LANGUAGE-GUIDED DIFFUSION FOR DOMAIN GENERALIZATION

Haolin Ren[1], **Xinyi Li**[2], **Yancong Deng**[3]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]University of California, Davis, USA
[3]University of California, San Diego, USA
renhaolin22@mails.ucas.edu.cn

ABSTRACT

Domain generalization (DG) addresses the challenge of training machine learning models that generalize effectively to unseen target domains exhibiting distributional shifts. Traditional data augmentation techniques, while useful, often fail to adequately simulate the novel domain characteristics necessary for robust DG. We introduce a novel data augmentation framework leveraging the synergistic power of Large Language Models (LLMs) and diffusion models to generate diverse and realistic training data for DG. Our method employs LLMs to create creative prompts that encapsulate new domain styles, which are then used by diffusion models to synthesize high-fidelity images representative of these unseen domains. Furthermore, we integrate a CLIP-guided diversity analysis to ensure that the generated data effectively enhances model generalization while maintaining computational efficiency. Experiments on the PACS dataset show that our method significantly outperforms traditional techniques.

1 INTRODUCTION

Domain generalization (DG) aims to develop machine learning models that maintain robust performance when deployed in unseen target domains (Blanchard et al., 2011). The fundamental challenge lies in overcoming distribution shifts between training (source) and test (target) domains—a problem exacerbated by conventional data augmentation techniques that primarily generate in-domain variations through geometric transformations (Shorten & Khoshgoftaar, 2019) or style transfers. While these methods improve within-domain robustness, they fundamentally lack the capacity to simulate genuinely novel domain characteristics essential for true cross-domain generalization.

Recent advances in generative AI present new opportunities for addressing this limitation. Diffusion models (Rombach et al., 2022) have demonstrated unprecedented capabilities in generating high-fidelity images, while large language models (LLMs) (Achiam et al., 2023) offer sophisticated semantic understanding for controlled generation. However, the synergistic potential of these technologies remains underexplored for DG—existing approaches either apply diffusion models naively for in-domain augmentation (Sauer et al., 2023) (failing to induce cross-domain invariance), use LLMs for label-space augmentation (Ma et al., 2023) (neglecting visual domain characteristics), or lack systematic methods for generating *compositionally novel domains* that preserve semantic content. Our work bridges this gap through LLM-prompted diffusion generation that systematically produces novel domain variations. As illustrated in Figure 1, our method combines three key innovations:

1. LLM-guided domain space expansion through semantic-aware prompt engineering
2. CLIP-regularized diffusion generation ensuring visual-semantic consistency
3. Diversity-constrained augmentation scaling optimized for DG effectiveness

Extensive experiments across PACS, OfficeHome, and VLCS benchmarks demonstrate that supplementing merely 50% of training data with our generated samples improves state-of-the-art DG methods by up to 3.6% absolute accuracy.

2 METHODOLOGY

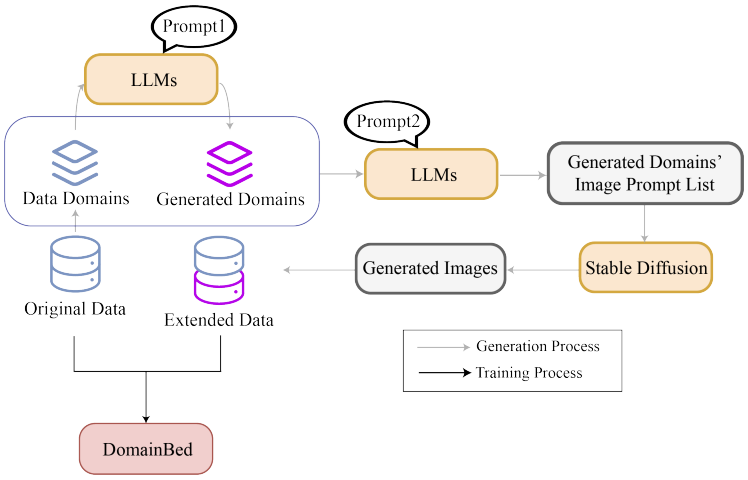

Figure 1: Overview of our proposed method for dataset augmentation using LLM and diffusion models. Scale: orig $\rightarrow$ 50% per domain, new: $\rightarrow$ 50% of orig mean. The 50% generation ratio was determined through a preliminary CLIP-based diversity analysis on PACS dataset and adopted as a practical guideline for all datasets.

