Text Meets Vision: A Deep Dive into CLIP Performance

Mingyue Zhao, Yuzhu Han, Skylar Shao

Summary

This project systematically evaluated the capabilities of OpenAI CLIP and OpenCLIP models for cross-modal retrieval and classification tasks. Motivated by the success of CLIP in zero-shot generalization, the study explored how model architecture, pretraining datasets, and prompt formulation affect performance. Both image-to-text classification and text-to-image retrieval experiments were conducted. Both image-to-text classification and text-to-image retrieval experiments were conducted. The image-to-text classification was performed across four diverse datasets: CIFAR-100, ImageNet-mini, Food101, and EuroSAT. For the text-to-image retrieval experiments, we additionally explored the COCO dataset along with these four datasets.

In classification tasks, vision transformer image encoders consistently outperformed convolutional encoders, and OpenCLIP models trained on larger datasets such as LAION-2B achieved higher zero-shot accuracy, particularly for general object recognition. However, linear probing significantly improved performance on domain-specific datasets like EuroSAT, highlighting the trade-offs between zero-shot flexibility and supervised adaptation.

In prompt-based retrieval experiments, variations in prompt style, subject specificity, and descriptive detail had a substantial impact on retrieval effectiveness. Both OpenAI CLIP and OpenCLIP demonstrated robustness to style and subject changes but remained highly sensitive to prompt specificity. Using human-written captions revealed that retrieval accuracy depends heavily on the concreteness of the language used.

Overall, the findings highlight the importance of dataset diversity, model architecture, and prompt design in maximizing the effectiveness of CLIP-based vision-language models. Future work will focus on quantitative retrieval evaluation, investigating model robustness to ambiguous prompts, and exploring few-shot tuning and automatic prompt optimization techniques.

Introduction

Recent advancements in vision-language models have significantly enhanced the ability of artificial intelligence systems to interpret and bridge visual and textual data. Among these, the Contrastive Language-Image Pretraining (CLIP) model proposed by OpenAI has gained substantial attention due to its remarkable zero-shot generalization capabilities (Radford et al., 2021). Unlike traditional computer vision models trained solely on fixed datasets, CLIP learns a joint representation space from large-scale web-collected image-text pairs, enabling it to perform effectively on unseen image classification tasks without explicit fine-tuning.

Building upon CLIP's success, subsequent research introduced OpenCLIP, an open-source implementation trained on publicly available datasets such as LAION-400M and LAION-2B (Ilharco et al., 2022). Comparing OpenAI's CLIP with OpenCLIP provides valuable insights into the effects of pretraining data diversity and scale on task performance. While prior work has primarily focused on evaluating zero-shot generalization on standard benchmarks, further exploration is needed to understand how model architecture, pretraining corpus, and prompt engineering influence cross-modal retrieval and classification performance across both general and specialized domains.

This project systematically evaluates and compare the effectiveness of both CLIP and OpenCLIP under two evaluation settings, zero-shot and linear probe, across multiple diverse datasets: CIFAR100, ImageNet-mini, Food101, EuroSAT, and COCO. Beyond replicating existing evaluations, this project specifically analyzes the strengths and limitations of different image encoders in varying domain contexts and investigate how different prompting strategies influence the performance of image retrieval tasks. By highlighting the sensitivity of these models to domain specificity and prompt design, our study aims to contributes to a deeper understanding of vision-language model applicability and guides future improvements.

Methodology

1. CLIP Architecture

CLIP is trained using contrastive pretraining. The complete workflow and core architecture are illustrated in Figure 1, covering both the training and inference phases.

