Decomposed evaluations of geographic disparities in text-to-image models

Abhishek Sureddy [*1] **Dishant Padalia** [*1] **Nandhinee Periyakaruppa** [*1] **Oindrila Saha** [1] **Adina Williams** [2]
Adriana Romero-Soriano [2 3 4 5] **Megan Richards** [**2] **Polina Kirichenko** [**2 6] **Melissa Hall** [**2]

Abstract

Recent work has identified substantial disparities in generated images of different geographic regions, including stereotypical depictions of everyday objects like houses and cars. However, existing measures for these disparities have been limited to either human evaluations, which are time-consuming and costly, or automatic metrics evaluating *full* images, which are unable to attribute these disparities to specific parts of the generated images. In this work, we introduce a new set of metrics, Decomposed Indicators of Disparities in Image Generation (Decomposed-DIG), that allows us to separately measure geographic disparities in the depiction of *objects* and *backgrounds* in generated images. Using Decomposed-DIG, we audit a widely used latent diffusion model and find that generated images depict objects with better realism than backgrounds and that backgrounds in generated images tend to contain larger regional disparities than objects. We use Decomposed-DIG to pinpoint specific examples of disparities, such as stereotypical background generation in Africa, struggling to generate modern vehicles in Africa, and unrealistically placing some objects in outdoor settings. Informed by our metric, we use a new prompting structure that enables a 52% worst-region improvement and a 20% average improvement in generated background diversity.[1]

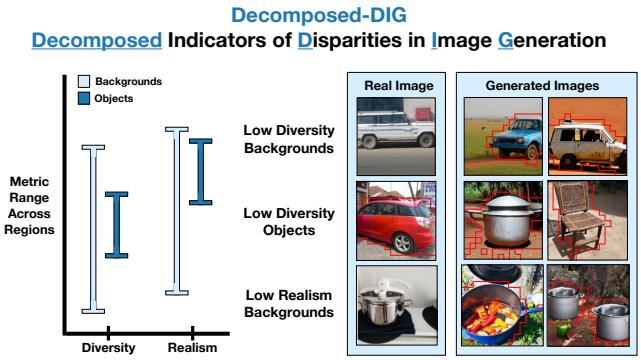

Figure 1: We introduce Decomposed-DIG, which decomposes measurements of geographic disparities in text-to-image generation between *object* and *background* representations. Using Decomposed-DIG, we identify generation patterns that contribute to geographic disparities.

1. Introduction

Recent advancements in text-to-image generative systems have driven immense progress both for visual content creation [15; 14] and training downstream discriminative models [6; 17; 18]. Despite this progress, there has also been increasing evidence that they produce content containing social biases that do not depict the real world accurately.

For example, recent works have identified that generated images often contain geographic biases, with large disparities in image realism and representation diversity among geographic regions [4]. These models can amplify harmful stereotypes, such as producing images that over-represent poverty and rudimentary infrastructure for Africa [1] or failing to produce adequate geographic representation for some regions [2]. To evaluate these disparities, some works propose automatic metrics that evaluate generated images *holistically*, either in comparison to real world images [4] or via group association rates [10]. However, recent work shows that humans interpret geographic representation via specific *components* of images, such as buildings presented in the background, natural vegetation, or stylization of central objects [5]. Therefore, it is important to understand which segments of generated images contribute to disparities to better inform the development of mitigations.

As a step in this direction, we introduce a decomposed evaluation protocol to disentangle and measure disparities between the target concept and its accompanying background in generated images. We (1) extend the precision- and coverage-based Disparity in Image Generation (DIG) indi-

[*]Joint first author [**]Directional lead [1]Department of Computer Science, University of Massachusetts Amherst [2]FAIR Labs, Meta AI [3]Mila-Quebec AI Institute [4]McGill University [5]Canada CIFAR AI Chair [6]New York University. Correspondence to: Melissa Hall <melissahall@meta.com>.

[1]All experiments were conducted by students at the University of Massachusetts Amherst on University of Massachusetts Amherst servers.

cators proposed in Hall et al. [4] and decompose them into object- and background-indicators and (2) enhance them by leveraging a state-of-the-art segmentation method [8]. We call this set of metrics "Decomposed-DIG."

Using Decomposed-DIG, we uncover a much more nuanced picture of geographic bias in model-generated images. In particular, we study images generated with the prompt `{object} in {region}` and find that:

- Generated images tend to exhibit higher realism for objects than backgrounds.
- The depiction of backgrounds in generated images has 1.7x larger disparities between geographic regions than the depiction of objects.
- Decomposed-DIG enables more precise characterization of bias modes of generative models, such as rarely including paved streets or buildings in backgrounds of images of Africa.

