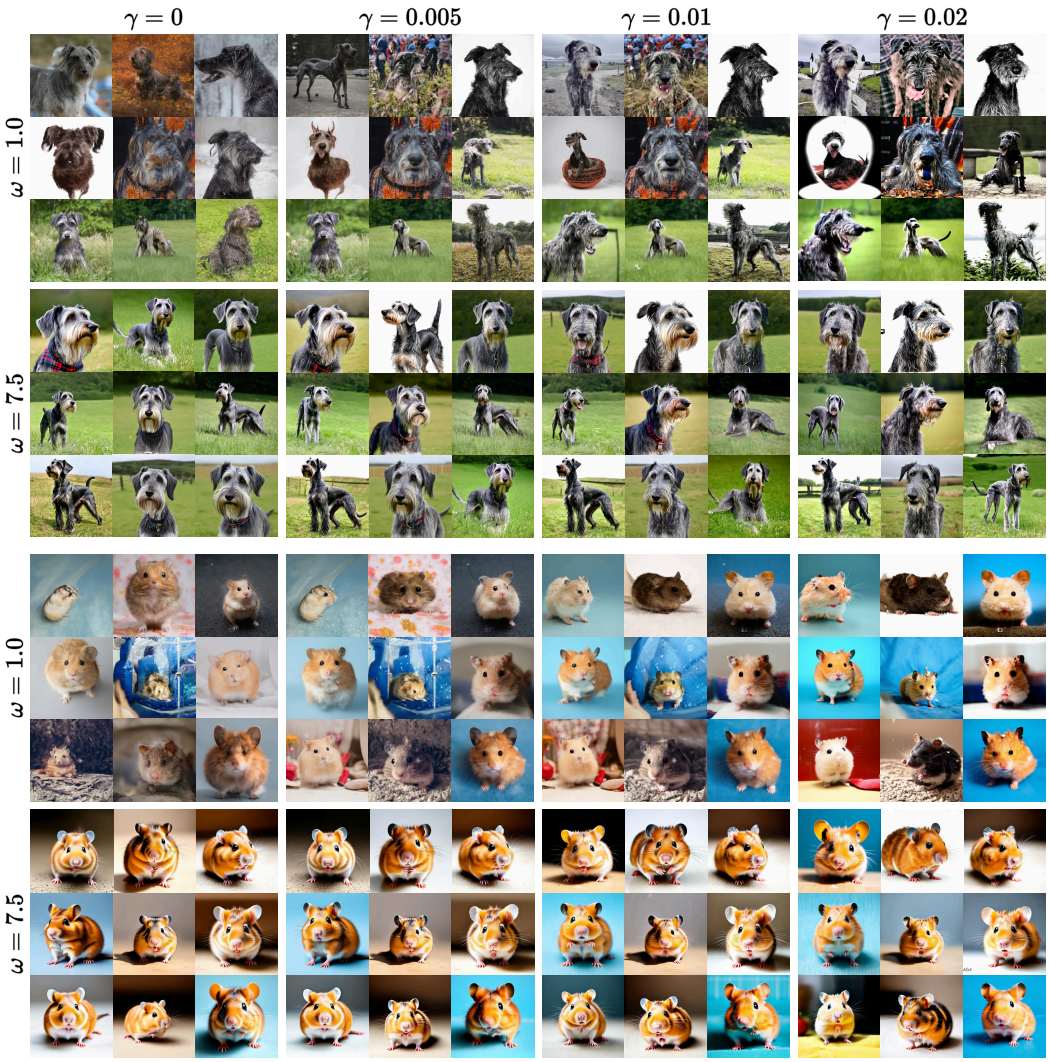


Figure 5: $\text{LDM}_{3.5\text{M}}$ on ImageNet-1k with different ω and $gamma$ values, $k = 32$. The classes are from top to bottom: `Irish wolfhound` and `hamster`

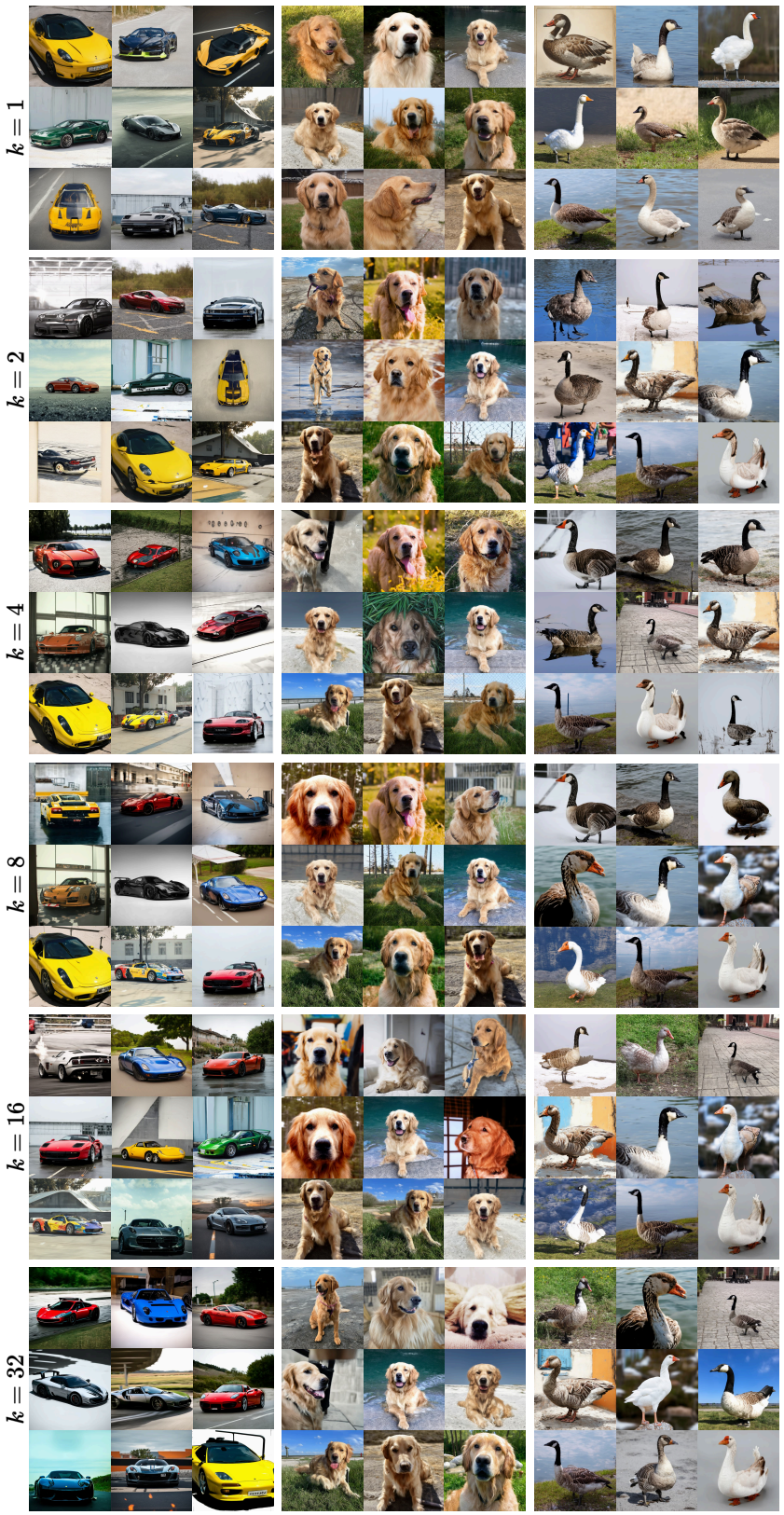
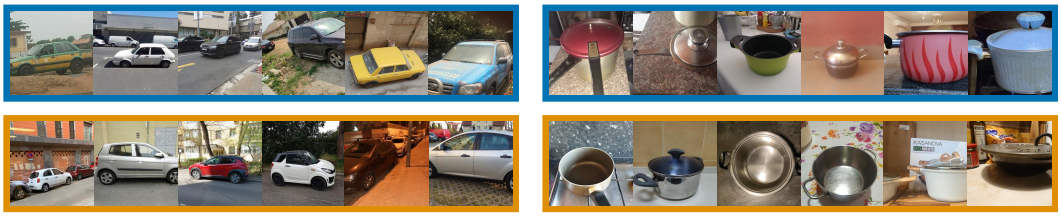


Figure 6: $\text{LDM}_{3.5\text{M}}$ on ImageNet-1k with different k values. ω and γ are picked based on the best F_1 score. The classes from left to right are: `sports car`, `golden retriever`, and `goose`.

(a) `{object}` in `{region}` baseline.

(b) c-VSG [25] generations.

(c) Our chamfer guidance.



(d) Samples from the GeoDE dataset.

Figure 7: Generated images and examples from the GeoDE dataset of cars (left) and cooking pots (right) using $\text{LDM}_{1.5}$ on GeoDE. Colors indicate images in Africa and Europe. Our Chamfer Guidance exhibits better-looking images with less saturated colors, increased subject quality and diversity in the backgrounds.