LLM as Dataset Analyst: Subpopulation Structure Discovery with Large Language Model

Yulin Luo[1*], Ruichuan An[1,2*], Bocheng Zou[1,3], Yiming Tang[1,4], Jiaming Liu[1], and Shanghang Zhang[1†]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2]Xi'an Jiaotong University [3]University of Wisconsin-Madison
[4]National University of Singapore
yulin@stu.pku.edu.cn

Abstract. The distribution of subpopulations is an important property hidden within a dataset. Uncovering and analyzing the subpopulation distribution within datasets provides a comprehensive understanding of the datasets, standing as a powerful tool beneficial to various downstream tasks, including Dataset Subpopulation Organization, Subpopulation Shift, and Slice Discovery. Despite its importance, there has been no work that systematically explores the subpopulation distribution of datasets to our knowledge. To address the limitation and solve all the mentioned tasks in a unified way, we introduce a novel concept of subpopulation structures to represent, analyze, and utilize subpopulation distributions within datasets. To characterize the structures in an interpretable manner, we propose the Subpopulation Structure Discovery with Large Language Models (SSD-LLM) framework, which employs world knowledge and instruction-following capabilities of Large Language Models (LLMs) to linguistically analyze informative image captions and summarize the structures. Furthermore, we propose complete workflows to address downstream tasks, named Task-specific Tuning, showcasing the application of the discovered structure to a spectrum of subpopulation-related tasks, including dataset subpopulation organization, subpopulation shift, and slice discovery. With the help of SSD-LLM, we can structuralize the datasets into subpopulation-level automatically, achieve average +3.3% worst group accuracy gain compared to previous methods on subpopulation shift benchmark Waterbirds, Metashift and Nico++, and also identify more consistent slice topics with a higher model error rate of 3.95% on slice discovery task for ImageNet. The code will be available at https://llm-as-dataset-analyst.github.io/.

Keywords: Subpopulation Structure Discovery · Large Language Model

* Equal Contribution.
† Corresponding Author.

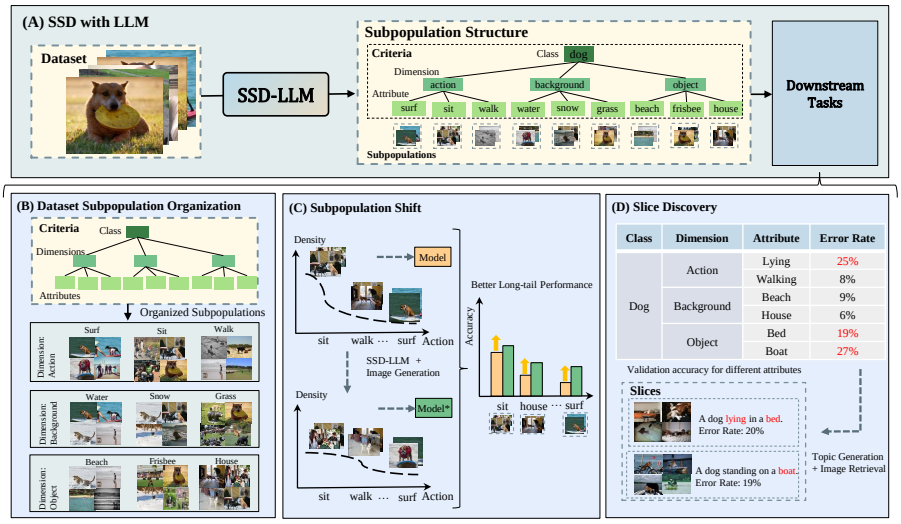

Fig. 1: (A) The Workflow of Subpopulation Structure Discovery with Large Language Models (SSD-LLM). SSD-LLM can further support several downstream tasks including: (B) Dataset Subpopulation Organization; (C) Subpopulation Shift; (D) Slice discovery.

1 Introduction

Subpopulation, defined by a set of data points that share common characteristics, is an important concept in machine learning [48]. Many tasks are subpopulation-related. For example, image clustering conditioned on text criteria [18] is to partition an image dataset into different subpopulations based on user-specified criteria, studying subpopulation shift [22, 48, 52] is to mitigate the negative impact of imbalanced subpopulation distributions in the training set on the model, slice discovery [2, 9] is aimed at identifying subpopulations model underperform.

Summarizing the commonalities of these tasks, we find that analyzing the subpopulation distribution is the key to solving all these problems. If the subpopulation distribution can be characterized, image clustering results under different criteria are naturally obtained [18], additional images can be supplemented to rare subgroups to balance the whole dataset [8], and slices can be easily discovered by statistics error rate on validation set [2]. Despite its importance, existing work [48] lacks systematic exploration of subpopulation distribution. To adjust the issue, for the first time, we propose the concept of subpopulation structure to represent, analyze, and utilize subpopulation distributions within datasets. By definition, a subpopulation structure is a set of hierarchical relations among several subpopulations determined by certain criteria.