Informed by our findings, we explore prompting as an early mitigation, finding that a prompt template that defines geographic information as an adjective, i.e. "European car," improves background diversity by 52% for the worst-performing region and by 20% on average, with slight improvement or little cost to background realism and object representation. We hope this work encourages more fine-grained, realistic evaluation of generative vision models and informs mitigations to better enable accurate, representative generations for all global regions.

2. Background

We build upon existing indicators for disparities in image generation (DIG) [4]. DIG utilizes a dataset of real images that capture the desired geographic representation as a reference dataset and thousands of images from the text-to-image model being evaluated. We focus on the Region Indicator, which uses manifold-based metrics precision [9] and coverage [12] to measure geographic disparities in terms of realism and representation diversity, measuring disparities in generations of *objects* in different *geographic regions*.

2.1. Datasets

Reference dataset of real images. For real images, we use the geographically representative dataset GeoDE [13], which contains images of 40 objects in 6 geographic regions. Depictions of objects take up at least 25% of each image, per dataset requirements. Following Hall et al. [4], we balance GeoDE to ensure the same number of images per object-region combination, yielding 27 objects and 29k images.

Dataset of generated images. We study a latent diffusion model (LDM) and use the prompt structure `{object} in {region}` as a template. We include all object-region combinations to match the distribution of images present in our filtered version of GeoDE.

2.2. DIG Indicators

The existing DIG Indicators [4] measure disparities in the realism and diversity of the generated images across different regions by utilizing precision [9] and coverage scores [12]. We summarize these metrics below and include their full definition in Appendix C.

Precision. Precision [9] approximates the *realism* of generated images by measuring the proportion of generated images that fall close to the set of real images. In particular, both the real and generated images are mapped to a shared feature space. A manifold of real images is constructed from the hypersphere of each real image, where the hypersphere is defined by the distance to the third-nearest neighbor of the respective image. Precision thus measures the proportion of generated image features that fall into the real image manifold. Intuitively, a higher precision value means that there are more generated images that are visually similar to a real image, i.e. more realistic.

Coverage. Coverage [12] approximates the *diversity* of generated images. We use the same image features described above and count the proportion of real images for which at least one generated image falls in its hypersphere. Thus, a higher coverage indicates that more real images have representation among the generated images.

3. Decomposed-DIG

We now introduce our benchmarking protocol Decomposed Indicators of Disparities in Image Generation ("Decomposed-DIG"). Assuming one has access to real and generated images (see Section 2.1), Decomposed-DIG requires three steps: (1) segment objects and backgrounds from images, (2) extract decomposed features, and (3) perform object- and background-specific measurements.

3.1. Object and background segmentation

To divide each image into object and background components, we perform segmentation using the Segment Anything Model (SAM; [8]). However, because SAM requires points, bounding boxes, or pre-existing masks as input prompts, we use the LangSAM library[2] which leverages GroundingDINO's [11] zero-shot bounding box object detection as inputs to SAM. We use bounding boxes corresponding to the specific object that we know appears in the real image or was used in prompting for the generated image. SAM then yields precise segmentation masks for object class. We consider the remaining image regions as image backgrounds.

In our experiments, we find that LangSAM works well for nearly all object classes in real images, exhibiting failures for a small subset of object classes such as "cleaning equipment," "hand soap," and "light fixture", which we remove from our evaluation (see Appendix C for more details).

[2] https://github.com/luca-medeiros/lang-segment-anything

Figure 2: Examples of real and generated images of objects in different regions (Europe in top row and Africa in the bottom row). We propose Decomposed-DIG to pinpoint geographic disparities related to the depiction of objects and backgrounds in generated images created with the prompt `{object} in {region}`. We then study an alternative prompt template that emphasizes the object more than the region: `{regional adjective} {object}` i.e. "European bag", which leads to higher background diversity. Red outlines show object/background decompositions.

3.2. Decomposed image features

With the object and background segmentation, we use a vision transformer (ViT) [3] for feature extraction (specifically, ViT base model with 16x16 patching pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K). Unlike CNN-based feature extractors such as InceptionV3 [16], the ViT's use of patches allows us to isolate the features corresponding to objects and backgrounds [7]. Using the segmentation mask from the previous step, we consider all the patches that contain at least one pixel of the object as object patches, and the remaining patches (which do not have any object pixels) as background patches (thus, together they make up the full image without overlap). For object-specific measurements, we mask out background patches by zeroing out the attention scores corresponding to these patches during the forward pass. Similarly, for background-specific measurements, we zero out the attention scores of the object patches. For full-image measurements, we use all the patches without masking. We use the CLS token features from the last layer of the ViT as the image features for each of the set-ups.

