TEXT MEETS VISION: A DEEP DIVE INTO CLIP PERFORMANCE

DATA586 Group 12
Mingyue Zhao, Yuzhu Han, Skylar Shao

CONTENT

- Background
- Model Architecture
- Datasets & Justification
- Experiments
- Conclusion
- References

BACKGROUND

CLIP stands for Contrastive Language–Image Pretraining. It's a powerful vision-language model developed by OpenAI. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task.

Why CLIP?

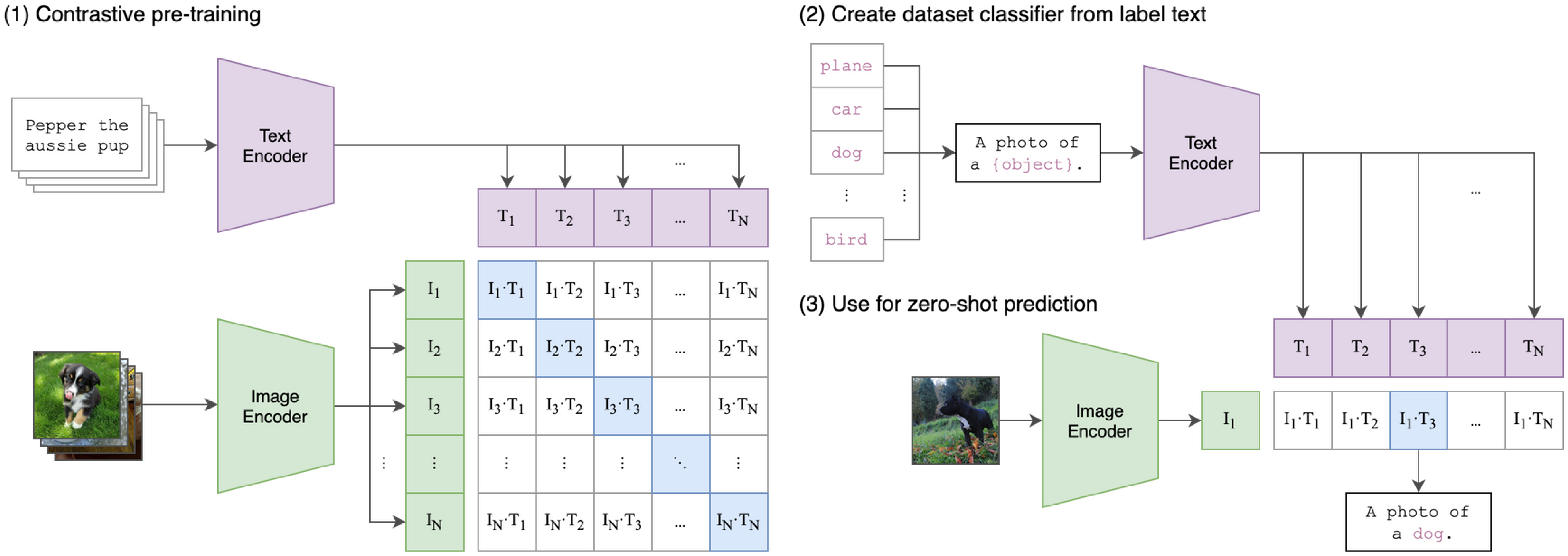
- bridging vision and language without task-specific training.
- align images and text in the same embedding space.

Project Goal

- Explore how well CLIP performs across multiple datasets
- Image-to-Text
  - Compare zero-shot **VS** linear probe **VS** Openclip
  - ResNET50 **VS** VIT32 **VS** VIT16
- Text-to-Image
  - prompt style change？
  - Subject noun change？
  - Descriptive adj. change?

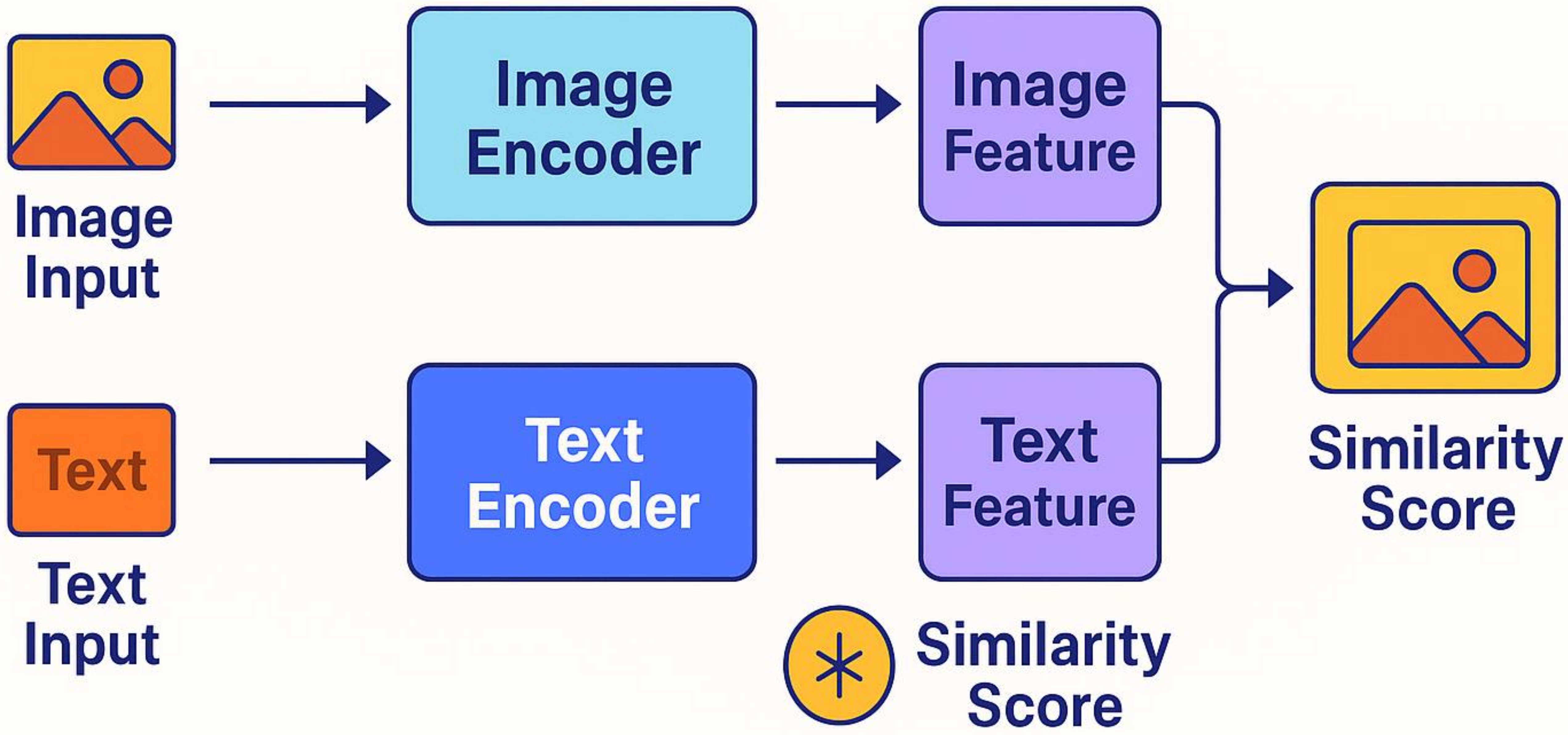
MODEL ARCHITECTURE

“A Vision + Language Model”



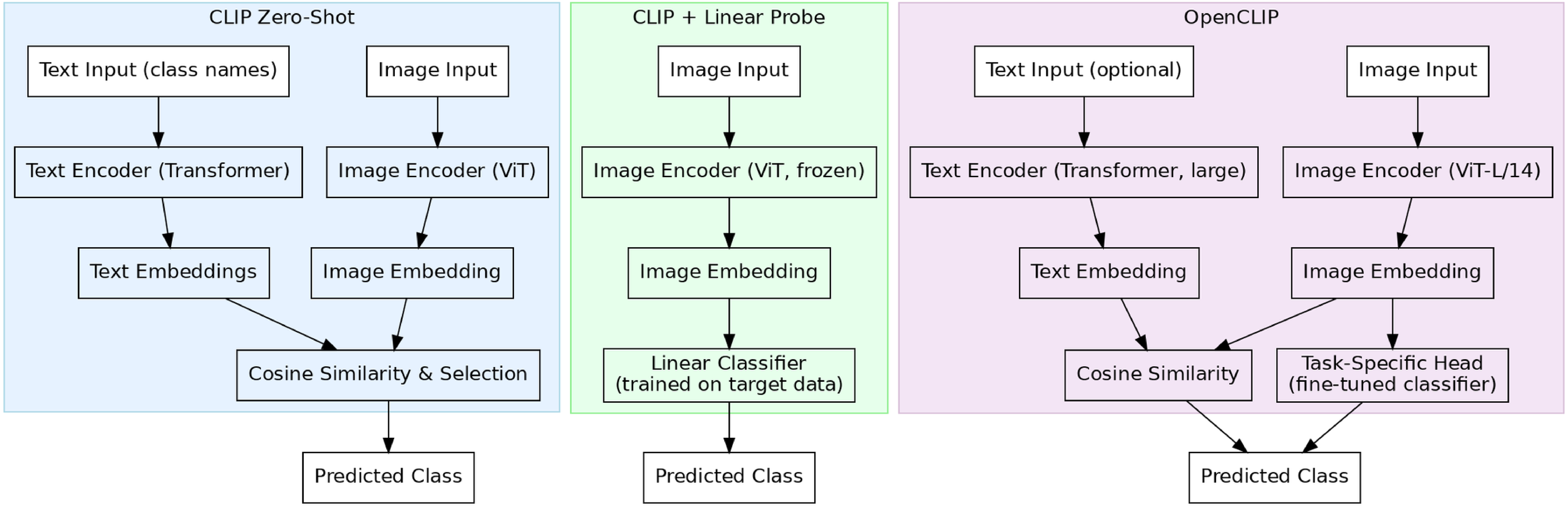
https://github.com/openai/CLIP

$$\text{similarity score} = \cos(\theta) = \frac{v_{img} \cdot v_{text}}{\|v_{img}\| \, \|v_{text}\|}$$



MODEL ARCHITECTURE

- CLIP Zero-Shot, CLIP + Linear Probe, and OpenCLIP

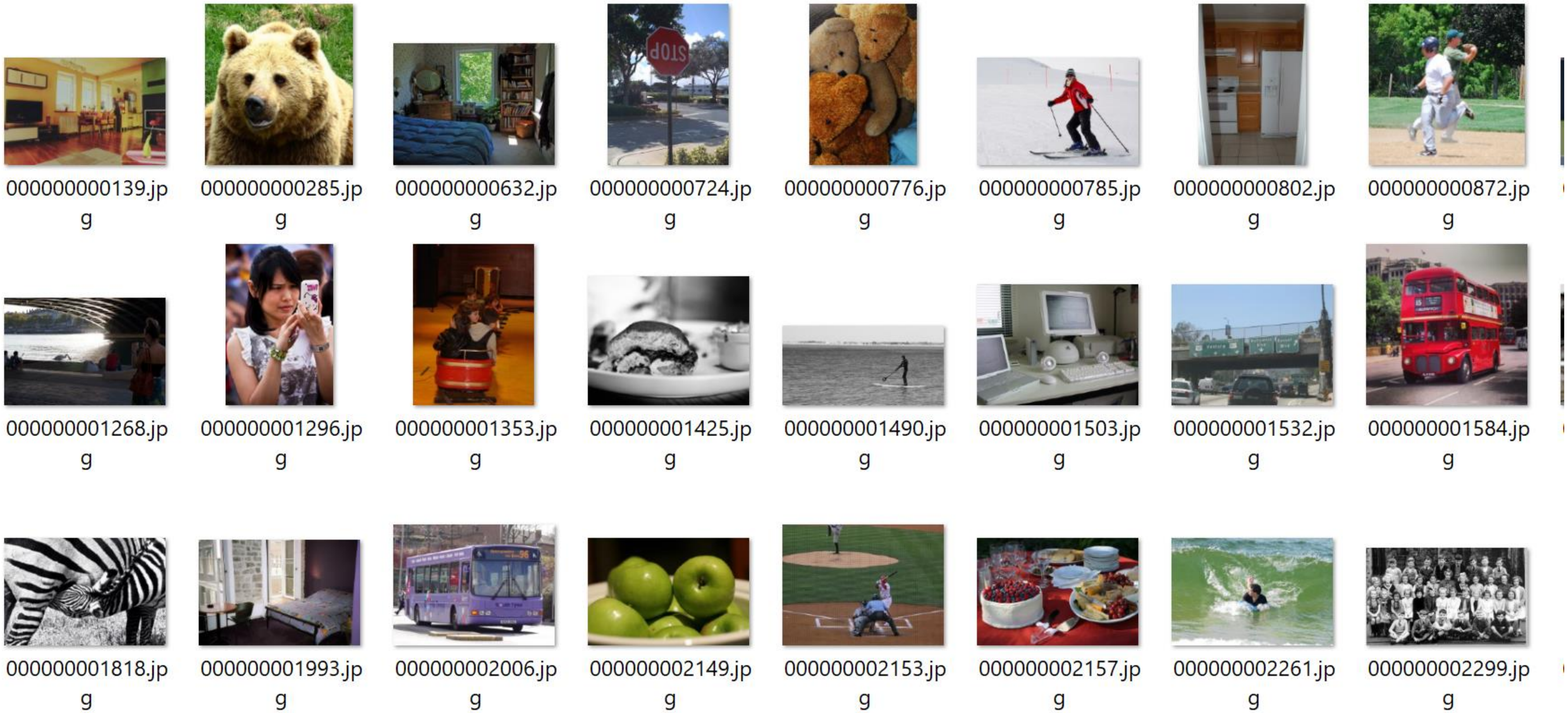


DATASETS

Dataset	Why we chose it
CIFAR100	General-purpose, small resolution, many classes
ImageNet-Mini	Benchmark-like, high diversity, test generalization
Food101	Fine-grained → test CLIP's zero-shot limit
EuroSAT	Remote sensing → test out-of-distribution generalization

