DIAGen: Semantically Diverse Image Augmentation with Generative Models for Few-Shot Learning

Tobias Lingenberg[*,1], Markus Reuter[*,1], Gopika Sudhakaran[†,1,2], Dominik Gojny[1], Stefan Roth[1,2], and Simone Schaub-Meyer[1,2]

[1]Department of Computer Science, Technical University of Darmstadt, Germany
[2]Hessian Center for AI (hessian.AI)
{tobias.lingenberg, markus.reuter}@stud.tu-darmstadt.de
{gopika.sudhakaran, stefan.roth, simone.schaub}@visinf.tu-darmstadt.de

Abstract. Simple data augmentation techniques, such as rotations and flips, are widely used to enhance the generalization power of computer vision models. However, these techniques often fail to modify high-level semantic attributes of a class. To address this limitation, researchers have explored generative augmentation methods like the recently proposed DA-Fusion. Despite some progress, the variations are still largely limited to textural changes, thus falling short on aspects like varied viewpoints, environment, weather conditions, or even class-level semantic attributes (*e.g.*, variations in a dog's breed). To overcome this challenge, we propose DIAGen, building upon DA-Fusion. First, we apply Gaussian noise to the embeddings of an object learned with Textual Inversion to diversify generations using a pre-trained diffusion model's knowledge. Second, we exploit the general knowledge of a text-to-text generative model to guide the image generation of the diffusion model with varied class-specific prompts. Finally, We introduce a weighting mechanism to mitigate the impact of poorly generated samples. Experimental results across various datasets show that DIAGen not only enhances semantic diversity but also improves the performance of subsequent classifiers. The advantages of DIAGen over standard augmentations and the DA-Fusion baseline are particularly pronounced with out-of-distribution samples.[1]

Keywords: Image Augmentation · Diffusion Models · Few-Shot Classification · Dataset Diversity.

1 Introduction

A common problem in the field of computer vision is the insufficient amount of real-world training data [24]. Collecting and annotating data at scale can

[1] Code is available at https://github.com/visinf/DIAGen, [*]Equal contribution, [†]Corresponding author.

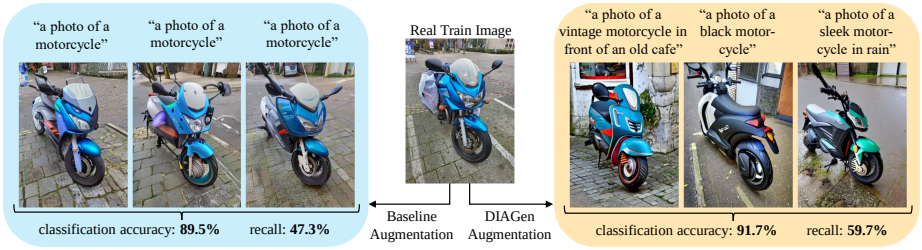

Fig. 1: Comparison of augmentation results between the baseline method DA-Fusion [47] *(left)* and our proposed approach DIAGen *(right)*, utilizing the same guiding image *(middle)* for the augmentation process. DIAGen demonstrates superior, semantically diverse image augmentations, as evidenced through more variation of object appearance and settings. This observation is supported by improvements in classification accuracy and recall as a diversity metric [19].

be difficult, expensive, and time-consuming [22]. To address this issue, data augmentation techniques are crucial in scenarios with very few labelled samples (few-shot) [30, 38], as they support generalization and robustness by introducing data variation. While offering valuable benefits, standard augmentation methods like rotations, flips, and scaling often fall short in providing semantic diversity beyond that of the original data [47]. This lack of diversity in training data negatively influences downstream applications, *e.g.*, for objects that typically occur in certain environments, such as a cow being correctly classified on a grassy background but not on a beach [4]. In response to the bottleneck of real data and its lack of diversity, researchers have turned to synthetic data generation as a promising alternative [22]. Recent advancements, such as DA-Fusion by Trabucco *et al.* [47], demonstrate the potential of synthetic data augmentation by using an off-the-shelf diffusion model.

However, the lack of diversity in synthetic data generation is still a known issue [22, 39]. Upon inspecting the images generated by DA-Fusion (see Figs. 1 and 6), it is evident that they appear to be very similar, primarily altering textural details and minor structural elements with no noticeable change in viewpoint, limiting its effect as an augmentation technique. We observe that DA-Fusion is constrained in achieving sufficient diversity due to a lack of control over how an image is augmented. To address this issue, we propose DIAGen (**D**iverse **I**mage **A**ugmentation with **Gen**erative Models), which builds on DA-Fusion and adds three components to it. Thereby, DIAGen enhances the semantic diversity of synthetic images while maintaining high quality, making it an effective augmentation technique to generate diverse training data that simulates a wide range of environments for applications such as autonomous vehicles and robotics. Thus, improving model robustness by enabling safer, more reliable behavior in real-world scenarios. Additionally, DIAGen can be applied to downstream tasks like relation detection [41, 44] to enhance the augmentation diversity of rare relations to mitigate biased prediction.

The **main contributions** of our work can be summarized as follows: *(i)* We introduce variations in the embedding space of learned class concepts by adding Gaussian noise, taking advantage of the semantic richness inherent in vector representations within the embedding space [25]. *(ii)* Inspired by the idea of He *et al*. [11], we guide the generation process of the diffusion model with varied class-specific text prompts. In contrast to [11], we obtain meaningful prompts by leveraging the world-knowledge of a text-to-text generative model, here *GPT-4* [1]. However, increasing the diversity may result in a reduced quality of some of the generated images, a challenge referred to as the fidelity-diversity trade-off [27]. *(iii)* To tackle this potential issue, we use a weighting mechanism for synthetic images, which was previously considered in the context of Generative Adversarial Networks (GANs) [50]. *(iv)* We show the effectiveness of our model by comparing the accuracy of a downstream classifier to DA-Fusion and standard augmentations across multiple datasets in few-shot settings.

2 Related Work

Synthetic Data for Few-Shot Learning. Numerous works have explored synthetic image generation using GANs [5, 16, 42, 53] and diffusion models [14, 15, 29, 35]. In few-shot learning, the small number of labelled images presents a challenge due to the inherently scarce class sampling. Collecting more real-world data is resource intensive [22], but synthetic data can utilize the knowledge of pre-trained generative model. While GANs have already been used for few-shot learning [3], diffusion models offer better results [11] due to their stability, high image quality, and flexibility during image generation [6].

