

FOLLOW-UP DIFFERENTIAL DESCRIPTIONS: LANGUAGE MODELS RESOLVE AMBIGUITIES FOR IMAGE CLASSIFICATION

Reza Esfandiarpour & Stephen H. Bach

Department of Computer Science

Brown University

Providence, RI 02906, USA

{reza_esfandiarpour, stephen_bach}@brown.edu

ABSTRACT

A promising approach for improving the performance of vision-language models like CLIP for image classification is to extend the class descriptions (i.e., prompts) with related attributes, e.g., using `brown sparrow` instead of `sparrow`. However, current zero-shot methods select a subset of attributes regardless of commonalities between the target classes, potentially providing no useful information that would have helped to distinguish between them. For instance, they may use color instead of bill shape to distinguish between sparrows and wrens, which are both brown. We propose Follow-up Differential Descriptions (FuDD), a zero-shot approach that tailors the class descriptions to each dataset and leads to additional attributes that better differentiate the target classes. FuDD first identifies the ambiguous classes for each image, and then uses a Large Language Model (LLM) to generate new class descriptions that differentiate between them. The new class descriptions resolve the initial ambiguity and help predict the correct label. In our experiments, FuDD consistently outperforms generic description ensembles and naive LLM-generated descriptions on 12 datasets. We show that differential descriptions are an effective tool to resolve class ambiguities, which otherwise significantly degrade the performance. We also show that high quality natural language class descriptions produced by FuDD result in comparable performance to few-shot adaptation methods. Code: <https://github.com/BatsResearch/fudd>

1 INTRODUCTION

What is the most distinguishing characteristic of a sparrow? It depends. To distinguish it from what? To distinguish it from a goldfinch, it is the brown color. But, to distinguish it from a wren, it is the conical bill (Fig. 1). Here, we propose a zero-shot approach to adapt the class representations of vision-language models based on other classes in an image classification task. We use natural language descriptions (called *prompts*) to provide visually differentiating information for target classes.


(A)	Color	Bill	(B)	Sparrow vs. Wren		Sparrow vs. Goldfinch	
Sparrow	Brown	Conical		Brown color	✓ Conical bill	✓ Brown color	Conical bill
Goldfinch	Yellow	Conical		<input checked="" type="checkbox"/> Sparrow	<input checked="" type="checkbox"/> Sparrow	<input checked="" type="checkbox"/> Sparrow	<input checked="" type="checkbox"/> Sparrow
Wren	Brown	Slender		<input checked="" type="checkbox"/> Wren	<input type="checkbox"/> Wren	<input type="checkbox"/> Goldfinch	<input checked="" type="checkbox"/> Goldfinch

Figure 1: A) Attributes for three different classes. B) Two sample classification tasks involving the wren class. The distinguishing characteristics of each class vary based on other classes. Our approach selects the class descriptions based on other classes in the dataset to provide the information that differentiates the target classes.

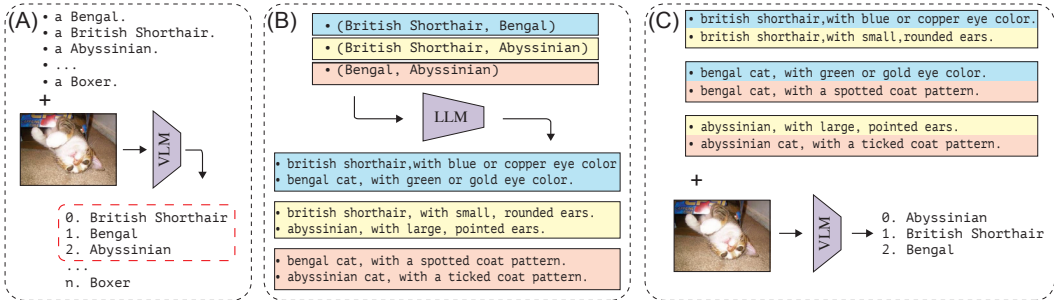


Figure 2: FuDD overview. A) Using the model’s initial prediction, we identify the potentially ambiguous classes. B) We use a large language model to generate class descriptions that differentiate the ambiguous classes. C) We use the new differential descriptions in a follow-up classification task to resolve the initial ambiguity and select the correct label.

Large Vision-Language Models (VLMs) use natural language as a source of supervision, which allows us to easily create new classifiers by describing the classes in natural language (e.g., class names) and labeling each image with the closest class. As a result, we can efficiently transfer the learned visual representations to downstream tasks by adapting the class descriptions (i.e., prompts) to the model’s pre-training distribution (Chen et al., 2023; Khattak et al., 2023; Menghini et al., 2023; Mirza et al., 2023; Novack et al., 2023; Patashnik et al., 2021; Radford et al., 2021; Zang et al., 2022; Zhang et al., 2023a).

The performance of prompting depends on careful prompt engineering to find class descriptions that provide the most helpful information for each task (Menon & Vondrick, 2023; Novack et al., 2023). Previous works have used class attributes for better zero-shot learning performance (Lampert et al., 2013; Parikh & Grauman, 2011; Romera-Paredes & Torr, 2015; Socher et al., 2013; Xian et al., 2018). Several recent works have adapted this idea for image classification with VLMs and propose to describe the classes by their attributes, such as color and shape (Menon & Vondrick, 2023; Pratt et al., 2022; Yang et al., 2023). Specifically, they prompt a large language model (LLM) with queries like `What does an image of a {class name} look like?`, and use the responses as the new class descriptions. The main idea is to enable the model to use the described attributes in addition to the class names to identify each class. This should lead to better class representations since models can better detect the attributes because of their prevalence during pre-training.

The additional information is helpful if it discriminates the class from all other classes in the target dataset. At least, the provided information should differentiate the class from other classes in the dataset that are frequently mistaken for it. Despite this crucial requirement, current zero-shot approaches generate descriptions solely based on the class itself without considering other classes in the dataset (Menon & Vondrick, 2023; Pratt et al., 2022). As a result, although the descriptions provide additional details about the class, they might not contain any information that differentiates the class from other classes in the dataset. Thus, current methods might generate class descriptions that are not helpful for the target task.

In this paper, we propose Follow-up Differential Descriptions (FuDD¹), a novel approach that tailors the class descriptions to each dataset and provides additional details that resolve the potential ambiguities in the target task. For each image, we first use a set of basic class descriptions as usual (Radford et al., 2021) and identify a subset of classes that, according to the VLM, are likely to be the true label and thus are considered ambiguous. Then, for each such ambiguous class, we generate a set of descriptions that include the details that differentiate it from other ambiguous classes. We rely on the extended world knowledge of pre-trained large language models to generate such differential descriptions at scale (Brown et al., 2020; Petroni et al., 2019). Finally, we use these differential descriptions in a follow-up classification task to resolve the initial ambiguity and predict the correct label. By customizing the class descriptions based on other classes in the dataset, FuDD aims to provide the most effective information for separating the classes of the target dataset.

¹Pronounced like food.

We evaluate our method on 12 datasets and show that FuDD consistently outperforms naive descriptions for all datasets. FuDD outperforms naive LLM-generated descriptions by 2.41 percentage points on average, with up to 13.95 percentage points improvement for the EuroSAT dataset. In our experiments, we show that not all descriptions resolve ambiguities, and effective class descriptions should provide differentiating information about ambiguous classes. Moreover, differentiating the highly-ambiguous classes is the most important factor, accounting for most of FuDD’s performance gains. In addition to GPT-3.5², we experiment with the smaller, publicly available Llama 2 model (Touvron et al., 2023) to study the impact of further fine-tuning, and find that the 7b-parameter model can provide helpful information for everyday concepts. It also benefits from further fine-tuning, especially for rare and abstract concepts in the EuroSAT and DTD datasets, with up to 23.41 percentage points boost in accuracy. Finally, we show that the performance when using high-quality class descriptions from FuDD is comparable to using few-shot methods, achieving performance competitive with 16-shot VLM adaptation methods (Yang et al., 2023; Zhou et al., 2022b) for some datasets. Our results uncover the potential of using natural language to tailor the class representations to each dataset by providing information that differentiates the ambiguous classes. These results motivate future work on creating effective class descriptions for each downstream task.

