CVBench: A Benchmarking Pipeline for Chat-centric Vision Understanding

King Abdullah University of Science and Technology

Taha Al-Nufaili, Yahya Al Malallah, Chen Zhao, Shuming Liu, Bernard Ghanem
AI Initiative Lab
KAUST

Motivation

Multimodal LLMs have impressive abilities of processing both textual and visual inputs, making them suitable for various vision-related tasks. However, objective performance comparison is challenging. There is no standardized tool for evaluating and comparing their performance in vision tasks.

Objective

We propose CVBench: A standardized evaluation pipeline for fair multimodal model comparisons in vision tasks. CVBench can take a multimodal model and automatically evaluate it using vision datasets and under different vision tasks.

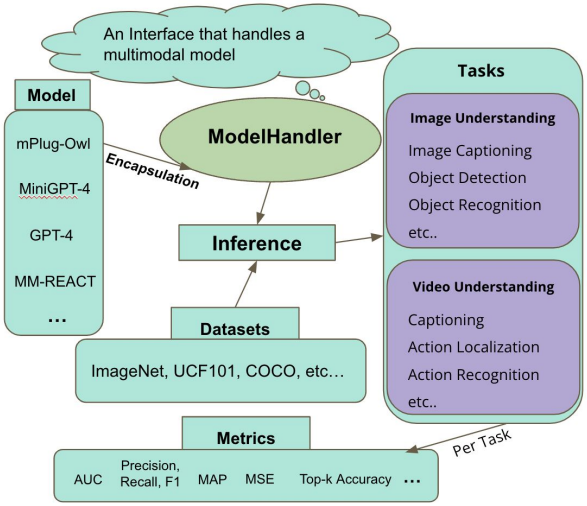

Figure 1. Pipeline Design

CVBench provides metrics like top-k accuracy, AUC, ROC, etc., to quantify model performance and facilitates objective model comparison using diverse vision datasets like UCF101 and ImageNet.

General Multimodal Architecture

Multimodal models extract vision and text embeddings and feed them into LLMs.

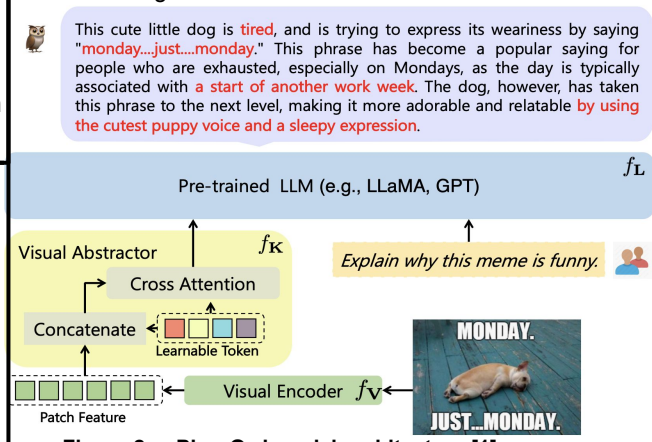

Figure 2. mPlug-Owl model architecture [1]

Pipeline Architecture

We engineered an Object-Oriented structure that can be generalized across different vision models on different datasets.

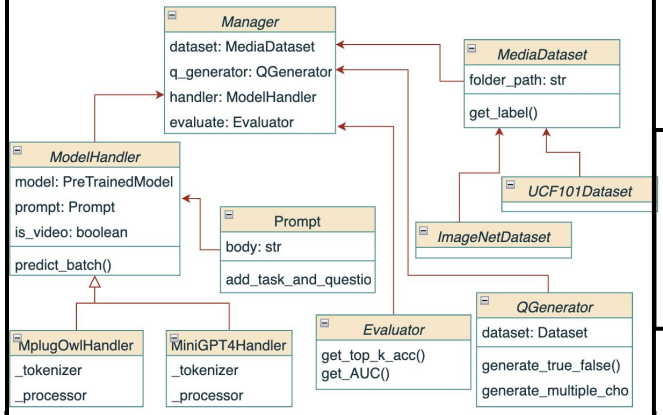

Figure 3. Pipeline Architecture

Evaluation

We use T/F, multiple choice, and open-ended questions for the inference. Then, to evaluate, we use pre-trained language models, like Bert, and NLP libraries, like SpaCy, to measure the similarity between an inference and a label.

Sample Runs on mPLUG-Owl:

Type	Expected	Result
True +	50	47
True -	4950	1276
False +	0	3674
False -	0	3
Total	5000	5000

Figure 4. True/False on ImageNet (5000 images, 100 categories)

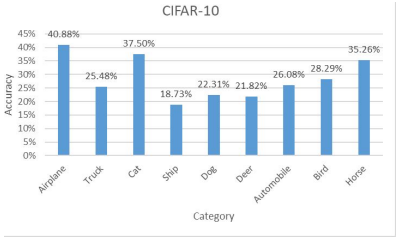

Figure 5. Multiple Choice on CIFAR-10 (10000 images, 10 categories)

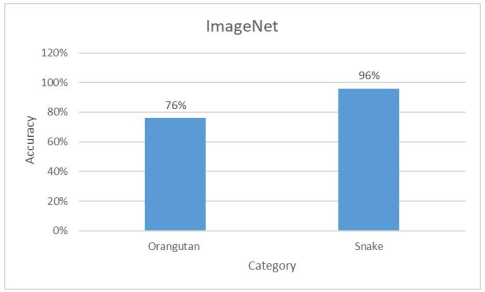

Figure 6.Open-ended on ImageNet. (100 images, 2 categories)

Conclusion

CVBench offers a standardized evaluation pipeline to facilitate fair comparisons of multimodal models in vision tasks. Researchers can utilize CVBench to assess their models across vision tasks, including object detection, action recognition, and captioning.

References

[1] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., & Huang, F. (2023). mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality.