Fine-tuning Vision-Language Models for Animal Behavior Analysis

Sepideh Mamooler Haozhe Qi Valentin Gabeff Syrielle Montariol
Antoine Bosselut* Alexander Mathis*
EPFL
{firstname.lastname}@epfl.ch

Abstract

Animal behavior analysis is fundamental to ethology, behavioral ecology, and neuroscience. Current methods typically use vision-only classifiers, which are task-specific and limited to closed-vocabulary classification paradigms. Vision-language models (VLMs) show strong video question-answering (VideoQA) performance across domains but remain underexplored for animal behavior. We present a novel framework that converts existing datasets into a comprehensive multi-task VideoQA dataset with code-based solutions without extra annotation. Fine-tuning InternVL3-8B on this dataset, we achieve up to 33.2 and 26.9 percentage point improvement over supervised vision-only baselines and zero-shot VLMs with 10× more parameters, respectively. Our systematic evaluation demonstrates the superiority of vision-language approaches and advances interpretable, code-based predictions for behavioral analysis.

1 Introduction

Animal behavior analysis underpins fields like neuroscience (Cisek & Green, 2024; Mathis et al., 2024), ethology (Anderson & Perona, 2014), and behavioral ecology (Tuia et al., 2022; Couzin & Heins, 2023). Modern research increasingly uses large video datasets with species and behavior annotations (Liu et al., 2023; Brookes et al., 2024; Rogers et al., 2023; Ma et al., 2023; Chen et al., 2023; Kholiavchenko et al., 2024; Duporge et al., 2025; Gabeff et al., 2025). To infer those variables, existing approaches typically train task-specific vision-only classification models (Feichtenhofer et al., 2019; Tong et al., 2022; Li et al., 2022; Tran et al., 2015; Carreira & Zisserman, 2017; Feichtenhofer, 2020) on a single dataset. However, these models cannot generalize to new behaviors without retraining and are limited to their trained tasks. Moreover, training these models on combinations of datasets with differing semantics is challenging—an area where language models could help due to the flexibility of language embeddings.

Vision-language models (VLMs) have shown strong video question answering (VideoQA) capabilities across domains like art, science, and sports (Zhu et al., 2025; Zhang et al., 2025; Lin et al., 2023; Li et al., 2024), but their use in animal behavior analysis is largely unexplored. A recent study by Sun et al. (2024) illustrates that a contrastive, large-scale VLM, VideoPrism (Zhao et al., 2024), outperforms specialist vision models in zero- and few-shot behavior classification, as demonstrated across mice, flies, and Kenyan wildlife recorded from drones. Jing et al. (2024) and Dussert et al. (2025) present broad zero-shot evaluations of generative VLMs across diverse tasks, highlighting significant room for improvement, partly due to a domain gap between typical internet training data and fine-grained animal behavior (Gabeff et al., 2024; Stevens et al., 2024). They focus on multiple-choice QA, which is impractical for real-world behavioral analysis. Santo et al. (2025) and Xu et al. (2025) show that species-specific fine-tuning for primates and mice enhances performance, suggesting the potential of adapted VLMs. However, these studies are limited to a few species and lack comparison with established vision-only baselines.

*Equal supervision

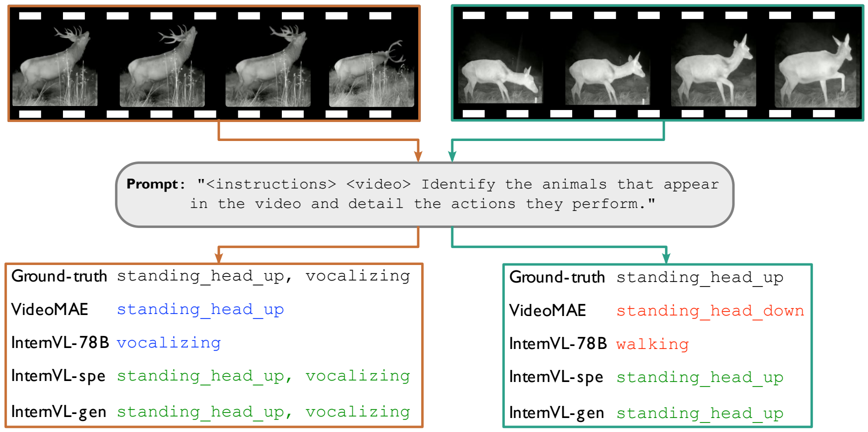

Figure 1: Examples from the MammAlps dataset with ground-truth annotation and prediction of our models vs baselines. Best viewed in color. We sampled 32 frames from the input videos. Correct predictions are indicated in green, incorrect ones in red, and partially correct ones in blue.

