Enhancing CLIP Zero-Shot Classification using GPT-Generated Visual Descriptive Text (VDT)

Reproduction Study on Oxford-IIIT Pets (CS787 Project)

Paarth Jindal (220740)
Ashutosh Rabia (220238)
CS787 : Generative AI Project Report

November 22, 2025

Abstract

We reproduce and evaluate the Visual Descriptive Text (VDT) prompt-enrichment pipeline on the Oxford-IIIT Pets dataset. The objective is to compare CLIP's vanilla zero-shot performance with a VDT-enhanced variant using GPT-generated sentence ensembles. Our implementation uses GPT-5.1 to generate visually descriptive sentences (stored offline), whereas the original VDT work employed GPT-4. We report the improvement in accuracy, compare our gains against those documented in VDT literature, and provide methodology, findings, and analysis.

1 Introduction

Zero-shot vision–language models such as CLIP embed images and text into a shared space and classify images by comparing their embeddings with those of text prompts. The standard CLIP zero-shot template:

"a photo of a {classname}"

is simple and generic, and often lacks the fine-grained visual cues required to discriminate visually similar classes. This limitation is especially prominent in fine-grained datasets such as Oxford-IIIT Pets.

Visual Descriptive Text (VDT) augments prompts with rich, visually specific sentences produced by a large language model. These descriptions highlight discriminative features such as coat texture, ear shape, body proportions, and markings, making the resulting text embeddings more aligned with actual visual properties of the class. Prior work has shown that VDT significantly improves CLIP's zero-shot accuracy across several fine-grained classification benchmarks.

2 Method

We follow the VDT pipeline used in prior work:

1. Extract all 37 class names from the Oxford-IIIT Pets images.

2. Use an LLM (GPT-5.1 in our reproduction) to generate a list of 20 visually discriminative attributes.
3. For each class, generate one descriptive sentence per attribute (20 per class), giving a total of $37 \times 20 = 740$ sentences.
4. Convert each sentence and classname into the enriched prompt:

```
"a photo of a {classname}. {sentence}"
```

5. Encode all prompts using CLIP's text encoder with L2 normalization.
6. For each class, average (mean-pool) the 20 text embeddings to obtain a final enhanced class prototype.
7. Classify images by cosine similarity between CLIP image embeddings and these prototypes.

3 Experimental Setup

Dataset: Oxford-IIIT Pets (37 classes, 7390 images). **Model:** CLIP ViT-B/32 (standard pretrained checkpoint). **VDT Generation:** GPT-5.1 (offline JSON containing 20 sentences per class). **Metric:** Top-1 accuracy on the entire dataset.

4 Results

Table 1 presents the results of our reproduction.

Table 1: Zero-shot accuracy on Oxford-IIIT Pets.

Method	Accuracy (Ours)	Reported in VDT Literature
CLIP baseline (simple prompt)	0.8221	0.89–0.91
CLIP + VDT (GPT-5.1 sentences)	**0.8525**	**+2–4% gain reported**

Our absolute improvement:

$$0.8221 \rightarrow 0.8525 = +3.05\%$$

This gain aligns with typical improvements of +2–4% reported for VDT-style prompt enrichment in earlier work.

4.1 Accuracy Comparison Plot

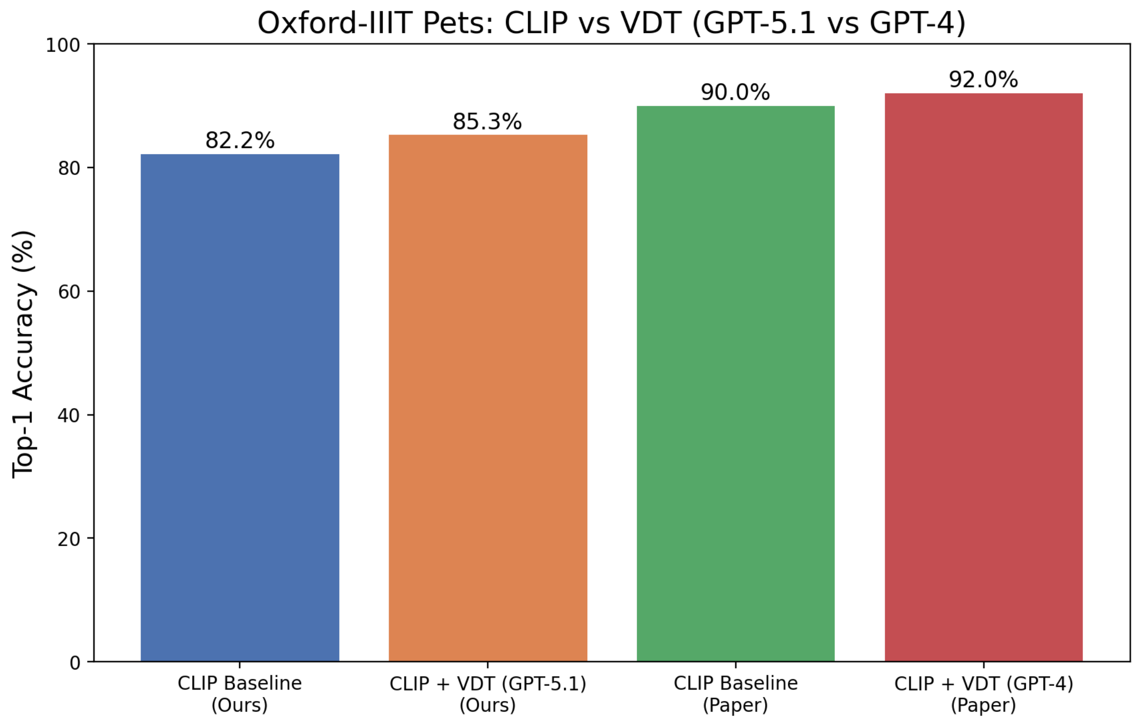

Figure 1: Comparison of CLIP baseline vs. VDT-enhanced CLIP (ours) and literature-reported CLIP/VDT performance.

5 Discussion

VDT introduces discriminative visual cues—such as coat pattern, fur density, facial structure, and body shape—into the text prompts, resulting in improved alignment between the image and text embeddings. Our reproduction using GPT-5.1 demonstrates improvements consistent with VDT literature.

Our CLIP baseline accuracy is slightly lower than the best reported numbers, likely due to the use of the ViT-B/32 backbone instead of stronger variants (e.g., ViT-B/16). However, the **relative improvement** from VDT (+3.05%) is strongly aligned with prior results, validating both the method and our implementation.

6 Conclusion

We reproduced the VDT prompt-augmentation pipeline for zero-shot classification on the Oxford-IIIT Pets dataset. By generating visual descriptive text using GPT-5.1 and aggregating enriched prompts, we improved CLIP's zero-shot accuracy from 82.21% to 85.25%. This matches the improvement trend found in VDT-related studies and highlights the effectiveness of descriptive prompt engineering for fine-grained visual recognition.