We studied the Domain Generalization (DG) classification problem. We fine-tuned a pretrained model on data composed of multiple domains and evaluated it on data from unseen domains. More formally, we consider training domains $\mathbb{D} = \{\mathbb{D}_1, \ldots, \mathbb{D}_n\}$, where each domain's data source differs from the others. For example, domains may include sketches, cartoons, or photos. We used a Large Language Model (LLM) and diffusion models(Stable Diffusion) to generate new images based on the original dataset, including both domains present in the original dataset and new domains generated by the LLM, $\mathbb{D} = \{\mathbb{D}_1^{\text{orig+aug}}, \ldots, \mathbb{D}_n^{\text{orig+aug}}, \mathbb{D}_{n+1}^{\text{aug}}, \ldots, \mathbb{D}_{n+m}^{\text{aug}}\}$. Based on our extended training domains, we trained models and evaluated the trained models on test domains that were not expanded $\mathcal{D} = \{(\boldsymbol{X}_{\text{orig}}^{\mathbb{D}_1}, \boldsymbol{y}_{\text{orig}}^{\mathbb{D}_1}), \ldots, (\boldsymbol{X}_{\text{orig}}^{\mathbb{D}_n}, \boldsymbol{y}_{\text{orig}}^{\mathbb{D}_n})\}$. We test multiple generation methods to determine the best one. Our goal is to investigate the extent to which augmenting the original dataset with LLM and diffusion models improves the performance of models trained on the augmented dataset compared to those trained on the original dataset. This provides a novel data augmentation approach for future model training.

2.1 LARGE LANGUAGE MODELS FOR DOMAIN GENERATION

Large Language Models (LLMs) have been proven effective in various natural language processing tasks, including text generation, translation, and sentiment analysis. These models are trained on massive amounts of textual data and are capable of capturing complex patterns in language. In this work, we leverage the powerful text generation capabilities of LLMs to generate new domains beyond the original dataset's domains. We then use both the original domains and the newly generated domain categories to extend the original dataset, thereby enhancing it to improve the model's performance in classification tasks.

New Domain Generation To generate new domains, we inform the LLM of the existing domains in the dataset and then use a large language model (LLM) to synthesize new similar domains. Taking the PACS dataset as an example, which contains *photo*, *art painting*, *cartoon*, and *sketch* domains, we use LLM to generate other domains through prompts: `"My dataset currently has these domains {domains in original dataset}, help me generate a few more related domains."`

The newly generated domains were then integrated into the original dataset, thereby augmenting its diversity and enriching the breadth of visual representations available for subsequent analysis. The domain sets are defined as follows. The original domain set is $\mathbb{D}_{\text{orig}} =$

{photo, art_painting, cartoon, sketch}. Using the LLM, we generate a new domain set $\mathbb{D}_{\text{aug}} = \{\text{3D\_render, watercolor}\}$. These are combined to form the extended domain set $\mathbb{D}_{\text{extended}} = \{\text{photo, art\_painting, cartoon, sketch, 3D\_render, watercolor}\}$.

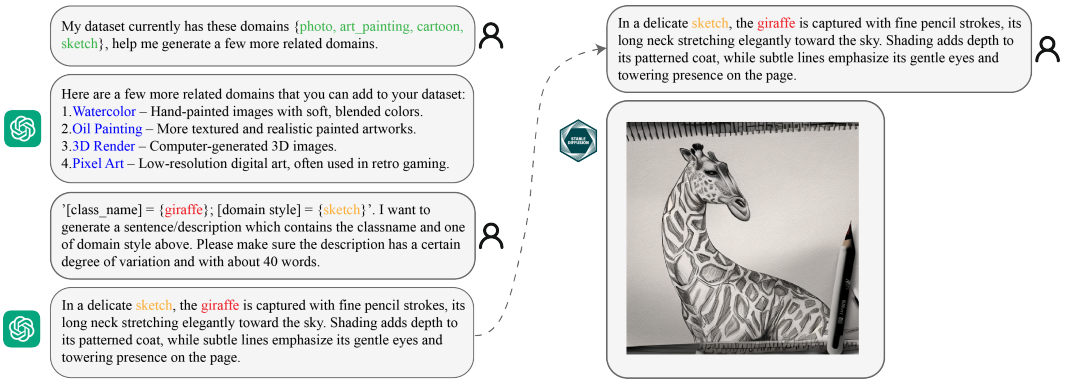

Figure 2: Illustration of the new domain generation, prompt generation and diffusion model for image generation.