Dataset	Video hours	Source	# Species	# Actions	# Activities	Train size	Test size
AnimalKingdom	50	YouTube	850	140	N/A	24.0K	6.0K
MammalNet	394	YouTube	173	12	N/A	13.3K	5.0K
MammAlps	8.5	Camera trap	5	19	11	4.2K	1.2K

Table 1: Source datasets statistics.

In this work, we present a novel method to convert existing animal behavior datasets with action, animal, and activity annotations into multi-task datasets with annotations in code format. Using this framework, we curate a comprehensive dataset to fine-tune InternVL3-8B (Zhu et al., 2025), a state-of-the-art 8B-parameter VLM. We fine-tune specialist VLMs on individual datasets and tasks, demonstrating substantial gains over vision-only baselines. Our generalist VLM, trained jointly on four tasks and three datasets, outperforms zero-shot VLMs that have 10× more parameters while producing code-based intermediate solutions that improve prediction reliability and trustworthiness.

2 Dataset Framework

We present a framework that transforms existing animal species and behavior classification datasets into multi-task question-answering datasets with code-based outputs, without requiring additional human or LLM-based annotation.

Source Datasets and Tasks. We used three complementary datasets (Table 1) covering a broad spectrum of animal behavior and taxa. **AnimalKingdom** (Ng et al., 2022) spans diverse taxa and action labels and is sourced from YouTube videos. **MammalNet** (Chen et al., 2023), also from YouTube, focuses on common behaviors shared across mammalian species. **MammAlps** (Gabeff et al., 2025) adds ecological diversity via camera trap footage from the Swiss Alps and includes hierarchical behavioral annotations (action and activity). We unified the source datasets into a single multi-task dataset with four core animal behavior tasks: **Animal Recognition** for identifying the species; **Action Recognition** for identifying actions; **Activity Recognition** for identifying higher-level behavior often comprising multiple actions; and **Joint Animal-Action Recognition** for identifying both species and actions simultaneously. We retained the original train-test splits for fair comparison with prior works. The combined dataset includes 69K unique videos and 152K annotations.

Input-Output Structure. Each input prompt includes task-specific instructions, output format, relevant definitions (e.g., action, activity), and the appropriate label spaces for

Dataset	Model	AnimalR	ActionR	ActivityR	Animal-ActionR
MammalNet	SlowFast (Chen et al., 2023)	43.0	39.4	-	22.8
	C3D (Chen et al., 2023)	44.4	40.3	-	24.6
	I3D (Chen et al., 2023)	43.4	41.2	-	24.0
	MViTV2 (Chen et al., 2023)	52.6	46.6	-	30.6
	InternVL3-8B-spe (ours)	76.9	66.3	-	49.3
	InternVL3-8B-gen (ours)	79.9	68.8	-	51.9
MammAlps	VideoMAE (Gabeff et al., 2025)	53.7/96.8	44.7/52.1	44.0/51.7	-
	InternVL3-8B-spe (ours)	-/96.5	-/56.0	-/53.5	-
	InternVL3-8B-gen (ours)	-/97.1	-/57.5	-/59.1	-
AnimalKingdom	I3D (Ng et al., 2022)	-	24.9/ -	-	-
	SlowFast (Ng et al., 2022)	-	24.4/ -	-	-
	X3D (Ng et al., 2022)	-	30.6/ -	-	-
	VideoMAE	71.2/56.2	53.5/52.7	-	14.0/15.3
	InternVL3-8B-spe (ours)	-/83.8	-/74.3	-	-/43.3
	InternVL3-8B-gen (ours)	-/88.9	-/79.4	-	-/48.5

Table 2: Comparison of fine-tuned VLMs with reported vision-only models. All models are fine-tuned on each dataset separately, except for InternVL3-gen. We report top-1 accuracy for MammalNet and mAP and Jaccard Index for MammAlps and AnimalKingdom in **mAP/Jaccard Index** format.

animals, actions, and activities, followed by video frames and the question. We provide example prompts and annotations in Appendix B.2.

To improve robustness to phrasing variations, we generated 10 question templates per task using ChatGPT. Given the large taxonomies, we applied a strategic sampling from the datasets' label space to ensure representative yet tractable label spaces (see Appendix B.1 for more details). The output follows a code-based format centered on a base function, `recognize`, which identifies entity instances under given conditions (e.g., `recognize(entity_type='action', condition='animal == dog')`), providing a unified interface across tasks. The annotations are a step-by-step solution in the form of code, derived directly from the original annotations, preserving accuracy while adding structure.

3 Experimental Setup

We evaluated on the test splits of the three source datasets (Sec. 2). For each predicted entity (action, animal, animal-action pair, activity), we checked for exact matches in the ground truth. We report F1 score, mean average precision (mAP), and to handle partial correctness in multi-label settings, Jaccard Index; for single-label cases, this reduces to top-1 accuracy. As zero-shot baselines, we used GPT-4o, InternVL3-8B, and InternVL3-78B. For supervised baselines, we included task-specific vision-only models trained on each source dataset, as reported in Ng et al. (2022); Chen et al. (2023); Gabeff et al. (2025). In all our experiments with VLMs, we used 32 uniformly sampled video frames. We fully fine-tuned InternVL3-8B following the official recommendations (Appendix A).

4 Results and Discussion

We considered two research questions: 1) Does a specialist VLM, fine-tuned on a single task and dataset, perform better than vision-only counterparts? 2) Does an 8B-scale generalist VLM, fine-tuned on a collection of tasks and datasets, perform better than large, state-of-the-art open-source and proprietary VLMs?

Vision-Only Baselines. For MammalNet (single-label), we report top-1 accuracy for various baselines (Chen et al. (2023)). However, for MammAlps and AnimalKingdom (multi-label), the authors reported mAP, which is not well-defined for generative models. Thus, for MammAlps, we computed the Jaccard Index and report it on top of mAP (Table 2). For AnimalKingdom – not having access to Ng et al. (2022)'s checkpoints – we fine-tuned VideoMAE (Tong et al., 2022) and report mAP and Jaccard Index. Given that VideoMAE

Base model	Training	AnimalR	ActionR	ActivityR	Animal-ActionR
GPT-4o		-	-	-	16.4
InternVL3-8B	zero-shot	43.1	30.9	46.6	10.3
InternVL3-78B		71.4	52.2	57.6	30.2
InternVL3-8B	vanilla	92.7±0.8	84.7±1.1	74.3±5.0	66.9±1.7
InternVL3-8B	code-based	92.7±1.5	85.5±1.0	74.0±3.8	67.1±1.1

Table 3: F1 score comparison of our generalist VLM, trained with and without code-based output format, with zero-shot baselines. We report the mean and std over three training runs, considering the weighted average performance over the three datasets for each run.

outperforms the reported baselines in terms of mAP, it can serve as a reference when comparing against VLMs using the Jaccard Index.

Specialist VLMs. First, we fine-tuned InternVL3-8B on each dataset and task individually. These specialised VLMs strongly outperformed vision-only baselines by up to 28 percentage points (Table 2, *InternVL3-8B-spe*). These results illustrate that fine-tuned VLMs outperform state-of-the-art fine-tuned vision-only models. Qualitative examples from MammAlps illustrate the performance gap (Figure 1).

Beyond these performance improvements, the biggest advantage of VLMs is that one can train generalist models across datasets and tasks with different semantic annotations, which we tackled next.

Generalist VLM. We considered strong zero-shot models as baselines and performed inference on the test sets of all datasets. Due to the high inference cost, we evaluated GPT-4o only on the joint animal-action task, as it's the most complex one (Table 3). Our generalist 8B-parameter VLM, fine-tuned on all datasets jointly, outperformed the zero-shot performance of InternVL3-78B by an average of 26.9 percentage points across all tasks. Moreover, the generalist VLM further outperformed specialist VLMs trained on each dataset separately (Table 2, *InternVL3-8B-gen*), highlighting the promise of merging behavioral datasets and task diversity in fine-tuning data. We also ablated the impact of code-based output format by fine-tuning InternVL3-8B on input-output pairs of our dataset without the code-based solutions (Table 3, *vanilla*). Code-based and vanilla formats led to similar performances, but the code-based format enables interpretation of the model's output by the user through structured and easily readable reasoning traces (see example in Appendix B.2).

5 Conclusion

In this work, we took a step toward systematically benchmarking VLMs for animal behavior tasks. We transformed existing animal behavior classification datasets into a multi-task videoQA dataset, which we used to fine-tune InternVL3-8B . Our results show that on all tasks and datasets, fine-tuned vision-language models significantly outperform the vision-only counterparts. Moreover, large models like GPT-4o underperform compared to much smaller fine-tuned VLMs, echoing prior findings on the limitations of zero-shot VLMs trained on general internet data (Gabeff et al., 2024; Santo et al., 2025; Xu et al., 2025), and highlighting the need for domain-specific adaptation. Finally, we show that VLMs can benefit from multi-task and multi-dataset training, generalizing knowledge drawn from very different environments and label spaces to outperform single-task, single-dataset VLMs. We also find that using a standardized code-based output format enables structured, step-by-step reasoning and greater transparency while keeping the performance high.

