Attributed Synthetic Data Generation for Zero-shot Image Classification

Shijian Wang[1]*, Linxin Song[2,3]*, Ryotaro Shimizu[2,3,4], Masayuki Goto[2], Hanqian Wu[1]
[1]Southeast University, China, [2]Waseda University, Japan
[3]ZOZO Research, Japan, [4]University of California San Diego, USA
shijian@seu.edu.cn, songlx.imse.gt@ruri.waseda.jp
ryotaro.shimizu@zozo.com, masagoto@waseda.jp, hanqian@seu.edu.cn

Abstract

*Zero-shot image classification is a challenging task aiming to classify real images without real training examples. Recent research has employed synthetic training images generated by text-to-image models to address the challenge. However, existing approaches heavily rely on simplistic prompt strategies, which limit the diversity of the synthetic images. In this paper, we propose **AttrSyn**, which leverages large language models to obtain attributed prompts. These prompts allow for the generation of more diverse attributed images (e.g., specifying attributes such as style and background). By conducting experiments on two fine-grained datasets, we demonstrate that AttrSyn significantly outperforms simple base prompts, regardless of the visual encoder and classifier settings.*

1. Introduction

Data scarcity poses a significant challenge in the field of image classification [3, 8, 14, 17], as the scarcity of labeled data hinders the development of robust image classification systems. Zero-shot image classification [7, 9, 13, 21], which refers to classifying real images without having access to real training examples, emerges as a crucial technique to this dilemma.

Since breakthroughs in text-to-image models, especially diffusion models [22], have enabled the generation of a vast number of high-quality synthetic images, recent works [4–6] are exploring the potential of employing synthetic images to tackle the challenges associated with zero-shot image classification. However, the majority of these efforts primarily focus on how synthetic images are utilized during the training stage, with relatively less attention paid to exploring the generation stage of synthetic images. Most relevant research [1, 2, 11, 16, 23] relies on employing simple class-conditional prompts for text-to-image models to generate synthetic images, which inherently limits the diversity of the synthetic images produced.

*These authors contributed equally to this work.

Diversity is critical to reducing the gap between synthetic and real images when used as training data. [4, 18]. Inspired by [25], which introduces an attribute-based text generation approach to enhance text classification, we propose **AttrSyn**, designed to generate synthetic images that encompass a greater degree of diversity. Specifically, with the assistance of large language models (LLMs), we obtain the attribute dimensions and their corresponding candidate values for a given dataset in an interactive semi-automatic manner. Following this, we randomly combine these attributes with the associated class name to create attributed prompts. These prompts are then fed into a text-to-image model, such as Stable Diffusion [15], to generate attributed synthetic images that boast a high level of diversity.

To evaluate the effectiveness of our AttrSyn method, we train classifiers with synthetic images and test their performances on two fine-grained image classification datasets. These datasets are particularly challenging for zero-shot classification. To further demonstrate the robustness of our method, we conduct a series of experiments under various vision encoder and classifier settings. The experimental results show that AttrSyn consistently outperforms the simple base prompt across all settings, yielding performance enhancements ranging up to a maximum of **9.33**.

We summarize our contributions as follows:

- We propose **AttrSyn**, a novel attributed synthetic image generation method to facilitate zero-shot image classification tasks.
- We highlight the significance of focusing on the upstream generation process of synthetic images.
- We demonstrate the superiority of **AttrSyn** over simple base prompts through experiments on two fine-grained datasets, highlighting its potential for improving zero-shot image classification performance.

2. Method

In this section, we present the details of AttrSyn, which leverages attribute-based synthetic data generation to achieve zero-shot image classification. Specifically, At-