3.3. Object- and background-specific measurements

With the object and background ViT features, we benchmark (1) object-specific ("Obj-only") and (2) background-specific ("BG-only") disparities in realism and diversity in generated images. We also maintain full image measurements ("Full-image") from previous work [4], including ViT tokens for the entire image.

Obj-only benchmarking. For object-specific measurements, we disaggregate real and generated images between geographic regions and calculate precision and coverage measurements by using only the ViT features corresponding to *object* segmentations. The object features coming from real images are used to approximate the manifold of real objects, whereas the object features coming from generated images are projected onto the real object manifold to compute precision and coverage. We include generated images that contain no object segmentations in the overall count of samples, which helps in penalizing generations that have consistency issues as observed in previous works [4; 5].

BG-only benchmarking. Similarly, for benchmarking of disparities in depicted backgrounds, we disaggregate between geographic regions and calculate precision and coverage using features corresponding to image *backgrounds*. In this case, generated images that do not have object segmentations are included in their entirety.

4. Results

Next, we leverage Decomposed-DIG to better understand the object and background components of the geographical disparities in the widely used LDM 1.5.[3]

4.1. Objects have better realism than backgrounds

We first compare the generations of objects and backgrounds, computing Decomposed-DIG metrics on Obj-only and BG-only portions of generated images, shown in Figure 3. We find lower precision in BG-only decompositions compared to Obj-only segments, across region subsets. This difference in precision suggests that generated backgrounds are less similar to real backgrounds than generated objects are to real objects. The examples in Figure 2 are consistent with this finding as generated backgrounds often portray settings that are not frequently found in real images, such as rural dirt scenes for Africa or large stone historic architecture for Europe. We further investigate the generation patterns that drive this difference in Section 4.3.

In addition, we find that *on average*, coverage is similar between Obj-only and BG-only set-ups. However, *per-*

[3]We focus on Obj-only and BG-only measurements in the main text and include Full-image results in Appendix D.

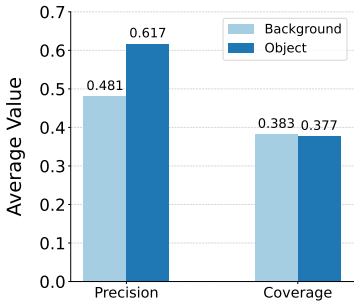

Figure 3: **In generated images, objects tend to have better realism (precision) than backgrounds, while representation diversity (coverage) is similar on average between objects and backgrounds.** Shown are precision and coverage measurements averaged over all regions for Obj-only and BG-only set-ups.

region disparities vary notably between the objects and backgrounds, which we discuss next.

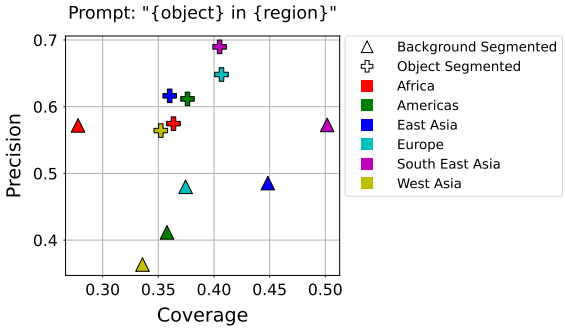

Figure 4: **Backgrounds in generated images have larger disparities in realism and diversity between geographic regions than objects.** We observe the larger variance of precision and coverage values for BG-only compared to Obj-only set-up.

4.2. Backgrounds have larger geographic disparities than objects

Using Decomposed-DIG, we find that backgrounds have much larger geographic disparities than objects (shown in Figure 4). In particular, BG-only measurements reveal that background coverage is *twice* as large for the best-performing region (Southeast Asia) than the worst performing region (Africa), while only 1.2x larger for object coverage. Furthermore, we see that per-region precision scores for backgrounds span about 1.5x the amount of per-region scores for objects, demonstrating larger region disparities for backgrounds. Finally, we observe that unlike the other regions, Africa shows changes only in coverage (not precision) between object and background setups. We discuss this further in Appendix D.

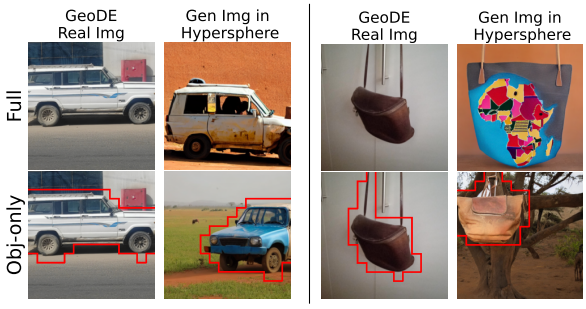

Figure 5: **When generating objects in Africa, the LDM struggles to depict full background diversity, in particular backgrounds with buildings and paved streets (left) and neutral indoor scenes (right).** Depicted are examples of real images where there are no generated images in the hypersphere of nearest neighbors for BG-only, but there are generated images in the hypersphere for Full-image and Obj-only (shown). Red outlines show object/background decompositions.