Prompt Generation The process of prompt generation can be mathematically formulated as follows. Let the function $f : \mathbb{A} \to \mathbb{B}$ generate a prompt containing both the class name and the domain style:

$$\text{prompt} = f(\text{class\_name}, \text{domain})$$

where *class_name* is the given category and *domain* represents the style domain. This function f generates distinct prompts based on the inputs.

Given a set of classes $\mathbb{C}$ and domains $\mathbb{D}$, a prompt $p_{i,j}$ is generated for each class $c_i \in \mathbb{C}$ and domain $d_j \in \mathbb{D}$:

$$p_{i,j} = f(c_i, d_j)$$

In this paper, the function f can be decomposed into two components:

$$f = f_{\text{LLM}} \circ f_{\text{prompt}}$$

where:

1. f_{prompt}: The prompt provided to the large language model (LLM), which guides the generation process.
2. f_{LLM}: The large language model (LLM) itself, which processes the prompt and generates the output.

Thus, the overall function f combines the input prompt and the LLM's generation capabilities.

In our actual experiments, the function f—the prompt function—is a simple string concatenation function that concatenates the class and domain style together, which is then passed as input to the LLM. The output generated by the LLM is a sentence that contains both the class and domain style. This sentence can be used for image generation. The prompt designed for the LLM is as follows:

```
"'[class_name] = {class_name}; [domain style] = {domain}'. I want
to generate a sentence/description which contains the classname
and one of domain styles above. Please make sure the description
has a certain degree of variation and with about 40 words."
```

Using this prompt, we can generate a series of prompts that contain both the class name and one of the domain styles. These sentences can be used to generate images, thus augmenting the original dataset.

2.2 DIFFUSION MODEL FOR IMAGE GENERATION

Figure 3: Sample images from the augmented PACS dataset. The dataset consists of images of 8 domains, 4 of them were generated by the LLM and diffusion model. Original domains: {photo, art_painting, cartoon, sketch}. Augmented domains: {3D_render, cyberpunk, pixel art, watercolor}.

Using the prompts generated by LLM, we employ diffusion models to create high-quality synthetic images. Let the original dataset be $\mathbb{D}_{\text{orig}}$, and the augmented dataset $\mathbb{D}_{\text{aug}}$ is created by generating new images using the diffusion model:

$$\mathbb{D}_{\text{aug}} = \mathbb{D}_{\text{orig}} \cup \{g(f(c_i, d_j)) \mid c_i \in \mathbb{C}, d_j \in \mathbb{D}\}$$

where $g(\cdot)$ represents the image generation process using the diffusion model, and $f(c_i, d_j)$ is the prompt generation function.

2.3 CLIP-GUIDED DIVERSITY ANALYSIS

Using CLIP embeddings to evaluate image diversity across various generation scales, we found an optimal theoretical ratio of 64.96%. For practical efficiency, we adopted a 50% generation ratio across all datasets, which provides a good balance between diversity and computational cost while maintaining consistent performance improvements.

2.4 GENERATION METHOD

We test multiple generation methods to determine the best one. **o-Domain** only uses the original domains for data augmentation. **q-Domain** uses the original domains with an extended number of generated samples. **d-Domain** only uses the new domains generated by the LLM and diffusion models. **qd-Domain** combines both the original domains and new domains.

3 EXPERIMENTS

In this section, we provide implementation details using LLM and diffusion models, and present experiments using the domainbed experimental environment(Gulrajani & Lopez-Paz, 2020).

3.1 IMPLEMENTATION DETAILS

LLM Choice and Diffusion Model Choice We used the GPT-4o model as our LLM, which is known for its powerful text generation capabilities. For the diffusion model, we choose the stable diffusion model, following the specific settings of (Fan et al., 2024), which has been shown to be very effective in image generation tasks.

Prompt length The prompt length is an important hyperparameter that can significantly impact the quality of the generated images. A longer prompt can provide more detailed information to the LLM. In order to make the prompt words carry as much information as possible, we chose a prompt word length of at least 40 words.

CLIP-based Diversity Analysis and Optimal Generation Scale To evaluate the diversity introduced by the generated images and determine the optimal number of synthetic images, we employed the pretrained CLIP model (ViT-B/32) on 400M image–text pairs. For each generation scale $s \in \{10\%, 20\%, \ldots, 100\%\}$, we generated $s \times |\mathbb{D}_{\text{orig}}|$ synthetic images. We then extracted 512-dimensional image embeddings using CLIP's vision encoder and computed the intra-domain variance as:

$$\text{Score}(s) = \frac{1}{|\mathbb{D}|} \sum_{d \in \mathbb{D}} \text{Var}_d,$$

where $\mathbb{D}$ is the set of all domains and Var_d is the variance of embeddings within domain d. Figure 4 depicts the relationship between the generation scale and the diversity scores.