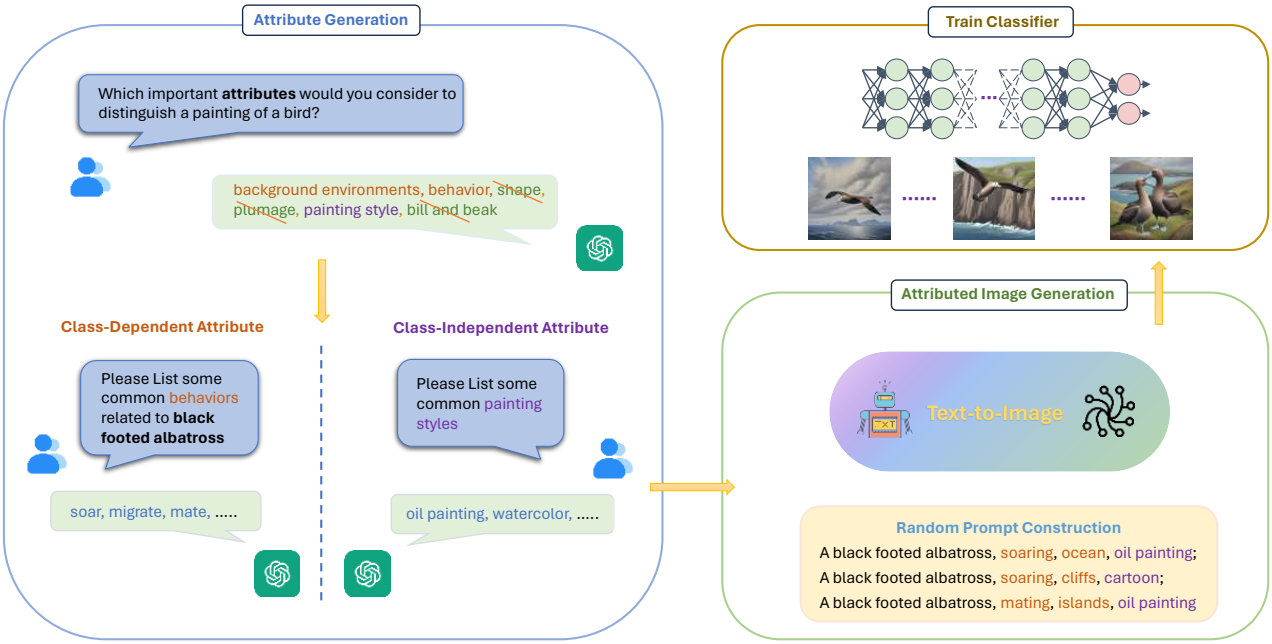


Figure 1. Overall workflow of AttrSyn.

trSyn employs Large Language Models (LLMs) to produce various and plausible attributes that enhance the prompts fed to the text-to-image model. These enhanced prompts enable more effective synthetic data for zero-shot image classification. The overall workflow is shown in Figure 1.

2.1. Attribute Generation

To obtain high-quality and diverse prompts, we adopt a semi-automated human-machine collaborative attribute generation approach inspired by [25]. The method initiates by defining several attribute dimensions for a given dataset. Since the process demands comprehensive knowledge of the dataset, we employ LLM to assist in initially determining these attribute dimensions. A human expert then interactively chooses the high-quality attribute dimensions that are suitable for the dataset. For example, in the case of CUB-200-Painting [20], a dataset comprising 200 distinct bird paintings, we employ "*Which important attributes would you consider to distinguish a painting of a bird?*" to query LLM and obtain "*background environments, behavior, shape, plumage, painting style, bill and beak*" as the response. The human expert then interactively selects "*background environments, behavior, painting style*" as the final attributes. To further query fine-grained diverse attributes, we query LLM to get diverse attribute values for each attribute dimension. We categorize attribute dimensions into class-dependent attributes and class-independent attributes and apply different query strategies.

Class-dependent attributes exhibit diverse attribute values among different classes, such as the "*background environment*" attribute for CUB-200-Painting. Class-independent attributes, such as "*painting style*" for CUB-200-Painting, can share the same values across all classes.

Class-Dependent Attributes. To prevent generating images that deviate from reality, it is crucial to ensure a strong correlation between class-dependent attribute values and their corresponding classes. For class-dependent attributes, we employ prompts, including the class name, to query LLM. Specifically, for the *black-footed albatross* class in CUB-200-Painting, we query LLM with the prompt "*Please list some common background environments related to black-footed albatross*" to get appropriate *background environment* values.

Class-Independent Attributes. To acquire class-independent attribute values applicable across various classes, we use a generic prompt to query LLM. For example, we use "*Please list some common painting styles*" as the prompt to get various potential painting styles.

2.2. Attributed Image Generation

After obtaining the attributes, it is important to prompt a text-to-image generative model efficiently to get attributed images that exhibit diversity across multiple attribute dimensions. Therefore we randomly sample values from each

Dataset	Domain	Task	# Train	# Test	Class
CUB-200-2011	Photo	Multi-class	5,994	5,794	200
CUB-200-Painting	Painting	Multi-class	—	3,047	200

Table 1. Statistics of datasets.

Dataset	# configurations / class	Attribute	# attribute value / class
CUB-200-2011	125	behavior	5
		background environment	5
		photo style	5
CUB-200-Painting	125	behavior	5
		background environment	5
		painting style	5

Table 2. Attribute dimensions for two datasets.

attribute dimension and concatenate them with the class name into attributed prompts. We then feed these prompts to the text-to-image model like Stable Diffusion to obtain fine-grained attributed images. For example, for the *black-footed albatross* class in the CUB-200-Painting dataset, one of its random attribute configurations is { "*behavior*"="*soaring*", "*background environment*"="*open ocean*", "*painting style*"="*oil painting*" }. We concatenate the class name and attribute values and use them to query the LLM. Subsequently, we use these attributed synthetic images to train an image classifier. AttrSyn greatly promotes the diversity of synthetic images, thereby achieving better zero-shot classification performance without real-world images available during the training stage.