2 RELATED WORK

There is an increasing body of work on adapting VLMs to a wide range of downstream tasks (Gao et al., 2021; Guo et al., 2023; Jia et al., 2022; Novack et al., 2023; Patashnik et al., 2021; Rao et al., 2022; Udandarao et al., 2022; Zeng et al., 2022; Zhang et al., 2021; 2022). Here, we describe the related work and highlight their differences with our method.

Prompt Tuning Prompt tuning is an efficient approach for few-shot adaptation of VLMs to downstream classification tasks. Instead of updating the model parameters, prompt tuning methods add learnable parameters to the input image or text (i.e., prompt) and learn these parameters through gradient descent for each dataset (Huang et al., 2022; Jia et al., 2022; Menghini et al., 2023; Nayak et al., 2022; Zhou et al., 2022a;b). For instance, CoOp adds a set of parameters to the class descriptions to represent the dataset context; and then uses a few labeled samples for training (Zhou et al., 2022b). Although prompt tuning methods achieve good accuracy, they require additional labeled examples, which limits their applications. On the other hand, our method is zero-shot and adapts to each dataset without any additional samples, with competitive performance to prompt-tuning methods in low-shot scenarios.

VLMs with Other Foundation Models One line of work uses the capabilities of other foundation models (Brown et al., 2020; Caron et al., 2021; Ramesh et al., 2021) in combination with VLMs to better adapt to downstream tasks (Chen et al., 2023; Gupta & Kembhavi, 2023; Menon & Vondrick, 2023; Mirza et al., 2023; Pratt et al., 2022; Surís et al., 2023; Zeng et al., 2022; Zhang et al., 2023b). For example, one can use the extended world knowledge of large language models (LLMs) in combination with VLMs to solve more complex visual tasks. Our approach is closely related to this line of work; we discuss the differences further in the next paragraph. Several other methods use text-to-image generation models (Rombach et al., 2022) on top of LLMs (Brown et al., 2020) to further improve the performance (Udandarao et al., 2022; Zhang et al., 2023a). For instance, SuS-X first uses an LLM to generate class descriptions and then uses a text-to-image generation model to generate synthetic images for each class based on these descriptions (Udandarao et al., 2022). Our experiments show that despite using no images, FuDD’s performance is comparable to SuS-X for most datasets while avoiding the complexities of text-to-image generation models.

Adaptation Through Description A specific approach for improving class representations without additional samples is to provide more informative class descriptions (Menon & Vondrick, 2023; Novack et al., 2023; Pratt et al., 2022; Roth et al., 2023). For example, WaffleCLIP adds high-level category names to class descriptions to avoid ambiguities caused by class names with multiple meanings (Roth et al., 2023). Another approach is to describe the classes with their attributes so the model can rely on attributes in addition to class names to identify images of each class (Menon & Vondrick, 2023; Pratt et al., 2022). For example, Menon & Vondrick (2023) propose to generate such class descriptions by querying an LLM about the most important attributes of each class. However, the generated descriptions might provide no useful information for separating the class from

²<https://openai.com/blog/openai-api>

other classes in the dataset. For example, the attribute color is not useful for separating sparrows and wrens since both are brown (Fig. 1). To address this issue, LaBo uses additional labeled examples to learn the importance of each attribute (Yang et al., 2023). Then, it selects a set of attributes that are the most discriminative for each class. Unlike LaBo, FuDD generates class descriptions that effectively separate the target classes in the first place, eliminating the need for further optimization. Despite using no labeled data, FuDD’s performance is comparable to few-shot LaBo for most datasets.

3 FOLLOW-UP DIFFERENTIAL DESCRIPTIONS (FUDD)

Here, we describe the components of our proposed method, FuDD. In Section 3.1, we explain how VLMs are used for image classification. In Section 3.2, we use the model’s initial predictions to identify potentially misrepresented classes that could lead to misclassifications, i.e., are ambiguous (Fig. 2a). In Section 3.3, we use large language models to generate class descriptions that explain the visually differentiating information for the ambiguous classes (Fig. 2b). In Section 3.4, we use these differential descriptions in a follow-up classification task to resolve the initial ambiguity (Fig. 2c).

3.1 BACKGROUND

Following previous work (Radford et al., 2021), given a set of classes, C , and a set of descriptions, D_c , for each class, we calculate the class embeddings as:

$$h_c = \frac{1}{|D_c|} \sum_{d \in D_c} \phi_T(d),$$

where ϕ_T is the VLM text encoder, and h_c is the embedding vector for class c . Since VLMs are trained to minimize the distance between related image-text pairs, we select the closest class to image embedding $\phi_I(x)$ as the label for image x , where ϕ_I is the VLM image encoder.

3.2 DETECTING AMBIGUOUS CLASSES

Enumerating the differences between all class pairs is prohibitive for large datasets with thousands of classes. Instead, we focus on a small subset of potentially ambiguous classes that can lead to misclassifications. For example, in Fig. 2a, the model is confident that boxer (a dog breed) is not the label. However, any of the three most similar classes (british shorthair, bengal, and abyssinian cat) is likely to be the true label. Therefore, differentiating visual information for these classes is sufficient for selecting the correct label. For an image x , we define the set of ambiguous classes C_A , as the k most similar classes:

$$C_A = \arg \max_{\{c_1, \dots, c_k\} \subseteq C} \sum_{c_i} \cos(\phi_I(x), h_{c_i}),$$

where ϕ_I is the VLM image encoder, and \cos is the cosine similarity operator.

3.3 DIFFERENTIAL DESCRIPTIONS

To help the model distinguish between ambiguous classes, we generate a set of class descriptions that explain their visual differences. We take advantage of the extended world knowledge of LLMs to generate such descriptions at scale. Despite being uni-modal, LLMs acquire knowledge of the visual world through massive pre-training datasets (Jiang et al., 2020; Petroni et al., 2019). For example, an LLM can learn how a sparrow looks by reading the related Wikipedia page.

For each pair of ambiguous classes, we condition the LLM to select the visually differentiating attributes and describe them for both classes. We use the in-context learning capabilities of LLMs (Brown et al., 2020) to guide the model to focus on visual characteristics by providing two fixed examples as part of the prompt. Similarly, we guide the LLM to generate descriptions that resemble photo captions, which is shown to better adapt to VLMs’ pre-training distributions (Radford et al., 2021) (refer to appendix for more details). We use the following prompt template:

Table 1: Accuracy of FuDD in comparison with baselines. B/32 and L/14* represent the ViT-B/32 and ViT-L/14@336px vision backbones. Δ Naive(k) is the improvement of FuDD with k ambiguous classes over the Naive LLM-generated descriptions proposed by Menon & Vondrick (2023).

Description	Cub		DTD		EuroSAT		FGVCAircraft		Flowers102		Food101	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Single Template	51.21	63.48	43.14	54.04	40.87	56.82	20.88	37.08	63.80	75.12	82.63	93.49
Template Set	51.52	64.07	42.71	55.32	46.76	54.27	21.15	38.31	63.44	74.14	83.16	93.77
Naive LLM	52.92	65.15	45.90	55.37	44.18	46.69	21.09	38.79	66.12	75.98	84.02	94.26
FuDD ($k=10$)	53.97	65.90	45.43	57.66	45.18	60.64	21.87	38.82	67.80	78.76	84.05	94.05
FuDD ($k= C $)	54.30	66.03	44.84	57.23	45.18	60.64	22.32	39.63	67.62	79.67	84.36	94.27
Δ Naive ($k=10$)	$\uparrow 1.05$	$\uparrow 0.75$	$\downarrow -0.47$	$\uparrow 2.29$	$\uparrow 1.00$	$\uparrow 13.95$	$\uparrow 0.78$	$\uparrow 0.03$	$\uparrow 1.68$	$\uparrow 2.78$	$\uparrow 0.03$	$\downarrow -0.21$
Δ Naive ($k= C $)	$\uparrow 1.38$	$\uparrow 0.88$	$\downarrow -1.06$	$\uparrow 1.86$	$\uparrow 1.00$	$\uparrow 13.95$	$\uparrow 1.23$	$\uparrow 0.84$	$\uparrow 1.50$	$\uparrow 3.69$	$\uparrow 0.34$	$\uparrow 0.01$