Former works like Metashift [22] and NICO++ [52] have constructed image datasets including the subpopulation information, which organizes the images with respect to some extra attributes, and can be viewed as a class-, attribute-, subpopulation-layer structure. The problem of such a structure is ignoring

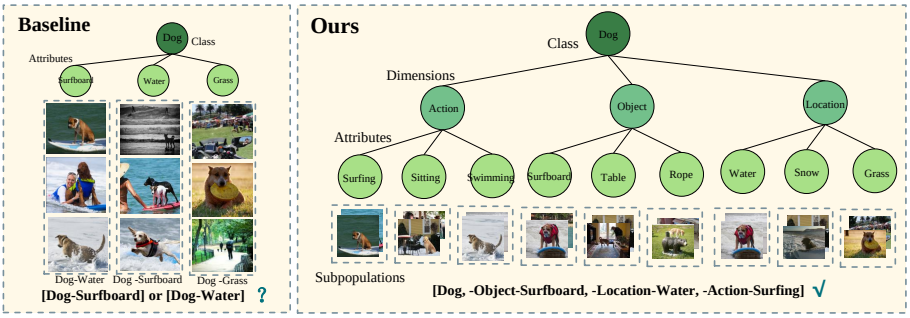

Fig. 2: Metashift has the same-level attributes *Surfboard*, *Water*, and *Grass* for class *Dog*, which is irrational due to the possible overlap. As an improvement, we take dimensions into consideration. The class *Dog* has dimensions including *Action*, *Co-occurrence Object*, *Location*, etc., and in dimension *Location*, it includes various attributes like *Water*, *Grass*, etc, which offers a more appropriate assignment for the samples.

the category of attributes (or *Dimension*), leading to attribute inconsistency and confusion. To solve this issue, we introduce a class-, dimension-, attribute-, and subpopulation-layer structure. The comparison of the two structures can be seen in Figure. 2. By articulating the classification dimensions, this improved structure provides more nuanced attribute assignments.

Identifying descriptive information within a dataset often requires large amounts of human manufacture [22, 52], which urges the need of automatic workflows to complete the subpopulation structure discovery. However, automatically identifying subpopulation structures within image datasets presents a significant challenge. The approach must be capable of extracting key information from images and summarizing essential content from extensive texts. Furthermore, it necessitates comprehensive world knowledge, enabling a broad understanding of various aspects of the datasets, including diverse categories, common attributes, and the relationships between dimensions and attributes.

Recently, Large Language Model (LLM) [17, 27, 46] and Multimodal Large Language Model (MLLM) [12, 25] have attracted wide attention due to their superior capacities. LLM has shown extensive world knowledge and remarkable abilities in summarization, instruction following [27], etc. MLLM extends the capabilities of LLM to handle visual inputs. By visual instruction tuning [25], MLLM can verbalize the rich information of images. Motivated by these, we propose a novel framework Subpopulation Structure Discovery with Large Language Model(SSD-LLM), illustrated in Figure. 1, to automatically uncover the structure. The core idea is to generate informative captions from images with MLLM, followed by analyzing and summarizing the subpopulation structure of datasets with LLM. Specifically, we design two elaborate prompt engineering components, Criteria Initialization and Criteria Refinement. The former utilizes a generate-and-select paradigm to summarize dimensions and attributes sequentially. The latter employs self-consistency as an indicator

to evaluate and refine the criteria. After obtaining complete criteria, each image is assigned to corresponding attributes according to its caption. The final subpopulation structures can be leveraged to finish various downstream tasks with the help of our proposed Task-specific Tuning. In this work, we focus on three application scenarios, i.e. dataset subpopulation organization, subpopulation shift, and slice discovery. We validate the effectiveness of SSD-LLM on these subpopulation-related tasks. For subpopulation shift, we achieve an improvement of +3.3% in worst group accuracy across three datasets compared to SOTA methods, and for slice discovery, we can identify more consistent slice topics with a higher model error rate of 3.95%.

Our contributions are summarized as follows:

- We introduce the concept of subpopulation structure to characterize subpopulation distribution in an interpretable manner for the first time.
- We propose class-dimension-attribute-subpopulation structure, reducing the attribute confusion of the current class-attribute-subpopulation structure.
- We propose Subpopulation Structure Discovery with Large Language Model (SSD-LLM) framework to uncover the underlying subpopulation structure of datasets automatically, with two elaborate prompt engineering components Criteria Initialization and Criteria Refinement.
- We provide methods for Task-specific Tuning, enabling the application of the structures across a spectrum of subpopulation-related tasks, including dataset subpopulation organization, subpopulation shift, and slice discovery.

2 Related Works

2.1 Hierarchical Structure of Image Datasets

Recent research has emphasized the need to organize datasets into hierarchical structures allowing for benchmarking various downstream tasks [22, 37, 43, 52]. Metashift [22] builds a collection of 12,868 sets of images related to 410 main subjects and their contexts. NICO++ [52], Waterbirds [43], and ImageNetBG [37] also propose methods for constructing various types of hierarchical datasets. However, the construction of these hierarchical datasets often requires manual annotation, hindering automatic construction. These approaches focus on just a single dimension, such as object context in Metashift, background in Waterbirds, and ImageNetBG, while more practical scenarios may involve multiple dimensions hidden within the comprehensive visual information.

2.2 Extract Information from Image Captions

Recent works such as ALIA [8], VeCAF [51], Bias2Text [15], and ICTC [19] explore utilizing caption models to obtain information from datasets. ALIA provides a method to augment datasets by generating variations of existing images through captioning and text-to-image models. While ALIA [8] supports dataset improvement, it lacks knowledge about attribute types, bias, or subpopulation

shift existence. Bias2Text [15] detects dataset bias by transforming images into descriptive captions and keywords. However, without large language model participation, Bias2Text fails to support classification dimension selection and can only differentiate images with basic keywords. More recently, ICTC [19] enables conditional image clustering using an LLM in a straightforward manner. Although ICTC clusters images when given the criterion, it requires human-assigned text prompts. Compared to ICTC, our prompt engineering paradigm supports scalable automatic subpopulation structure dataset organization without human criteria assignment and can generate comprehensive criteria tailored to datasets for various downstream tasks.

2.3 LLM Prompt Engineering

As the popularity of LLMs has surged, prompt engineering, the process of crafting and refining prompts to guide LLMs towards desired outputs [17,46] has also shown more and more importance. Various prompt engineering methods [28,35] and principles have emerged, and researchers or engineers have explored their applications in a diverse range of downstream tasks [4,10,27,33,34]. In particular, in-context learning has emerged as a pivotal technique, validated both experimentally [7, 21] and theoretically [6, 30, 47]. This approach involves providing the LLM with context information relevant to the task at hand, enabling it to generate more accurate and relevant responses. Least-to-most prompting [53] breaks down complex tasks into smaller, more manageable steps, enhancing LLM's reasoning skill by querying the LLM with more simplified sub-questions. Self-consistency [45] proposes to ensemble multiple responses to the LLM given the same prompt to get enhanced results, suggesting that consistent responses as an indicator for correct problem solving. Self-refining [31] demonstrates that we can use LLMs to refine their outputs by themselves with careful designing of prompts. In this work, we leverage a combination of prompt engineering techniques, including in-context learning, chain-of-thought, self-consistency, and self-refining to tackle subpopulation structure discovery effectively.

3 Method

In this section, we introduce our proposed method, subpopulation structure discovery with large language models(SSD-LLM). We describe the overall pipeline of SSD-LLM in Section 3.1, outlining how the paradigm automatically discovers the latent subpopulation structures inside the dataset. The process begins with captioning the images in the dataset with an MLLM, detailed in Section 3.2, and proceeds with Criteria Initialization with an LLM in Section 3.3. The paradigm then refines the initialized criteria through a recursive self-refinement procedure, detailed in Section 3.4. Finally, images are assigned to attributes, completing the subpopulation structure discovery process, as elaborated in Section 3.5. The section concludes with a discussion of how to apply the method to various downstream tasks, as presented in Section 3.6.

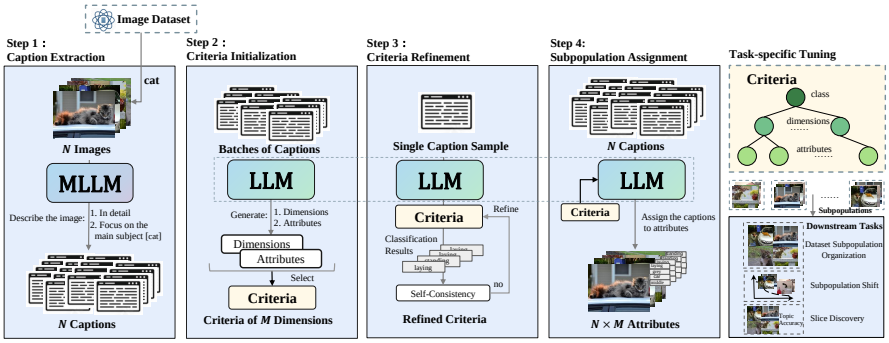

Fig. 3: Subpopulation Structure Discovery with Large Language Model (SSD-LLM). (Step 1) Multimodality Large Language Model (MLLM) extracts informative captions from images. (Step 2) LLM initializes the criteria with a sample-based generate-and-select paradigm. (Step 3) LLM refines the criteria using self-consistency as an indicator. (Step 4) LLM assigns each caption with specific attributes according to the refined criteria, uncovering the intrinsic subpopulation structures hidden in the dataset. The resulting criteria and subpopulations are used in several downstream tasks.

3.1 Overview

To automatically discover the subpopulation structures, we propose a novel prompt engineering paradigm that effectively leverages the capabilities of both multimodal large language models (MLLMs) and large language models (LLMs). Our proposed method comprises four key steps (Figure. 3). First, we transform the images into information-rich captions that capture the main information in the images using MLLMs. Second, we employ a novel sample-based prompt engineering method to guide an LLM to produce criteria consisting of dimensions and corresponding attributes organizing the dataset. Third, we prompt the LLM to refine this generated criteria. Last, we assign all the images in the dataset to specific attributes accordingly, uncovering the intrinsic subpopulation structures in the dataset, and paving the way for further analysis about the dataset. Detailed descriptions of each step are provided below. For notations, consistent with various former works [30, 46], we denote the operation of getting responses from the language models as LLM and $MLLM$, and use $[,]$ to represent the concatenation operation of two pacts of texts.