DATASETS

Coco



DATASETS

ImageNet-mini

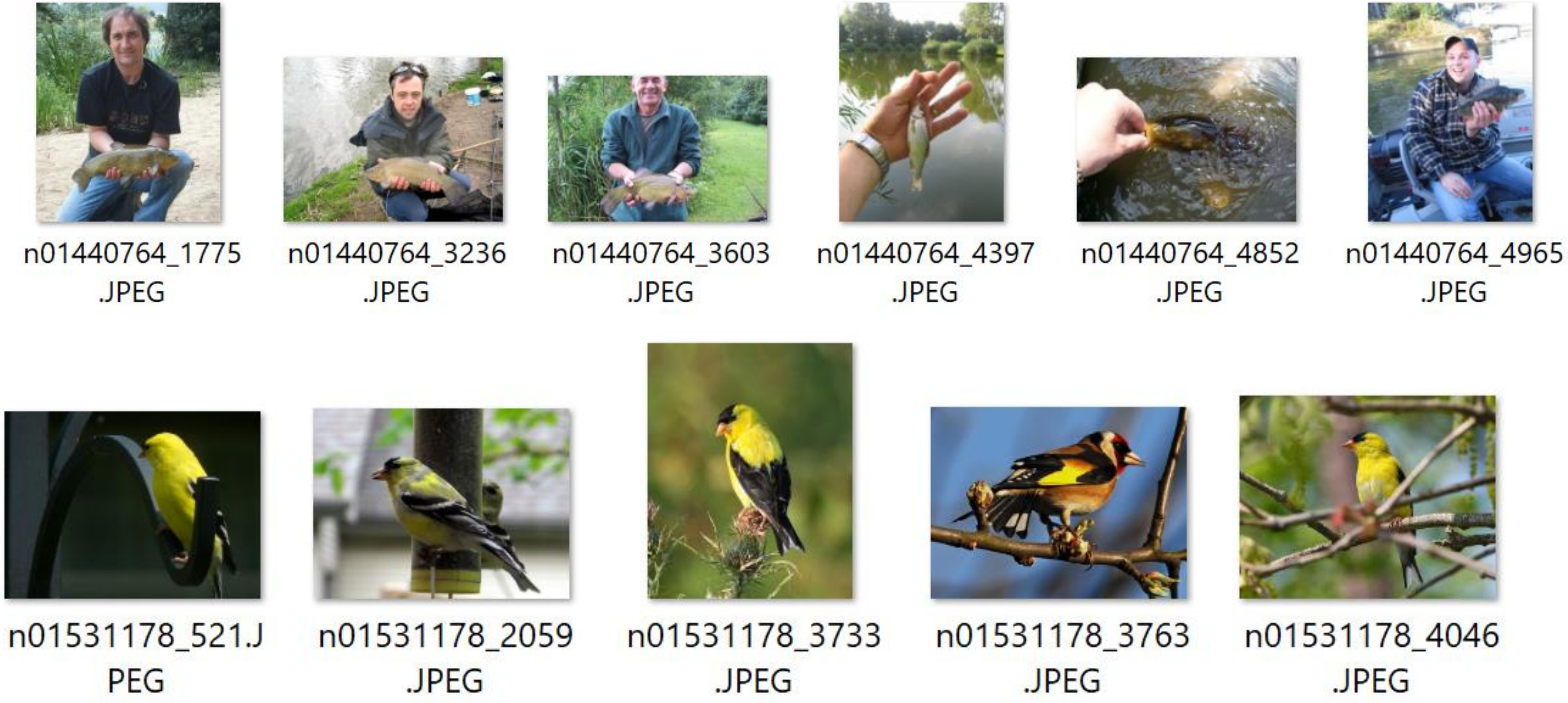


IMAGE-TO-TEXT CLASSIFICATION

Experiments- Image-to-text

Hyperparameter tuning: Image Encoder

Model	Architecture	Patch/Kernal 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 kernels	WIT (WebImageText, 400 M)
ViT-L-14	Vision Transformer (Large)	14×14 patches	LAION-2B
ConvNeXt-Base	ConvNeXt-Base CNN	7×7 kernels	LAION-400M

