You May Speak Freely: Improving the Fine-Grained Visual Recognition Capabilities of Multimodal Large Language Models with Answer Extraction

Supplementary Material

A. Labeling Species Answer Extraction Data

Motivation. For labeling answer extraction data, we prepare a user interface built with Label Studio which shown in Fig. 5. On the right contains information on the inputs to the model, specifically the question, image, and ground truth label. On the left is main interface, which consists a free-form response from a model to the example displayed on the right, as well as the choice predicted by the model via *nlg2choice*. Labeling consists of two pieces: (1) identification and highlighting of text spans containing answers to the input question and (2) matching the highlighted text span to a class within the given schema.

Text span labeling. For the text span labeling task, labelers read through the text to identify the first incidence of a predicted species. Before highlighting the species found within the text, the prediction is confirmed to be a real species through a brief internet search. If there are any issues with highlighting the predicted species, a text description of the issue is written in the input text box at the bottom and the example is skipped.

Choice assignment. Once the span is labeled, we next wish to either assign the span to a class within the dataset or tag the span as a schema failure. For expediency, the labelers are also equipped with the nlg2choice prediction and a tool for sorting most like class names with fuzzy matching score. When the nlg2choice predict matches the text span, the prediction is "Consistent with nlg" and the labelers check the button. Otherwise, "Inconsistent with nlg" is checked and a natural language explanation is put in the bottom text box. When the nlg2choice is inconsistent, "answer={`species`}" is put within that text box.

B. Zero-Shot 4-Way VQA Performance

Answer extraction does not consistently increase performance. Following recent work [36], we evalute *nlg2choice* on popular fine-graiend classification datasets as a 4-way VQA task. iNat-Animal and iNat-Plant are the rows of iNat21 [39] when subset to rows which whose kingdom matches "Animalia" and "Plantae." ImgNet-Animal and ImgNet-Artifact are the rows of ImageNet-1K [32] subset to rows which have an anceter of "Animal" and "Artifact" in the WordNet hierarchy [26]. The choices for each example are from the top 3 SigLIP [48] predictions and the correct class. For the *nlg2choice* prompt we use the same prompt as the previous work. We report the open-ended question performance in Tab. 8 where we find that it slightly underperforms or has not effect on the lettering approach. Specifically, we find that accuracy changes by **-0.03**, **+0.58**, and **+0.30** on average for Qwen-2.5VL, Llama-3.2V, and Intern3VL, respectively.

Method	CUB200	iNat-Animal	iNat-Plant	ImgNet-Animal	ImgNet-Artifact
Qwen-2.5VL-7B					
A/B/C/D	65.50	**41.33**	**41.61**	**85.20**	80.01
nlg2choice$_{open}$	**67.43** (**+1.93**)	40.19 (-1.14)	40.03 (-1.58)	84.03 (-1.17)	**81.81** (**+1.80**)
Llama-3.2-Vision-11B					
A/B/C/D	65.52	**32.44**	**31.88**	79.93	**75.89**
nlg2choice$_{open}$	**68.79** (**+3.27**)	31.52 (-0.92)	31.57 (-0.31)	**81.39** (**+1.46**)	75.29 (-0.60)
Intern3VL-8B					
A/B/C/D	**50.52**	35.40	**36.39**	77.50	69.41
nlg2choice$_{open}$	47.30 (-3.22)	**35.86** (**+0.46**)	35.64 (-0.75)	**81.91** (**+4.41**)	**70.02** (**+0.61**)

Table 1. **4-way VQA accuracy across fine-grained classifcation tasks.**

Decreasing the choice set improves performance. We are also able to compare the 4-way VQA setting of CUB200 directly to the many-way prediction displayed in Tab. 2. We see that subsetting the full choice set to four choices incurs a performance improvement of **+7.55**, **+19.20**, and **+28.83** for Qwen-2.5VL, Llama-3.2V, and Intern3VL, respectively.