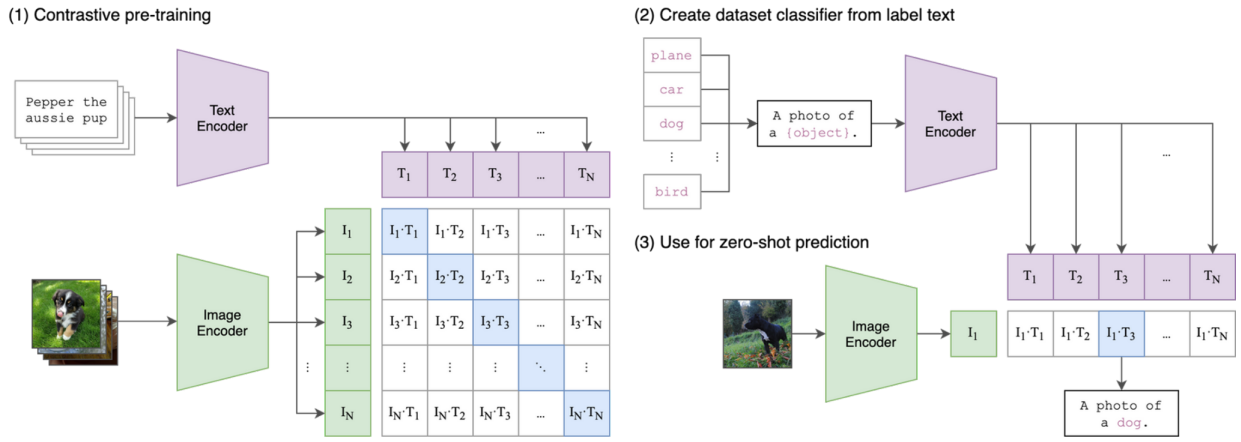

Figure 1 The Architecture of CLIP Model (OpenAI, 2021)

In the training phase, CLIP processes large batches of image-text pairs sourced from the internet. Each image, such as a puppy photo labeled “Pepper the Aussie pup,” is converted into a feature vector by the image encoder (for example, ResNet or Vision Transformer architectures). Simultaneously, the corresponding text descriptions are transformed into feature vectors via the text encoder, typically employing a Transformer-based model. Both image and text embeddings are projected into a shared embedding space, enabling direct comparisons between visual and textual data. CLIP computes a similarity matrix by calculating the dot product between every possible pair of image and text embeddings within the batch. The primary training objective is contrastive: maximize the similarity scores for correctly matched image-text pairs (located along the diagonal of the matrix) while minimizing similarity for all mismatched pairs (off-diagonal elements).

During the inference phase, the trained CLIP model can perform zero-shot classification without additional fine-tuning. Class labels (e.g., “dog,” “plane,” “bird”) are reformulated into natural language prompts (such as “a photo of a dog”) and encoded as reference text vectors. When an unseen image is introduced, it passes through the image encoder to produce an embedding vector, which is subsequently compared with all pre-generated label embeddings. The class

corresponding to the text embedding with the highest similarity score is chosen as the final prediction, demonstrating CLIP's capability for versatile and efficient cross-modal understanding.

2. Project Structure

The project is structured into two main components: image-to-text classification and text-to-image retrieval, in order to comprehensively evaluate the bidirectional capabilities of CLIP-based models. In addition to reproducing established zero-shot classification evaluations, this project extends prior work by systematically analyzing linear probing performance, model architecture differences, and prompt sensitivity in retrieval tasks. Conducting both evaluation directions enables a thorough analysis of model strengths and limitations, providing a deeper understanding of generalization behavior across diverse tasks.

2.1 Datasets

Five datasets were selected to evaluate model performance across both general and specialized domains:

- **CIFAR-100**: Contains 100 categories of everyday objects captured in small-scale images, widely used for evaluating general object recognition.
- **ImageNet-mini**: A subset of ImageNet including various general object categories, suitable for assessing model generalization.
- **Food101**: A fine-grained classification dataset featuring 101 categories of food images.
- **EuroSAT**: Consists of satellite images covering multiple land-use and land-cover classes, providing a domain-specific evaluation setting.
- **COCO (Common Objects in Context)**: A richly annotated dataset of images and captions, extensively used for text-to-image retrieval experiments.

The selection of diverse datasets enables the evaluation of model performance across different classification and retrieval scenarios.

2.2 Image-to-Text Classification

Three evaluation strategies were applied to perform image-to-text classification: zero-shot inference using OpenAI CLIP, linear probing on OpenAI CLIP features, and zero-shot inference using OpenCLIP. Their core differences are summarized as follows:

