Unlabeled Data Improves Fine-Grained Image Zero-shot Classification with Multimodal LLMs

Yunqi Hong, Sohyun An, Andrew Bai, Neil Y.C. Lin, Cho-Jui Hsieh
University of California, Los Angeles

Introduction & Motivation

Multimodal Large Language Models (MLLMs) have recently shown strong zero-shot performance on general image classification tasks. However, **fine-grained classification remains a major challenge**.

- Fine-grained classification: distinguishing between visually similar subcategories such as bird species, car models, or flower types.
- These tasks require the model to notice **subtle, localized visual cues**, yet MLLMs often **overlook fine details**.

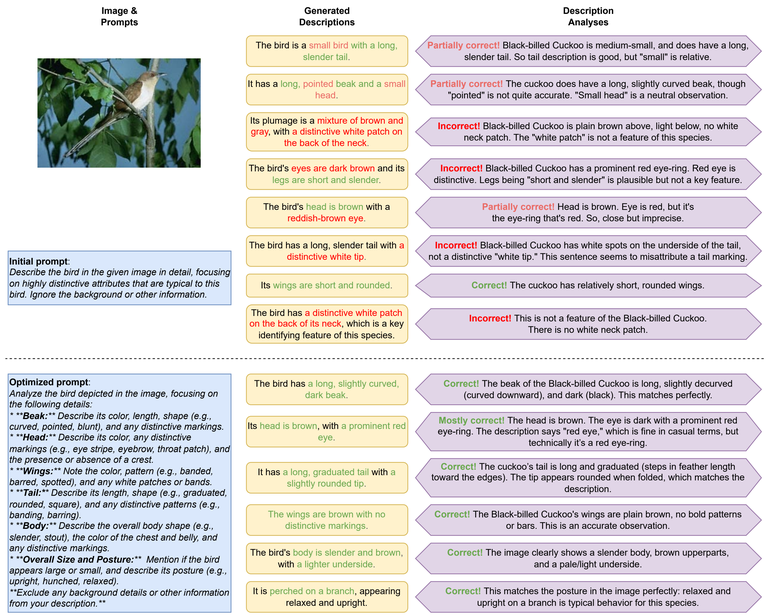
Existing approaches depend on:

- Large labeled datasets;
- Computationally expensive fine-tuning.

Large amounts of unlabeled images are easily accessible at test time, but current MLLMs do not know how to leverage such data to self-improve.

➢ **Can an MLLM enhance its own fine-grained classification accuracy using only unlabeled data and without any model updates?**

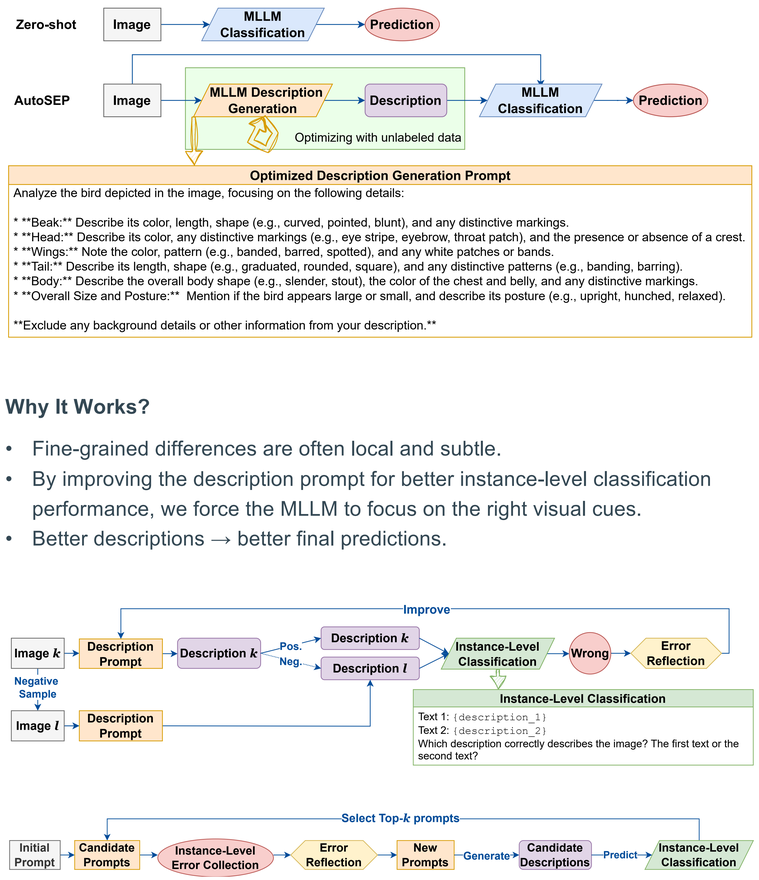
No finetune, no white-box access, no ground-truth label



Our Method: AutoSEP

AutoSEP (Automatic Self-Enhancing Prompt Learning): a fully unsupervised, black-box prompt learning framework.