Recently, Trabucco *et al*. [47] attempted to generalize their pipeline based on a diffusion model, DA-Fusion, to unseen concepts by integrating Textual Inversion [8]. This method uses three to five real images to learn new visual concepts, creating pseudo word vectors for a text-to-image model. Textual Inversion is ideal for few-shot learning, enhancing the text encoder's vocabulary with new concepts. DA-Fusion then uses these embeddings to generate synthetic images with Stable Diffusion [34]. The denoising process of the diffusion model is conditioned on the text prompt and guided by a real training image [23]. DIAGen builds upon the work of Trabucco *et al*. [47] to increase the semantic diversity of the synthetic images further.

Diversity in Datasets. The quality of machine learning models heavily relies on the diversity of their training data. A lack of diversity can lead to biases and poor performance [18], particularly in few-shot scenarios [17]. Creating synthetic data comes with a challenging trade-off: balancing fidelity for accurate representation and diversity for increased coverage [27]. While previous methods have improved synthetic data quality, they only address coverage implicitly. That said, there have been attempts to explicitly focus on the aspect of diversity. Wang *et al*. [49] demonstrated promising results by using a diversity measurement-based meta-learner. He *et al*. [11] utilized a text-to-image diffusion model and inserted

different descriptive image prompts to achieve a higher coverage for zero- and few-shot learning. Although we use a similar idea, our method can generalize to unseen concepts and provide explicit control over how image prompts are generated, by instructing our LLM with a meta prompt.

Due to the importance of a high visual quality and coverage of synthetic datasets, many metrics have been proposed to assess these properties [40]. The most common metrics are the Fréchet Inception Distance (FID) [13] and the Inception Score (IS) [37], which rely on a pre-existing classifier (InceptionNet [43]). However, both summarise the comparison of the two distributions (real and synthetic) into a single number, overlooking the distinction between fidelity and diversity [36]. To address this, more refined metrics like precision and recall [2, 36], density and coverage [27], and the Vendi score [7] have been developed. In our work, we use an improved version of precision and recall [19] due to its wide acceptance in the text-to-image community and its high agreement with human perception [40].

Out-of-Distribution Generalization. Conventional machine learning algorithms often assume that training and test data come from the same distribution. However, in real-world applications, this assumption often fails to hold due to unforeseen distributional shifts. This can lead to a drastic decline in real-world performance [12, 21, 26, 28]. Especially in safety-critical applications, these out-of-distribution (OOD) scenarios need to be handled with the same quality and confidence as identically distributed data. While there have been advancements in OOD detection to reject these samples or hand them over to human users (*e.g.*, in the case of autonomous driving) [51], our goal is to investigate whether an increased dataset diversity allows for the implicit handling of these cases.

Recent advancements in large-scale models designed to encapsulate extensive world knowledge have enabled a broader coverage of OOD examples. Tong and Dai [45] demonstrate the promising performance of pre-trained text-to-image diffusion models for OOD generalization. However, despite advancements, studies indicate that large-scale text-generation models are still not as effective at handling OOD cases compared to identically distributed data [33, 48, 52]. Building on these developments, we leverage two distinct large-scale models trained on text and image modalities, harnessing their comprehensive world knowledge.

3 Methodology

The input to our model DIAGen is a small dataset, $R = \{R_n \mid 1, \ldots, N\}$ of N real images, containing only a few images per class. The output of the pipeline (see Fig. 2) is an expanded labelled dataset that includes both the real images as well as M corresponding synthetic images $S_{n,m}$ for each real image R_n. The goal is to augment the small given dataset in a semantically diverse way to enable the training of a downstream application with better generalization.

Before detailing our method to increase semantic diversity, we first lay out, how we define diversity. We focus on improving the intra-class diversity that

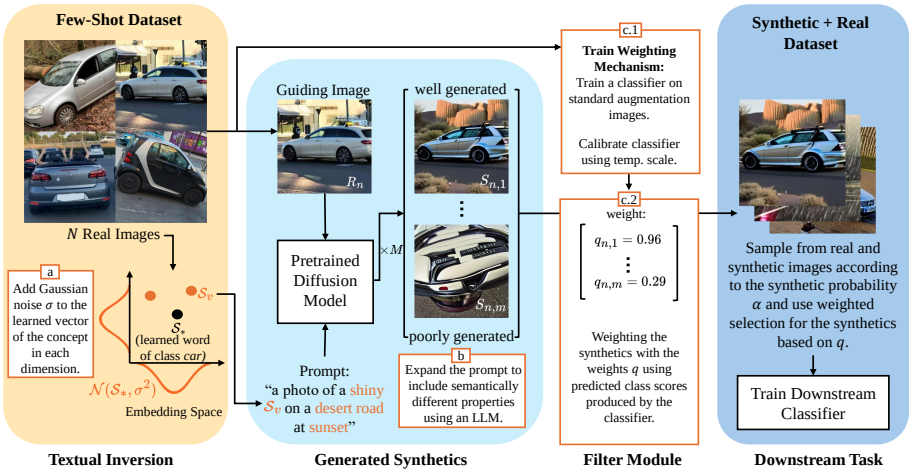

Fig. 2: DIAGen's image generation pipeline based on DA-Fusion [47]. Our contributions include: *a)* Varying the learned class concept in the embedding space by applying Gaussian noise. *b)* Using varied class-specific prompts generated by an LLM. *c)* Training and utilizing a classifier trained on real images as a weighting mechanism. All real images combined with the generated synthetic ones are then used to train an arbitrary downstream model. The ratio of real to synthetic images can be controlled by the synthetic probability hyperparameter α.

represents the variance within the data points of a class. We aim to improve two different aspects of diversity: First, we address the different semantically meaningful contexts in which the class can occur. Here we use the three categories of diverse settings as proposed by Kattakinda and Feizi [18], namely weather conditions, time of day, and environment. Second, we enhance the diversity of object appearance itself, *e.g.*, by changing the type of a motorcycle (*cf*. Fig. 1).

3.1 Embedding Noise

The diffusion model that DIAGen builds upon is conditioned on a text prompt [34] containing the learned pseudo word vector for a specific class. The word vector of a class, *e.g.*, *car* to give a concrete example, is learned with Textual Inversion [8] and the resulting embedding vector $\mathcal{S}_*$ is inserted into the prompt "a photo of a $\mathcal{S}_*$". Following Mikolov *et al.* [25], who observed that directions in embedding spaces represent semantic meaning, *e.g.*, $king - man + woman = queen$, and that vectors that are very close to each other also have very similar meaning, we propose adding noise on top of the learned class concept vectors (see Fig. 2, contribution a). We hypothesize that varying $\mathcal{S}_*$ of an object yields images of similar object types, since their representations are likely to be close together in the embedding space. This may result in scenarios where the representation of *oldtimer* is next to the embedding vector of our learned representation of *car*.