4.3. Characterizing generation patterns contributing to geographic disparities

Informed by our quantitative findings, we next use Decomposed-DIG to qualitatively pinpoint generation patterns that contribute to geographic disparities, investigating low diversity backgrounds, low diversity objects, and low realism backgrounds.

Low Diversity Backgrounds: Generated backgrounds of Africa rarely include neutral, grey colored scenes. We examine real images pertaining to low diversity background generations, i.e. real images contain generated images in their hyperspheres with the Obj-only set-up but not with the BG-only set-up. This reveals examples where the generated images do not cover the full diversity of real backgrounds. For example, in Figure 5 we show examples of real images of a car and a bag in Africa with a boxy, SUV-like appearance in a neutral background featuring grey concrete and an unmarked building. When we consider the full image in application of the original DIG measurements, multiple generated images depicting a similar object but starkly different backgrounds are included in the real image's hypersphere. However, when we utilize Decomposed-DIG and split up the analysis between object and background, we find that there are no generated image backgrounds that fall in the real image's background hypersphere. Thus, we identify our first failure mode: the LDM struggles to depict realistic neutral scenes featuring plain grey backgrounds in the generation of objects in Africa.

Low Diversity Objects: Generated cars in Africa do not include red sedans. We next inspect real images pertaining to low diversity object generations, i.e. with generated images in their hyperspheres in the BG-only set-up but not

Figure 6: **The LDM struggles to depict full object diversity, particularly more modern vehicles (such as sedans) in Africa**. Shown are real images with the generated images that are in their hypersphere of nearest neighbors for BG-only measurement, but not Full-image or Obj-only. Red outlines show object/background decompositions.

the Obj-only set-up. These serve as examples where the generations do not cover the full diversity of real objects. Figure 6 shows examples of real images of red sedans in Africa that have generated images in their hypersphere for the BG-only set-up but not the Full-image or Obj-only measurement. This shows the utility of Decomposed-DIG in allowing us to pinpoint that the reason for the full image's empty hypersphere is not due to lack of background representation in the generated images but rather object depiction. And indeed, when we manually inspect all images of generated cars in Africa, we find no examples of red sedans and more than 90% are dirty and rusty (similar to those in Figure 2).

Low Realism Backgrounds: Generated backgrounds in Europe unrealistically portray cooking pots with outdoor backgrounds. Finally, we examine generated images whose BG-only measurements never fall in the manifold of the real images. This allows us to identify cases of poor generation realism. Using the BG-only set-up, we inspect random examples of generated images created with the prompt `cooking pot in Europe` that do not fall in the real image manifold. We observe that these generations often place the cooking pots in outdoor scenes, along brick walls or on a stony surface with foliage and dirt behind them, while real images of cooking pots depict them indoors. Randomly selected examples are included in Figure 7, with more in Appendix Figure 14.

5. Early mitigations via new prompt template

Informed by our findings, we explore prompting as an early mitigation for regional disparities in generation. Because our analysis discovered backgrounds to be a driver of disparities, we study whether prompts that use adjective descriptors ("European bag") rather than noun descriptors that could be construed as an instruction to depict the object *in* the region ("bag in Europe") reduce disparities in generations. We repeat experiments using the new prompting template.

Table 1 shows results of the new prompting method in comparison with the original method, and Appendix Figure 12 shows a side-by-side comparison. We find that adjective-based prompting substantially improves background diversity (coverage), by 52% for the worst region and 20% on average. When we inspect random qualitative examples shown in Figure 2, we see that the backgrounds tend to be more varied, with fewer stereotypical region representations not seen in the real dataset and more neutral and plain scenes. We see that this comes with a small improvement in worst-group background realism (precision) and a small drop of 3% in average precision. In addition, the new prompting template leads to small improvements in realism and diversity in the depiction of objects on average.

Figure 7: **Generated images of cooking pots in Europe often unrealistically place the pots outdoors.** Depicted are generated images that, when measured with BG-only setting, never fall in the manifold of real images. Red outlines show object/background decompositions.