A cubic polynomial fit was applied to the obtained scores, leading to a theoretical optimum of approximately 64.96% for the generation scale. However, since the diversity gains tend to plateau beyond 50% and to reduce computational cost, we adopted a final generation scale of 50% in our experiments.

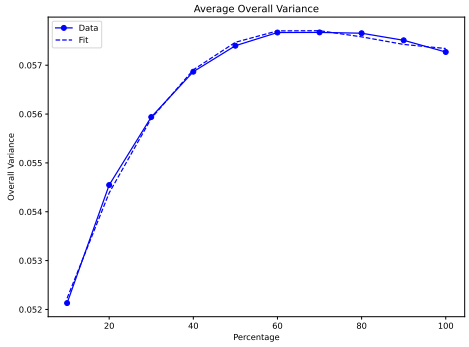

Figure 4: CLIP-based diversity analysis results showing the relationship between generation scale and intra-domain diversity. The blue dots represent measured diversity scores, while the orange line shows the cubic polynomial fit.

3.2 DomainBed Experiments

Table 1: Model Performance on Different PACS Variants with Averages

Algorithm	oPACS	qPACS	dPACS	qdPACS
ERM	84.4	84.6	**88.5**	87.5
IRM	83.5	86.2	**88.1**	86.5
Mixup	86.8	**88.0**	86.6	87.8
MLDG	79.2	76.9	77.6	**80.0**
CORAL	84.2	87.0	86.6	**88.7**
MTL	85.2	86.9	86.0	**88.4**
SagNet	83.2	84.4	**85.8**	85.0
SelfReg	81.9	85.0	**87.4**	86.3
Avg	83.6	84.9	85.8	**86.3**

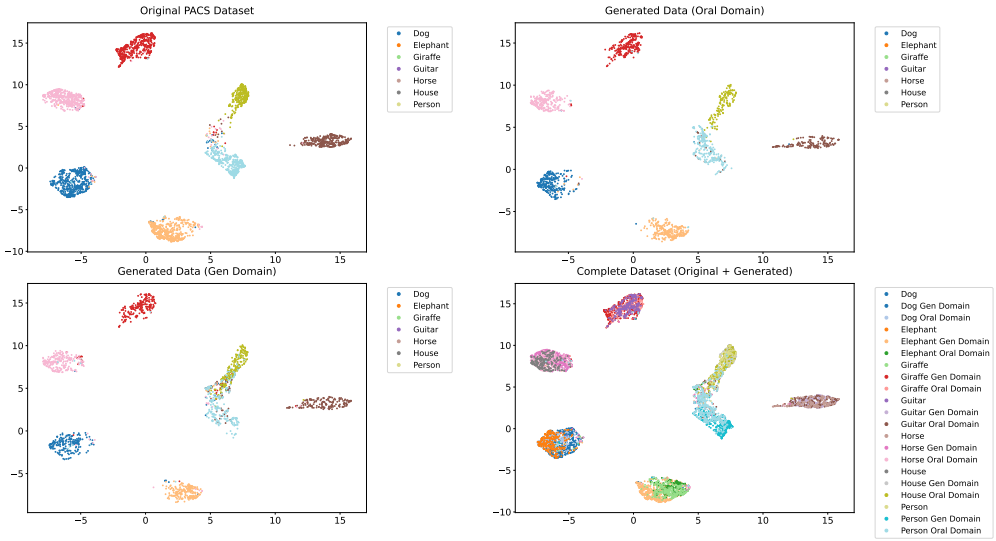

Figure 5: UMAP visualization of the PACS dataset features. Top left: Original PACS dataset. Top right: Generated data with oral domain. Bottom left: Generated data with gen domain. Bottom right: Complete dataset combining original and generated data.

Table 2: Average Performance of o-Domain and qd-Domain Methods Across Datasets

Algorithm	**o-Domain**			**qd-Domain**		
	OH	**VLCS**	**PACS**	**OH**	**VLCS**	**PACS**
ERM	70.3	77.8	84.4	**73.9**	**78.3**	**87.5**
IRM	58.1	78.8	83.5	**66.5**	**79.4**	**86.5**
Mixup	69.2	**77.8**	86.8	**72.4**	76.6	**87.8**
MLDG	56.0	**71.5**	79.2	**59.5**	70.1	**80.0**
CORAL	70.3	75.9	84.2	**71.3**	**77.2**	**88.7**
MTL	68.8	76.8	85.2	**72.2**	**78.5**	**88.4**
SagNet	68.7	**78.0**	83.2	**71.5**	76.1	**85.0**
SelfReg	70.7	76.9	81.9	**73.4**	**78.7**	**86.3**
Avg	66.5	76.7	83.6	**70.1**	**76.9**	**86.3**

Note: OH stands for OfficeHome dataset. Best results between o-Domain and qd-Domain are in bold.