The generation of a noisy embedding vector $\mathcal{S}_v$ can be formulated as

$$\mathcal{S}_v = \mathcal{S}_* + \mathcal{N}(0, \sigma^2)\,, \tag{1}$$

where $\mathcal{S}_*$ is the original word embedding vector obtained with Textual Inversion [8] and $\mathcal{N}(0, \sigma^2)$ is a noise sample of the same dimensions from a Gaussian distribution with zero mean and variance σ^2. Further details on the visual impact of this noise and the selection of the hyperparameter σ^2 are in Appendix D.

3.2 LLM Prompting

To achieve more explicit control over image generation beyond simply adding noise to the class embedding, we utilise a large language model (LLM) to provide textual guidance for the diffusion model (see Fig. 2, contribution b). Specifically, we employ *GPT-4* [1], known for its robustness and extensive knowledge acquired from internet-scale data. We also tried the smaller model Llama2 (7B) [46] and observed a similar performance.

Due to the different functioning and training data of language and image models, the covered knowledge also differs. This is beneficial in scenarios where the diffusion model has rarely seen a concept and hence has no contextual knowledge of the concept. An LLM such as GPT-4 can provide additional meaningful context so that the resulting synthetic images exhibit high semantic diversity.

We instructed GPT-4 to dynamically generate a certain number of prompts in the following style:

> a photo of a ⟨adjective⟩ $\mathcal{S}_v$ ⟨location and/or weather preposition⟩ ⟨weather⟩ ⟨location⟩ ⟨time of day with preposition⟩

As mentioned earlier, $\mathcal{S}_v$ denotes the learned embedding vector of a *class* after adding noise, which can be treated as a new pseudo-word. Every placeholder enclosed in brackets is optional and may be completed by GPT-4 to generate prompts of varying lengths and complexity. For instance, the final prompt of *class: dog* could be "a photo of a ⟨fluffy⟩ $\mathcal{S}_v$", for *class: plane* "a photo of a $\mathcal{S}_v$ ⟨flying above a city at night⟩", and for *class: spoon* "a photo of an ⟨antique⟩ $\mathcal{S}_v$ ⟨on a wooden table⟩". For more details see Appendix E.

3.3 Weighting Mechanism

Similar to DA-Fusion [47], the extent to which the generated images can deviate from the guiding image is controlled by a strength hyperparameter t_0. This parameter, ranging from 0 to 1, relates to the time step of inserting the guiding image during the diffusion model's denoising process. When $t_0 \to 0$, the generated images closely resemble the guiding image. While increasing t_0 enhances the image diversity, this increased freedom also leads to a higher probability of generating synthetic images that do not match the intended class label, which can result in either distorted class representations or entirely unrelated concepts.

We select a higher value for t_0 (see Appendix B) than Trabucco *et al.* [47] to encourage diversity at the potential cost of class fidelity.

To address the issue of poorly matching synthetic images and thus increase the class fidelity, we implement a weighting mechanism (see Fig. 2, contribution c). This module operates by estimating a class confidence score q for each generated synthetic image using a classifier trained on the original data. To enhance the significance of these confidence scores, we apply temperature scaling [9], a well-established method for calibrating probabilistic models. The temperature T scales the logit output vector of the classifier $\mathbf{z}$ before calculating the softmax function $\mathbf{q} = \text{softmax}\left(\frac{\mathbf{z}}{T}\right)$. From the scaled class scores $\mathbf{q}$, we pick the entry $q \in [0, 1]$ corresponding to the class of the guiding image, which serves as confidence score. The value of T is optimized with respect to the cross-entropy loss on the validation set. This process does not affect the overall accuracy of the classifier but refines the confidence estimates.

Instead of using a binary threshold on the confidence score q and filtering out images whose confidence is below that threshold, we use a weighting scheme following Rebbapragada and Brodley [32], since this retains the full dataset and reduces the impact of filtering errors. Moreover, it avoids having to optimize the threshold as a sensitive hyperparameter.

The probability to select a specific real or synthetic image is defined as

$$P_{\text{real},n} = \frac{1}{N}(1-\alpha) \qquad \text{and} \qquad P_{\text{syn},n,m} = \frac{1}{N}\left(\alpha \frac{q_{n,m}}{\sum_{j=1}^{M} q_{n,j}}\right). \tag{2}$$

That is, each real image R_n with $n \in \{1, \ldots, N\}$ is chosen with probability $P_{\text{real},n}$ during training, where α is a hyperparameter called synthetic probability. A synthetic image $S_{n,m}$ with $m \in \{1, \ldots, M\}$, which was generated based on the real image R_n, is selected with probability $P_{\text{syn},n,m}$.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our method, four datasets were utilized. A consistent set of hyperparameters was used across all datasets to maintain the model's off-the-shelf property and to allow for direct comparison.

First, the *FOCUS* dataset [18] was chosen, which contains 21K images of 10 different classes in common and uncommon settings, altering the time of day, weather condition, and location. This broad distribution makes FOCUS a well-suited dataset for our experiments, enabling the evaluation of DIAGen's ability to reproduce the distribution of the data only knowing very few images per class.

Second, we test our model on the *MS COCO* dataset [20]. This dataset comprises common objects in context, which implies that objects occur in different settings and have a variety of appearances. This dataset allows for a direct comparison to the baseline DA-Fusion, as it was also used in Trabucco *et al.* [47].

Third, we introduce our own dataset, *Custom COCO*, which is based on a subset of 23 classes from MS COCO [20]. In contrast to MS COCO, our dataset ensures that each image contains only one selected class, making it more suitable for single-label classification. The primary motivation for creating an alternative to MS COCO is to address the common data leakage issues in publicly available datasets [47]. Large pre-trained generative models, such as Stable Diffusion [34] utilized by DIAGen, are likely to be trained on instances from benchmark datasets such as MS COCO. Therefore, the diffusion model may have already observed validation and test images. To mitigate this, our Custom COCO dataset is based on custom-collected images, ensuring that none of them exists elsewhere on the internet. Thus, guaranteeing that the model is exposed to entirely novel images, eliminating the risk of prior exposure during training.

Fourth, to better evaluate the diversity of synthetic images produced by DIAGen, we generated an additional test set *Uncommon Settings* for the same classes as in Custom COCO, however in uncommon settings. This test set aims to measure the ability to classify out-of-distribution (OOD) samples. Our test set includes 247 uncommon scenarios, like *a chair in space* or *a bicycle at the bottom of the sea*. These test images were collected by conducting internet searches using a number of unusual locations that we had previously compiled.

4.2 Experimental Setup

We compare DIAGen's results against two baselines: DA-Fusion was chosen as the first baseline since DIAGen is built upon this model. We use the original experimental setup of Trabucco *et al.* [47]. Secondly, we compare DIAGen to standard augmentations, given their widespread use for data augmentation tasks. For this, we used a combination of rotations, flips, scale adjustments, and crops. More details on the experimental setup including all values for the hyperparameters can be found in the supplemental material.