	ImageNet		ImageNet V2		Oxford Pets		Places365		Stanford Cars		Stanford Dogs	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Single Template	62.04	74.85	54.77	68.79	84.98	92.86	39.10	40.70	60.37	78.06	58.01	73.61
Template Set	63.37	76.54	55.91	70.85	84.55	92.70	40.91	42.54	60.38	79.12	57.79	74.01
Naive LLM	63.52	76.37	55.96	70.47	83.76	93.08	40.58	41.43	59.63	77.90	57.86	74.02
FuDD ($k=10$)	64.05	76.70	56.62	70.60	86.92	93.40	42.12	43.95	60.86	78.25	60.03	75.99
FuDD ($k= C $)	64.19	77.00	56.75	71.05	89.34	93.51	42.17	44.09	61.46	78.96	60.28	76.34
Δ Naive ($k=10$)	$\uparrow 0.53$	$\uparrow 0.33$	$\uparrow 0.66$	$\uparrow 0.13$	$\uparrow 3.16$	$\uparrow 0.32$	$\uparrow 1.54$	$\uparrow 2.52$	$\uparrow 1.23$	$\uparrow 0.35$	$\uparrow 2.17$	$\uparrow 1.97$
Δ Naive ($k= C $)	$\uparrow 0.67$	$\uparrow 0.63$	$\uparrow 0.79$	$\uparrow 0.58$	$\uparrow 5.58$	$\uparrow 0.43$	$\uparrow 1.59$	$\uparrow 2.66$	$\uparrow 1.83$	$\uparrow 1.06$	$\uparrow 2.42$	$\uparrow 2.32$

For the following objects, generate captions that represent the distinguishing visual differences between the photos of the two objects. Generate as many captions as you can.

Object 1: {class name 1}

Object 2: {class name 2}

Following the provided samples, the model generates several responses similar to:

Visual characteristic: Bill color

Caption 1: A photo of a black-footed albatross, with a yellow bill.

Caption 2: A photo of a laysan albatross, with a pink bill.

Given a pair of classes c_1 and c_2 , we define the pairwise differential descriptions for class c_1 , $D_{c_1}^{c_2}$, as all the values for `Caption 1` in the LLM response, and similarly define $D_{c_2}^{c_1}$. As a result, $D_{c_1}^{c_2}$ contains all the descriptions that visually distinguish c_1 from c_2 . For each ambiguous class c , we combine all its pairwise descriptions to obtain the set of differential descriptions D'_c

$$D'_c = \bigcup_{c_i \in C_A \setminus \{c\}} D_c^{c_i}.$$

The new set of differential descriptions, D'_c , contains all the information necessary for separating class c from other ambiguous classes.

3.4 FOLLOW-UP CLASSIFICATION

Since this visually differentiating information resolves the initial ambiguity, after the first round of classification based on the original class descriptions, we create a follow-up classification task with only the ambiguous classes, C_A , and the new differential descriptions, D'_c . Finally, we follow the steps in Section 3.1 to predict the label.

4 EXPERIMENTS

In this section, we show the effectiveness of FuDD through extensive experiments. We show that FuDD outperforms both generic and naive LLM-generated description ensembles. We design further

Table 2: Accuracy of differential and non-differential descriptions for ambiguous classes. B/32 and L/14* represent the ViT-B/32 and ViT-L/14@336px vision backbones. Δ is the improvement of differential over non-differential descriptions.

Descriptor	CUB		DTD		FGVCAircraft		Flowers102		Food101	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Differential	53.62	65.79	45.37	56.91	22.17	39.06	67.62	79.54	84.17	94.34
Non-Differential	52.28	64.38	42.82	56.44	22.14	36.90	65.73	77.74	83.92	94.02
Δ	$\uparrow 1.35$	$\uparrow 1.42$	$\uparrow 2.55$	$\uparrow 0.47$	$\uparrow 0.03$	$\uparrow 2.16$	$\uparrow 1.89$	$\uparrow 1.81$	$\uparrow 0.25$	$\uparrow 0.32$

	Oxford Pets		Places365		Stanford Cars		Stanford Dogs	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Differential	87.24	93.68	42.45	44.26	60.90	79.39	60.31	75.96
Non-Differential	86.24	93.62	41.73	43.98	60.74	78.55	59.30	75.41
Δ	$\uparrow 1.01$	$\uparrow 0.06$	$\uparrow 0.73$	$\uparrow 0.28$	$\uparrow 0.16$	$\uparrow 0.85$	$\uparrow 1.01$	$\uparrow 0.55$

analytical experiments to show that not all semantic information resolves class ambiguities, and effective class descriptions should provide information that differentiates the ambiguous classes. Additionally, we find that describing the differences between highly ambiguous classes is the most important, accounting for most of FuDD’s performance gains.

Datasets. We evaluate our method on 12 image recognition datasets. We use the CUB200-2011 (Wah et al., 2011) (fine-grained bird species), Describable Textures Dataset (DTD) (Cimpoi et al., 2014) (texture classification), EuroSAT (Helber et al., 2019) (satellite image classification), FGVCAircraft (Maji et al., 2013) (aircraft model classification), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), ImageNet (Deng et al., 2009), ImageNetV2 (Kornblith et al., 2019), Oxford IIIT Pets (Parkhi et al., 2012), Places365 (Zhou et al., 2017), Stanford Cars (Krause et al., 2013), and Stanford Dogs (Khosla et al., 2011) datasets.

Setup. We use an instruction-tuned GPT-3 model (Brown et al., 2020; Ouyang et al., 2022), `gpt-3.5-turbo-0301`, which is available through OpenAI API³ as our LLM, and CLIP (Radford et al., 2021) as our VLM (refer to appendix for results with other VLMs). **ImageNet Descriptions:** Because of the large number of classes in ImageNet, to accommodate the API limitations, we cache the pairwise descriptions only for ambiguous classes detected by the ViT-B/32 backbone. In all experiments, we limit the available differential descriptions to these cached values (refer to appendix for details).

Baselines. We use three baselines. **Single Template:** to adapt to CLIP’s pre-training distribution, we adopt `A photo of a {class name}.` as the class description (Radford et al., 2021). **Template Set:** we use the 80 generic templates proposed by Radford et al. (2021) to study the benefits of FuDD’s better semantic information beyond simple prompt ensembling. **Naive LLM:** we follow Menon & Vondrick (2023) to create naive LLM-generated descriptions with the same LLM as ours, which uses a prompt like `What are useful features for distinguishing a {class name} in a photo? with a few in-context examples.`

4.1 RESULTS

The benefits of using FuDD’s semantic information exceed simple description ensembling (Table 1). When descriptions are provided for all classes ($k=|C|$), FuDD outperforms the generic template set on 11 out of 12 datasets with ViT-B/32 (base) and ViT-L/14@336px (large) backbones. Moreover, FuDD is more effective than naive LLM-generated descriptions at resolving class ambiguities. On average, FuDD outperforms naive LLM-generated descriptions by 1.44% and 2.41% with base and large vision backbones, respectively, with up to 13.95% improvements on EuroSAT. Notably, when using the base vision backbone, naive LLM-generated descriptions perform worse than the generic template set on the FGVCAircraft, Oxford Pets, Places365, and Stanford Cars datasets. On the other

³<https://openai.com/blog/openai-api>

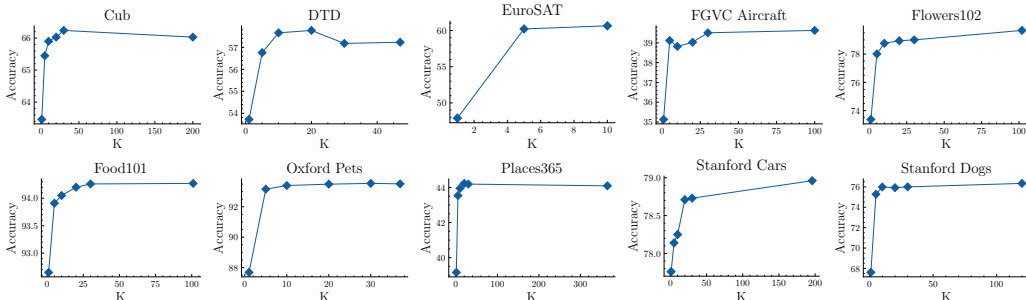


Figure 3: Impact of differential descriptions for k most ambiguous classes with ViT-L/14@336px. $k=1$ is accuracy with a single template. Providing differentiating details for the most ambiguous classes accounts for most of FuDD’s gains, with diminishing gains for less ambiguous classes.

hand, FuDD improves accuracy compared to the generic template set by providing differentiating information that resolves ambiguities. Importantly, we observe similar improvements with $k=10$, where FuDD only describes the differences between the 10 most ambiguous classes. This emphasizes the significant impact of class ambiguities on accuracy, which allows computational efficiency, that we will discuss in more detail later.