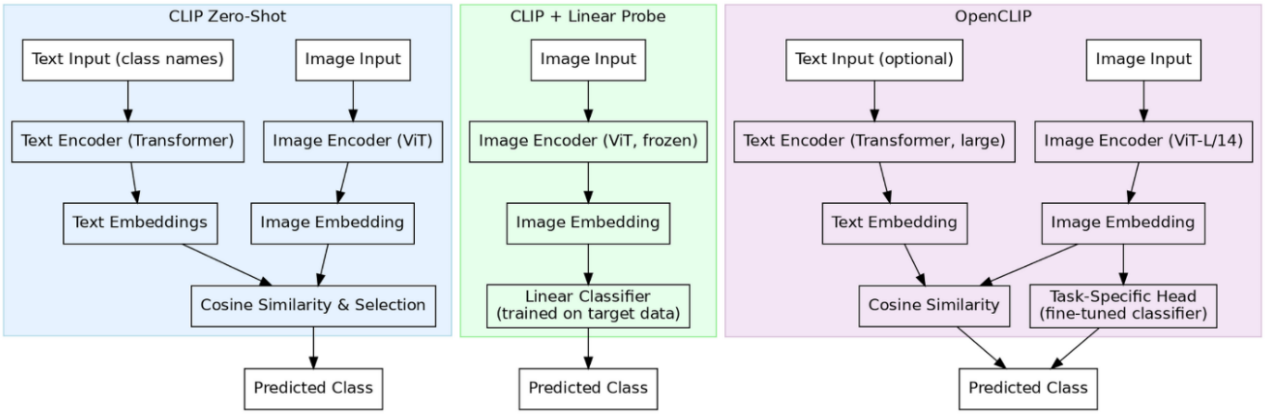

Figure 2 The Core Architectural Differences of CLIP Zero-shot, Linear Probe and Open Clip.

- **Zero-shot**: Classification is performed without any additional training, by directly matching images to pre-encoded text prompts via cosine similarity.
- **Linear probe**: A logistic regression classifier is trained on fixed CLIP image embeddings to adapt pretrained features to specific downstream tasks.
- **OpenCLIP**: An open-source variant of CLIP trained on alternative, larger datasets such as LAION-2B, offering insights into the effect of pretraining data scale and diversity.

Experiments were conducted across four datasets: CIFAR-100, ImageNet-mini, Food101, and EuroSAT.

Multiple backbone architectures were evaluated (Table 1):

Table 1 Summary of Model Configurations.

Model	Architecture	Patch/Kernel Size	Pre-training (OpenAI CLIP / OpenCLIP)
ViT-B/32	Vision Transformer (Base)	32×32 patches	WIT (WebImageText, 400 M pairs) LAION-2B (2B pairs)
ViT-B/16	Vision Transformer (Base)	16×16 patches	WIT (WebImageText, 400 M) LAION-400M (400 M pairs)
Resnet50	ResNet-50 CNN	7×7 convolutions	WIT (WebImageText, 400 M)
ViT-L-14	Vision Transformer (Large)	14×14 patches	LAION-2B
ConvNeXt-Base	ConvNeXt-Base CNN	7×7 convolutions	LAION-400M

Performance was assessed using Top 1 and Top 5 classification accuracies for zero-shot evaluations. For linear probing, both training accuracy and test accuracy were recorded to evaluate the effectiveness of feature adaptation and to monitor potential overfitting. These metrics enabled a comparative analysis across different model backbones and pretraining sources.

2.3 Text-to-Image Retrieval

The second experiment of the project focused on evaluating text-to-image retrieval performance, specifically analyzing the sensitivity of CLIP models to prompt formulation. Several types of prompt variations were explored:

- **Prompt style changes**: Modifying the prompt structure (e.g., “a photo of” vs. “an illustration of”) to assess retrieval domain sensitivity.
- **Subject noun changes**: Altering the main subject (e.g., “apples” vs. “oranges”) to test semantic sensitivity.
- **Descriptive adjective changes**: Introducing or modifying adjectives (e.g., “a flying white bird”) to evaluate retrieval refinement based on textual specificity.

In addition to the ImageNet-Mini dataset, the COCO dataset was extensively used for retrieval experiments due to its rich and detailed caption annotations. Evaluations included both

controlled prompt engineering tests and experiments using human-written captions from COCO to assess retrieval robustness under realistic and fine-grained linguistic descriptions.

Through these experiments, the project provides a detailed examination of how CLIP and OpenCLIP models respond to prompt variations, and how linguistic precision impacts visual retrieval effectiveness across different datasets.

Experiment

1. Image-to-text Classification

Image-to-text classification experiments were conducted using publicly available pretrained CLIP-based model checkpoints. For zero-shot evaluations, a standardized prompt template ("a photo of a [class name]") was applied across all datasets to ensure consistency and minimize linguistic variability. Linear probing experiments trained a logistic regression classifier on frozen CLIP embeddings, with both training and testing accuracies recorded to assess feature adaptation and monitor potential overfitting. All evaluations followed the original model input preprocessing and resolution settings to maintain comparability across architectures and datasets.

The performance of the evaluated models across different datasets and configurations is summarized in Tables 2–5.

Table 2 Classification Accuracies on the CIFAR-100 Dataset.