The model's effectiveness was evaluated on a downstream classifier, comparing its behaviour on four datasets: FOCUS [18], MS COCO [20], Custom COCO, and Uncommon Settings. The downstream classifier accuracy serves as the primary metric for our studies, following the work of Ravuri and Oriol [31].

To ensure relevance for few-shot learning, we trained on small, varying dataset sizes containing 2, 4, and 8 examples per class. Furthermore, to increase the reproducibility and reliability of our findings, we used 3 different seeds to alter the selection of the images in the training split and calculated the mean.

4.3 Classification Accuracy

Fig. 3 shows the downstream classifier accuracy for DIAGen, the baseline DA-Fusion, and standard augmentations. We plot the accuracy over different few-shot dataset sizes, limiting the size to 2, 4, and 8 examples for each class.

We observe a consistent improvement in validation and test accuracy, by as much as +5% points across the four datasets when compared to DA-Fusion. The gain of DIAGen against standard augmentations is even more evident, reaching

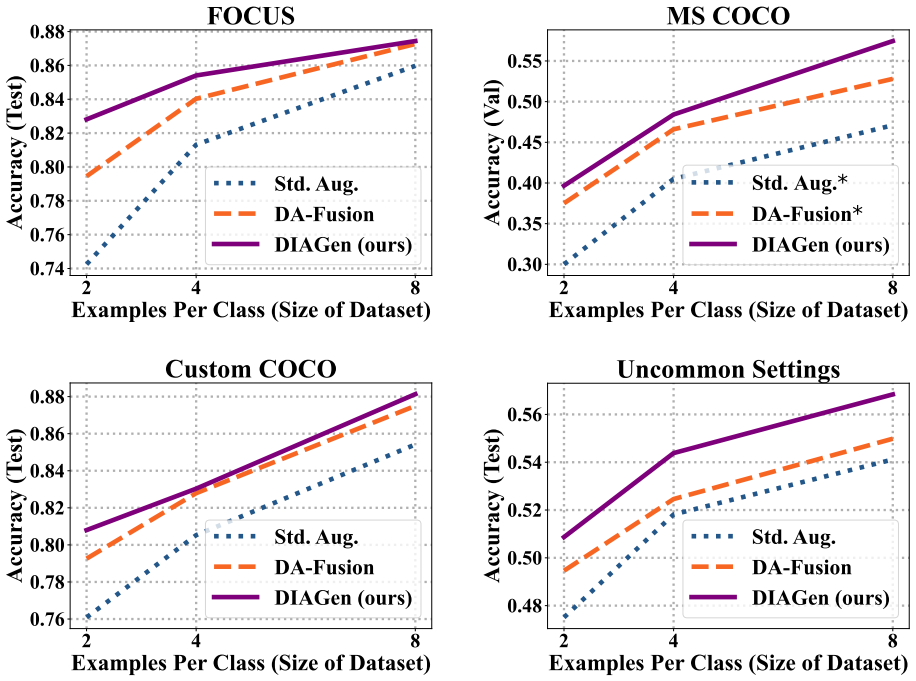

Fig. 3: Downstream classification accuracy of DIAGen, DA-Fusion [47], and standard augmentations on four datasets: *(a)* FOCUS, *(b)* MS COCO, *(c)* Custom COCO dataset, and *(d)* training on Custom COCO with evaluation on Uncommon Settings test set. Runs marked with * are taken from Trabucco *et al*. [47].

up to +10.5% points. These results highlight the effectiveness of DIAGen, especially in limited data scenarios. In few-shot learning situations where training examples are scarce, DIAGen introduces additional semantic diversity as we further analyse below, thereby strengthening the model's generalization ability.

By using Uncommon Settings, which targets edge cases of real-world object occurrences, we measure how effectively each method can cover a broad range of real-world scenarios. An analysis of the results on the Uncommon Setting test set (see Fig. 3, bottom-right) reveals a significantly higher accuracy of our DIAGen, with gains of approximately +2% points compared to DA-Fusion and +3% points compared to standard augmentations, across all dataset sizes. While the training dataset remains identical to Custom COCO, the test set now includes samples from a distribution entirely different from the training data. This supports the hypothesis that our augmentation technique improves semantic diversity, particularly in generalizing to edge cases and uncommon scenarios.

4.4 Ablation Study

We now analyze the contributions of the three components in our DIAGen pipeline: embedding noise, LLM prompts, and weighting mechanism. We conduct an ablation study by running each module independently to assess their individual impact. Fig. 4 illustrates the accuracy gains attributed to each component relative to the DA-Fusion baseline.

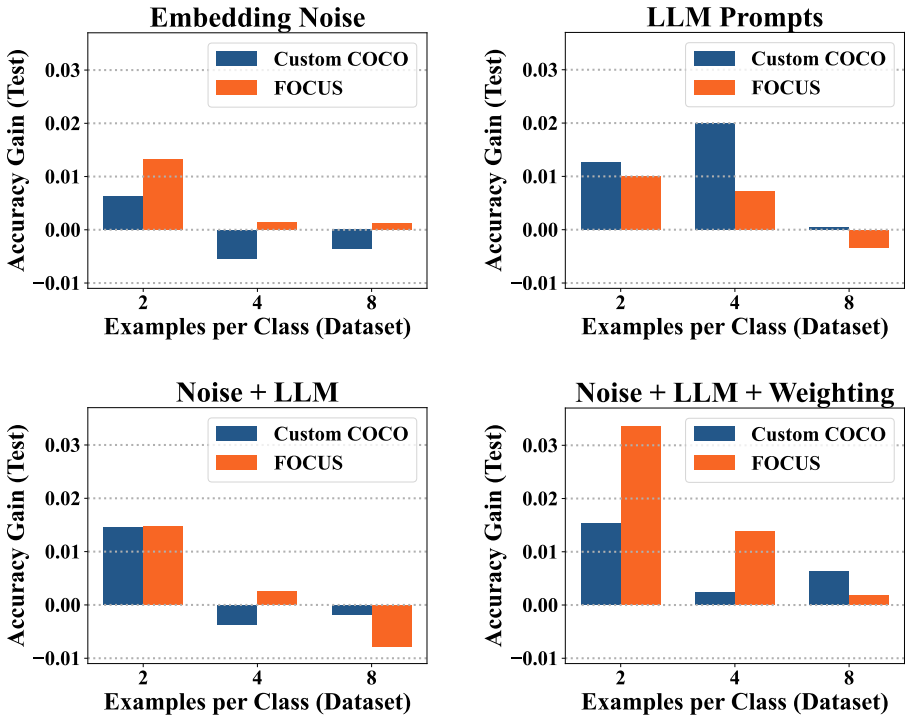

Fig. 4: Ablation study of the three proposed components, showcasing their distinct contribution to the classification accuracy. We illustrate the accuracy gains over DA-Fusion [47] solely utilizing embedding noise *(top left)*, employing only the LLM prompt module *(top right)*, and combining both *(bottom left)*. Also shown are the improvements by adding the weighting mechanism *(bottom right)*. Latter corresponds to the full DIAGen method.