Importance of Differential Descriptions To study the importance of differentiating information, we compare differential descriptions with non-differential descriptions, which describe characteristics that do not separate the ambiguous classes. We select non-differential descriptions from attributes not used by differential descriptions. To control for the number of descriptions, we augment the descriptions with approx. 80 prefixes like image and snapshot, with minimal impact on semantic information (refer to appendix for details). As shown in Table 2, non-differential descriptions perform worse than differential descriptions. Non-differential descriptions lead to lower accuracy by at least 1% for six datasets, with up to 2.16% and 2.55% for FGVC Aircraft and DTD. These results confirm that not all semantic information resolves class ambiguities, and effective class descriptions like FuDD should provide the necessary information to differentiate the ambiguous classes.

Role of Class Ambiguities Since FuDD mainly focuses on ambiguous classes, here, we examine the importance of resolving class ambiguities. Figure 3 plots the accuracy of FuDD for the k most ambiguous classes for different values of k , which corresponds to varying levels of ambiguity (Section 3.2). $k = 1$ is accuracy with a single template. We find that describing the differences between the five most ambiguous classes accounts for most of FuDD’s performance gains, with diminishing benefits for less ambiguous classes. We can thus get most of the benefit while avoiding high computational costs, especially in the case of diverse datasets and open-set problems.

5 PUBLICLY AVAILABLE LANGUAGE MODELS

The inaccessibility of proprietary LLMs like GPT-3.5 hinders further research into fine-tuning LLMs for visual classification. Here, we use publicly available LLMs to generate the descriptions and study the impact of fine-tuning. Specifically, we fine-tune the 7b-parameter Llama 2 model (Touvron et al., 2023) on the descriptions generated by GPT-3.5 for the ImageNet dataset. We find that even the original Llama 2 model provides useful semantic information for visual classification. Moreover, fine-tuning improves Llama 2 performance, especially for rare concepts like satellite images, achieving comparable performance to GPT-3.5.

Table 3: The percentage of descriptions in a random sample that are correct and visually differentiating for EuroSAT.

Model	Correct	Useful
Llama 2	82.26	19.35
Llama 2 FT	90.52	48.28

As reported in Table 4, the original Llama 2 provides helpful semantic information for everyday objects like flowers, outperforming the generic template set. However, it struggles to describe the visual differences for datasets with rare objects like EuroSAT or abstract concepts like DTD. Although the fine-tuned model is not trained on test datasets, it learns the structure of the task and provides differ-

Table 4: FuDD’s accuracy with Llama 2 generated descriptions before and after fine-tuning. B/32 and L/14* represent the ViT-B/32 and ViT-L/14@336px vision backbones.

Model	Cub		DTD		EuroSAT		FGVCAircraft	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Llama 2 ($k=10$)	53.14	64.12	41.91	54.47	27.71	37.78	21.03	38.64
Llama 2 ($k= C $)	53.33	64.65	41.91	54.89	27.71	37.78	20.76	37.86
Llama 2 FT ($k=10$)	53.45	64.81	42.66	56.54	39.14	61.19	22.14	38.31
Llama 2 FT ($k= C $)	53.37	64.43	43.14	56.17	39.14	61.19	22.44	39.57

Model	Flowers102		Food101		Oxford Pets		Stanford Cars	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Llama 2 ($k=10$)	66.03	77.88	83.54	94.00	87.19	93.19	60.27	77.53
Llama 2 ($k= C $)	66.16	78.00	84.08	94.15	89.34	93.24	60.48	77.95
Llama 2 FT ($k=10$)	65.98	77.67	84.32	94.28	86.05	92.67	60.07	78.20
Llama 2 FT ($k= C $)	66.55	77.51	84.52	94.25	87.49	92.31	61.05	79.06

entiating visual information for these datasets. Using the ViT-L/14@336px backbone, fine-tuning improves the accuracy by 3.25% on average, with up to 23.41% for EuroSAT.

To better understand the impact of fine-tuning, we manually evaluate a random subset of pairwise differential descriptions before and after fine-tuning. Through visual inspection, for each pair, we check 1) if each description is correct and 2) if the pair helps differentiate the images of the two classes. Although fine-tuning helps with both measures, it significantly improves the usefulness of the descriptions: as shown in Table 3, after fine-tuning, 48% of pairwise descriptions help differentiate the two classes, compared to only 19% before fine-tuning. As illustrated in Fig. 5, unlike the original model, the fine-tuned model describes attributes that are more diverse and focused on low-level visual features rather than higher-level semantic concepts. For a more robust analysis, Fig. 4 plots the top-5 most common attributes before and after fine-tuning for EuroSAT. Similarly, the original model describes a limited set of attributes with mostly high-level semantic information, while the fine-tuned model generates a diverse set of visually differentiating attributes based on the input classes.

6 FUDD VS. OTHER AUXILIARY INFORMATION

To put the importance of differential descriptions in perspective, we compare FuDD against other approaches for VLM adaptation. We show that FuDD provides more helpful information through differential descriptions compared to simple heuristics like using high-level category names. We find that FuDD can better use the potential of natural language descriptions and achieve comparable performance to other methods that use text-to-image generation models or labeled samples in low-shot settings.

Few-Shot Description Selection We compare FuDD against LaBo, an alternative method that uses few-shot learning to select a subset of naive LLM-generated class descriptions that are more discriminative (Yang et al., 2023). As reported in Table 5, although FuDD is zero-shot and uses no

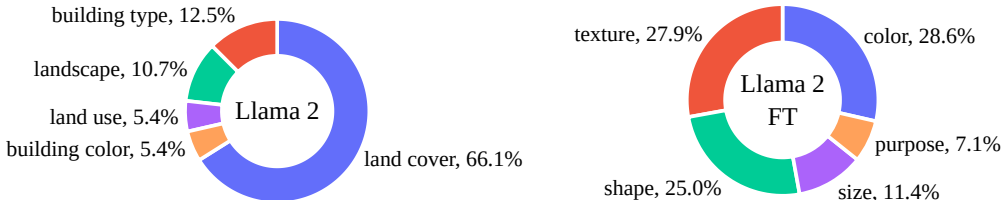


Figure 4: Top-5 most common attributes described by Llama 2 before and after fine-tuning. The fine-tuned model describes a more diverse and visually differentiating set of attributes.

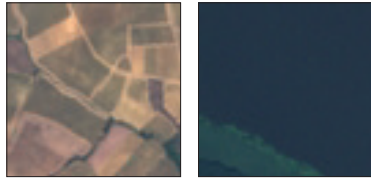
Table 5: Accuracy of FuDD compared to LaBo (Yang et al., 2023), which uses few-shot learning to select the most effective LLM-generated descriptions for each class with ViT-L/14 backbone. #S is the number of labeled samples.

Method	Cub		DTD		Aircraft		Flowers102		Food101		ImageNet	
	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S
FuDD($k=10$)	64.26	0	56.76	0	37.38	0	78.61	0	93.28	0	75.67	0
FuDD($k= C $)	64.14	0	57.07	0	37.89	0	79.48	0	93.40	0	75.99	0
LaBo	54.19	1	55.26	2	37.71	2	82.05	1	92.45	Full	72.60	16

labeled images, it performs better than 16-shot and full-shot (training on all samples) LaBo on ImageNet and Food101, respectively. For the other four datasets, FuDD’s performance is comparable to LaBo in low-shot scenarios. Unlike LaBo, FuDD encourages the descriptions to be discriminative as part of the generation process, eliminating the need for further optimization.