3. Experiment

3.1. Dataset

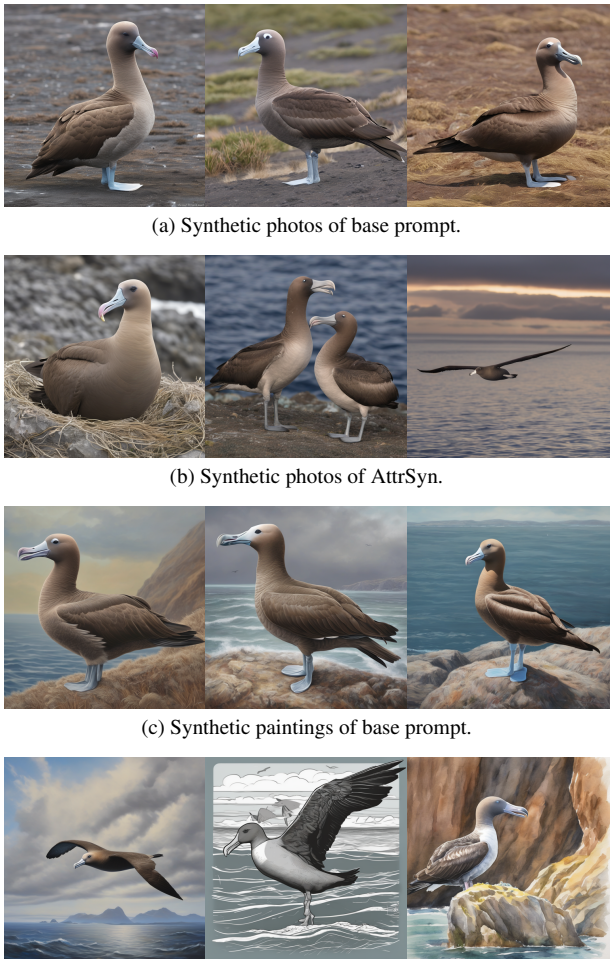
The challenge posed by fine-grained image classification within the zero-shot setting is notably difficult. To verify the effect of AttrSyn, we consider the following two fine-grained image classification datasets.

- **CUB-200-2011** [19]: The CUB-200-2011 dataset contains 200 different categories of bird photos, capturing variations in appearance, pose, and background. Each image is associated with a corresponding class label.
- **CUB-200-Photo** [20]: The CUB-200-Painting dataset contains 200 different categories of bird paintings, and these categories are consistent with CUB-200-2011.

We summarize the statistics of used datasets in Table 1, from which we can see that the train set size of CUB-200-2011 is nearly 6,000 and CUB-200-Painting doesn't split the train set. Thus we generate **6,000** training images for both datasets.

3.2. Images Generation

Text-to-Image Model. Considering the quality of synthetic images and the alignment with prompts, we choose *stable-diffusion-xl-base-1.0* [12] as the text-to-image model to generate images.



(a) Synthetic photos of base prompt.

(b) Synthetic photos of AttrSyn.

(c) Synthetic paintings of base prompt.

(d) Synthetic paintings of AttrSyn.

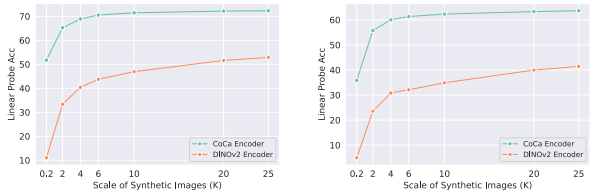
Figure 2. Visualization for the black-footed albatross class.

Base Prompt. [18] introduced a bag of tricks aimed at enhancing the diversity of synthetic images. In practice, we adopt the core tricks of their method as our baseline, which includes the domain and class name into the prompt, called the base prompt. Specifically, for CUB-200-2011, our base prompt template is "*a {class name} bird, photo*", and we use "*a {class name} bird, painting*" as the base prompt for CUB-200-Painting.

AttrSyn. We employ LLM to produce the corresponding attribute dimensions and attribute values for the two datasets, then filter them with human-machine collaborative strategies. The high-quality attribute dimensions that we obtained are shown in Table 2, from which we can see that each class of these two datasets has 125 different attribute prompt configurations.

Classifier	Method	CUB-Photo		CUB-Painting	
		CoCa	DINOv2	CoCa	DINOv2
LR	Base Prompt	69.68	40.08	58.42	27.24
	AttrSyn	**70.56** ↑0.88	**43.75** ↑3.67	**61.34** ↑2.92	32.10 ↑4.86
SVM	Base Prompt	67.81	27.99	53.27	16.48
	AttrSyn	69.85 ↑2.04	29.01 ↑1.02	56.06 ↑2.79	18.21 ↑1.73
MLP	Base Prompt	68.88	39.11	59.14	23.33
	AttrSyn	69.58 ↑0.70	41.30 ↑2.19	61.04 ↑1.90	**32.66** ↑9.33

Table 3. Performances of classifiers trained with the synthetic datasets. We present the performance gain compared to the base prompt set in green.