Embedding noise leads to major improvements when only 2 examples per class are used for training. Although the positive effect of adding noise on its own decreases with more examples per class, its combination with the other components yields significant benefits. We attribute the synergy of the combined method to the ability of the embedding noise and LLM to increase diversity at the expense of class fidelity, a trade-off that the weighting mechanism mitigates by

assigning a lower weight to low-quality images. Weighting is effective in refining the augmentation process, as visualized by comparing its results with the runs only utilizing noise and LLM prompts in Fig. 4.

In contrast to embedding noise, using LLM prompts alone yields promising results, significantly improving the accuracy in case of 2 and 4 examples per class. Interestingly, although DIAGen proves effective across all tasks and dataset sizes, the use of LLM prompts alone outperformed the combined application in specific scenarios (4 examples per class with Custom COCO). This observation suggests that DIAGen holds the potential to achieve even better results through task-specific fine-tuning by activating different components of its pipeline. Our findings show that while each component can independently improve the accuracy, their true strength emerges in combination.

To further verify that the observed improvements in DIAGen are not merely due to hyperparameter adjustments (see Appendix C), we conducted an experiment directly comparing DIAGen with DA-Fusion using an identical set of hyperparameters (see Fig. 5). The results clearly show that DIAGen's performance gains stem from our contributions, rather than from changes in hyperparameters alone. In fact, relaxing the hyperparameters within the DA-Fusion model proves counterproductive, often resulting in reduced performance.

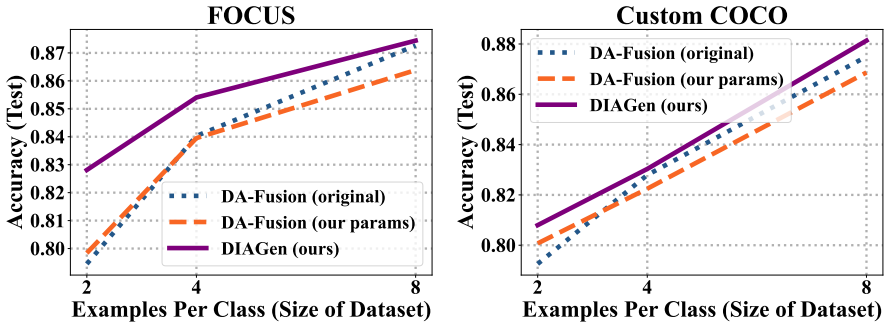

Fig. 5: Direct comparison of downstream classifier accuracy between DIAGen (ours) and the baseline method DA-Fusion (our parameters), using the same hyperparameters for a fair evaluation. For reference, the original DA-Fusion method with its parameters from Trabucco *et al.* [47] is also included.

4.5 Diversity Analysis

While it is important to evaluate the results of the downstream application, we also consider the overall quality and especially the semantic diversity of the synthetic dataset by exploring alternative metrics. If we visually compare the two datasets generated by DA-Fusion and DIAGen, we clearly observe a higher level of diversity with our method (see Fig. 6). At first glance, the DA-Fusion images

Fig. 6: Qualitative comparison of synthetic images generated by the DA-Fusion [47] baseline *(left)* and our model DIAGen *(right)*. Each row was created using the same guiding image.

all look very similar. Small changes can only be observed in fine textural details, such as the imprint on the bottle in Fig. 6 (left). DIAGen, on the other hand, achieves a higher degree of diversity in its images. As shown in Fig. 6(right), it generates varied cars, like an oldtimer and a silver car, in different styles and settings, whereas DA-Fusion produces nearly identical cars. DIAGen's images are both accurately labeled and semantically diverse.

Table 1: Averaged precision and recall [19] between the two distributions of the real images and the synthetic ones generated by the baseline DA-Fusion [47] respectively our model DIAGen. Training of the models was done using dataset sizes of 2, 4, and 8 examples per class for each of the three datasets.

		FOCUS			Custom COCO			MS COCO		
		(2)	(4)	(8)	(2)	(4)	(8)	(2)	(4)	(8)
DA-Fusion [47]	Prec.	98.67	**97.67**	93.67	**89.33**	**78.00**	**82.67**	**89.31**	**89.69**	**86.08**
DIAGen (ours)	(%)	**99.33**	96.00	**95.33**	71.33	60.67	58.00	83.75	81.56	81.03
DA-Fusion [47]	Rec.	8.00	9.33	7.00	20.67	21.67	47.33	3.69	8.03	15.19
DIAGen (ours)	(%)	**25.67**	**26.00**	**20.00**	**58.00**	**57.67**	**59.67**	**36.06**	**39.25**	**41.39**

To objectively quantify the diversity enhancement, we use the precision and recall metrics for the real and synthetic dataset distributions as defined by Kynkäänniemi *et al.* [19] (see Appendix F). When interpreting the results, it is important to consider the dataset sizes. Our Custom COCO dataset is rela-

tively small, with less than 50 images per class collected by us, leading to a high data bias. In contrast, the MS COCO and FOCUS datasets contain significantly more images per class. Notably, the FOCUS dataset was collected with an emphasis on including uncommon settings. As a result, the real image distributions are likely to differ significantly among these three datasets.

The results in Tab. 1 show a significant recall improvement, with up to a 37.3% increase across all datasets and training samples, reflecting greater image diversity. These results align with observed increases in image diversity. For precision, the FOCUS dataset shows only minor differences, indicating that DIAGen generates diverse yet class-consistent images. However, Custom COCO exhibits a notable precision drop. This can be attributed to the small size of the Custom COCO dataset, which does not adequately represent the real-world data distribution of its classes, as stated above. For instance, DIAGen produces an oldtimer as a *car*, which is a valid real-world representation of *cars*, but Custom COCO does not contain any oldtimer images. While Custom COCO addresses data leakage, its distribution is not fully representative of each class and this "good" sample is considered to be outside the real distribution, which lowers the precision score. For this reason, the precision score for Custom COCO has limited significance. This argumentation is backed by the precision results for MS COCO, where we observe a smaller drop in precision compared to DA-Fusion. This difference is attributed to the fidelity-diversity trade-off.

Overall, these results underline that DIAGen notably enhances diversity in synthetic images.