Model	CLIP - Zero shot	CLIP - Linear Prob	OpenCLIP - Zero shot
ViT-B/32	Top1: 64.18% Top5: 88.15%	Train: 99.35% Test: 73.20%	Top1: 75.89% Top5: 93.86%
ViT-B/16	Top1: 68.04% Top5: 89.14%	Train: 98.90% Test: 78.05%	Top1: 71.61% Top5: 92.47%
Resnet50	Top1: 40.62% Top5: 72.30%	Train: 67.95% Test: 58.75%	
ViT-L-14			Top1: 82.39% Top5: 96.47%
ConvNeXt-Base			Top1: 46.50% Top5: 73.79%

Table 3 Classification Accuracies on the ImageNet-Mini Dataset.

Model	CLIP - Zero shot	CLIP - Linear Prob	OpenCLIP - Zero shot
ViT-B/32	Top1: 61.89% Top5: 85.24%	Train: 100% Test: 37.96%	Top1: 61.89% Top5: 85.24%
ViT-B/16	Top1: 62.17% Top5: 87.48%	Train: 100% Test: 44.59%	Top1: 62.22% Top5: 85.44%
Resnet50	Top1: 54.45% Top5: 81.01%	Train: 74.86% Test: 16.82%	
ViT-L-14			Top1: 70.20% Top5: 90.39%
ConvNeXt-Base			Top1: 49.27% Top5: 74.28%

Table 4 Classification Accuracies on the Food101 Dataset.

Model	CLIP - Zero shot	CLIP - Linear Prob	OpenCLIP - Zero shot
ViT-B/32	Top1: 82.06% Top5: 96.89%	Train: 99.43% Test: 86.79%	Top1: 82.10% Top5: 96.62%
ViT-B/16	Top1: 87.78% Top5: 98.44%	Train: 99.75% Test: 91.19%	Top1: 85.91% Top5: 97.62%
Resnet50	Top1: 77.57% Top5: 95.41%	Train: 89.76% Test: 84.10%	
ViT-L-14			Top1: 90.78% Top5: 98.76%
ConvNeXt-Base			Top1: 70.74% Top5: 89.68%

Table 5 Classification Accuracies on the EuroSAT Dataset.

Model	CLIP - Zero shot	CLIP - Linear Prob	OpenCLIP - Zero shot
ViT-B/32	Top1: 32.18% Top5: 82.87%	Train: 98.14% Test: 95.65%	Top1: 42.24% Top5: 90.03%
ViT-B/16	Top1: 35.80% Top5: 80.89%	Train: 98.23% Test: 95.94%	Top1: 32.12% Top5: 83.49%
Resnet50	Top1: 17.75% Top5: 76.40%	Train: 93.88% Test: 93.20%	
ViT-L-14			Top1: 12.25% Top5: 88.29%
ConvNeXt-Base			Top1: 26.53% Top5: 74.33%

The evaluated CLIP-based models demonstrated strong zero-shot generalization for general object recognition tasks, consistent with the findings reported in Radford et al. (2021). Vision transformer encoders, such as ViT-B/32 and ViT-B/16, consistently outperformed CNN-based encoders like ResNet-50, aligning with trends observed in the original CLIP study. OpenCLIP models pretrained on the larger and more diverse LAION corpus generally outperformed OpenAI CLIP models, particularly on datasets like CIFAR-100, emphasizing the impact of pretraining dataset scale and diversity.

Linear probe applied to CLIP models led to notable improvements in test accuracy compared to zero-shot inference. On most datasets, test accuracy after linear probing exceeded zero-shot Top 1 results, although zero-shot Top 5 accuracy often remained competitive with or even higher than linear probe test accuracy. The strong zero-shot Top 5 accuracy reflects the generalization and robustness of CLIP models across unseen categories. On ImageNet-mini, however, high training accuracy accompanied by relatively low test accuracy suggested potential overfitting and reduced generalization.

Dataset characteristics also influenced model performance. On general datasets such as CIFAR-100 and ImageNet-mini, transformer-based models with larger backbones, such as ViT-L-14, achieved the highest accuracies. Although Food101 focuses on food categories, its distinctive visual features and simpler structure led to even higher zero-shot accuracies compared to

CIFAR-100 and ImageNet-mini. In contrast, performance on EuroSAT was significantly lower for both CLIP and OpenCLIP models under zero-shot evaluation. Linear probing substantially improved accuracy on EuroSAT, suggesting the limitations of zero-shot transfer in specialized domains. Additionally, performance decreased across vision transformer models as patch size decreased, indicating that while smaller patches enhance feature extraction for general objects, they are less effective in capturing coarse spatial structures necessary for specialized tasks.