3.2 Caption Extraction

To begin our approach, we leverage the powerful image captioning capabilities of the MLLM to transform the images into informative and detailed captions. Instead of briefly describing the images, we prompt the MLLM to generate more detailed captions centered around the main subject CLS. The prompt we used in this step is stated as follow:

P_1 = *"Describe the image of the subject CLS in detail."*

Step 1 Caption Extraction

Input: Dataset: D_{img}, MLLM
Output: Image Captions: C

```
for i in range(NumOfIterations)
    img = D_img.sample()
    c = MLLM(img, P_1)
    C.append(c)
end for
```

Step 2-1 Dimension Generation

Input: Captions: C, LLM
Output: dimensions: $Dims$

```
for i in range(NumOfIterations)
    c = C.sample(NumOfSamples)
    S.append(LLM([P_2^1,c]))
end for
Dims = LLMSummary(S)
```

3.3 Criteria Initialization

To discover the hidden subpopulation structures within the dataset, we employ an LLM to delve into the information-rich captions generated in the previous step. Our objective is to identify certain criteria that effectively partition the images into several distinct subgroups. Beyond simply dividing the dataset into subgroups, we articulate the classification dimension for the partition and record all the resulting attributes generated from the classification process. Along with the class information and the resulting subpopulations, this criteria naturally form a four-layer structure, class-, dimension-, attribute-, and subpopulation-. Noticing criteria encompass multiple dimensions and their corresponding attributes, we adopt a generate-and-select paradigm with the LLM to discover the dimensions and the attributes sequentially.

To determine the dimensions and attributes, we employ an iterative sampling approach, repeatedly prompting the LLM to propose dimensions and attributes based on batches of image captions. In each iteration, the LLM generates candidate dimensions and attributes, which are subsequently processed through an LLM summarization process. This sample-and-summarize approach effectively addresses the challenges when processing large datasets. Since the number of dimensions that can differentiate images in a dataset is relatively small, and these dimensions have an appearance in numerous images, our sample-based approach effectively identifies relevant dimensions. The prompts we used in this step are stated as follow, omitting Chain-of-thought examples for simplicity.

P_2^1 = *"Suggest some dimensions that can di erentiate the following image captions."*

P_2^2 = *"Suggest a complete criterion to di erentiate the following image captions by the given dimension."*

3.4 Criteria Refinement

To further refine the criteria and ensure its effectiveness in classifying image captions across the dataset, we implement a recursive refining process. This approach proposes a novel method for identifying image captions requiring further refinement utilizing the self-consistency of LLM responses as an indicator [45].

Step 2-2 Attribute Generation

Input: Dimensions: $Dims$, Captions: C, Large language model: LLM
Output: Initialized criteria: $Criteria$

```
for dim in Dims do
    for i in range(NumOfIterations) do
        c = C.sample(NumOfSamples)
        S.append(LLM([P_2^2, dim, c]))
    end for
    Attributes = LLMSumarry(S) *list of attributes
    Criteria[dim] = Attributes
    S.reset()
end for
```

This choice stems from our empirical observation that if an image can be accurately classified according to a particular dimension, it should consistently be classified into the same attribute multiple times. Inconsistent responses, however, suggest that the current criteria require further refinement, either to merge redundant attributes or to include new attributes. The prompts we used in this step are stated as follow and the pseudocode is included in the appendix:

P_3^1 = *"Classify the caption by the criteria listed below."*

P_3^2 = *"We are unable to classify the following image caption using the provided criteria due to attributes redundancy or misappearance. If redundancy, please prune the redundant attributes. If missing, please suggest an additional attribute that would match the image caption."*

3.5 Subpopulation Assignment

Equipped with the comprehensive criteria, we proceed to systematically assign each image to the specific attributes of each dimension. Images assigned to the same attributes across all dimensions form distinct subgroups within the dataset, revealing the intrinsic subpopulation structures hidden within the data. These subpopulation structures can then be leveraged to perform various downstream tasks, completing our overall pipeline for employing an LLM to analyze the image dataset. The prompt we used in this step is stated as follow:

P_4 = *"Please assign following caption to one attribute of given dimension."*

Step 4 Subpopulation Assignment

Input: Captions: C, Large language model: LLM, criteria: $Criteria$
Output: Further assignments for each caption c.

```
for c in C do
    for [dim, Attributes] in Criteria do
        c.assign(LLM([P_4,c,dim,Attributes]))
    end for
end for
```

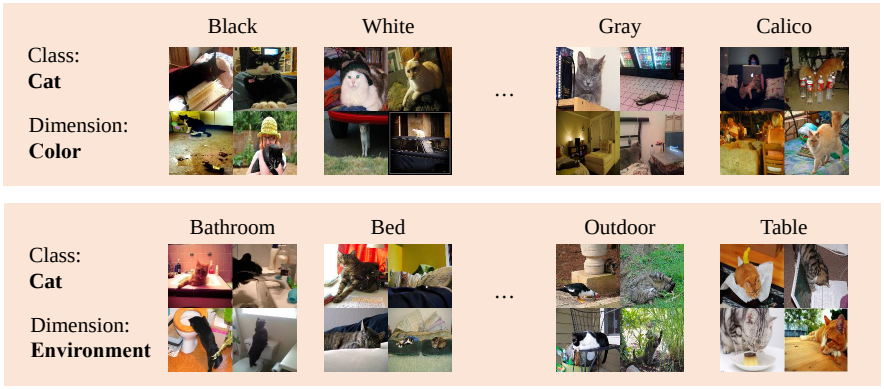

Fig. 4: A visualization of organised subpopulations in a dataset of cats.