C. Prompt Variations

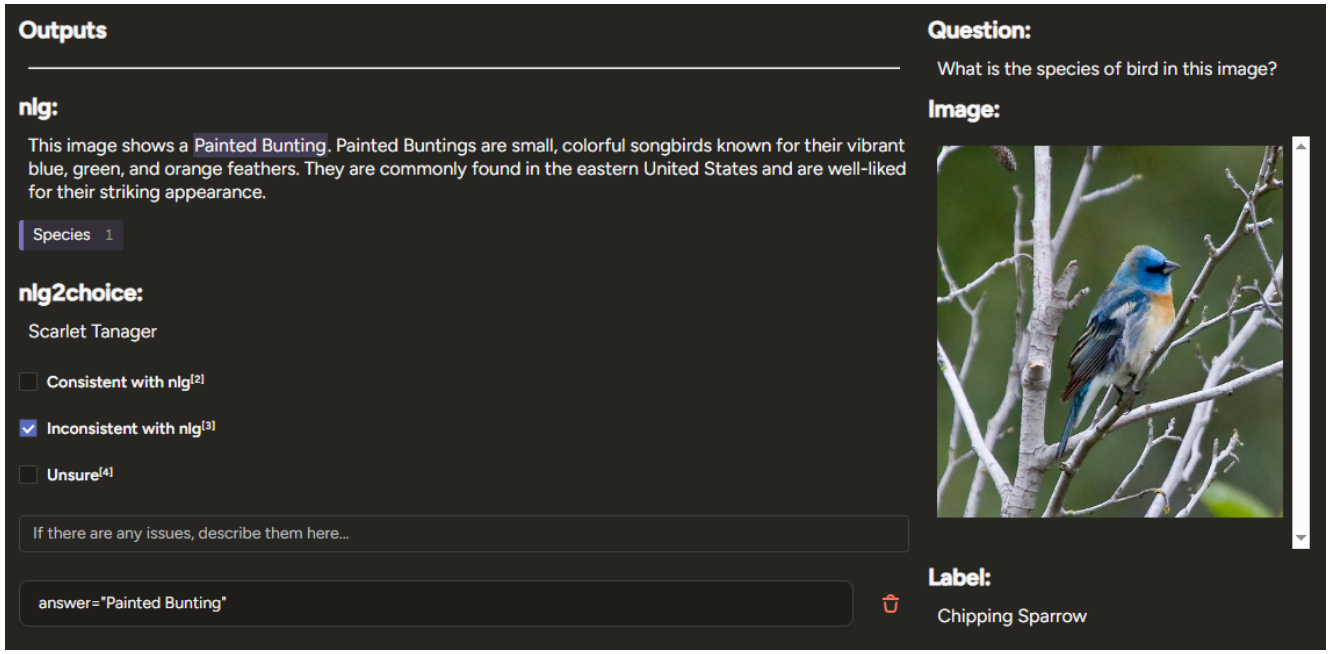

Figure 1. **Labeling interface for efficacy of the answer extraction stage of *nlg2choice*.** Labelers are asked to highlight the name predicted in free-form responses from various models and whether the name occurs easily within the dataset choice list. Above shows an interesting example: for the Chipping Sparrow on the right, the model freely responded with "Painted Bunting" and answer extraction made an additional mistake by predicting "Scarlet Tanager."

```
"What { type } is this { domain }?"
"What is the { type } of this { domain }?"
"What is the { type } of the { domain }?"
"What is the { type } of the { domain } in this image?"
"What is the { type } of the { domain } in the image?"
"Identify this { domain }'s { type }."
"Name the { type } shown in the image."
"Which { domain } { type } is pictured here?"
"Classify the { type } of this { domain }."
"What { domain } { type } does the photo depict?"
"Determine the { type } of the { domain } in view."
"Provide the common name of this { domain }."
"To which { type } does this { domain } belong?"
"Label the { type } of the { domain } shown."
"Recognize and state this { domain }'s { type }."
```

Table 2. **Base template variations generated by Sec. 3.1**

Dataset	Type	Domain
CUB200	species	bird
Flowers	species	flower
Aircrafts	variant	aircraft
Cars	year, make, and model	car
Foods	name	food
NABirds	species	bird
iNat-Birds	species	bird

Table 3. **Dataset variables for filling out prompt templates.**