Overall, the results demonstrate that vision transformer architectures outperform CNN-based encoders for image-to-text classification. OpenCLIP models, pretrained on the larger and more diverse LAION dataset, consistently outperformed CLIP models in zero-shot settings, highlighting the importance of pretraining corpus quality. Although strong zero-shot performance was achieved on general object datasets, accuracy declined on domain-specific tasks such as EuroSAT, where minimal supervised adaptation through linear probing provided substantial gains.

2. Prompt-Based Image Retrieval

This section explores how variations in prompt phrasing affect the retrieval performance of CLIP-based models. Experiments were conducted using both CLIP and OpenCLIP pretrained checkpoints, and results are discussed based on four focused questions.

2.1 Effect of Prompt Style on Retrieval Results

Prompt style, such as specifying “photo” versus “drawing,” strongly influenced retrieval outcomes. As shown in Figure 3-5, prompts containing "photo" retrieved realistic photographic images, while prompts containing “drawing” retrieved stylized or artistic representations. This demonstrates that prompt style can guide the model towards specific visual domains.

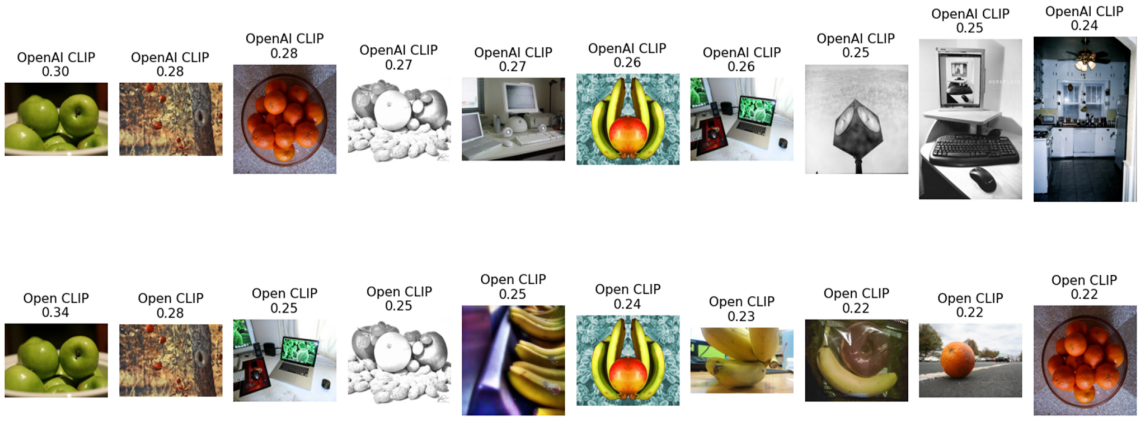

Figure 3 Retrieval results for “a photo of apples”.