3.6 Task-specific Tuning for Downstream Tasks

Once we have identified the criteria and subpopulations within the dataset, we can leverage this information to tackle several downstream tasks effectively. This includes organizing the subpopulations, which can reveal valuable insights into the data, such as potential subpopulation biases and the presence of long-tail attributes. In practice, the subpopulation structures can be used to improve model performance on various tasks if combined with extra operations, including handling subpopulation shifts and slice discovery.

Dataset Subpopulation Organization Organizing subpopulations within a dataset can provide beneficial information about the numbers of data points in each subpopulation and can reflect dataset biases and help us identify some long-tail subgroups. We conduct experiments on the task of image clustering conditioned on human-specified criteria to evaluate the quality of subpopulation organization. In specific, when organising the subpopulations of a given image dataset, we first select out the relevant dimensions and then attach attributes assigned by SSD-LLM directly to the images accordingly. Our method, SSD-LLM, automates subpopulation organization within datasets, and thus holds the potential to revolutionize the way hierarchical datasets are constructed.

Subpopulation Shift Subpopulation shift [48] stands as a common challenge in machine learning, occuring when the proportion of some subpopulations between training and deployment changes, and is shown to be of significant influence to model performances [39, 48]. SSD-LLM, combined with image generation, offers a solution to better handle the scenarios of subpopulation shifts. In our experiments, after we apply SSD-LLM to the datasets, we collect statistics the number of images contained in each subpopulation and utilize diffusion model to generate images for underrepresented subpopulations. Subsequently, we sample attributes from the subpopulation structure for each underrepresented subpopulation and employ LLM to make complete sentences based on these words as the input prompt of a diffusion model. The diffusion model generates images augmented

Dataset	Criterion	SCAN*	IC|TC	Ours
Stanford 40 Action	Action	0.346	0.747	**0.817**
	Location	0.357	0.671	**0.705**
	Mood	0.276	0.746	**0.768**
Place365	Place	0.332	-	**0.696**
PPMI	Musical Instruction	0.598	0.934	**0.955**
Cifar10	Object	0.839	0.911	**0.921**
STL10	Object	0.798	0.986	**0.988**

Table 1: Quantitative results of Dataset Subpopulation Organization. Method labeled with * is evaluated by having a human provided ground truth labels, cause the method itself is an unsupervised learning paradigm.

to the image dataset which further helps to achieve balanced classes and attributes. Moreover, we propose to harness an LLM to suggest extra dimensions and attributes based on the current sets in this task for enriched subpopulation structure, generating more diverse images.

Slice Discovery Slice discovery is a task aiming at uncovering subpopulations within a dataset where a machine learning model consistently exhibits poor performances. These subpopulations with underperformances, or slices, provide valuable insights into the model's limitations, potential biases [14], and how to improve the performances. SSD-LLM conducts slice discovery for an image dataset with the help of the assigned attributes. In detail, we first calculate the error rates on all subpopulations discovered with the SSD-LLM. Then we identify out the subpopulations with the highest error rate and use the LLM to summarize out discriptions based on the attributes of the subpopulations in the form of texts representing the slice topics, completing the task of slice discovery.

4 Experiments

We now present experimental results demonstrating the effectiveness of SSD-LLM. In particular, we present the main settings and results in this section and defer extra details, including various visualization result, to appendix. In our experiments, we mainly use LLaVA1.5 [26] for MLLM and GPT-4 [1] for LLM.

We conduct experiments on Dataset Subpopulation Organization, Subpopulation Shift, and Slice Discovery. Our superior performance underscores the method's efficacy in identifying and analyzing subgroups, further affirming its utility in addressing related challenges. Moreover, it illustrates that the unified paradigm can effectively address a variety of downstream tasks.

4.1 Evaluation on Dataset Subpopulation Organization

Setup The data assignment process facilitated by SSD-LLM, can be considered a form of clustering. To evaluate the quality of these identified subpopulations,

Type	Method	Average Accuracy				Worst Group Accuracy			
		Waterbirds	Metashift	Nico++	Average	Waterbirds	Metashift	Nico++	Average
Vanilla	ERM	84.1	91.2	76.3	83.7	69.1	82.1	17.8	56.3
Subgroup Robust Methods	GroupDRO	86.9	91.5	74.0	84.1	73.1	83.1	12.2	56.1
	JTT	88.9	91.2	77.5	85.9	71.2	82.6	15.6	56.5
	LfF	86.6	80.4	77.5	81.5	75.0	72.3	15.6	54.3
	LISA	89.2	91.4	75.0	85.2	77.0	79.0	18.9	58.3
Imbalanced Learning	Resample	86.2	92.2	77.3	85.2	70.0	81.0	16.7	55.9
	Reweight	86.2	91.5	73.8	83.8	71.9	83.1	12.2	55.7
	Focal	89.3	91.6	73.1	84.7	71.6	81.0	16.7	56.4
	CBLoss	86.8	91.4	76.3	84.8	74.4	83.1	12.2	56.6
	BSoftmax	88.4	91.3	74.2	84.6	74.1	82.6	16.7	57.8
Traditional Data Augmentation	Mixup	89.2	91.4	73.0	84.5	77.5	79.0	14.4	57.0
	RandAug	86.3	90.9	72.0	83.1	71.4	80.9	16.7	56.3
Diffusion	Class Prompt	85.9	91.5	78.0	85.1	71.3	82.7	18.5	57.5
	Class-Attribute	89.1	91.4	78.6	86.4	73.5	83.8	18.8	58.7
	CiP	88.0	91.1	78.3	85.8	73.5	82.4	19.3	58.4
LLM+Diffusion	**SSD-LLM (Ours)**	**90.5**	**93.0**	**80.4**	**88.0**	**79.1**	**84.8**	**22.1**	**62.0**