	Obj-only				BG-only			
	Avg. Prec.	Worst Prec.	Avg. Covg.	Worst Covg.	Avg. Prec.	Worst Prec.	Avg. Covg.	Worst Covg.
Orig.	0.617	0.564	0.377	**0.352**	**0.481**	0.363	0.383	0.278
New	**0.665**	**0.609**	**0.390**	0.344	0.466	**0.389**	**0.461**	**0.423**
Delta	**0.048**	**0.045**	**0.013**	-0.008	-0.015	**0.026**	**0.078**	**0.145**
Delta (%)	**8%**	**8%**	**3%**	-2%	-3%	**7%**	**20%**	**52%**

Table 1: **Prompting with region adjectives improves background diversity with little cost to object realism or diversity.** Shown are Decomposed-DIG values for the prompt templates `{object} in {region}` (**Orig.**) and `{region adjective} {object}` (**New**). Worst refers to the lowest performing region, primarily Africa and West Asia (see Figure 12).

6. Conclusion

In this work, we introduce the Decomposed-DIG benchmark and use it to identify that a widely used LDM contributes to geographic disparities largely through the depiction of backgrounds. We identify specific failures like the lack of generation of red sedans for Africa and placing cooking pots outdoors for Europe. Finally, we experiment with an early attempt at mitigations via a new prompting strategy, demonstrating that Decomposed-DIG can be useful in informing mitigations that can lead to up to a 52% improvement in representation diveristy of backgrounds for the worst-performing region. We hope that this work paves the way for more fine-grained analysis and mitigation strategies addressing disparities in text-to-image generation.

References

[1] Bansal, H., Yin, D., Monajatipoor, M., and Chang, K.-W. How well can text-to-image generative models understand ethical natural language interventions? In *EMNLP (Short)*, 2022.

[2] Basu, A., Babu, R. V., and Pruthi, D. Inspecting the geographical representativeness of images from text-to-image models. *arXiv preprint arXiv:2305.11080*, 2023.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Hall, M., Ross, C., Williams, A., Carion, N., Drozdzal, M., and Soriano, A. R. Dig in: Evaluating disparities in image generations with indicators for geographic diversity, 2023.

[5] Hall, M., Bell, S. J., Ross, C., Williams, A., Drozdzal, M., and Soriano, A. R. Towards geographic inclusion in the evaluation of text-to-image models, 2024.

[6] Hemmat, R. A., Pezeshki, M., Bordes, F., Drozdzal, M., and Romero-Soriano, A. Feedback-guided data synthesis for imbalanced classification, 2023.

[7] Jain, S., Salman, H., Wong, E., Zhang, P., Vineet, V., Vemprala, S., and Madry, A. Missingness bias in model debugging, 2022.

[8] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Gershick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[9] Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models, 2019.

[10] Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H. B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.-Y., Fei-Fei, L., Wu, J., Ermon, S., and Liang, P. Holistic evaluation of text-to-image models, 2023.

[11] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[12] Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models, 2020.

[13] Ramaswamy, V. V., Lin, S. Y., Zhao, D., Adcock, A. B., van der Maaten, L., Ghadiyaram, D., and Russakovsky, O. Beyond web-scraping: Crowd-sourcing a geodiverse dataset. In *arXiv preprint*, 2023.

[14] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022.

[15] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.

[17] Tian, Y., Fan, L., Isola, P., Chang, H., and Krishnan, D. Stablerep: Synthetic images from text-to-image models make strong visual representation learners, 2023.

[18] Yeo, T., Atanov, A., Benoit, H., Alekseev, A., Ray, R., Akhoondi, P. E., and Zamir, A. Controlled training data generation with diffusion models. *arXiv preprint arXiv:2403.15309*, 2024.

A. Social Impacts

In this work, we introduced Decomposed-DIG to better understand patterns of geographic disparities in text-to-image generative models. The use of our metrics may have broader impacts in guiding the priorities of realism and representation diversity disparity mitigation efforts, especially along the axes of object and background representation. In addition, the nature of these metrics necessarily focuses analysis on certain geographic categorizations and groups. While these criteria and groups have been shown to closely correspond to human prioritization when evaluating geographic representation [5; 2], we hope that our work does not preclude research into other methods for understanding geographic disparities and ensuring inclusivity in text-to-image geographic capabilities.

B. Limitations

While Decomposed-DIG is useful in providing insights specific to object and background realism and representation diversity in generated images, we discuss possible limitations of our method.

First, it is possible that the feature extractor used for precision and coverage measurements is better for either backgrounds or objects, potentially biasing our results. However, we did not observe issues along these lines in our qualitative review of experimental results.

In addition, we can only apply Decomposed-DIG for cases where the segmentation model and object detection pipeline works well. While we took steps to mitigate this risk by evaluating with the real images (for which we have ground truth information) and removing images and classes with significant failures, there may be unique failures for generated images. We analyze the small set of failure cases that we observed with the segmentation pipeline further in Appendix C.

Finally, automatic metrics are not sufficient in identifying the full breadth of human preference in image realism and representation diversity. Thus, Decomposed-DIG is best used accompanied by qualitative study.

C. Additional Methodology Details

In this Section, we discuss additional details about our proposed methodology and related background material.

C.1. Precision and coverage definitions

C.1.1. Precision or Realism

Precision measures how close the generated images are to the real images from the reference dataset. We calculate the precision by following the approach in Kynkäänniemi et al. [9] and Hall et al. [4], where we determine the proportion of the generated images that lie within the manifold, i.e. k-th nearest neighbor distance of at least one of the real images in a pre-defined feature space. The existing work [4] calculates this k-th nearest neighbor distance in the InceptionV3 [16] feature space. The formula for calculating the precision is given by previous works [4; 9] as:

$$\mathrm{P}(\mathcal{D}_r, \mathcal{D}_g) = \frac{1}{|\mathcal{D}_g|} \sum_{i=1}^{|\mathcal{D}_g|} \mathbb{1}_{h_g^{(i)} \in \text{manifold}(\mathcal{D}_r)}, \tag{1}$$

P where $\mathcal{D}_r = \{h_r^{(j)}\}$ is dataset of reference images features, $\mathcal{D}_g = \{h_g^{(i)}\}$ is dataset of generated images features.

C.1.2. Coverage or Diversity

Diversity is a measure of how representative the generated images dataset is of the reference dataset. In other words, it is the proportion of the reference images that have at least one of the generated images within their manifold in a pre-defined feature space. We use this definition of coverage proposed in Naeem et al. [12], to measure the diversity of generated images. We compute the coverage following the prior works [4; 12] as follows:

$$\mathrm{C}(\mathcal{D}_r, \mathcal{D}_g) = \frac{1}{|\mathcal{D}_r|} \sum_{j=1}^{|\mathcal{D}_r|} \mathbb{1}_{\exists\, i \text{ s.t. } h_g^{(i)} \in \text{manifold}\left(h_r^{(j)}\right)}. \tag{2}$$

Figure 8: Illustration of the image segmentation using LangSAM: full images (a, d), segmented objects (b, e) and segmented backgrounds (c, f).

C.2. Language-based object and background segmentation

To build our Decomposed-DIG metrics, we first perform an automatic object-background segmentation using LangSAM,[4] a state-of-the-art segmentation tool. LangSAM automatically segments images in two stages. The first stage uses GroundingDino [11] to generate a bounding box for an image given the object prompt from the text prompt we provide (e.g. "bag"). The second stage uses the bounding box to prompt a segmentation model SAM [8].

Once we have the segmentation masks, we fill in the removed components of the image. For object-segmented images, the background is whitened. For background-segmented images, the object is whitened. Figure 8 illustrates the resulting object and background segmented images using LangSAM.

C.2.1. Data Filtering

The real image dataset, GeoDE [13] has 27 classes of objects with at least 170 images in each of the 6 regions. We found 3 classes in GeoDE had over 100 segmentation failures out of 1020 images (170 $\times$ 6 regions), and we excluded these classes from analyses. The excluded classes are: "cleaning equipment", "hand soap" and "light fixture". Hence, we used only the resulting 24 classes for our analysis, namely: tree, hat, cooking pot, jug, toothbrush, bag, waste container, bicycle, chair, stove, dog, car, plate of food, spices, medicine, hairbrush comb, toy, light switch, front door, storefront, lighter, dustbin, toothpaste toothpowder, and candle.

C.3. Feature extractors

Initially, we experimented with convolutional neural network (CNN) based models like Inception-v3 [16]. However, a significant drawback of CNNs is that they consider the entire image for feature extraction. In segmented images, large portions of the image are replaced with a specified pixel value, affecting the resulting features that are extracted. With patch-based models, we avoid this problem by only selecting patches within the segmentation.

D. Additional findings

In this Section, we include additional experimental results.

D.1. Effects of new prompting strategy

Figure 9 shows results averaged over all regions comparing old prompting structure – `{object} in {region}`, and the new prompting structure – `{regional adjective} {object}`. We also include object-specific measurements with the old and new prompting strategies in Figures 10 and 11 respectively.

D.2. Full image results

We show a side-by-side comparison of both prompting strategies for Full-image, Obj-only, and BG-only measurements in Figure 12.