5 Conclusion

We introduced DIAGen, an off-the-shelf image augmentation technique designed to increase semantic diversity in datasets with few labeled examples per class. DIAGen expands the DA-Fusion framework [47] by incorporating three key components: The first two modules of DIAGen focus on increasing diversity in the augmentation process by *(i)* introducing noise to the class representations in the embedding space, and *(ii)* enriching text prompts with semantically meaningful content, leveraging the capabilities of an LLM. The last module is designed to complement these strategies to keep a high class fidelity by *(iii)* using a weighting mechanism to reduce the influence of suboptimal generated images using a classifier. These components help balance fidelity and diversity in synthesized images. The resulting model improves classification accuracy across various datasets and enhances recall as a diversity metric. It is particularly effective in enabling downstream models to generalize to uncommon scenarios and edge cases, making it valuable for augmenting data in few-shot settings.

Acknowledgements. This project is partially funded by the European Research Council (ERC) under the EU Horizon 2020 programme (grant agreement No. 866008) and the State of Hesse, Germany, through the cluster project "The Third Wave of Artificial Intelligence (3AI)".

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical report. arXiv:2303.08774 [cs.CL] (2023)
2. Alaa, A., Van Breugel, B., Saveliev, E.S., van der Schaar, M.: How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In: ICML. pp. 290–306 (2022)
3. Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks. arXiv:1711.04340 [stat.ML] (2017)
4. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
5. Besnier, V., Jain, H., Bursuc, A., Cord, M., Pérez, P.: This dataset does not exist: training models from generated images. In: ICASSP. pp. 1–5 (2020)
6. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. NeurIPS **34**, 8780–8794 (2021)
7. Friedman, D., Dieng, A.B.: The Vendi Score: A diversity evaluation metric for machine learning. TMLR (2023)
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
11. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? In: ICLR (2023)
12. Henriksson, J., Berger, C., Ursing, S.: Understanding the impact of edge cases from occluded pedestrians for ML systems. In: SEAA. pp. 316–325 (2021)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. NeurIPS **30** (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
15. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR **23**(47), 1–33 (2022)
16. Jahanian, A., Puig, X., Tian, Y., Isola, P.: Generative models as a data source for multiview representation learning. In: ICLR (2022)
17. Jiang, S., Zhu, Y., Liu, C., Song, X., Li, X., Min, W.: Dataset bias in few-shot image recognition. TPAMI **45**(1), 229–246 (2022)
18. Kattakinda, P., Feizi, S.: Focus: Familiar objects in common and uncommon settings. In: ICML. pp. 10825–10847 (2022)
19. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. NeurIPS **32** (2019)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
21. Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P.: Towards out-of-distribution generalization: A survey. arXiv:2108.13624 [cs.LG] (2021)

22. Man, K., Chahl, J.: A review of synthetic image data and its use in computer vision. Journal of Imaging **8**(11), 310 (2022)
23. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)
24. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: IIPhDW. pp. 117–122 (2018)
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. NeurIPS **26** (2013)
26. Murphy, K.P.: Probabilistic machine learning: Advanced topics. MIT press (2023)
27. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: ICML. pp. 7176–7185 (2020)
28. Nagarajan, V., Andreassen, A., Neyshabur, B.: Understanding the failure modes of out-of-distribution generalization. In: ICLR (2021)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171 (2021)
30. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621 [cs.CV] (2017)
31. Ravuri, S., Vinyals, O.: Classification accuracy score for conditional generative models. NeurIPS **32** (2019)
32. Rebbapragada, U., Brodley, C.E.: Class noise mitigation through instance weighting. In: ECML. pp. 708–715 (2007)
33. Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., Liu, P.J.: Out-of-distribution detection and selective generation for conditional language models. In: ICLR (2023)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
35. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH. pp. 1–10 (2022)
36. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. NeurIPS **31** (2018)
37. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. NeurIPS **29** (2016)
38. Shijie, J., Ping, W., Peiyi, J., Siping, H.: Research on data augmentation for image classification based on convolution neural networks. In: CAC. pp. 4165–4170 (2017)
39. Singh, K., Navaratnam, T., Holmer, J., Schaub-Meyer, S., Roth, S.: Is synthetic data all we need? Benchmarking the robustness of models trained with synthetic images. In: CVPR Workshops (2024)
40. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, E., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. NeurIPS **36** (2024)
41. Sudhakaran, G., Dhami, D.S., Kersting, K., Roth, S.: Vision relation transformer for unbiased scene graph generation. In: ICCV. pp. 21882–21893 (2023)
42. Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: Oasis: only adversarial supervision for semantic image synthesis. International Journal of Computer Vision pp. 2903–2923 (2022)
43. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)

44. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR. pp. 3716–3725 (2020)
45. Tong, J., Dai, L.: Out-of-distribution with text-to-image diffusion models. In: PRCV. pp. 276–288 (2023)
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv.2302.13971[cs.CL] (2023)
47. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: ICLR Workshop: Mathematical and Empirical Understanding of Foundation Models (2024)
48. Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., et al.: On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. In: ICLR Workshop: Trustworthy and Reliable Large-Scale Machine Learning Models (2023)
49. Wang, L., Zhang, S., Han, Z., Feng, Y., Wei, J., Mei, S.: Diversity measurement-based meta-learning for few-shot object detection of remote sensing images. In: IGARSS. pp. 3087–3090 (2022)
50. Xue, Y., Zhou, Q., Ye, J., Long, L.R., Antani, S., Cornwell, C., Xue, Z., Huang, X.: Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification. In: MICCAI. pp. 387–396 (2019)
51. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. arXiv:2110.11334 [cs.CV] (2021)
52. Yang, L., Song, Y., Ren, X., Lyu, C., Wang, Y., Liu, L., Wang, J., Foster, J., Zhang, Y.: Out-of-distribution generalization in text classification: Past, present, and future. arXiv:2305.14104 [cs.CL] (2023)
53. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: DatasetGAN: Efficient labeled data factory with minimal human effort. In: CVPR. pp. 10145–10155 (2021)

A Limitations and Future Work

DIAGen exploits the pre-trained knowledge of a diffusion model and an LLM. However, in some scenarios where these models have limited exposure to certain objects during training, this knowledge can be insufficient. For example, if the diffusion model has rarely or never encountered images of an oldtimer car, it may struggle to produce realistic images. Therefore, evaluating how well DIAGen performs on datasets containing rarely seen classes remains a future work direction. Additionally, testing DIAGen on datasets with more visually similar classes, to cover a different level of granularity, could provide further insights into its robustness.

Furthermore, our method relies on high quality training images, where objects are well captured in the scene. This is necessary to learn meaningful class embeddings with Textual Inversion. Ensuring this quality can be a constraint in certain scenarios. Also as described by Man *et al.* [21], a common problem of generating large synthetic datasets within reasonable time is that significant computational power is required. This is also true for our method, which may be limiting for some users.