Table 2: Comparison of methods for image classification with subpopulation shifts.

we assess the clustering accuracy by comparing images against secondary labels that reflect subgroup attributes derived via SSD-LLM. Full information for datasets, text criterion, and model selection can be found in appendix.

Comparison Methods SCAN [41] is a two-stage clustering method that decouples feature learning and clustering. IC|TC [19] is a new paradigm for image clustering that supports human interaction. It utilizes the given Text Criteria to accurately control the quality of the clustering results.

Results and Analysis In Table. 1, we report the average accuracy achieved by each method based on the predefined textual criteria. When compared to IC|TC, SSD-LLM demonstrates competitive performance. It is important to note that IC|TC incorporate artificial judgment in its process, which leads to poor scalability when handling large datasets. In contrast, our approach is fully automated. We conduct visualization of our organized subpopulations, as shown in Figure. 4. We can visually observe that the pictures and dimensions are indeed consistent, indicating the effectiveness of our mining and assigning process.

4.2 Evaluation on Subpopulation Shift

Setup We evaluate subpopulation shifts on three commonly used image datasets, Metashift (Cats vs Dogs), Waterbirds (Landbirds vs Waterbirds), and NICO++. We choose *Average Accuracy* and *Worst Group Accuracy* as evaluation metrics. To ensure a fair comparison, following [48], we conduct a random search of 16 trials over a joint distribution of all hyperparameters. We then use the validation set to select the best hyperparameters for each algorithm, fix them, and rerun the experiments under five different random seeds to report the final average results. To make the evaluation more realistic, we consider the model selection setting *Attributes are unknown in both training and validation*.

Comparison Methods Following recent benchmarking efforts [48], we compare SSD-LLM with several types of methods: (1) *vanilla:* ERM [42], (2) *Subgroup Robust Methods:* GroupDRO [38], LfF [32], JTT [24], LISA [49], (3) *Imbalanced*

Method|Categories	Boat	Bird	Car	Cat	Dog	Truck	Topic Error Rate
ImageNet	4.33	0.81	11.33	11.14	0.69	11.71	6.72
General Prompt	47.82	12.11	43.55	14.22	10.19	12.65	23.42
GPT-Suggest	57.55	12.87	43.59	12.71	16.34	28.12	28.53
Domino(Bert)	76.26	42.26	54.21	33.89	24.50	29.54	43.44
B2T	77.62	30.04	58.17	**36.36**	19.80	33.47	42.58
SSD-LLM (Ours)	**79.31**	**45.67**	**60.34**	32.97	**26.48**	**39.57**	**47.39**

Table 3: Results of slice discovery on Imagenet-1K with various SDMs.

Learning: ReSample [13], ReWeight [13], Focal [23], CBLoss [5], Bsoftmax [36], (4) *Traditional Data Augmentation:* Mixup [50], RandAug [3], (5) *Di usion:* Class Prompt [40], Class-Attribute Prompt [40], CiP [20].

Results and Analysis From Table 2, SSD-LLM surpasses previous methods, with a +1.6% improvement in average accuracy and +3.3% in worst group accuracy across three datasets. The analysis is as follows: (1) Despite being based on conventional ERM, data-based approaches show competitive performance compared to model-based algorithms, highlighting their potential. (2) For diffusion-based methods, class-attribute prompts outperform class prompts, underscoring the importance of understanding dataset imbalanced attributes for effective image generation. However, the need for pre-identifying these attributes emphasizes the value of SSD-LLM, which automates attribute discovery and provides detailed annotations, enhancing performance. (3) The superior performance of CiP over Class Prompt underscores the significance of diverse text prompts. (4) Our method's superior performance results from a comprehensive analysis of subpopulation imbalances within the dataset. The strategic text diversity achieved through attribute sampling and LLM sentence-making effectively addresses subpopulation shifts.

4.3 Evaluation on Slice Discovery

Setup In contrast to typical slice discovery tasks, we redefine evaluation pipeline following [11]. In this study, we evaluate bugs found of ImageNet models. Specifically, a classification is deemed incorrect when an image containing target object is erroneously identified by the model as containing an unrelated object.

Comparison Methods Domino [9] represents a state-of-the-art method in slice discovery, which effectively clusters errors identified in the validation set and characterizes them through captions generated automatically. B2T [16] is a recently proposed framework which identifies and interprets visual biases in vision models using keyword extraction from captions of mispredicted images.