[4]https://github.com/luca-medeiros/lang-segment-anything

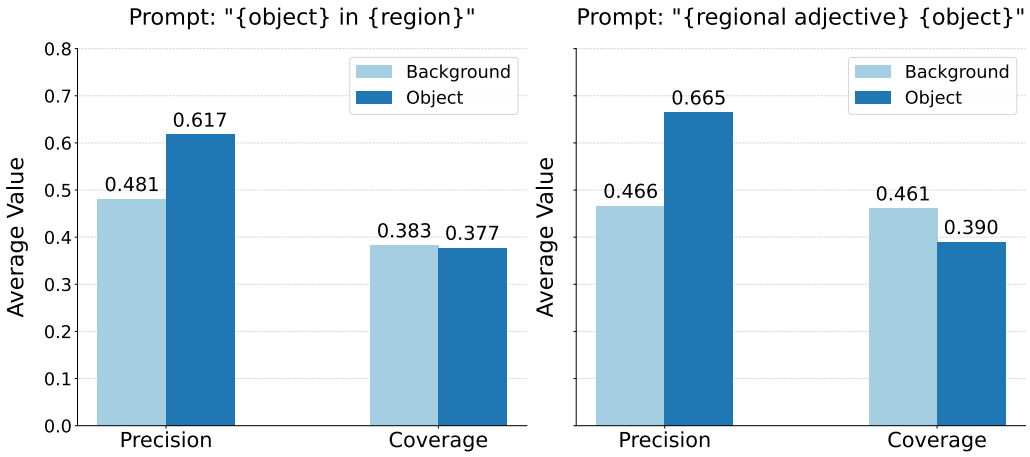

Figure 9: Precision and coverage measurements averaged over all regions for Obj-only and BG-only set-ups for {`object`} `in` {`region`} generation prompt (left) and {`regional adjective`} {`object`} prompt (right).

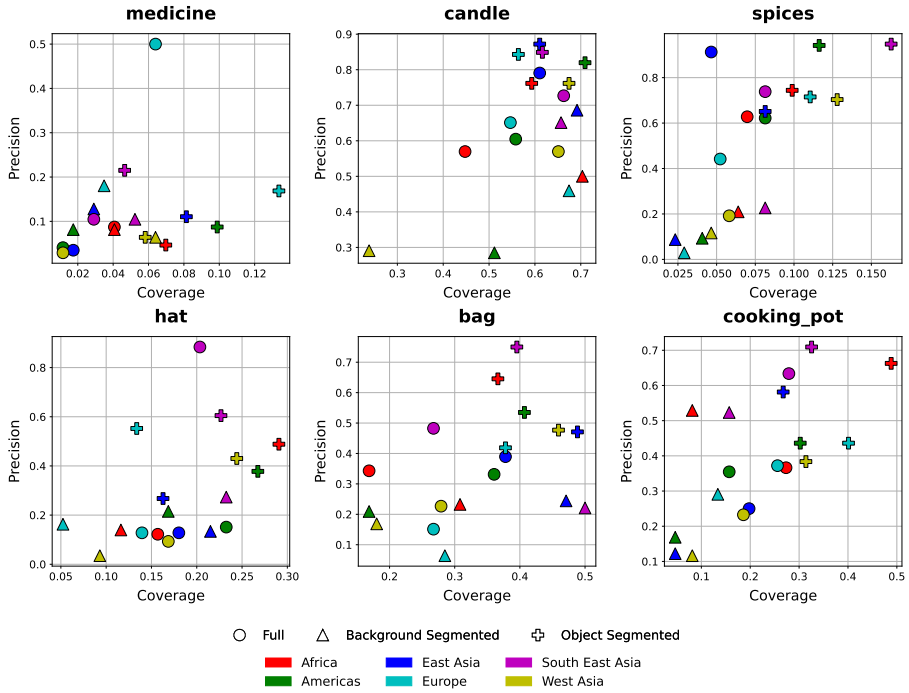

Figure 10: Object-wise Decomposed-DIG with prompt template {`object`} `in` {`region`}. Note axes scale differences between plots.