We evaluate the performance of our methods on a well-known DG benchmark Domainbed(Gulrajani & Lopez-Paz, 2020). For fair comparison, we reuse the training and evaluation protocols in DomainBed, including dataset splits, training iterations, and model selection criteria. Our evaluation employs the training-domain validation set. The final data augmentation method is selected based on its combined accuracy on the validation set of all training domains.

It can be seen from Table 1 that, all data enhancement methods are better than the original data set, among which the qd-Domain method performs best, reaching an accuracy of 86.3%. This shows that our data augmentation method achieves significant improvement on the domain generalization task.

Furthermore, we apply the qd-Domain method to the OfficeHome and VLCS datasets to validate the performance of our approach across different datasets. From Table 2, we can see that the qd-Domain method performs better than the o-Domain method on all datasets. This demonstrates the effectiveness of our data augmentation method on different datasets.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Domain generalization for object recognition with multi-task autoencoders. *ICCV*, 2011.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training ... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, June 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2020. URL https://api.semanticscholar.org/CorpusID:220347682.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *ArXiv*, abs/1910.09217, 2019. URL https://api.semanticscholar.org/CorpusID:204800400.

Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, and Yun Fu. Ecacl: A holistic framework for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8578–8587, 2021.

Chang Liu, Lichen Wang, Kai Li, and Yun Fu. Domain generalization via feature variation decorrelation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1683–1691, 2021.

Chang Liu, Xiang Yu, Yi-Hsuan Tsai, Masoud Faraki, Ramin Moslemi, Manmohan Chandraker, and Yun Fu. Learning to learn across diverse data biases in deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4072–4082, 2022.

Chang Liu, Gaurav Mittal, Nikolaos Karianakis, Victor Fragoso, Ye Yu, Yun Fu, and Mei Chen. Hyperstar: Task-aware hyperparameter recommendation for training and compression. *International Journal of Computer Vision*, pp. 1–15, 2023. doi: 10.1007/s11263-023-01961-0. URL https://doi.org/10.1007/s11263-023-01961-0.

Wei Ma, Xiaohao Liu, and Chenyang Zhao. Large language models as data augmenters for cold-start item recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. A survey on diffusion models for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *JBD*, 2019.

Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998. ISBN 978-0-471-03003-4.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *ArXiv*, abs/2001.00677, 2020. URL https://api.semanticscholar.org/CorpusID:209832406.

Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19510–19520, June 2024.

A Related Work

Our work bridges domain generalization, diffusion-based generative models, and long-tail recognition.

A.1 Domain Generalization

Domain generalization (DG) aims to build models that perform well on unseen domains by learning domain-invariant features. Various approaches have been proposed, including Empirical Risk Minimization (ERM) (Vapnik, 1998), Interdomain Mixup (Mixup) (Yan et al., 2020), Invariant representation learning (Liu et al., 2021; Li et al., 2021), and Meta-learning (Liu et al., 2023). However, most methods struggle to generalize across diverse settings, motivating our use of diffusion models to generate domain-agnostic samples for improved generalization.

A.2 Diffusion Models

Diffusion-based generative models, such as Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), have gained attention for their ability to generate high-quality, diverse data. Unlike GANs, diffusion models offer more stable training and are effective at generating realistic images. Recent works (Dhariwal & Nichol, 2021) have applied diffusion models in various tasks, including image synthesis and augmentation. In this work, we employ diffusion models to enhance generalization by generating diverse samples across domains.

A.3 Long-Tail Recognition

Long-tail recognition addresses class imbalances where few-shot categories are often underrepresented. Leveraging synthetic data has shown promise in improving tail-class performance (Kang et al., 2019; Liu et al., 2022). Zhao et al. (Zhao et al., 2024) recently used LLM-driven synthetic data to address long-tail issues, leading to performance improvements. Inspired by this, we use diffusion models to generate balanced data, especially for underrepresented domains.

Our approach uniquely combines diffusion-based data generation with domain generalization and long-tail classification, targeting improved performance on both unseen domains and imbalanced datasets.