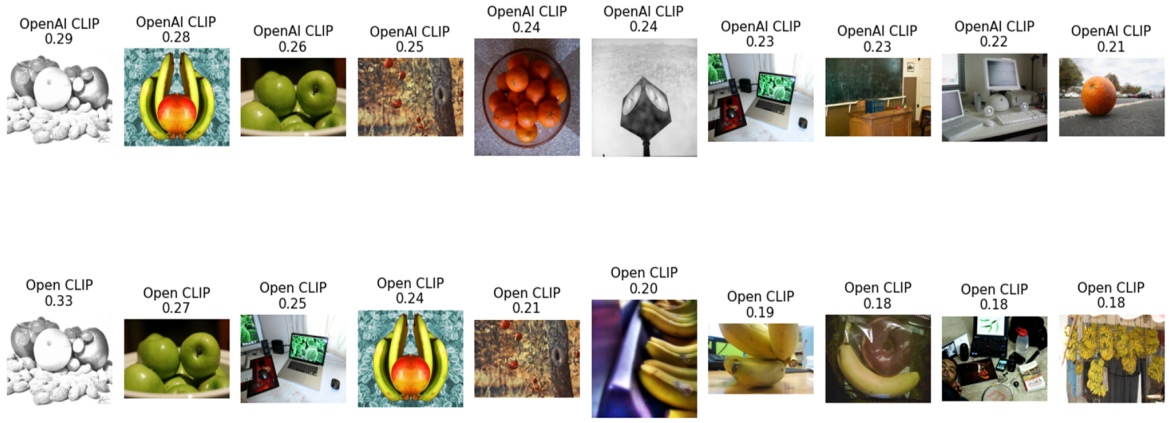

Figure 4 Retrieval results for “a drawing of apples”.

```
Evaluating: 'a photo of a {}': 100%|██████████| 157/157 [13:00<00:00,  4.97s/it]
Prompt: 'a photo of a {}' -> Accuracy: 61.71%
Evaluating: 'a sketch of a {}': 100%|██████████| 157/157 [13:00<00:00,  4.97s/it]
Prompt: 'a sketch of a {}' -> Accuracy: 58.91%
Evaluating: 'a drawing of a {}': 100%|██████████| 157/157 [13:06<00:00,  5.01s/it]
Prompt: 'a drawing of a {}' -> Accuracy: 61.10%
Evaluating: 'a cartoon of a {}': 100%|██████████| 157/157 [13:46<00:00,  5.27s/it]
Prompt: 'a cartoon of a {}' -> Accuracy: 61.82%
Evaluating: 'an artistic rendering of a {}': 100%|██████████| 157/157 [13:02<00:00,  4.98s/it]
Prompt: 'an artistic rendering of a {}' -> Accuracy: 63.17%
Evaluating: 'a low-resolution photo of a {}': 100%|██████████| 157/157 [13:01<00:00,  4.98s/it]
Prompt: 'a low-resolution photo of a {}' -> Accuracy: 63.23%
Evaluating: '{}': 100%|██████████| 157/157 [13:02<00:00,  4.98s/it]
No prompt (only category name) -> Accuracy: 55.15%
```

Figure 5 Different prompts on the satellite dataset.

2.2 Impact of Subject Noun Variation

Changing the subject noun within prompts resulted in distinctly different retrieval outputs. For instance, prompts using “apples” versus “oranges” led to retrievals centered around the specified object (Figure 6-7). This confirms that CLIP models maintain strong alignment between textual

input and visual content, successfully differentiating object categories with minimal textual changes.

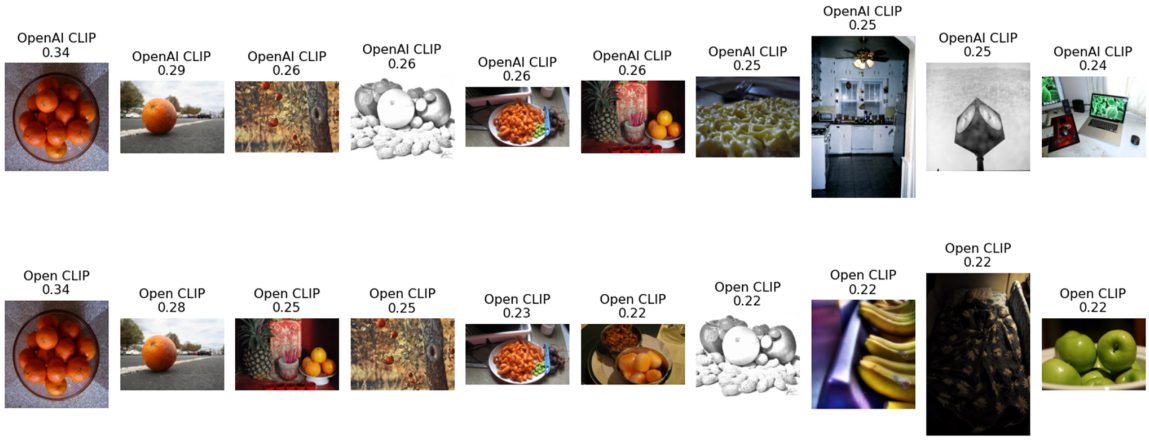

Figure 6 Retrieval results for “a photo of oranges”.

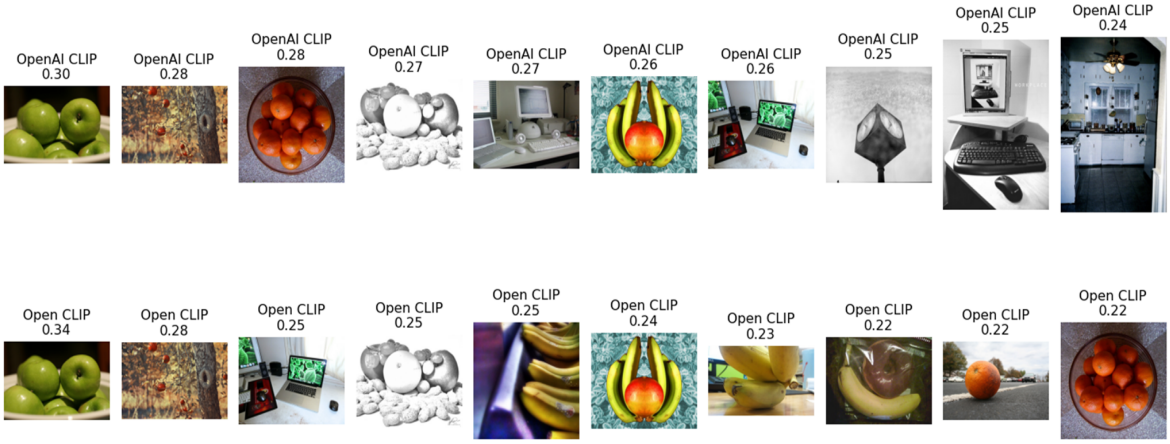

Figure 7 Retrieval results for “a photo of apples”.