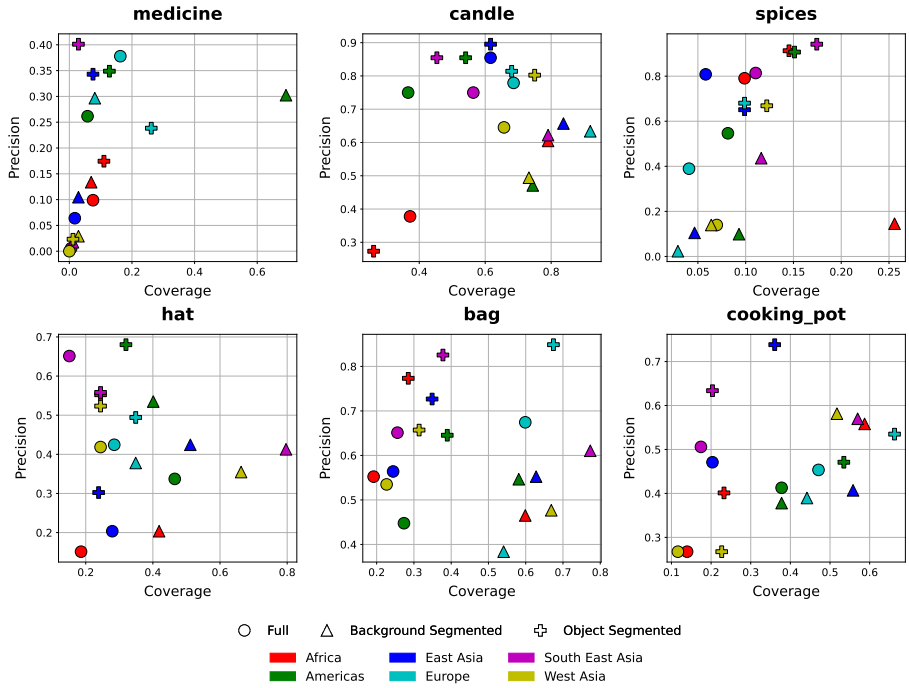

Figure 11: Object-wise Decomposed-DIG with prompt template {`regional adjective`} {`object`}. Note axes scale differences between plots.

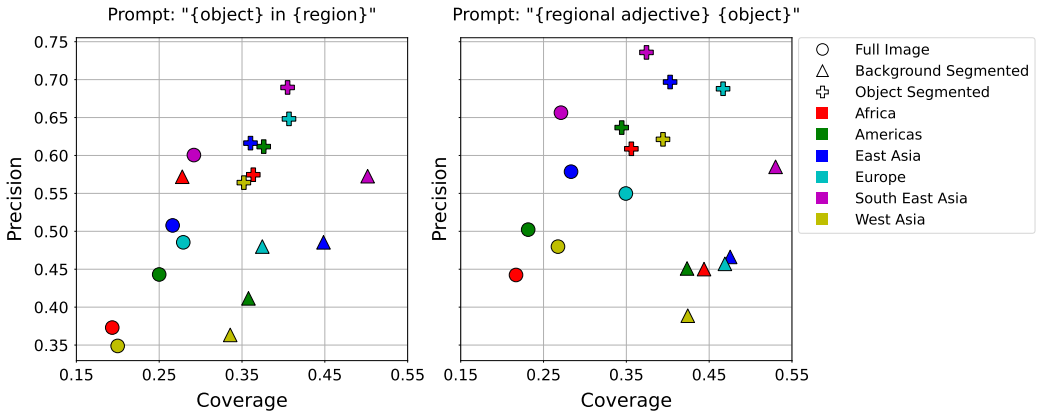

Figure 12: Decomposed-DIG metrics comparing the two prompting strategies.

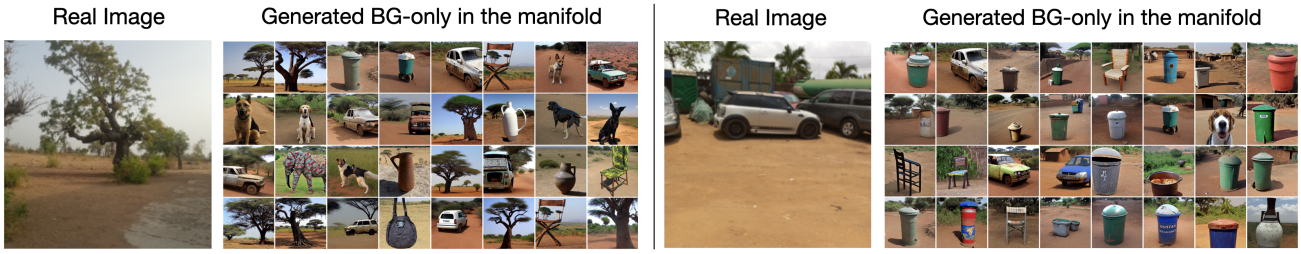

Figure 13: Examples of the reference image from African region which have many generated images falling into the real image manifold in BG-only setting.

Figure 14: Additional examples of the generated images for `cooking pot in Europe` template, illustrating the low realism for background generation.

D.3. Analyzing the precision BG-only score for Africa

From Figure 4, we observe that Africa has high precision in BG-only setting. We conjecture that this is because most of the generated image backgrounds are stereotypical and resemble sandy and plain desert areas, while most of the images in the reference dataset (GeoDE) have diverse set of backgrounds such as paved roads, kitchen floors, and grass areas. There are very few images of Africa in the reference dataset which contain stereotypical backgrounds and, as a result, most of the generated images fall into the manifolds of BG-only setting of these few images. As an illustration, the real images in Figure 13 have more than 50 generated images included in its manifold in BG-only setting. However, other more realistic backgrounds do not have any associated generated image in their manifold. Hence, we hypothesize that a minority of real images led to high precision and low coverage scores for BG-only measurement in Africa.

D.4. Additional visual examples

Figure 14 contains additional examples of generated images placing cooking pots outdoors.