Results and Analysis We evaluate the effectiveness of our method in slice discovery on 6 representative superclasses in Imagenet. As shown in Table. 3, our SSD-LLM overcomes all other SDMs, including Domino [9], by a significant margin. We find topics given by Domino tend to encounter two unsatisfactory cases: loss of semantics, and missing class information. These cases are also discussed

Methods	Index	Criteria Initialization		Criteria	Accuracy		
		Dimension	Attribute	Refinement	Action	Location	Mood
SSD LLM	A	10 × 20	20 × 20	×	75.6	67.3	71.6
	B	5 × 20	20 × 20	×	65.2	60.0	66.2
	C	20 × 20	20 × 20	×	75.9	67.1	71.3
	D	10 × 20	10 × 40	×	70.3	68.9	69.5
	E	10 × 20	40 × 10	×	80.2	70.0	74.4
	F	10 × 20	20 × 20	✓	81.7	70.5	76.8

Table 4: Ablation study on the number of samples (NUM of rounds×NUM per round), and Criteria Refinement (with or without).

in [11], where they reasoned this phenomenon into the inherent difficulties of automatic SDMs. However, evidence suggests that our SSD-LLM can handle these problems, while keeping a high error rate and maintaining automation. Specifically, SSD-LLM achieves an average error rate of 47.39%, surpassing Domino [9] by 3.95%. Furthermore, when we trace back to origin dataset, the discovered slice is also very consistent(detailed visualizations included in appendix). Interestingly, we find the data mining process of SSD-LLM is just the same as human data scientists, who take up hypotheses and improve model performance by viewing batches of bad subpopulations. Experiments show that our method overcome the inherent difficulties while maintaining automaton, paving the way for data-centric methods.

4.4 Ablation Study

Hyperparameters of Criteria Initialization/Refinement The N × M in Table. 4 represents the rounds and samples of suggestions. Setting A serves as our baseline. For A, B, and C, results show that insufficient suggested dimensions lead to decreased performance, while enough dimension samples lead to stable performance, as the ICTC task requires only the most appropriate match. Too few suggestions may fail to find suitable dimensions, resulting in mismatched attribute generation. For A, D, and E, results highlight the importance of the number of captions in summarizing attributes. Too many captions can cause some to be overlooked, reducing total identified attributes and performance. The improvement from A to F demonstrates that Criteria Refinement enhances attribute comprehensiveness and final performance.

Diffusion Generation Strategy From Table. 5, we analyze key components for employing diffusion models to address subpopulation shift. For A→B, we confirm that managing imbalance attributes within datasets helps solve the task. For B→D, we illustrate that a comprehensive subpopulation structure benefits the task. For D→E, balanced subpopulation sampling improves data quality and model training. For E→F, LLM-generated prompts produce more reasonable images, enhancing results. For F→G, LLM suggests additional attributes to enrich subpopulation structures, generating more diverse images and improving model generalization. For F→H, we verify the scaling capability of our SSD-LLM.

Methods	Index	Attribute Mode	Sample Mode	SD Prompt Mode	Number	Average Accuracy	Worst Group Accuracy
ERM	——	——	——	——	×1	84.1	69.1
SSD-LLM	A	GT Attribute Unknown	random	Direct	×1	85.9	71.3
	B	GT Attribute Known	random	Direct	×1	89.1	73.5
	C	GT Attribute Known	weighted	Direct	×1	89.3	73.8
	D	SSD-LLM Attribute	random	Direct	×1	89.5	76.2
	E	SSD-LLM Attribute	weighted	Direct	×1	89.8	77.4
	F	SSD-LLM Attribute	weighted	LLM Sentence	×1	90.1	78.3
	G	SSD-LLM & LLM Suggest	weighted	LLM Sentence	×1	90.5	79.1
	H	SSD-LLM Attribute	weighted	LLM Sentence	×2	91.1	79.2

Table 5: Ablation study on the sample mode and SD prompt mode.

5 Discussion

Although our method, SSD-LLM, has already shown effectiveness in various settings, the current exploration of the algorithm is limited to image datasets. Besides, this approach may bear the potential bias from the MLLMs and LLMs. However, it may be reduced from extra human-in-the-loop guidiances if available.

For future works, we suggest the following promising directions:

- Structure Format The four-layer subpopulation structure can be expanded to more suitatble structures according to specific task requirements.
- Downstream Tasks SSD-LLM can have more applications in various computer vision and multimodality tasks, e.g. object detection and VQA.
- Dataset Construction The subpopulation structure obtained from SSD-LLM holds the potential to guide dataset construction with better fairness [44] or further supporting the construction of unbiased datasets [29].
- Technical Extensions The core procedures of SSD-LLM, using LLM to conduct group-level summarizations, can be extended to more types of contents including patterns of model hallucinations.

6 Conclusion

In this work, we present the first systematic exploration of subpopulation structure discovery. We provide a precise definition of subpopulation structure and introduce a fine-grained criteria to determine the structures. We propose SSD-LLM for automatic subpopulation structure discovery incorporating elaborate prompt engineering techniques . SSD-LLM can be combined with subsequent operations to better tackle several downstream tasks, including dataset subpopulation organization, subpopulation shift and slice discovery, with minor Task-specific Tuning. We conduct extensive experiments to verify our proposed methods, demonstrating the remarkable effectiveness and and generality of SSD-LLM.

Acknowledgements

Shanghang Zhang is supported by the National Science and Technology Major Project of China (No. 2022ZD0117801).

