

CATALOG: A Camera Trap Language-guided Contrastive Learning Model

Julian D. Santamaria¹, Claudia Isaza¹, Jhony H. Giraldo²

¹ SISTEMIC, Faculty of Engineering, Universidad de Antioquia-UdeA, Medellín, Colombia.

² LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France.

{julian.santamaria, victoria.isaza}@udea.edu.co, jhony.giraldo@telecom-paris.fr

This supplementary material includes some examples of the predefined templates and detailed examples of specific prompts used to generate the category descriptions with the LLM [1]. Additionally, it provides the prompt used to create textual descriptions of animals present in cropped images using LLaVA [3].

A. Templates

In this section, we present some examples of templates designed for the camera-trap image recognition task. These templates have been adapted from the ImageNet templates used in CLIP [5] and are shown below:

- a photo captured by a camera trap of a {}.
- a camera trap photo of the {} captured in poor conditions.
- a cropped camera trap image of the {}.
- a camera trap image featuring a bright view of the {}.
- a camera trap image of the {} captured in clean conditions.
- a camera trap image of the {} captured in dirty conditions.
- a camera trap image with low light conditions featuring the {}.
- a black and white camera trap image of the {}.
- a cropped camera trap image of a {}.
- a blurry camera trap image of the {}.
- a camera trap image of the {}.
- a camera trap image of a single {}.
- a camera trap image of a {}.
- a camera trap image of a large {}.
- a blurry camera trap image of a {}.
- a pixelated camera trap image of a {}.
- a camera trap image of the weird {}.
- a camera trap image of the large {}.
- a dark camera trap image of a {}.
- a camera trap image of a small {}.

For each template, we replace “{ }” by the specific category in \mathcal{C}^D .

B. Prompts

In this section, we provide the prompts for the LLM and LLaVA used in CATALOG.

B.1. Prompt LLM

The prompt utilized to get the LLM description of the animal species follows a structure based on the methodology discussed in [4] and is shown below.

You are an AI assistant specialized in biology and providing accurate and detailed descriptions of animal species. We are creating detailed and specific prompts to describe various species. The goal is to generate multiple sentences that capture different aspects of each species’ appearance and behavior. Please follow the structure and style shown in the examples below. Each species should have a set of descriptions that highlight key characteristics.

Example Structure:

Badger:

- a badger is a mammal with a stout body and short sturdy legs.
- a badger’s fur is coarse and typically grayish-black.
- badgers often feature a white stripe running from the nose to the back of the head dividing into two stripes along the sides of the body to the base of the tail.
- badgers have broad flat heads with small eyes and ears.
- badger noses are elongated and tapered ending in a black muzzle.
- badgers possess strong well-developed claws adapted for digging burrows.
- overall badgers have a rugged and muscular appearance suited for their burrowing lifestyle.

B.2. Prompt LLaVA

The prompt used in LLaVA aligns with the approach employed in [2] and is structured as follows:

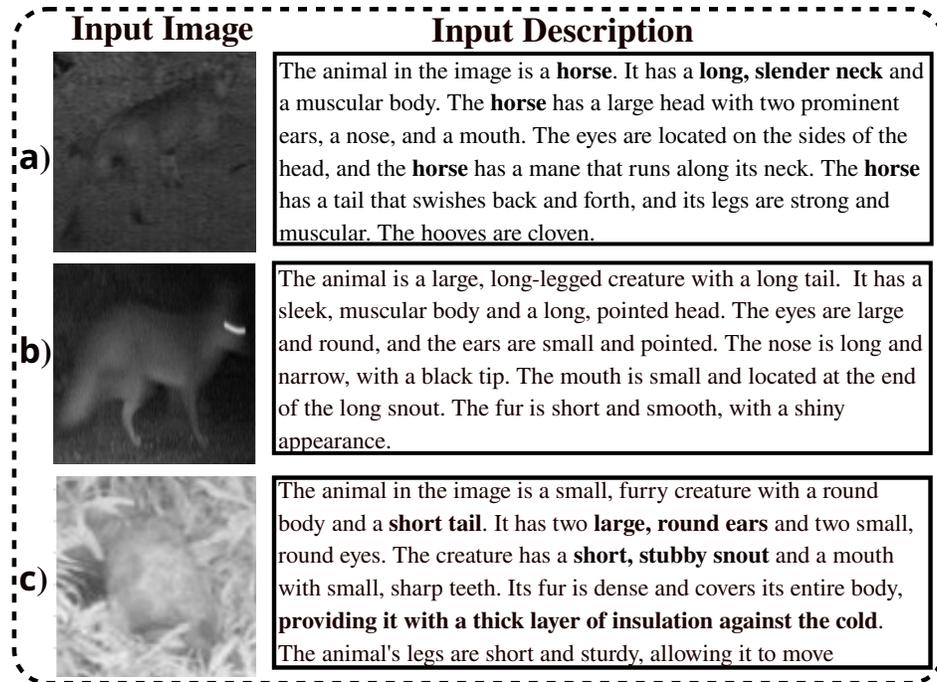


Figure 1. Failure cases of the CATALOG model for camera trap image classification. **case a**: The VLM generates an incorrect description with details that do not match the input image. **case b**: A blurry image results in a vague and unhelpful description. **case c**: When the input image is very unclear, the VLM generates a random and irrelevant description. These examples show that when the descriptions are wrong or not informative, the model makes incorrect predictions.

SYSTEM: You are an AI assistant specialized in biology and providing accurate and detailed descriptions of animal species. \n << image >> \n

USER: You are given the description of an animal species. Provide a very detailed description of the appearance of the species and describe each body part of the animal in detail. Only include details that can be directly visible in a photograph of the animal. Only include information related to the appearance of the animal and nothing else. Make sure to only include information that is present in the species description and is certainly true for the given species. Do not include any information related to the sound or smell of the animal. Do not include any numerical information related to measurements in the text in units: m cm in inches ft feet km/h kg lb lbs. Remove any special characters such as unicode tags from the text. Return the answer as a single paragraph.

C. Analysis of CATALOG’s Errors

In this section, we provide a more detailed analysis of the limitations of the CATALOG model. Fig. 1 illustrates how sensitive the model is to the input descriptions generated for the VLM. These descriptions provide additional information to help the model make better classifications. However, when the descriptions are wrong or unclear, the model makes incorrect predictions.

To clarify the model’s sensitivity to input descriptions, Fig. 1 presents three examples:

- In **case a**, the VLM hallucinates by adding details that are not in the image. For example, it describes the animal as a horse, even though the input image does not match this description. This leads to a completely incorrect prediction.
- In **case b**, the input image is blurry and hard to interpret. The VLM generates a vague description with little useful information. As a result, the model does not get enough context to make the correct prediction.
- In **case c**, the image is so unclear that the VLM creates a random description that has no connection to the input. This random description further confuses the model, leading to a wrong prediction.

These examples show that the CATALOG model depends on the accuracy and quality of the descriptions created by the VLM. When the descriptions are not reliable or informative, the model struggles to classify the input correctly. Improving the robustness of the VLM is crucial for handling noisy or unclear inputs.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan,

- Pranav Shyam, Girish Sastry, Amanda Asbell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020. [1](#)
- [2] Zalan Fabian, Zhongqi Miao, Chunyuan Li, Yuanhan Zhang, Ziwei Liu, Andrés Hernández, Andrés Montes-Rojas, Rafael Escucha, Laura Siabatto, Andrés Link, et al. Multimodal foundation models for zero-shot animal species recognition in camera trap images. *arXiv preprint arXiv:2311.01064*, 2023. [1](#)
- [3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2024. [1](#)
- [4] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. [1](#)
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Asbell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. [1](#)