Finally, the augmentation outputs produced by DIAGen, like those of DA-Fusion [46], are limited to image-label pairs, making them unsuitable for other tasks such as object detection or segmentation. Addressing this limitation and extending the method to support a broader range of tasks is a potential direction for future work.

B Extended Implementation Details

As an extension of the experimental setup described, we present additional technical details of the implementation. During the Textual Inversion process, following Trabucco *et al.* [46], the images are uniformly cropped and resized to a resolution of 512×512. A training batch size of 4 is used, and there are a total of 1000 optimization steps. The learning rate is set to $5 \cdot 10^{-4}$. We use *Compvis/stable-diffusion-v1-4* by Rombach *et al.* [33] as our pre-trained diffusion model.

In the generation process, $M = 10$ synthetic images are produced for each real image. For textual guidance we created 10 prompts for each class, ensuring that each prompt is used once for each guiding image. We are varying the number of used guiding images (2, 4, 8 examples per class) to simulate different dataset sizes.

As described in Sec. D, suitable variations in the noise level are 0.005, 0.01 and 0.025. We set the strength parameter t_0 to 0.7. A higher strength parameter results in a synthetic image that deviates further from the guiding image and, therefore, crucially controlling fidelity and diversity. For the parameter guidance scale, we use a value of 15. This value determines the conditioning of the diffusion model to the text prompt. We increased it to see a larger effect of our generated prompts in the synthetic images. It is recommended to avoid setting it too high as

it may result in surreal and noisy images. An ablation study on these important hyperparameters can be found in Sec. C.

In the downstream classifier training, following Trabucco *et al.* [46], we run 50 epochs with 200 iterations per epoch, presenting for each evaluated method the same total amount of training examples. We use a split of both, real and synthetic images as train data. The parameter controlling this split ratio, called the synthetic probability α, is set to 70% (higher than the 50% used by Trabucco *et al.* [46]) so that the training process can benefit more from the diversity introduced through the synthetic data. The model with the highest validation accuracy after training is then used for the evaluation on the test set. The downstream model's pre-trained backbone is *ResNet-50* [10]. We are only fine-tuning the last linear layer, as it was done by Trabucco *et al.* [46].

The experiments were conducted on PyTorch, using Nvidia RTX A6000 GPUs and took 280h (Textual Inversion) + 460h (image generation) + 160h (training downstream classifier) for all presented experiments on all datasets.

C Hyperparameter Ablation

The objective of our hyperparameter ablation study is to identify a set of hyperparameters that consistently yield robust results across all tested datasets and dataset sizes. We systematically evaluated different values for the strength t_0 (see Fig. 7) and guidance scale gs (see Fig. 8) to determine their impact on the performance and to select the best values.

The decision to select a specific hyperparameter configuration is based on the values of DA-Fusion [46] (in blue) and the idea of changing the hyperparameters in the direction that implies more diversity. From the tested variants, we choose the configuration $t_0 = 0.7$ and $gs = 15$, which represents a moderate deviation from the original DA-Fusion values and seems to achieve strong and robust results across all runs.

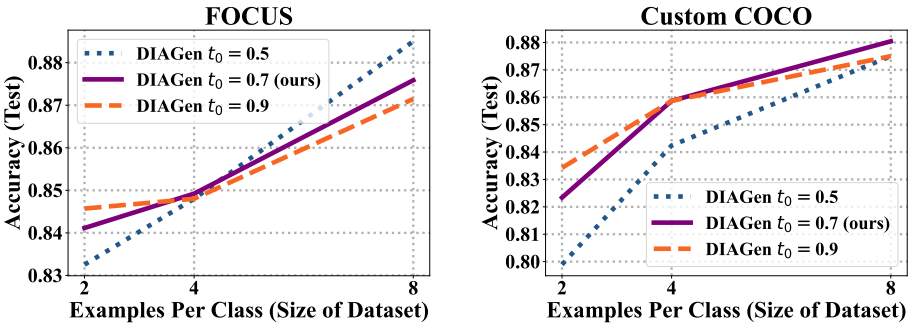

Fig. 7: Variations of the hyperparameter strength t_0. Higher values lead to images that deviate further from the guiding image ($t_0 = 0.7$ is selected for DIAGen).

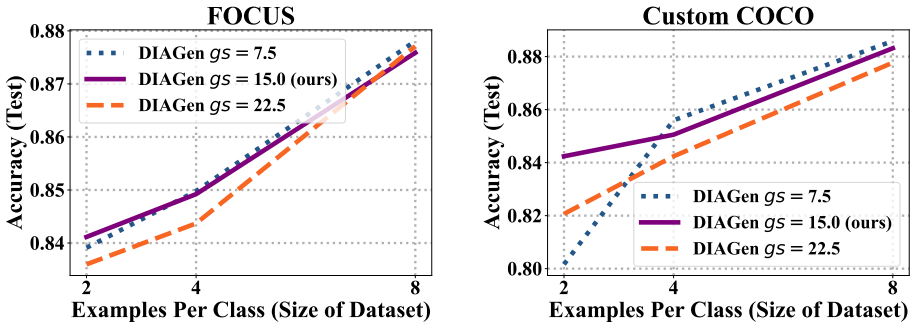

Fig. 8: Variations of the hyperparameter guidance scale, controlling the conditioning to the text prompt ($gs = 15$ is selected for DIAGen).

D Extended Noise Ablation

$\sigma^2 = 0$ $\sigma^2 = 0.005$ $\sigma^2 = 0.01$ $\sigma^2 = 0.025$

Fig. 9: Qualitative analysis presenting the effects of noise in the word embedding space. Gaussian noise with different variances σ^2 was added to the word vectors $\mathcal{S}_*$ of *car* and *clock*. The results show that semantic meaning can be influenced by small variations in the embedding space. As the noise level increases, noticeable changes occur, such as the shift in color of a car or the transformation of a clock from digital to analog.

The embedding noise module of DIAGen has a hyperparameter controlling the amount of Gaussian noise added to the word vectors (see Sec. 3.1). One challenge was to choose appropriate values for the variance σ^2, where a value

that is too low would not make enough difference to the original embedding $\mathcal{S}_*$ and a value that is too high would result in noisy embedding vectors $\mathcal{S}_v$ that no longer reflect the learned class concept.

Used are three qualitatively verified values $\sigma_0^2 = 0.005$, $\sigma_1^2 = 0.01$ and $\sigma_2^2 = 0.025$, which are in the acceptable range for all tested classes (see Fig. 9). We have chosen three values that alternate throughout the generation process in order to better analyze the influence of this hyperparameter. We observed a class-dependent upper boundary $\sigma^2 > 0.05$, where the variance is getting too high to maintain the correct class. For the final model we therefore suggest a value $\sigma^2 \in [0.01, 0.025]$, which can be further optimized. It may also be necessary to use different variances σ^2 for different classes. Overall, our observations show the sensitivity of the variance hyperparameter.

