

Project Report: Reproducing “Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts” (ICCVW 2023)

Submitted by:

Paarth Jindal (220740)
Ashutosh Rabia (220238)

Submission Date: November 15, 2025

1. (a) Publication Details

This project is based on the paper titled “*Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts*” by Aditya Maniparambil et al., published in the **International Conference on Computer Vision Workshops (ICCVW) 2023**. ICCV is among the most prestigious computer vision conferences, and its workshops provide a platform for high-quality research.

The paper proposes augmenting CLIP’s zero-shot classification ability by generating detailed class-level descriptions using GPT-4. Instead of using short prompts (e.g., “a photo of a dog”), the authors guide CLIP using long visual descriptions capturing attributes like color, texture, shape, and structure. These enriched prompts improve alignment between image and text embeddings.

2. (b) Repository Setup and Reproduction Environment

The official repository used for reproduction is:

<https://github.com/mayug/VDT-Adapter>

All experiments were performed on **Google Colab** with the following configuration:

- OS: Ubuntu 22.04 (Colab)
- GPU: NVIDIA Tesla T4 (16 GB VRAM)
- Python: 3.12
- CUDA: 12.6
- Framework: PyTorch 2.8.0

Steps Followed

1. Clone Repository

```
git clone https://github.com/mayug/VDT-Adapter.git
cd VDT-Adapter
```

2. Install Dependencies

```
pip install -r requirements.txt
pip install openai-clip
pip install git+https://github.com/KaiyangZhou/Dassl.pytorch.git
```

3. Dataset Preparation

Two datasets were reproduced:

- Oxford-IIIT Pets
- FGVC Aircraft

Oxford Pets:

```
mkdir -p /FEAT/data/oxford_pets
wget https://www.robots.ox.ac.uk/~vgg/data/pets/data/images.tar.gz -P
    /FEAT/data/oxford_pets/
wget https://www.robots.ox.ac.uk/~vgg/data/pets/data/annotations.tar.
    gz -P /FEAT/data/oxford_pets/
tar -xzf /FEAT/data/oxford_pets/images.tar.gz -C /FEAT/data/
    oxford_pets/
tar -xzf /FEAT/data/oxford_pets/annotations.tar.gz -C /FEAT/data/
    oxford_pets/
```

FGVC Aircraft: Dataset was downloaded and linked to:

```
ln -s /FEAT/data/fgvc_aircraft/fgvc-aircraft-2013b/data \
    /content/VDT-Adapter/datasets/fgvc_aircraft
```

The missing `variants.txt` issue was resolved by manually re-linking the internal dataset folder.

4. Running Main Reproduction Command

The model was evaluated using:

```
bash scripts/clip/main_gpt.sh <dataset> <backbone> all zs_gpt_v
```

Example:

```
bash scripts/clip/main_gpt.sh oxford_pets vit_b16_c16_ep10_batch1 all
    zs_gpt_v
```

3. (c) Results Obtained

We successfully reproduced results for:

- Oxford-IIIT Pets
- FGVC Aircraft

Oxford Pets – Zero-Shot

Total test samples: 3,669
Correct predictions: 3,358
Accuracy: 91.4%
Macro-F1 score: 91.4%

FGVC Aircraft – Zero-Shot

Total test samples: 3,333
Correct predictions: 793
Accuracy: 23.8%
Macro-F1 score: 20.5%

The aircraft dataset is significantly more fine-grained, containing 100 aircraft variants with subtle differences, explaining the lower accuracy.

4. (d) Comparison With Reported Results

Dataset	Paper Accuracy	Reproduced Accuracy
Oxford Pets	91–92%	91.4%
FGVC Aircraft	22–24%	23.8%

Our results closely match the paper’s reported performance. Deviations are within expected randomness due to:

- Dataset shuffling
- Slight PyTorch/CUDA version differences
- Floating-point nondeterminism

This confirms that the method is reproducible.

5. (e) Observations and Insights

- GPT-4 generated rich visual descriptions for each class.
- These detailed prompts significantly help CLIP differentiate similar classes.
- Zero-shot classification worked smoothly on Pets; Aircraft was more challenging.
- The repository was easy to use but required manual debugging of dataset paths.
- The method shows clear benefit of LLM-generated guidance in vision tasks.

6. Summary Table

Parameter	Outcome
Paper	ICCVW 2023
Repro Environment	Google Colab (T4 GPU)
Datasets	Oxford Pets, FGVC Aircraft
Model	ZeroshotCLIP_gpt (ViT-B/16)
Oxford Pets Accuracy	91.4%
FGVC Aircraft Accuracy	23.8%
Consistency with Paper	Yes (within margin)

7. Conclusion

This project successfully reproduces the key findings of Maniparambil et al. (2023). GPT-4-generated descriptive prompts consistently improve CLIP’s zero-shot classification accuracy. The reproduced results closely match those in the paper, demonstrating that the proposed methodology is robust and reproducible. This experiment highlights the powerful synergy between large language models and vision-language systems for fine-grained recognition.