Experiments- Image-to-text

CIFAR100

	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%

Experiments- Image-to-text

ImageNet-mini

	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%

Experiments- Image-to-text

FOOD101

	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%

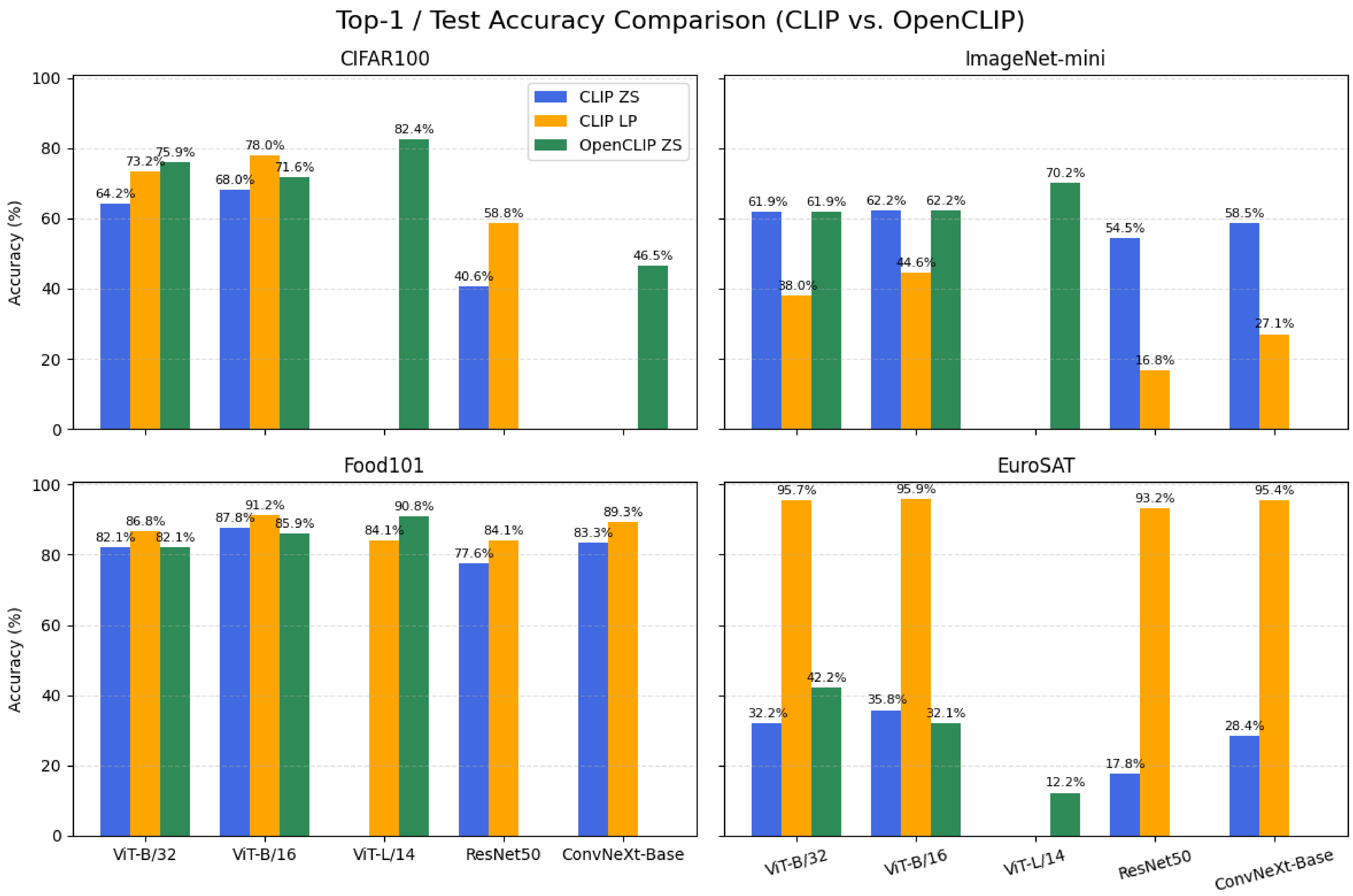
Experiments- Image-to-text

EuroSAT

	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%

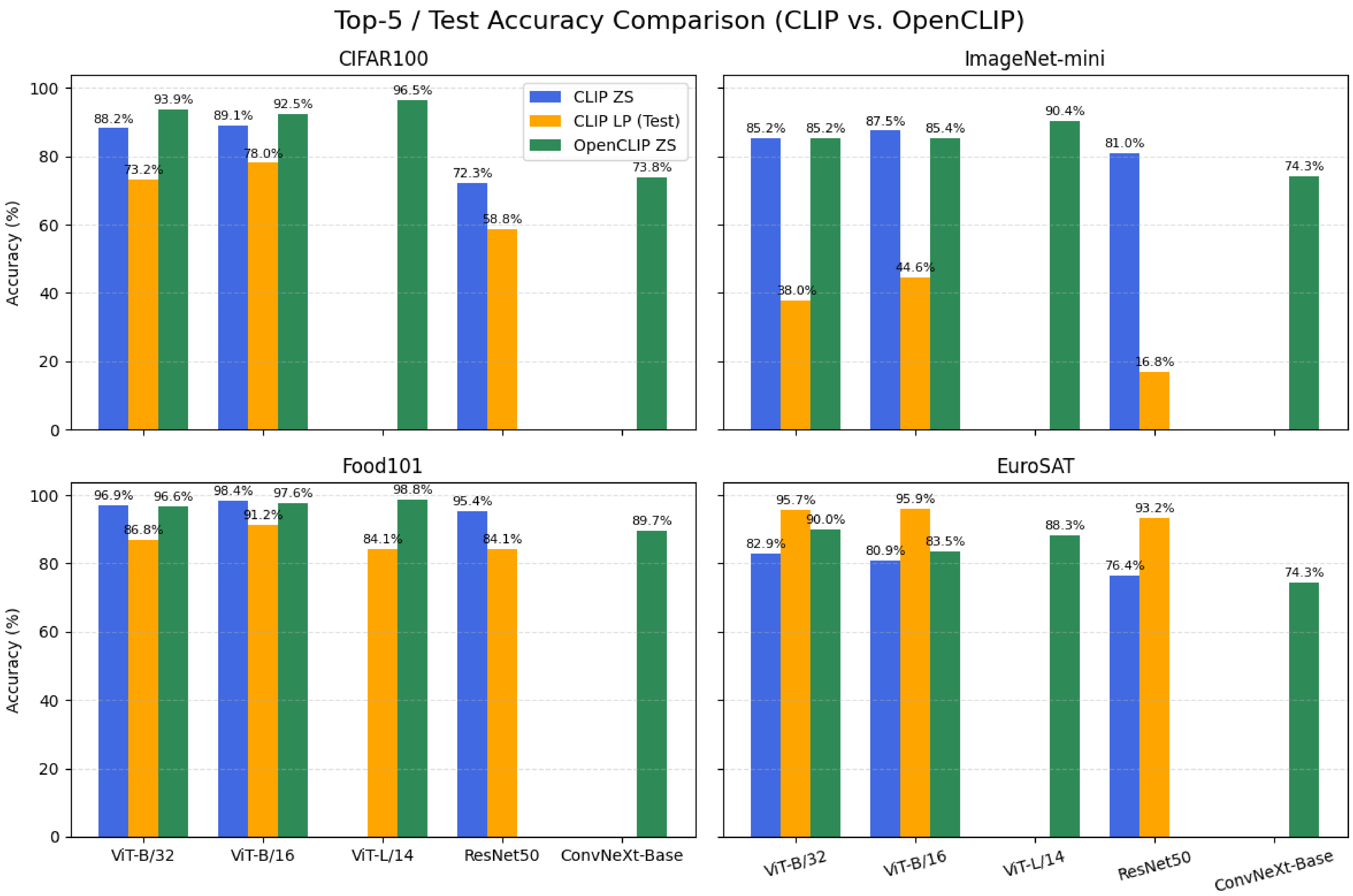
Experiments- Image-to-text

Comparison



Experiments- Image-to-text

Comparison



Experiments- Image-to-text

Key findings:

- **Robust zero-shot performance**: The top5 accuracy matches or outperforms linear-probe baselines across multiple tasks without task-specific fine-tuning
- **Dataset-dependent accuracy:** Perform better on general object datasets but shows reduced performance on specialized or fine-grained domains (e.g., satellite image in EuroSAT)
- **Pre-training impacts behavior:** Different pre-training data (WIT vs. LAION) affect output accuracy, even under the same natural-language prompts

GET PICTURES FROM PROMPTS

Research Questions

1. Does prompt style (adjective: photo vs. drawing) affect retrieval results?

- Example: "a photo of an orange" vs. "a drawing of an orange" .

2. What is the impact of changing the subject noun?

- Example: "apple" vs. "orange".