Ultimately, our approach supports the integration of VLMs into broader scientific pipelines, providing ethologists, ecologists and neuroscientists with strong tools to scale up complex animal behavior analysis tasks (Ye et al., 2023; Mathis et al., 2024; Xu et al., 2025).

Acknowledgments

We thank Andy Bonnetto for their feedback on the dataset framework. This work was funded by Swiss SNF grant (320030-227871), EPFL's SV-ENAC I-PhD program (G.V.) and Boehringer Ingelheim Fonds PhD stipends (H.Q. and S.M.). We also gratefully acknowledge the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Center for Imaging, Sony Group Corporation, and a Meta LLM Evaluation Research Grant.

References

David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, Oct 2014. URL https://doi.org/10.1016/j.neuron.2014.09.005.

Otto Brookes, Majid Mirmehdi, Colleen Stephens, Samuel Angedakin, Katherine Corogenes, Dervla Dowd, Paula Dieguez, Thurston C Hicks, Sorrel Jones, Kevin Lee, et al. Panaf20k: a large video dataset for wild ape detection and behaviour recognition. *International Journal of Computer Vision*, 132(8):3086–3102, 2024.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13052–13061, 2023.

Paul Cisek and Andrea M Green. Toward a neuroscience of natural behavior. *Current opinion in Neurobiology*, 86:102859, 2024.

Iain D Couzin and Conor Heins. Emerging technologies for behavioral research in changing environments. *Trends in Ecology & Evolution*, 38(4):346–354, 2023.

Isla Duporge, Maksim Kholiavchenko, Roi Harel, Scott Wolf, Daniel I Rubenstein, Margaret C Crofoot, Tanya Berger-Wolf, Stephen J Lee, Julie Barreau, Jenna Kline, et al. Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos. *International Journal of Computer Vision*, pp. 1–12, 2025.

Gaspard Dussert, Vincent Miele, Colin Van Reeth, Anne Delestrade, Stéphane Dray, and Simon Chamaillé-Jammes. Zero-shot animal behaviour classification with vision-language foundation models. *Methods in Ecology and Evolution*, 2025.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, 132(9):3770–3786, 2024.

Valentin Gabeff, Haozhe Qi, Brendan Flaherty, Gencer Sumbul, Alexander Mathis, and Devis Tuia. Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13854–13864, 2025.

Yinuo Jing, Ruxu Zhang, Kongming Liang, Yongxiang Li, Zhongjiang He, Zhanyu Ma, and Jun Guo. Animal-bench: Benchmarking multimodal video models for animal-centric video understanding. *Advances in Neural Information Processing Systems*, 37:78766–78796, 2024.

Maksim Kholiavchenko, Jenna Kline, Michelle Ramirez, Sam Stevens, Alec Sheets, Reshma Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, et al. Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 31–40, 2024.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. Lote-animal: A long time-span dataset for endangered animal behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20064–20075, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xiaoxuan Ma, Stephan Kaufhold, Jiajun Su, Wentao Zhu, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. *Advances in Neural Information Processing Systems*, 36:27501–27531, 2023.

Mackenzie W. Mathis, Adriana Perez Rotondo, Edward F. Chang, Andreas Savas Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187:5814–5832, 2024. URL https://api.semanticscholar.org/CorpusID:273378461.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19023–19034, 2022.

Mitchell Rogers, Gaël Gendron, David Arturo Soriano Valdez, Mihailo Azhar, Yang Chen, Shahrokh Heidari, Caleb Perelini, Padriac O'Leary, Kobe Knowles, Izak Tait, et al. Meerkat behaviour recognition dataset. *arXiv preprint arXiv:2306.11326*, 2023.

Giulio Cesare Mastrocinque Santo, Patrícia Izar, Irene Delval, Victor de Napole Gregolin, and Nina ST Hirata. Fine-tuning video-text contrastive model for primate behavior retrieval from unlabeled raw videos. *arXiv preprint arXiv:2505.05681*, 2025.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19412–19424, 2024.