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 preprint arXiv:2303.08774 (2023)
2. Chen, M., Li, Y., Xu, Q.: Hibug: On human-interpretable model debug. In: NeurIPS (2023)
3. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
4. Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L.: Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 (2023)
5. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
6. Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., Wei, F.: Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (2023)
7. Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Sui, Z.: A survey for in-context learning. arXiv preprint arXiv:2301.00234 (2022)
8. Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J.E., Darrell, T.: Diversify your vision datasets with automatic diffusion-based augmentation. arXiv preprint arXiv:2305.16289 (2023)
9. Eyuboglu, S., Varma, M., Saab, K., Delbrouck, J.B., Lee-Messer, C., Dunnmon, J., Zou, J., Ré, C.: Domino: Discovering systematic errors with cross-modal embeddings. arXiv preprint arXiv:2203.14960 (2022)
10. Fu, Y., Peng, H., Khot, T., Lapata, M.: Improving language model negotiation with self-play and in-context learning from ai feedback. arXiv preprint arXiv:2305.10142 (2023)
11. Gao, I., Ilharco, G., Lundberg, S., Ribeiro, M.T.: Adaptive testing of computer vision models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4003–4014 (2023)
12. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
13. Japkowicz, N.: The class imbalance problem: Significance and strategies. In: Proc. of the Int'l Conf. on artificial intelligence. vol. 56, pp. 111–117 (2000)
14. Johnson, N., Cabrera, A., Plumb, G., Talwalkar, A.: Where does my model underperform? a human evaluation of slice discovery algorithms. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing **11**(1), 65–76 (Nov 2023). `https://doi.org/10.1609/hcomp.v11i1.27548`, `http://dx.doi.org/10.1609/hcomp.v11i1.27548`
15. Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., Shin, J.: Bias-to-text: Debiasing unknown visual biases through language interpretation. arXiv preprint arXiv:2301.11104 (2023)
16. Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., Shin, J.: Bias-to-text: Debiasing unknown visual biases through language interpretation. arXiv preprint arXiv:2301.11104 (2023)

17. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. NeurIPS **35**, 22199–22213 (2022)
18. Kwon, S., Park, J., Kim, M., Cho, J., Ryu, E.K., Lee, K.: Image clustering conditioned on text criteria. In: ICLR (2023)
19. Kwon, S., Park, J., Kim, M., Cho, J., Ryu, E.K., Lee, K.: Image clustering conditioned on text criteria. arXiv preprint arXiv:2310.18297 (2023)
20. Lei, S., Chen, H., Zhang, S., Zhao, B., Tao, D.: Image captions are natural prompts for text-to-image models. arXiv preprint arXiv:2307.08526 (2023)
21. Li, X., Qiu, X.: Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. arXiv preprint arXiv:2305.05181 (2023)
22. Liang, W., Zou, J.: Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. arXiv preprint arXiv:2202.06523 (2022)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
24. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: International Conference on Machine Learning. pp. 6781–6792. PMLR (2021)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
27. Liu, J., Hu, T., Zhang, Y., Gai, X., Feng, Y., Liu, Z.: A chatgpt aided explainable framework for zero-shot medical image diagnosis. arXiv preprint arXiv:2307.01981 (2023)
28. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
29. Liu, Z., He, K.: A decade's battle on dataset bias: Are we there yet? (2024), `https://arxiv.org/abs/2403.08632`
30. Luo, Y., Tang, Y., Shen, C., Zhou, Z., Dong, B.: Prompt engineering through the lens of optimal control. arXiv preprint arXiv:2310.14201 (2023)
31. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651 (2023)
32. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems **33**, 20673–20684 (2020)
33. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442 (2023)
34. Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M.: Communicative agents for software development. arXiv preprint arXiv:2307.07924 (2023)
35. Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., Chen, H.: Reasoning with language model prompting: A survey. arXiv preprint arXiv:2212.09597 (2022)
36. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in neural information processing systems **33**, 4175–4186 (2020)

37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
38. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
39. Santurkar, S., Tsipras, D., Madry, A.: Breeds: Benchmarks for subpopulation shift (2020)
40. Shipard, J., Wiliem, A., Thanh, K.N., Xiang, W., Fookes, C.: 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. pp. 769–778 (2023)
41. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: European conference on computer vision. pp. 268–285. Springer (2020)
42. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)
43. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
44. Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., Russakovsky, O.: Revise: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision **130**(7), 1790–1810 (2022)
45. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
47. Xie, S.M., Raghunathan, A., Liang, P., Ma, T.: An explanation of in-context learning as implicit bayesian inference. arXiv preprint arXiv:2111.02080 (2021)
48. Yang, Y., Zhang, H., Katabi, D., Ghassemi, M.: Change is hard: A closer look at subpopulation shift (2023)
49. Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., Finn, C.: Improving out-of-distribution robustness via selective augmentation. In: International Conference on Machine Learning. pp. 25407–25437. PMLR (2022)
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
51. Zhang, R., Cai, Z., Yang, H., Liu, Z., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., Chang, B., Du, Y., et al.: Vecaf: Vlm-empowered collaborative active finetuning with training objective awareness. arXiv preprint arXiv:2401.07853 (2024)
52. Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., Cui, P.: Nico++: Towards better benchmarking for domain generalization (2022)
53. Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al.: Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625 (2022)