3. How do descriptive adjectives refine retrieval?

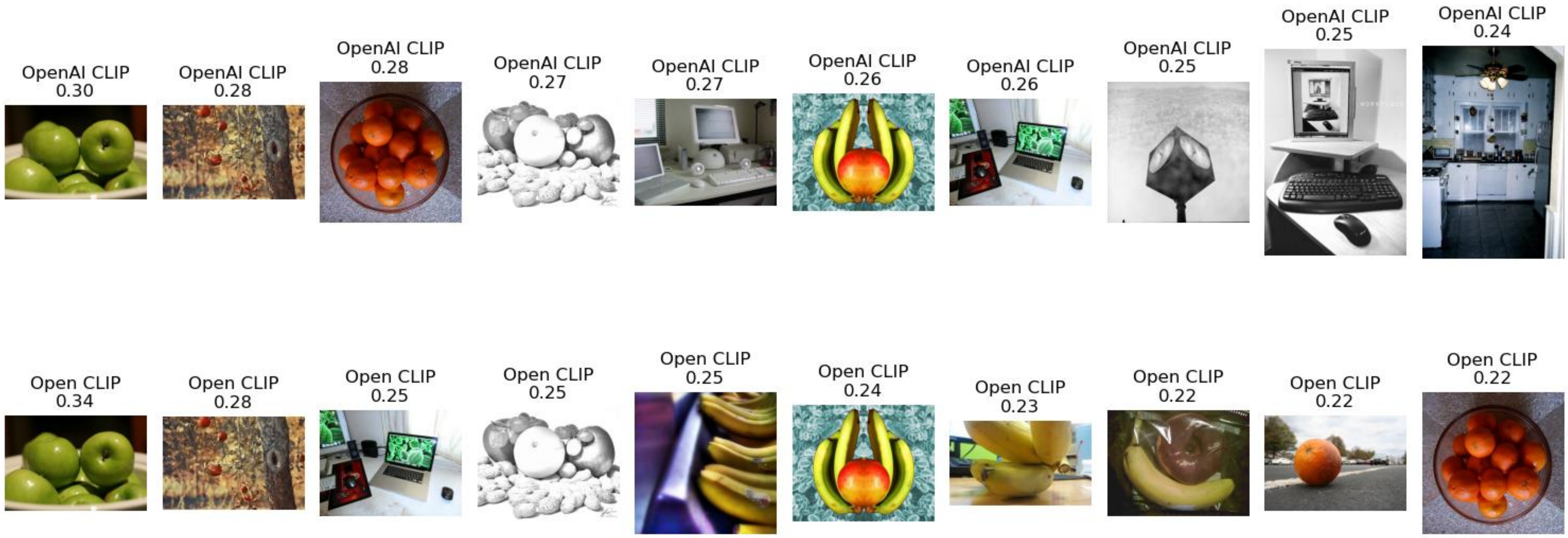
- Example:
  - "a photo of a bird"
  - "a photo of a white bird"
  - "a photo of a flying white bird"

4. Can CLIP find the original image from its human caption?

- Use COCO captions as prompts and see if the correct image is top-1.

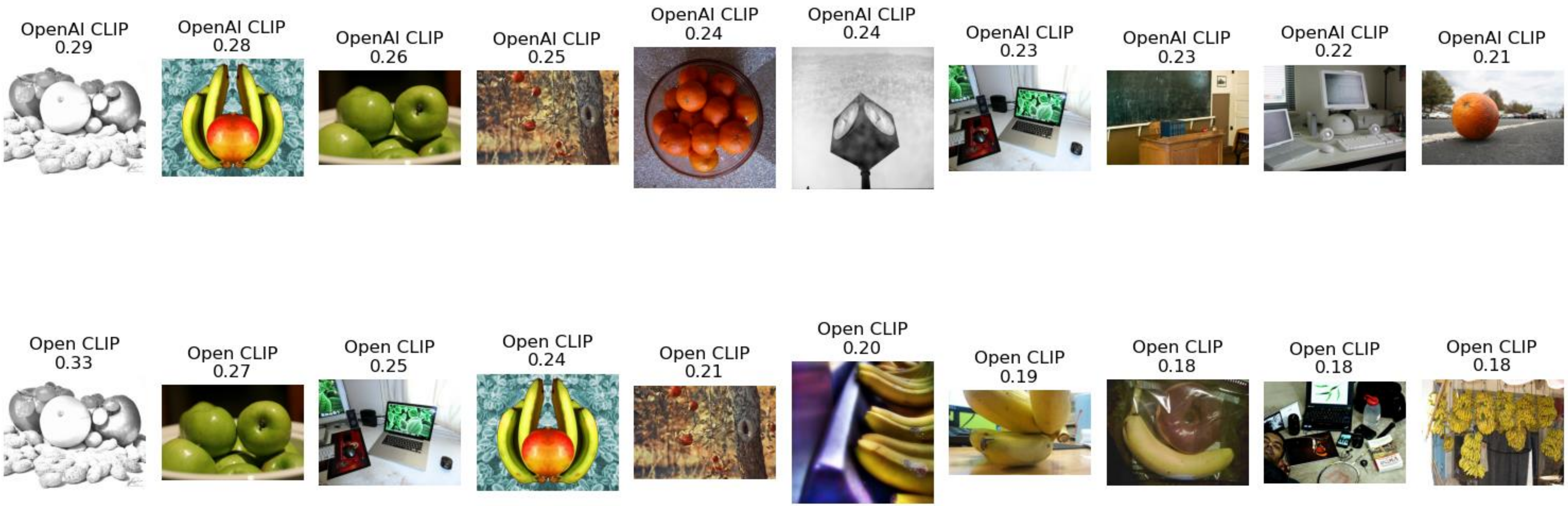
1. DOES PROMPT STYLE AFFECT RETRIEVAL RESULTS?

1. a photo of apples



Coco

2. a drawing of apples



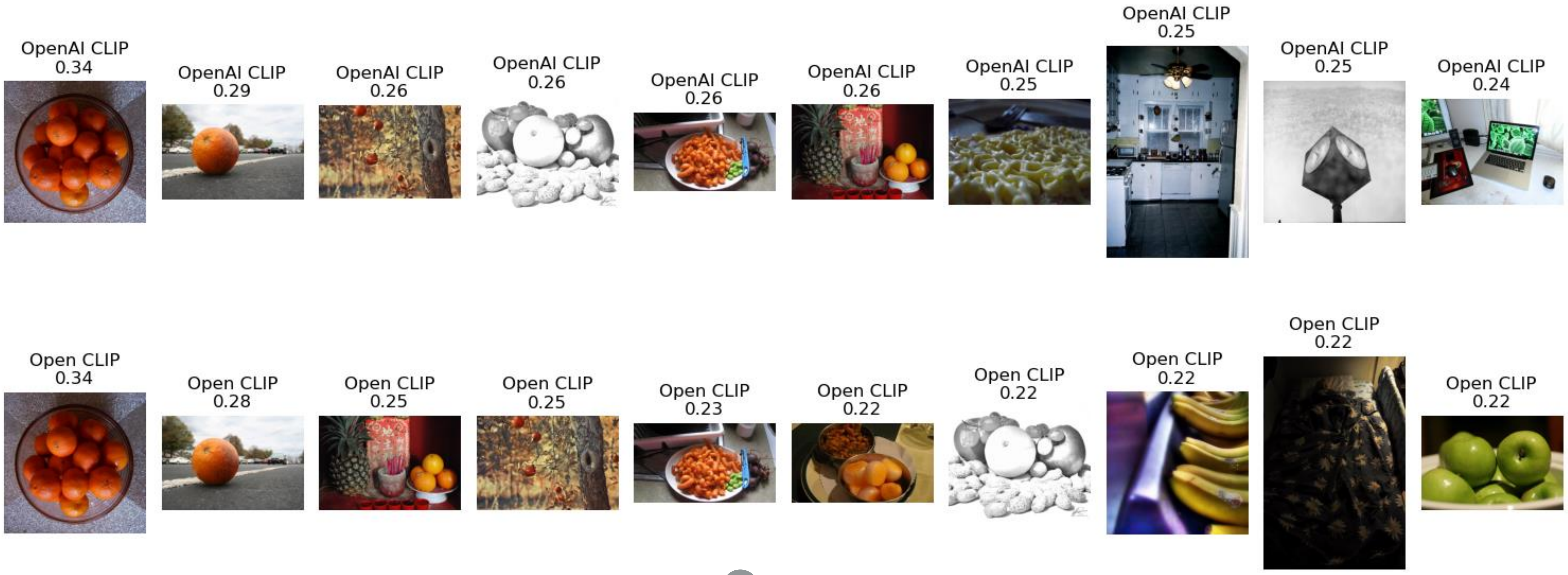
Coco

3. Different prompts on satellite dataset

```
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%
```

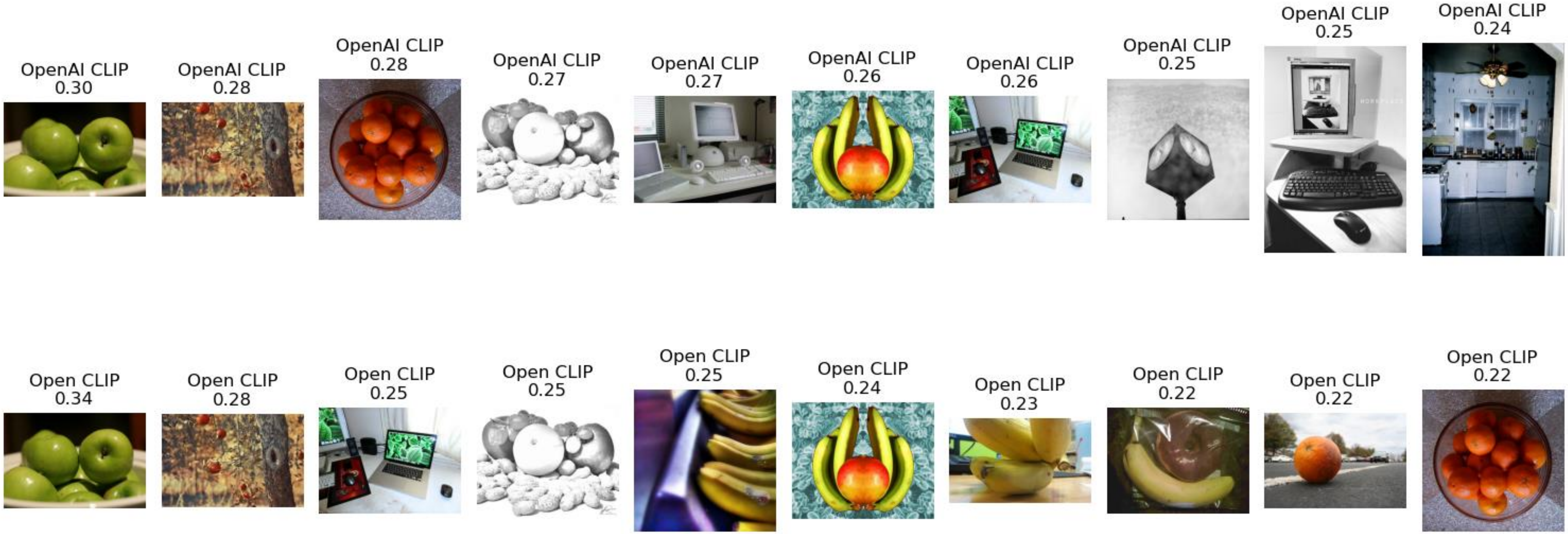
2. WHAT IS THE IMPACT OF CHANGING THE SUBJECT NOUN?