High Level Concepts WaffleCLIP (Roth et al., 2023) uses high-level category names to address class ambiguities by specifying the dataset context. As shown in Table 6 in appendix, FuDD performs better than WaffleCLIP for seven of the eight datasets. Although high-level category information is helpful, the additional details provided by FuDD are necessary to resolve more complex class ambiguities beyond what is caused by similar class names.

Additional Images In Table 6 in appendix, we compare FuDD with CoOp (Zhou et al., 2022b), which uses additional labeled images to learn a set of parameters as part of class descriptions. Without using any images, FuDD performs better than 16-shot CoOp on Food101 and Oxford Pets datasets and better than 4-shot CoOp on the ImageNet dataset. On the other five datasets, FuDD’s performance is comparable to CoOp in low-shot settings. We also compare FuDD with SuS-X (Udandarao et al., 2022), which avoids the additional labeled images by using a pre-trained LLM (Brown et al., 2020) and a text-to-image generation model (Rombach et al., 2022) to generate additional images for each class. As reported in Table 6 in appendix, despite using no images, FuDD achieves a performance comparable to SuS-X by only relying on the LLM-generated descriptions. FuDD uses the potential of natural language more effectively through differential descriptions and achieves comparable performance without additional labeled data or complexities of using text-to-image generation models.



Class name	- Permanent crop land	- Lake or sea
Llama 2	- A type of land cover with cropland	- A type of land cover with water
Llama 2 FT	- Brownish color - Rough texture - With geometric shape - Shows agricultural fields and irrigation systems	- Blueish color - Smooth texture - With natural, irregular shape - Shows islands and coastlines

Figure 5: Descriptions generated by Llama 2 before and after fine-tuning.

7 CONCLUSION

In this work, we introduce FuDD, a novel zero-shot approach that uses natural language to provide vision-language models with differentiating information about classes in downstream image recognition tasks. FuDD identifies a potentially ambiguous subset of classes and uses a large language model to generate visually differentiating descriptions that resolve the ambiguity. We show that not all information helps resolve class ambiguities, and effective descriptions should provide discriminative information about the ambiguous classes. Well-designed class descriptions, such as the ones produced by FuDD, can achieve comparable performance to few-shot prompt tuning methods in low-shot settings. Our results uncover the potential of natural language for tailoring the class representations to each dataset by providing differentiating information about ambiguous classes. These results motivate future work on creating effective natural language class descriptions for each downstream task.

ACKNOWLEDGEMENTS

This material is based on research sponsored by Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL) under agreement number FA8750-19-2-1006. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL) or the U.S. Government. We gratefully acknowledge support from Google and Cisco. Disclosure: Stephen Bach is an advisor to Snorkel AI, a company that provides software and services for data-centric artificial intelligence.

REFERENCES

- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.
- Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.
- Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.
- 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*, pp. 248–255. Ieee, 2009.
- Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.
- Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 746–754, 2023.
- Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

- Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>. If you use this software, please cite it as below.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.
- Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.
- Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.
- Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.
- Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.
- Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=jlAajNL8z5cs>.
- M Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Mateusz Kozinski, Horst Possegger, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *arXiv preprint arXiv:2305.18287*, 2023.
- Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*, 2022.
- Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.
- Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pp. 26342–26362. PMLR, 2023.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Devi Parikh and Kristen Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR 2011*, pp. 1681–1688. IEEE, 2011.
- Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.
- Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.
- Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091, 2022.
- 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*, pp. 10684–10695, 2022.
- Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pp. 2152–2161. PMLR, 2015.
- Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. *arXiv preprint arXiv:2306.07282*, 2023.
- Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.
- Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198*, 2022.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

- Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.
- Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023.
- Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.
- Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aweek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.
- Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.
- Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8552–8562, 2022.
- Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15211–15222, 2023a.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.
- Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Table 6: Accuracy of FuDD compared to other approaches that use high-level category names, WaffleCLIP (Roth et al., 2023), labeled samples, CoOp (Zhou et al., 2022b), and text-to-image generation models, SuS-X (Udandarao et al., 2022), with ResNet-50 He et al. (2016) backbone. #S is the number of labeled samples. *SuS-X-SD uses synthetically generated images.

Method	Cub		DTD		EuroSAT		Aircraft		Flowers102		Food101	
	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S
FuDD($k=10$)	49.45	0	43.51	0	39.42	0	19.77	0	67.39	0	80.65	0
FuDD($k= C $)	49.26	0	43.51	0	39.42	0	19.92	0	68.76	0	80.95	0
WaffleCLIP	48.34	0	39.25	0	35.08	0	-	-	-	-	81.38	0
CoOp	-	-	44.39	1	50.63	1	18.68	2	68.12	1	74.67	16
SuS-X-SD*	49.10	2	51.00	4	47.69	15	19.92	79	67.32	31	77.02	34

	ImageNet		ImageNetV2		Oxford Pets		Places365		Stanford Cars	
	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S	Acc.	#S
FuDD($k=10$)	60.69	0	53.19	0	86.86	0	40.64	0	56.62	0
FuDD($k= C $)	60.78	0	53.60	0	87.52	0	40.69	0	56.77	0
WaffleCLIP	60.12	0	52.89	0	85.80	0	39.03	0	-	-
CoOp	59.99	4	-	-	87.01	16	-	-	55.59	1
SuS-X-SD*	61.65	36	-	-	85.09	71	-	-	57.14	5

A DIFFERENT VISION ENCODERS

In general, the benefits of FuDD over the generic template set are more significant for smaller models (Fig. 6a). However, larger vision encoders can better take advantage of the nuanced information provided by FuDD beyond naive LLM-generated descriptions (Fig. 6b). We believe as image-text representations improve, VLMs can better take advantage of available semantic information, making natural language class descriptions even more important for vision tasks in the future.

B LLM PROMPTING DETAILS

We use `gpt-3.5-turbo-0301` to generate the differential descriptions. `gpt-3.5-turbo-0301` is a GPT-3 model that is fine-tuned to follow instructions (Brown et al., 2020; Ouyang et al., 2022). In this form, the model is given a sequence of user and assistant messages and is expected to generate the next assistant message. In our experiments, we encode two fixed sample outputs as assistant messages and ask the model to generate similar output for a given pair of classes. Specifically, we use the template messages in Table 9, and replace `class name 1` and `class name 2` with the desired classes to get the corresponding pairwise differential descriptions.

ImageNet Descriptions. As mentioned in Section 4, because of the large number of classes in the ImageNet dataset, it is not possible to generate the differential descriptions for all class pairs. Since

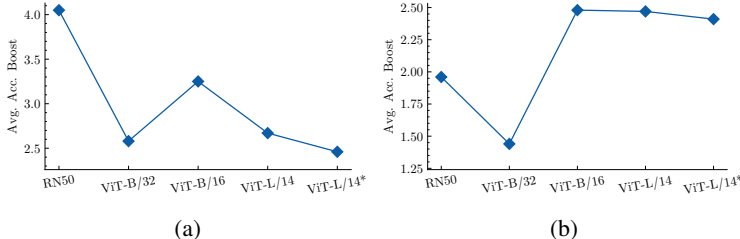


Figure 6: FuDD’s average accuracy boost for different vision backbones compared to a) generic template set and b) naive LLM-generated descriptions

Table 7: Total API costs for each dataset in Dollars

Dataset	Cost
Cub	15.47
DTD	0.84
EuroSAT	0.03
FGVCAircraft	3.85
Flowers	4.00
Food101	3.93
ImageNet	60.24
Pets	0.52
Places365	51.63
Stanford Cars	14.85
Stanford Dogs	5.55

differential descriptions are the most effective for ambiguous classes, we generate and cache all the pairwise descriptions for ambiguous classes and limit the available pairwise differential descriptions to this set. Following Section 3.2, we use ViT-B/32 vision backbone with $k = 5$ to detect the most ambiguous classes. For all pairs in each set of ambiguous classes, we generate and cache the corresponding pairwise differential descriptions, $D_c^{c_i}$, as explained in Section 3.3. In all our experiments, we use the cached pairwise differential descriptions, $D_c^{c_i}$. If the cache does not exist for a class pair, we use a single template description instead, i.e., $D_c^{c_i} = \{\text{A photo of a class name.}\}$.