2.3 Influence of Descriptive Adjectives

Adding descriptive adjectives to prompts refined retrieval specificity. The retrieval output (Figure 8-10) indicates that CLIP and OpenCLIP can leverage descriptive language for more fine-grained search. Besides, both of those models can accurately retrieve images based on multiple simple keywords (“bird, white, flying”) without requiring full sentences (Figure 11) This shows that a few well-chosen words are sufficient for fine-grained retrieval.

Figure 8 Retrieval results for “a photo of a bird”.

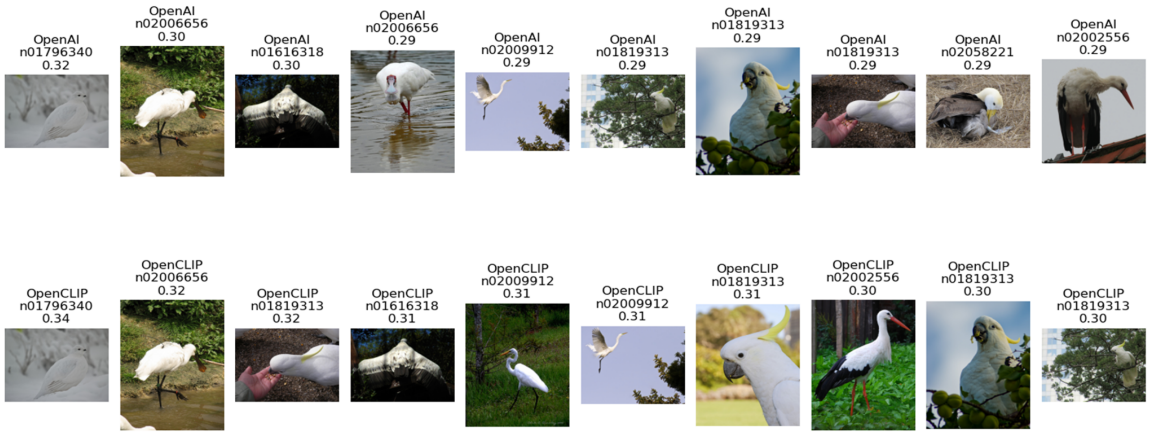

Figure 9 Retrieval results for “a photo of a white bird”.

Figure 10 Retrieval results for “a photo of a flying white bird”.

Figure 11 Retrieval results for “bird, white, flying”.

2.4 Retrieval Performance Using Human-Written Captions

1) Prompt: “Three teddy bears, each a different color, snuggling together.”

Figure 12 Original.

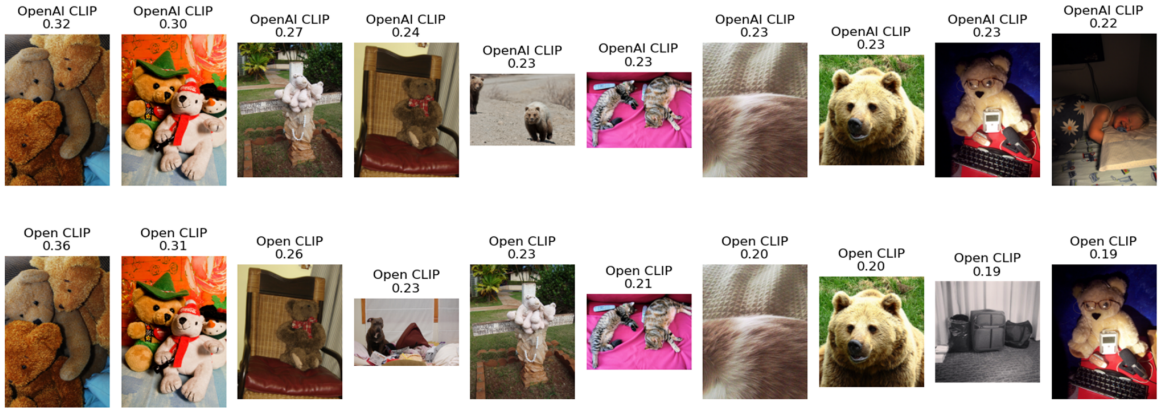

Figure 13 Results.

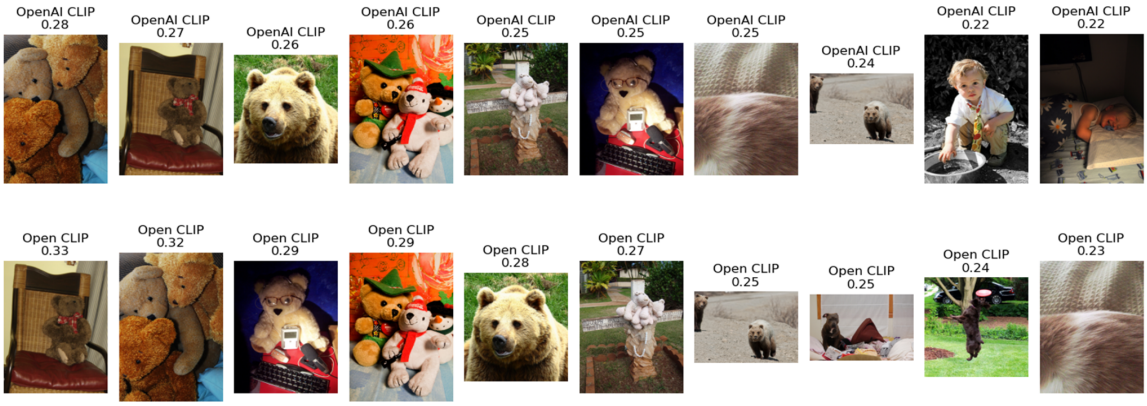

Figure 14 Addition: Teddy bear.

2) Prompt: “The people are posing for a group photo.”

Figure 15 Original.

Figure 16 Results.

In the first case (Figure 12-14), using COCO’s human-written captions as prompts, both models successfully retrieved the correct image as top 1. This is because the image was unique in the dataset, and the caption provided a detailed visual description. However, in the second case (Figure 15-16), where many images depicted “people posing for a group photo,” retrieval was

more challenging. Although OpenCLIP found the correct image at top 1, OpenAI CLIP did not. This outcome shows that retrieval performance depends heavily on caption specificity: clear and detailed visual descriptions yield better results, while abstract or contextual captions lower accuracy.

Overall, these retrieval experiments demonstrate that CLIP models are highly sensitive to prompt phrasing, with small changes in wording significantly affecting retrieval performance. Both OpenAI CLIP and OpenCLIP provide robust frameworks for natural language-based image retrieval, showing resilience to style and subject variations while remaining highly sensitive to prompt specificity. Interestingly, the two models sometimes favor different styles or concepts, highlighting subtle differences in how they understand prompts. Using human-written captions is effective when the description directly corresponds to the visual content.

Conclusion and Discussion

This study evaluated the capabilities of OpenAI CLIP and OpenCLIP models for cross-modal retrieval and classification tasks through both image-to-text and text-to-image experiments. By leveraging CLIP and OpenCLIP models, the experiments enabled flexible zero-shot evaluation across diverse domains without additional fine-tuning. A comprehensive bidirectional analysis revealed the models' sensitivity to prompt phrasing and their adaptability to different user intents. Additionally, by comparing zero-shot performance with linear probing, the study quantified trade-offs between generalization capacity and the benefits of lightweight supervised adaptation.

While OpenAI CLIP and OpenCLIP models demonstrated strong zero-shot performance, several limitations remain. The datasets used, although diverse, were not large or comprehensive enough to fully test generalization across different domains. Retrieval evaluation relied mainly on manual inspection, without using quantitative metrics like precision or recall, which introduced potential subjectivity. This study also focused mostly on prompt-level variations and did not systematically explore multimodal ambiguities or adversarial examples. Additionally, although linear probing improved domain-specific performance, it requires labeled data and extra training, which deviates from the original zero-shot learning approach.

Future work can focus on systematic quantitative evaluations, investigating prompt ambiguity and model failures, comparing against traditional supervised classifiers, and expanding to video and multi-modal retrieval. Few-shot tuning and automatic prompt optimization will also be explored to further enhance model adaptability and robustness.

References

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2022). *OpenCLIP (Version 0.1)* [Computer software]. Zenodo. https://github.com/mlfoundations/open_clip

OpenAI. (2021). *CLIP* [Computer software]. GitHub. https://github.com/openai/CLIP

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision* (arXiv:2103.00020). arXiv. https://arxiv.org/abs/2103.00020