1. a photo of oranges



Coco

2. a photo of apples

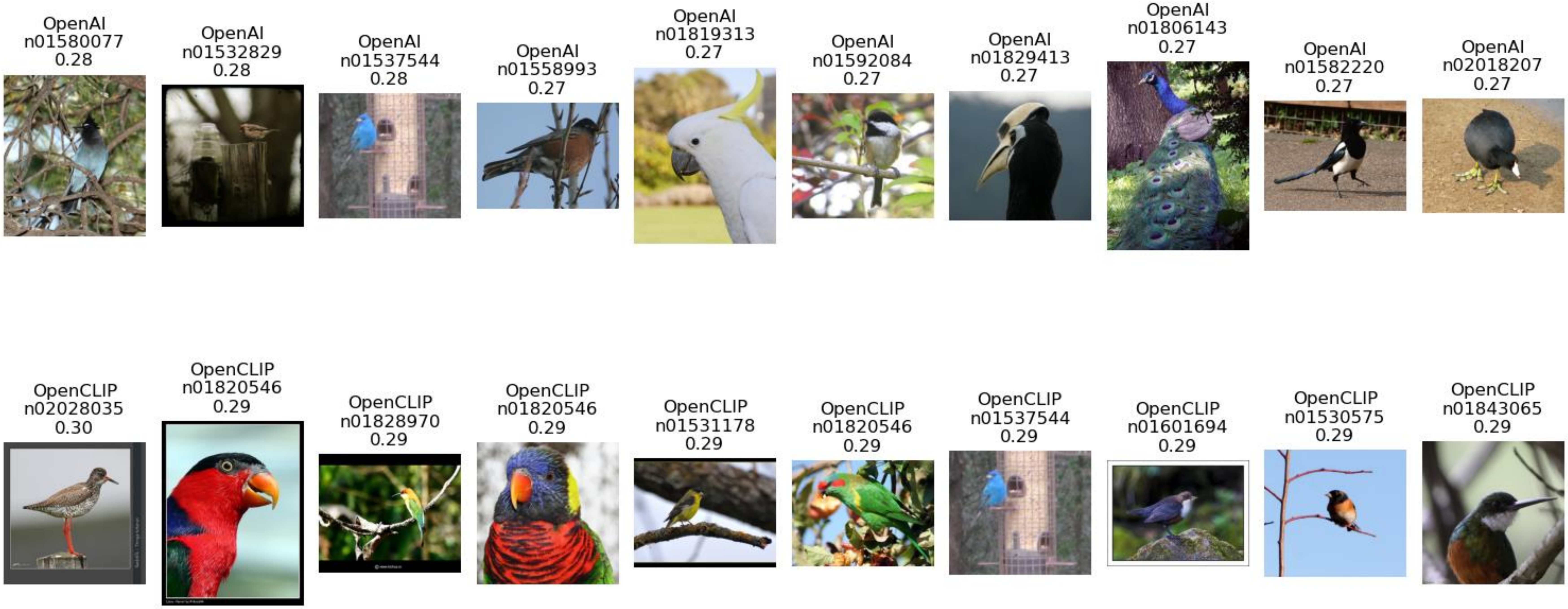


Coco

3. HOW DO DESCRIPTIVE ADJECTIVES REFINE RETRIEVAL?

BIRDS

1. a photo of bird

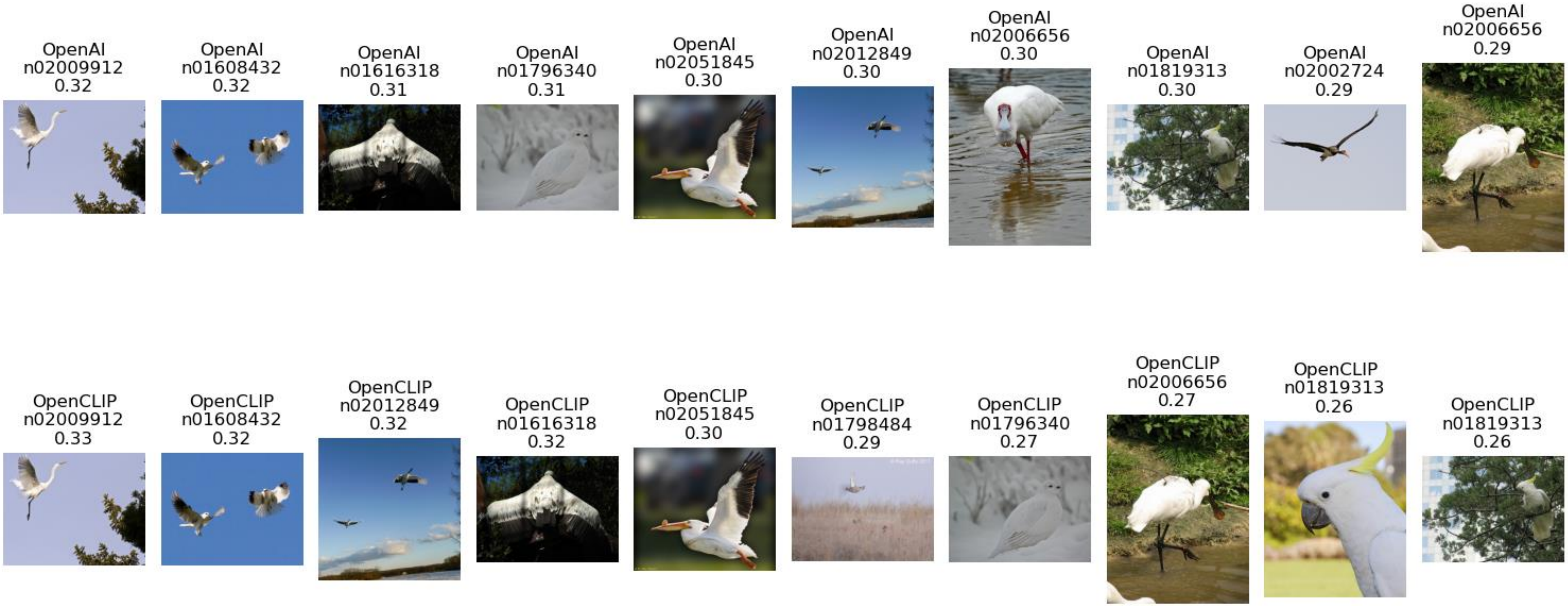


ImageNet-Mini

2. a photo of white bird

ImageNet-Mini

3. a photo of flying white bird



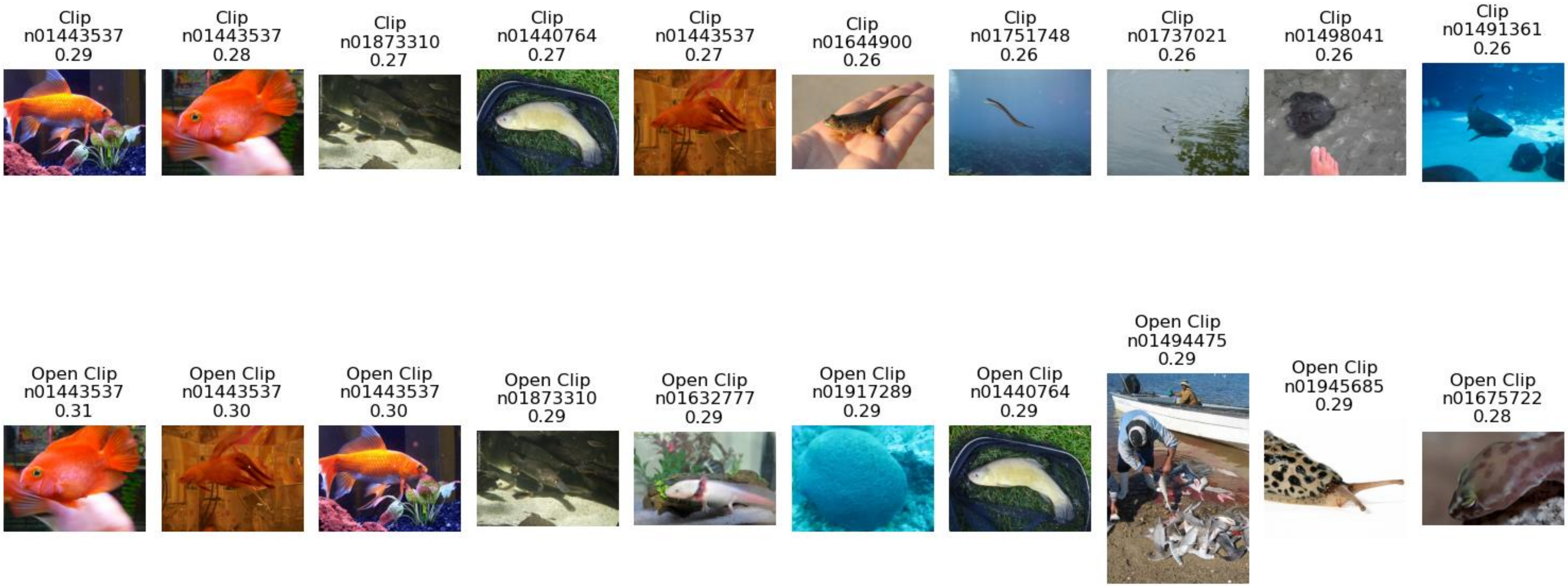
ImageNet-Mini

4. bird, white, flying

ImageNet-Mini

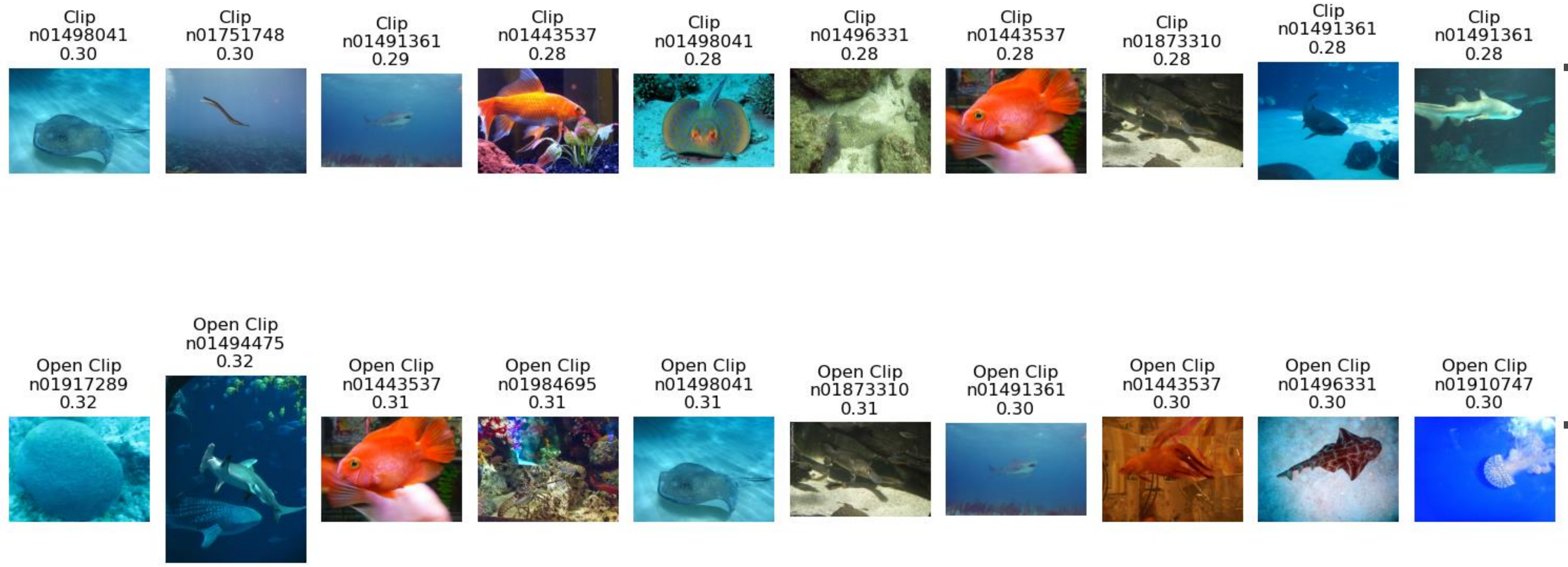
FISH

1. fish



ImageNet-Mini

2. a photo of ocean fish



ImageNet-Mini

3. a photo of river fish

ImageNet-Mini

4. CAN CLIP FIND THE ORIGINAL IMAGE FROM ITS HUMAN CAPTION?

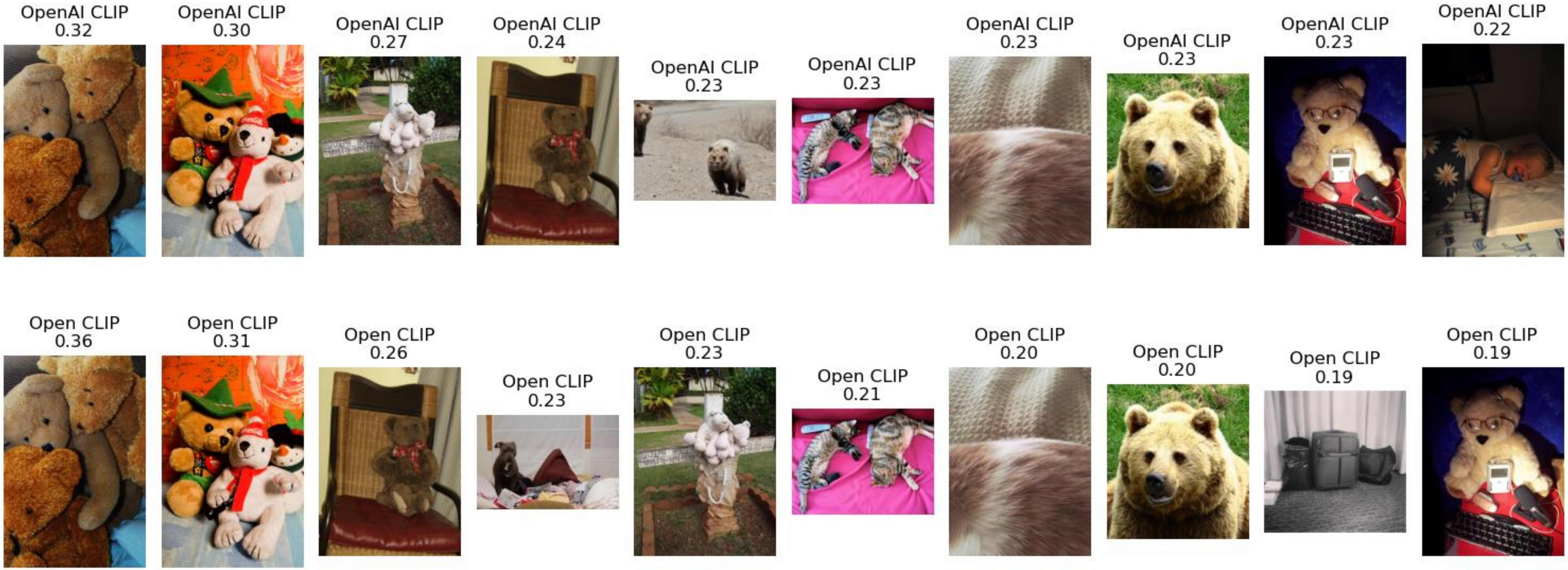
1. Three teddy bears, each a different color, snuggling together.

Original:

Coco

1. Three teddy bears, each a different color, snuggling together.

Results:



Coco

1. Three teddy bears, each a different color, snuggling together.

Addition: Teddy bear

Coco

HOWEVER, IF YOU HAVE MANY SIMILAR PICTURES...

2. The people are posing for a group photo.

Original:

Coco

2. The people are posing for a group photo.

Results:

OpenAI CLIP
0.26

OpenAI CLIP
0.26

OpenAI CLIP
0.26

OpenAI CLIP
0.26

OpenAI CLIP
0.25

OpenAI CLIP
0.25

OpenAI CLIP
0.25

OpenAI CLIP
0.25

OpenAI CLIP
0.25

OpenAI CLIP
0.25

Open CLIP
0.25

Open CLIP
0.25

Open CLIP
0.25

Open CLIP
0.25

Open CLIP
0.24

Open CLIP
0.23

Open CLIP
0.23

Open CLIP
0.22

Open CLIP
0.22

Open CLIP
0.22

Coco

3. A meal is lying on a plate on a table.

Original:

Coco

3. A meal is lying on a plate on a table.

Results:

Coco

Key findings:

- **Prompt engineering is powerful:** Small changes in prompt wording can dramatically affect results.
- **Model differences:** OpenCLIP and OpenAI CLIP sometimes favor different image styles or concepts.
- **Practical:** Good prompts yield high accuracy in image search without any fine-tuning.

Summary:

- CLIP models transfer well across tasks with minimal training.
- Zero-shot is effective, but **linear probe offers significant gains**.
- Prompt design & model size are key to performance.

Future Improvements:

- Current datasets are **not large or diverse enough** to fully test generalization.
- **No numerical metrics** were used for retrieval tasks; evaluation was **manual**.
- Explore **few-shot tuning** strategies for better adaptation.
- Investigate **automatic prompt optimization** techniques.

REFERENCE

- **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. In Proceedings of the International Conference on Machine Learning (ICML)*.
https://github.com/openai/CLIP

- **OpenCLIP**

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. (2021).. *OpenCLIP* (Version 0.1) [Computer software]. Zenodo.
https://github.com/mlfoundations/open_clip

THANK YOU