E Extended LLM Ablation

In the main paper we demonstrated the impact of the LLM contribution on the accuracy of the downstream classifier (see Sec. 3.2 and 4.4). To gain a deeper understanding of the influence of the class-specific prompts on the resulting synthetic image, we intend to conduct a more qualitative analysis.

The final instruction provided to *GPT4*, which was utilized to generate the results presented in the main paper, was as follows:

> Create prompts for me that have the following structure:
> "a photo of a [adjective] <classname> [location and/or weather preposition] [weather] [location] [time of day with preposition]"
> The <classname> is replaced with the actual classname, e.g. 'car'
> All the attributes in [..] are optionals. This means example prompts for car could be:
> 'a photo of a red car' (adjective optional)
> 'a photo of a car on a road' (location optional)
> 'a photo of a car in snow' (weather optional)
> 'a photo of a car at night' (time of day optional)
> 'a photo of a huge car in a tunnel' (adjective and location optionals)
> 'a photo of a green car on a foggy bridge at daytime' (all optionals)
> 'a photo of a car' (no optional)
> If you use adjectives, they should be visual. So don't use something like 'interesting'.
> Also vary the number of optionals that you use.
> Can you give me {num_prompts} prompts of this structure for class {name} please.

The {num_prompts} and {name} parameters are user-defined variables that specify the number of resulting class prompts and the class name included in the prompts. We varied the number of optional arguments as image quality sometimes improved with shorter prompts, while other times, longer, more descriptive prompts yielded better results. Fig. 10 illustrates some of this behaviour.

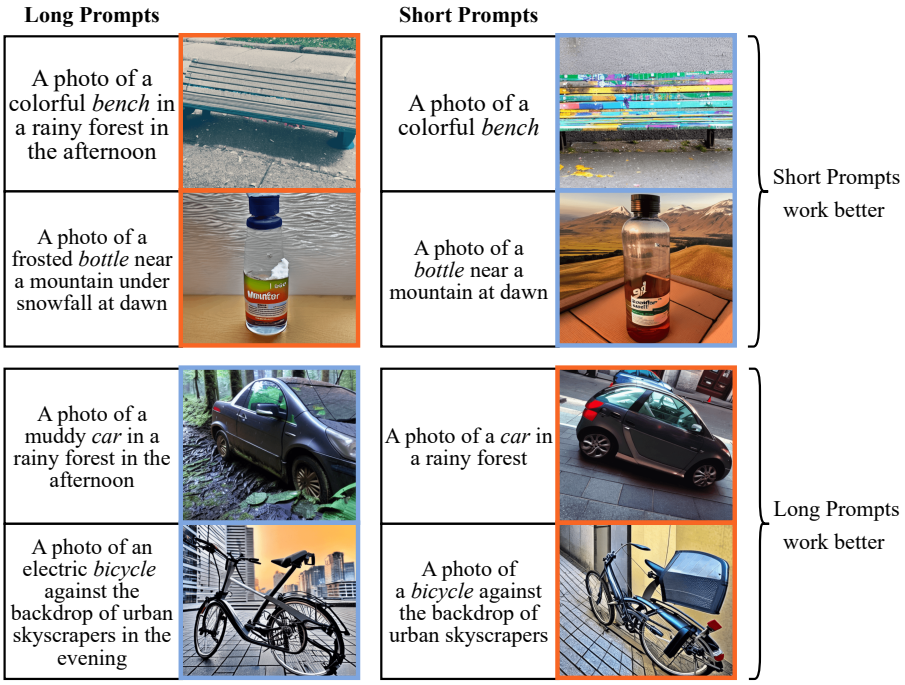

Fig. 10: Qualitative ablation of long and short class prompts to the diffusion model. Images from one row were generated using the same guiding image. In the top two rows the diffusion model was not able to integrate the content of the long prompts into the resulting images, while the aspects "colorful" and "near a mountain at dawn" can be identified in the synthetics generated with the shorter prompts. The opposite can be observed for the bottom two rows.

Depending on the guiding image, the corresponding synthetic images better represent the content from shorter or longer prompts. We observed this characteristic throughout the entire synthetic dataset which led to the decision to vary the number of optional arguments in the prompt.

To verify that our chosen instruction provided to the LLM leads to different prompt lengths, we compared the average word count of our LLM prompt and a modified version that omitted the phrase "also vary the number of optionals that you use" and the example prompts. The average class prompt resulting from our LLM instruction was 28.6% shorter. Since we use multiple prompts of different lengths for each guiding image, we increase the robustness to generate at least some synthetic images that reflect the content of the class prompt.

F Precision and Recall

In order to determine whether our efforts to enhance intra-class diversity have been successful, we use *precision and recall* of the distributions of the real and synthetic dataset (see Tab. 1). This metric was originally proposed by Sajjadi *et al.* [35] and improved by Kynkäänniemi *et al.* [18]. It constructs two independent manifolds for representations of real images $R_1, ..., R_N$ and synthetic images $S_{1,1}..., S_{N,M}$. *Improved precision and recall* is defined by

$$\text{precision} := \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}_{\text{manifold}(R_1,...,R_N)}(S_{n,m}) \tag{3}$$

$$\text{recall} := \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\text{manifold}(S_{1,1},...,S_{N,M})}(R_n) \tag{4}$$

where $\mathbb{I}_{(\cdot)}$ is the indicator function. For a given set A and an element y the indicator function is given as follows:

$$\mathbb{I}_A(y) := \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{if } y \notin A \end{cases} \tag{5}$$

The manifold of an image set $X_1, \ldots, X_P$ is defined by

$$\text{manifold}(X_1, \ldots, X_P) := \bigcup_{i=1}^{P} \text{Sph}(X_i, \text{NND}_k(X_i)), \tag{6}$$

where $\text{Sph}(x, r)$ denotes the sphere in $\mathbb{R}^D$ around datapoint x with radius r and D as the number of dimensions in the feature space. $\text{NDD}_k(X_i)$ expresses the distance from x to the k^{th} nearest neighbour (kNN) among $\{X_1 \ldots X_P\}$ excluding itself. According to Stein *et al.* [39] we use *DINOv2* to extract the features from the images and $k = 5$ for the kNN algorithm. Precision measures the similarity of generated instances to the real ones. Recall measures the generator's ability to replicate all instances from the real dataset. This enables to distinguish between fidelity (precision) and diversity (recall) of a synthetic dataset.