(a) CUB-Photo Scale (b) CUB-Painting Scale

Figure 3. Comparisons on scales of synthetic images.

Synthetic Images. To match the scale of the original training setting and maintain class balance, we generate 30 images per class for both datasets. To qualitatively compare the performance of AttrSyn and the base prompt, we sample images of *black-footed albatross* from the two datasets under the two methods, respectively. The visualization results are in Figure 2, from which we can see that our AttrSyn method shows significantly higher diversity than the base prompt for both two datasets.

3.3. Training with Synthetic Images

To quantitatively evaluate the effect of synthetic images on zero-shot image classification, we test the model performance trained with them. Specifically, we use a vision encoder pre-trained on extensive images to extract the vision embeddings. Subsequently, we train a classifier based on these embeddings to test.

Vision Encoder. To evaluate the effect of our proposed method across various vision encoders, we conducted experiments utilizing two cutting-edge vision encoders: CoCa [24] and DinoV2 [10].[1]

Classifier. To demonstrate the robustness of our method, we use three different classifiers: Logistic Regression (LR), SVM, and MLP in the experiments and compare their effects. For MLP, when we use CoCa as the vision encoder, we use a 3-layer MLP with an input dimension of 768, an output dimension of 200, and a hidden layer dimension of 512. When employing DinoV2 as the vision encoder, we use a 4-layer MLP with an input dimension of 1,536, an output dimension of 200, and hidden layer dimensions of 768 and 512. We train the MLP with a constant learning rate of 5×10^{-4} for up to 400 epochs with an early stopping strategy and split the synthetic images in an 8:2 ratio for training and validation.

[1]The specific checkpoint is *CoCa-ViT-L-14-laion2B-s13B-b90k* and *DinoV2-gaint*.

3.4. Experimental Results

Main Results. The performances of classifiers trained on synthetic images generated by the base prompt and AttrSyn are shown in Table 3, from which we can see that for both datasets, AttrSyn outperforms the base prompt under any vision encoder and classifier settings, yielding a performance increment ranging from 0.70 to 9.33.

Impact of Different Vision Encoders. It can be seen from Table 3 that in our experiments, employing CoCa as the vision encoder has higher performance than DinoV2. We hypothesize that the CoCa checkpoint pre-trained on a larger image dataset enables more generalized vision features, and can perform better on synthetic images.

Impact of Synthetic Data Scales. To evaluate the impact of varying scales of synthetic training data on test performance, we generated synthetic images in quantities of 200, 2k, 4k, 6k, and 25k for both CUB-200-2011 and CUB-200-Painting datasets. Subsequently, we evaluate the performance of logistic regression using CoCa and DINOv2 as encoders, respectively. The experimental result curve is shown in Figure 3, from which we can see that as the scale of synthetic training images increases, the test performances improve but eventually reach a plateau. In the future, we consider it significant to investigate efficient selection strategies of synthetic training images that can effectively complement the AttrSyn.

4. Conclusion

In conclusion, we explore the generation of synthetic images that are more effective during the training stage to facilitate zero-shot image classification. We underscore the importance of focusing on the data generation stage and introduce AttrSyn, a novel synthetic image generation method that leverages large language models to generate attributed prompts. These prompts enable the generation of attributed images with greater diversity. The experiments conducted on two fine-grained image classification datasets demonstrate the effectiveness of our method across various configurations of vision encoders and classifiers, highlighting its potential in zero-shot image classification tasks. Furthermore, experiments on the scaling of synthetic images reveal the significance of exploring strategies for synthetic image selection that complement AttrSyn effectively.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 126:961–972, 2018. 1

[2] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850, 2019. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*, 2023. 1

[5] Shreyank N Gowda. Synthetic sample selection for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 58–67, 2023.

[6] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1

[7] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4525–4534, 2017. 1

[8] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[9] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 1

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[11] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE international conference on computer vision*, pages 1278–1286, 2015. 1

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[13] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 1

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[16] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1

[17] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[18] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 769–778, 2023. 1, 3

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds200-2011 dataset. http://www.vision.caltech.edu/visipedia/CUB-200-2011.html, 2011. 3

[20] Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Progressive adversarial networks for fine-grained domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9213–9222, 2020. 2, 3

[21] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019. 1

[22] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 1

[23] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *2022 international conference on robotics and automation (ICRA)*, pages 6496–6503. IEEE, 2022. 1

[24] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 4

[25] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2