C DIFFERENTIAL VS. NON-DIFFERENTIAL DETAILS

For the non-differential experiments in Table 2, we select a set of descriptions that describe a set of details for each class that does not differentiate it from other ambiguous classes. For some class c , to create non-differential descriptions, we first collect all the available differential descriptions, which is equal to the differential descriptions with $K = |C|$. Next, we create the normal set of differential descriptions, D'_c , with $K = 10$ as explained in Section 3. As a result of our description generation method, we know the corresponding attribute for each of the differential descriptions. Now, we filter the set of all available differential descriptions to exclude all the descriptions that their corresponding attribute is similar to the attributes explained by descriptions in D'_c . The remaining descriptions do not include any attribute that helps to separate the ambiguous classes.

We consider two attributes to be similar if they share a common word. For instance, `color` and `coat color` are similar since they share the word `color`. We split the attributes by white space and use simple string matching to check for this criteria. Because of the lack of diversity in the available descriptions for DTD, Oxford Pets, and Stanford Dogs datasets, using this criteria leads to a small number of remaining non-differential descriptions for each class. Therefore, for fair comparisons, we relax the criteria for these three datasets and compare attributes without splitting by white space, i.e., `color` and `coat color` are not considered similar for these three datasets.

As mentioned by previous work, the number of descriptions for each class also impacts the accuracy (Radford et al., 2021; Roth et al., 2023). We use description augmentation to control for the number of prompts for both differential and non-differential descriptions. Description augmentation creates a large number of descriptions from the original class descriptions with minimal impact on semantic information. Specifically, we create an augmented set of descriptions for each class description by replacing the original prefix (e.g. `a photo of a`) with a set of similar prefixes like `an image of a` and `a snapshot of a`. We use the prefixes used in Radford et al. (2021). Now to calculate the augmented description embedding, we average over the embeddings of all the descriptions in the corresponding augmented description set.

D COSTS

Although FuDD queries the LLM more than naive approaches, the API calls are very affordable and do not hinder wider adoption of FuDD. On average, one input prompt and model response

is 380 and 199 tokens, respectively. With OpenAI pricing at the time of writing (\$0.001/1k and \$0.002/1k tokens for input and response), the cost is \$0.78 per 1000 queries, leading to affordable prices as reported in Table 7. FuDD also accommodates datasets with a large number of classes like ImageNet by recognizing the more significant role of ambiguous classes, reducing the costs for ImageNet dataset from \$388 to \$60 (see Appendix B for details).

In addition, as studied extensively in Section 5, we can use off-the-shelf or fine-tuned LLMs like Llama 2 to generate differential descriptions using in-house hardware to avoid API costs or accommodate other issues like working with private and sensitive data.

E ADDITIONAL VLMS

To further evaluate FuDD, we repeat our main experiments with different VLMS. We choose OpenCLIP (Cherti et al., 2023; Ilharco et al., 2021) because of its superior performance to CLIP. For example, OpenCLIP with ViT-L/14 backbone trained on `datacomp_xl_s13b_b90k` improves the performance of its CLIP counterpart by 4 percentage points on ImageNet dataset using Single Template descriptions. We also run experiments using BLIP-2 (Li et al., 2023) because it is trained using an entirely different strategy with a combination of image-text contrastive loss, image-text matching loss, and generation loss. To calculate image-text similarity using BLIP-2, we follow a similar procedure to Li et al. (2023). As reported in Table 8, these other VLMS can also use the additional information provided by FuDD and improve the performance beyond naive LLM-generated descriptions.

Table 8: FuDD accuracy using OpenCLIP and BLIP2 models. DComp is the checkpoint trained on datacomp_xl_s13b_b90k and Laion is the checkpoint trained on laion2b_e16 dataset.

Description	Cub				DTD			
	ViT-B-32		ViT-L-14		ViT-B-32		ViT-L-14	
	DComp	Laion	DComp	BLIP2	DComp	Laion	DComp	BLIP2
Single Template	<u>72.63</u>	<u>63.93</u>	85.31	23.52	54.95	51.12	63.19	53.24
Template Set	72.52	63.12	85.00	27.36	55.80	52.71	64.79	55.53
Naive	<u>73.63</u>	63.79	85.66	27.51	58.99	56.86	66.22	55.48
FuDD ($k=10$)	73.42	63.91	<u>85.69</u>	<u>28.51</u>	58.24	55.43	<u>67.23</u>	56.70
FuDD ($k= C $)	73.82	64.19	86.16	28.77	57.45	54.95	67.77	56.86
	EuroSAT				FGVCAircraft			
	ViT-B-32		ViT-L-14		ViT-B-32		ViT-L-14	
	DComp	Laion	DComp	BLIP2	DComp	Laion	DComp	BLIP2
Single Template	38.03	41.73	61.14	52.14	29.94	26.31	51.82	14.19
Template Set	41.02	41.44	61.62	51.01	30.75	<u>26.46</u>	<u>51.88</u>	14.46
Naive	49.28	45.37	69.72	63.63	<u>31.41</u>	25.77	52.09	<u>16.35</u>
FuDD ($k=10$)	55.74	57.27	74.05	70.84	31.59	26.67	50.89	15.54
FuDD ($k= C $)	55.74	57.27	74.05	70.84	31.38	26.43	51.25	16.44
	Flowers102				Food101			
	ViT-B-32		ViT-L-14		ViT-B-32		ViT-L-14	
	DComp	Laion	DComp	BLIP2	DComp	Laion	DComp	BLIP2
Single Template	72.13	67.49	80.60	55.91	85.89	81.31	<u>94.49</u>	85.07
Template Set	72.06	67.67	81.12	57.85	85.75	81.24	94.39	86.74
Naive	73.18	66.53	79.90	60.81	85.49	81.48	94.05	87.71
FuDD ($k=10$)	<u>73.98</u>	<u>69.00</u>	83.35	61.60	<u>86.40</u>	<u>81.50</u>	94.43	88.50
FuDD ($k= C $)	75.05	69.91	<u>83.02</u>	<u>61.21</u>	86.49	81.86	94.51	88.63
	ImageNet				ImageNet V2			
	ViT-B-32		ViT-L-14		ViT-B-32		ViT-L-14	
	DComp	Laion	DComp	BLIP2	DComp	Laion	DComp	BLIP2
Single Template	68.44	65.20	78.83	60.93	60.33	56.91	71.96	56.20
Template Set	69.13	65.61	79.15	66.07	60.76	57.36	72.05	60.60
Naive	68.60	65.42	79.03	66.15	60.06	57.19	71.92	61.00
FuDD ($k=10$)	<u>69.13</u>	<u>65.91</u>	<u>79.25</u>	<u>67.31</u>	<u>60.94</u>	<u>57.70</u>	<u>72.08</u>	<u>61.28</u>
FuDD ($k= C $)	69.35	66.20	79.50	68.55	61.28	57.90	72.39	62.53
	Oxford Pets				Places365			
	ViT-B-32		ViT-L-14		ViT-B-32		ViT-L-14	
	DComp	Laion	DComp	BLIP2	DComp	Laion	DComp	BLIP2
Single Template	89.40	87.54	94.74	76.70	41.52	41.88	43.21	43.72
Template Set	88.47	87.49	93.51	76.91	43.20	42.84	44.73	43.67
Naive	89.86	<u>89.07</u>	94.79	81.17	42.24	42.61	44.06	43.51
FuDD ($k=10$)	<u>90.71</u>	89.04	<u>94.90</u>	81.96	42.79	<u>43.16</u>	44.98	<u>45.30</u>
FuDD ($k= C $)	90.95	90.00	95.18	83.51	<u>43.13</u>	43.55	<u>44.89</u>	45.52
	Stanford Cars				Stanford Dogs			
	ViT-B-32		ViT-L-14		ViT-B-32		ViT-L-14	
	DComp	Laion	DComp	BLIP2	DComp	Laion	DComp	BLIP2
Single Template	88.42	86.82	93.67	79.97	63.92	59.76	79.23	47.70
Template Set	88.66	87.02	<u>93.71</u>	<u>80.48</u>	64.93	59.50	79.22	47.76
Naive	87.38	86.76	93.56	80.08	65.02	<u>60.12</u>	79.22	50.31
FuDD ($k=10$)	88.35	<u>86.87</u>	93.77	80.25	<u>65.13</u>	59.84	<u>79.39</u>	<u>52.42</u>
FuDD ($k= C $)	<u>88.45</u>	86.84	93.65	81.06	65.96	60.22	80.29	52.91


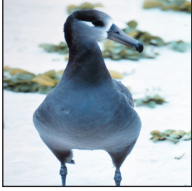



<p>Class 1</p>	<p>Class 0: Tennessee Warbler</p> 
<p>Black-footed Albatross</p> 	<p>Attribute: size 0: A photo of a tennessee warbler, a small songbird that is only about 4 inches long. 1: A photo of a black-footed albatross, a large seabird with a wingspan of up to 7 feet.</p> <p>Attribute: coloration 0: A photo of a tennessee warbler, a bright yellow bird with olive-green wings and back. 1: A photo of a black-footed albatross, a dark-colored bird with a white head and underparts.</p> <p>Attribute: bill shape 0: A photo of a tennessee warbler, a bird with a small, pointed bill. 1: A photo of a black-footed albatross, a bird with a large, hooked bill.</p>
<p>Mangrove Cuckoo</p> 	<p>Attribute: bill length 0: A photograph of a tennessee warbler, a type of bird, with a short, pointed bill. 1: A photograph of a mangrove cuckoo, a type of bird, with a long, curved bill.</p> <p>Attribute: tail length 0: A photograph of a tennessee warbler, a type of bird, with a short, square tail. 1: A photograph of a mangrove cuckoo, a type of bird, with a long, graduated tail.</p> <p>Attribute: wing bars 0: A photograph of a tennessee warbler, a type of bird, with two white wing bars. 1: A photograph of a mangrove cuckoo, a type of bird, with no wing bars.</p>
<p>Caspian Tern</p> 	<p>Attribute: size 0: A photograph of a tennessee warbler, a type of bird, which is smaller in size. 1: A photograph of a caspian tern, a type of bird, which is larger in size.</p> <p>Attribute: bill color 0: A photograph of a tennessee warbler, a type of bird, with a black bill. 1: A photograph of a caspian tern, a type of bird, with a red bill.</p> <p>Attribute: wing shape 0: A photograph of a tennessee warbler, a type of bird, with short, rounded wings. 1: A photograph of a caspian tern, a type of bird, with long, pointed wings.</p>
<p>Black-throated Sparrow</p> 	<p>Attribute: head pattern 0: A photograph of a tennessee warbler, a type of bird, with yellow and gray head pattern. 1: A photograph of a black-throated sparrow, a type of bird, with black and white striped head pattern.</p> <p>Attribute: eye ring 0: A photograph of a tennessee warbler, a type of bird, with no eye ring. 1: A photograph of a black-throated sparrow, a type of bird, with white eye ring.</p> <p>Attribute: breast color 0: A photograph of a tennessee warbler, a type of bird, with yellow breast color. 1: A photograph of a black-throated sparrow, a type of bird, with black and white striped breast color.</p>

Figure 7: Various pairwise differential descriptions for sample class “Tennessee Warbler” and four different classes.

Table 9: Prompt template for generating pairwise differential descriptions. We use two fixed sample pairs of classes to guide the model to generate the appropriate output.

Type	Content
System	You are a knowledgeable teacher.
User	<p>For the following objects, generate captions that represent the distinguishing visual differences between the photos of the two objects. Generate as many captions as you can.</p> <p>Object 1: Field sparrow Object 2: Acadian flycatcher</p>
Assistant	<p><START> Visual characteristic: Bill shape Caption 1: A photograph of a field sparrow, a type of bird, with cone bill shape. Caption 2: A photograph of a acadian flycatcher, a type of bird, with all-purpose bill shape.</p> <p>Visual characteristic: Breast color Caption 1: A photograph of a field sparrow, a type of bird, with buff breast color. Caption 2: A photograph of a acadian flycatcher, a type of bird, with white breast color.</p> <p>Visual characteristic: Back pattern Caption 1: A photo of a field sparrow, a type of bird, with striped back pattern. Caption 2: A photo of a acadian flycatcher, a type of bird, with solid back pattern. <END></p>
User	<p>For the following objects, generate captions that represent the distinguishing visual differences between the photos of the two objects. Generate as many captions as you can.</p> <p>Object 1: Cornet Object 2: Flute</p>
Assistant	<p><START> Visual characteristic: Shape Caption 1: A photo of a cornet, a type of musical instrument, with a conical bore. Caption 2: A photo of a flute, a type of musical instrument, with a cylindrical bore. <END></p>
User	<p>For the following objects, generate captions that represent the distinguishing visual differences between the photos of the two objects. Generate as many captions as you can.</p> <p>Object 1: {class name 1} Object 2: {class name 2}</p>

Table 10: Sample descriptions generated by GPT-3.5 for Cub



Classes	Descriptions
 <p>Caspian Tern</p>	<p>Attribute: size</p> <p>0: a photograph of a caspian tern, a type of bird, which is larger in size.</p> <p>1: a photograph of a tennessee warbler, a type of bird, which is smaller in size.</p> <p>Attribute: bill color</p> <p>0: a photograph of a caspian tern, a type of bird, with a red bill.</p> <p>1: a photograph of a tennessee warbler, a type of bird, with a black bill.</p>
 <p>Tennessee Warbler</p>	<p>Attribute: wing shape</p> <p>0: a photograph of a caspian tern, a type of bird, with long, pointed wings.</p> <p>1: a photograph of a tennessee warbler, a type of bird, with short, rounded wings.</p> <p>Attribute: tail shape</p> <p>0: a photograph of a caspian tern, a type of bird, with a forked tail.</p> <p>1: a photograph of a tennessee warbler, a type of bird, with a square tail.</p>

Table 11: Sample descriptions generated by GPT-3.5 for DTD

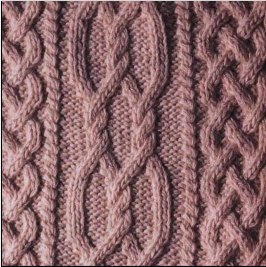

Classes	Descriptions
 <p>Interlaced</p>	<p>Attribute: texture</p> <p>0: a photo of an interlaced surface, with a woven texture.</p> <p>1: a photo of a stratified surface, with a layered texture.</p> <p>Attribute: pattern</p> <p>0: a photo of an interlaced surface, with a criss-cross pattern.</p> <p>1: a photo of a stratified surface, with a horizontal pattern.</p>
 <p>Stratified</p>	<p>Attribute: material</p> <p>0: a photo of an interlaced surface, made of woven fibers.</p> <p>1: a photo of a stratified surface, made of layered sediment or rock.</p>

Table 12: Sample descriptions generated by GPT-3.5 for EuroSAT



Classes	Descriptions
 Industrial or commercial building	<p>Attribute: color</p> <p>0: a satellite image of industrial or commercial buildings, with a mix of grey, white, and black colors.</p> <p>1: a satellite image of a river, with blue and green colors.</p> <p>Attribute: texture</p> <p>0: a satellite image of industrial or commercial buildings, with a rough and angular texture.</p> <p>1: a satellite image of a river, with a smooth and flowing texture.</p>
 River	<p>Attribute: shape</p> <p>0: a satellite image of industrial or commercial buildings, with rectangular and square shapes.</p> <p>1: a satellite image of a river, with a winding and curvy shape.</p> <p>Attribute: pattern</p> <p>0: a satellite image of industrial or commercial buildings, with a grid-like pattern.</p> <p>1: a satellite image of a river, with a meandering pattern.</p>

Table 13: Sample descriptions generated by GPT-3.5 for Flowers



Classes	Descriptions
 Bird of paradise	<p>Attribute: color</p> <p>0: a photo of a bird of paradise, a type of flower, with bright orange and blue colors.</p> <p>1: a photo of a globe thistle, a type of flower, with muted blue and green colors.</p> <p>Attribute: shape</p> <p>0: a photo of a bird of paradise, a type of flower, with a unique bird-like shape.</p> <p>1: a photo of a globe thistle, a type of flower, with a spherical shape.</p>
 Globe thistle	<p>Attribute: texture</p> <p>0: a photo of a bird of paradise, a type of flower, with smooth and glossy petals.</p> <p>1: a photo of a globe thistle, a type of flower, with spiky and rough texture.</p>

Table 14: Sample descriptions generated by GPT-3.5 for Food101



Classes	Descriptions
 <p>Hot dog</p>	<p>Attribute: type of food 0: a photo of a hot dog, a type of fast food, with a sausage in a bun. 1: a photo of sashimi, a type of japanese cuisine, with raw fish slices.</p> <p>Attribute: cooking method 0: a photo of a hot dog, a type of fast food, with a grilled sausage. 1: a photo of sashimi, a type of japanese cuisine, with raw fish slices.</p>
 <p>Sashimi</p>	<p>Attribute: serving style 0: a photo of a hot dog, a type of fast food, served with ketchup and mustard. 1: a photo of sashimi, a type of japanese cuisine, served with soy sauce and wasabi.</p> <p>Attribute: texture 0: a photo of a hot dog, a type of fast food, with a chewy texture. 1: a photo of sashimi, a type of japanese cuisine, with a soft and tender texture.</p>

Table 15: Sample descriptions generated by GPT-3.5 for ImageNet



Classes	Descriptions
 <p>Minivan</p>	<p>Attribute: number of wheels 0: a photo of a minivan, which has four wheels. 1: a photo of a rickshaw, which has three wheels.</p> <p>Attribute: size 0: a photo of a minivan, which is larger in size and can accommodate more passengers. 1: a photo of a rickshaw, which is smaller in size and can accommodate fewer passengers.</p>
 <p>Rickshaw</p>	<p>Attribute: propulsion 0: a photo of a minivan, which is powered by an engine. 1: a photo of a rickshaw, which is powered by human pedaling.</p> <p>Attribute: type of vehicle 0: a photo of a minivan, which is a modern automobile. 1: a photo of a rickshaw, which is a traditional asian vehicle.</p>

Table 16: Sample descriptions generated by GPT-3.5 for Oxford Pets



Classes	Descriptions
 <p>German shorthaired</p>	<p>Attribute: body type</p> <p>0: a photo of a german shorthaired, a type of dog, with a muscular and athletic body type.</p> <p>1: a photo of a sphynx, a type of cat, with a slender and sleek body type.</p> <p>Attribute: coat</p> <p>0: a photo of a german shorthaired, a type of dog, with a short and dense coat.</p> <p>1: a photo of a sphynx, a type of cat, with no coat or hair.</p>
 <p>Sphynx</p>	<p>Attribute: ears</p> <p>0: a photo of a german shorthaired, a type of dog, with floppy ears.</p> <p>1: a photo of a sphynx, a type of cat, with large and pointed ears.</p> <p>Attribute: facial features</p> <p>0: a photo of a german shorthaired, a type of dog, with a snout and a prominent nose.</p> <p>1: a photo of a sphynx, a type of cat, with a flat face and no visible nose bridge.</p>

Table 17: Sample descriptions generated by GPT-3.5 for Stanford Cars



Classes	Descriptions
 <p>2012 Hyundai Veracruz SUV</p>	<p>Attribute: body type</p> <p>0: a photo of a 2012 hyundai veracruz suv, a large vehicle with a high ground clearance and a boxy shape.</p> <p>1: a photo of a 2009 spyker c8 convertible, a sleek and low-slung sports car with a convertible top.</p> <p>Attribute: number of doors</p> <p>0: a photo of a 2012 hyundai veracruz suv, a vehicle with four doors.</p> <p>1: a photo of a 2009 spyker c8 convertible, a vehicle with two doors.</p>
 <p>2009 Spyker C8 Convertible</p>	<p>Attribute: wheel design</p> <p>0: a photo of a 2012 hyundai veracruz suv, with standard alloy wheels.</p> <p>1: a photo of a 2009 spyker c8 convertible, with unique and intricate spoke wheels.</p> <p>Attribute: grille design</p> <p>0: a photo of a 2012 hyundai veracruz suv, with a large and prominent grille.</p> <p>1: a photo of a 2009 spyker c8 convertible, with a small and distinctive grille.</p>

Table 18: Sample descriptions generated by GPT-3.5 for Stanford Dogs

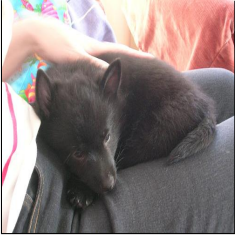

Classes	Descriptions
 <p>Schipperke</p>	<p>Attribute: size</p> <p>0: a photo of a schipperke dog, a small breed of dog.</p> <p>1: a photo of a saint bernard dog, a large breed of dog.</p> <p>Attribute: coat color</p> <p>0: a photo of a schipperke dog, with black coat color.</p> <p>1: a photo of a saint bernard dog, with white and brown coat color.</p>
 <p>Saint bernard</p>	<p>Attribute: ear shape</p> <p>0: a photo of a schipperke dog, with pointed ears.</p> <p>1: a photo of a saint bernard dog, with droopy ears.</p> <p>Attribute: tail length</p> <p>0: a photo of a schipperke dog, with a short tail.</p> <p>1: a photo of a saint bernard dog, with a long tail.</p>

Table 19: Sample descriptions generated by GPT-3.5 for FGVC Aircraft





Classes	Descriptions
 <p>Airbus A319</p>	<p>Attribute: size</p> <p>0: a photo of an airbus a319 aircraft, a commercial airliner that is much larger than a cessna 172.</p> <p>1: a photo of a cessna 172 aircraft, a small, single-engine plane that is much smaller than an airbus a319.</p> <p>Attribute: wing shape</p> <p>0: a photo of an airbus a319 aircraft, with swept-back wings.</p> <p>1: a photo of a cessna 172 aircraft, with straight wings.</p>
 <p>Cessna 172</p>	<p>Attribute: engine placement</p> <p>0: a photo of an airbus a319 aircraft, with engines mounted under the wings.</p> <p>1: a photo of a cessna 172 aircraft, with a single engine mounted on the nose of the plane.</p> <p>Attribute: cockpit windows</p> <p>0: a photo of an airbus a319 aircraft, with a large cockpit window that extends over the top of the plane.</p> <p>1: a photo of a cessna 172 aircraft, with a small, single-piece windshield in the cockpit.</p>

Table 20: Sample descriptions generated by GPT-3.5 for Places365

Classes	Descriptions
 <p data-bbox="410 1031 456 1056">Pier</p>	<p data-bbox="597 743 915 768">Attribute: location</p> <p data-bbox="597 772 1328 827">0: a photo of a pier, which is located near a body of water.</p> <p data-bbox="597 831 1312 886">1: a photo of a plaza, which is located in a city or town center.</p> <p data-bbox="597 919 899 945">Attribute: purpose</p> <p data-bbox="597 949 1360 1003">0: a photo of a pier, which is used for docking boats and ships.</p> <p data-bbox="597 1008 1360 1062">1: a photo of a plaza, which is used for public gatherings and events.</p>
 <p data-bbox="402 1430 467 1455">Plaza</p>	<p data-bbox="597 1092 883 1117">Attribute: design</p> <p data-bbox="597 1121 1328 1176">0: a photo of a pier, which is typically long and narrow with a flat surface.</p> <p data-bbox="597 1180 1295 1264">1: a photo of a plaza, which is typically open and spacious with various features like fountains, benches, and sculptures.</p> <p data-bbox="597 1297 980 1323">Attribute: surroundings</p> <p data-bbox="597 1327 1328 1402">0: a photo of a pier, which is surrounded by water and may have views of the ocean or other bodies of water.</p> <p data-bbox="597 1407 1360 1491">1: a photo of a plaza, which is surrounded by buildings and may have views of city streets and architecture.</p>