1. **Generates image descriptions** that capture fine-grained details. This description highlights key visual details the model should attend to.
2. Evaluates these descriptions using an **instance-level scoring function** based only on unlabeled data.
3. **Iteratively improves** the description prompt, allowing the MLLM to self-enhance over time and focus on the right discriminative features.



Why It Works?

- Fine-grained differences are often local and subtle.
- By improving the description prompt for better instance-level classification performance, we force the MLLM to focus on the right visual cues.
- Better descriptions → better final predictions.

Experiment Results

Overall Performance

- AutoSEP consistently improves fine-grained classification across all tested MLLMs.
- +13% average improvement over standard zero-shot classification.
- +3% improvement over the best unsupervised baselines.

	CUB_vireo			*iNat_butterfly*		
Optimization-free						
Vanilla zero-shot	49.21±3.36	75.73±1.83	55.28±3.05	61.14±2.11	68.00±1.46	40.25±4.65
Zero-shot with descriptions	51.46±4.11	86.07±0.90	50.79±3.72	56.57±3.45	71.14±3.88	48.25±2.48
Zero-shot with majority vote	49.21±1.80	76.69±2.01	55.06±3.01	62.00±2.14	66.07±1.86	46.56±2.59
Few-shot with random labels	45.62±8.03	57.08±18.2	43.82±7.78	45.43±13.8	64.00±10.7	32.13±8.47
Few-shot with MLLM labels	44.94±8.56	72.36±10.5	46.97±7.63	51.43±3.26	82.57±14.3	42.13±12.2
Multiple images display [16]	56.18±7.95	68.54±13.6	63.82±5.34	55.71±1.56	72.29±12.2	32.25±3.44
K-means clustering	40.22±9.25	57.08±11.2	57.53±15.0	58.57±5.89	59.14±5.39	38.38±6.92
Optimization-based						
With random labels	41.24±6.85	72.81±3.13	48.31±4.44	53.24±1.72	79.86±4.43	45.13±1.71
With majority vote	53.71±5.41	87.19±1.83	51.24±2.72	62.29±1.63	61.07±1.89	51.50±3.36
SPO	52.81±9.59	70.34±3.73	51.69±1.59	53.14±2.24	79.71±9.14	45.26±2.47
AutoSEP (Ours)	**59.18**±0.53	**87.42**±1.10	**65.45**±1.24	**66.57**±1.42	**82.71**±2.43	**55.38**±1.00

Effectiveness of Instance-Level Classification Signal

- Classification performance **steadily improves across iterations**.
- Both instance-level and class-wise accuracy increase together.
- **Strong positive correlation** between instance-level and class-wise accuracy supports using instance-level retrieval as the optimization signal.

Table 2: Correlation between instance-level classification and class-wise classification.

Datasets	*CUB_cuckoo*	*CUB_oriole*	*CUB_vireo*	*iNat_butterfly*	*iNat_lupine*
Gemini 1.5 Flash	0.72±0.11	0.47±0.24	0.70±0.19	0.63±0.22	0.53±0.02
Qwen2-VL-72B-Instruct	0.63±0.13	0.38±0.15	0.64±0.10	0.84±0.11	0.70±0.17

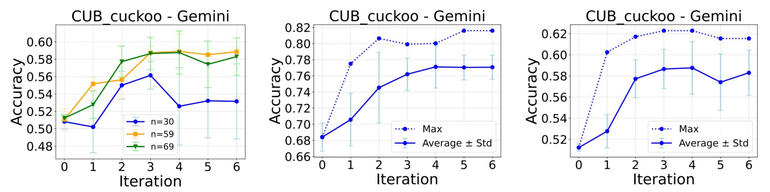


Figure 4: Classification accuracy of Gemini with various number of samples for optimization.

(a) Instance-level Classification

(b) Class-wise Classification