Jennifer J Sun, Hao Zhou, Long Zhao, Liangzhe Yuan, Bryan Seybold, David Hendon, Florian Schroff, David A Ross, Hartwig Adam, Bo Hu, et al. Video foundation models for animal behavior analysis. *bioRxiv*, pp. 2024–07, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1): 1–15, 2022.

Teng Xu, Taotao Zhou, Youjia Wang, Peng Yang, Simin Tang, Kuixiang Shao, Zifeng Tang, Yifei Liu, Xinyuan Chen, Hongshuang Wang, et al. Mousegpt: A large-scale vision-language model for mouse behavior analysis. *arXiv preprint arXiv:2503.10212*, 2025.

Shaokai Ye, Jessy Lauer, Mu Zhou, Alexander Mathis, and Mackenzie Mathis. Amadeusgpt: a natural language interface for interactive animal behavioral analysis. *Advances in neural information processing systems*, 36:6297–6329, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024.

Jinguo Zhu, W Wang, Z Chen, Z Liu, S Ye, L Gu, H Tian, Y Duan, W Su, J Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. *URL https://arxiv. org/abs/2504.10479*, 9, 2025.

Appendix

A Model training

We followed the official recommendations when finetuning InterVL from `https://github.com/OpenGVLab/InternVL`. We used AdamW optimizer (Loshchilov & Hutter, 2017) with `beta1`=0.9, `beta2`=0.999, and `epsilon`=$1e-08$, and cosine learning rate scheduler. We trained for one epoch with 4 H200 GPUs. We trained three times with seeds equal to 42, 83, and 105.

B Dataset Framework

B.1 Input Prompt Label Space

For AnimalKingdom, which contains 140 action and 850 animal classes, we constructed a 15-class label space for each sample as follows: For **actions** we included all actions present in the target video, sample maximum 10 additional actions from the same behavioral category, and maximum 4 actions from other categories. For **animals** we included all animal species in the target video, sample a maximum of 10 additional species sharing the same taxonomic parent class, and a maximum of 4 species from other taxonomic groups. This sampling strategy leverages AnimalKingdom's behavioral and taxonomic hierarchies to create challenging yet focused label spaces that maintain biological relevance.

For MammalNet, we applied the same animal sampling strategy for species label space, and included all 12 action classes for behaviors. MammAlps, having smaller label spaces across all categories, used complete label spaces without sampling.

B.2 Prompt Examples

Here we provide examples of inputs and annotations with and without code format for the joint animal-action recognition task from the MammAlps dataset. For the experiments with InternVL3-78B and GPT-4o , we added an additional example of the output format exactly following the code-based annotations.

Prompt

You are an assistant specialized in analyzing animal videos. Your task is to answer questions about the animals and their behaviors in a given video. Instruction:
You are provided with the following base function, which you can use to decompose the main question into subtasks and solve them step by step:

```
def recognize(entity_type: str, condition: Optional[str]) -> List[str]:
    Returns all unique entities of the specified
          type detected in the video (e.g., 'animal', 'action', 'activity').
    If condition is provided, returns all
         entities of the specified type that appear when the given condition is true.

    Example:
        >>> recognize(entity_type='animal')
        ['dog', 'cat']
        >>> recognize(entity_type='action')
        ['bark', 'run']
        >>> recognize(entity_type='action', condition='animal == dog')
        ['bark', 'run']
```

In addition to these base function, you may use standard Python functions such as average, max, min, sum, len, sorted, etc., as needed to help you answer the questions.
Output format:
Your final output should be 'Final answer:' followed by the list of [animal, action]s recognized in the video, formatted as `List[List[str]]`.
An action is a discrete, often well-defined motor event or behavior performed by an animal, typically characterized by a specific goal or function. You should use the following label space to identify animals, and actions:
Animals label space: `hare, fox, wolf, roe_deer, red_deer`
Actions label space: `bathing, walking, jumping, grazing, sniffing, scratching_hoof, standing_head_down, defecating, scratching_antlers, scratching_body, laying, vocalizing, looking_at_camera, drinking, standing_head_up, shaking_fur, running, unknown, urinating`
[video]
Identify the animals that appear in the video and detail the actions they perform.

Annotation with code format

```
step 1
animals = recognize(entity_type='animal')
output:List[str]: ['red_deer']

step 2
action_per_animal = dict(
    animal: recognize(entity_type='action', condition='animal == animals[i]'
)
for i in range(len(animals)))
output:Dict[str, List[str]]: {red_deer: walking}

step 3
animal_action_pairs = [
    [animal, action]
    for animal, actions in zip(animals, action_per_animal)
    for action in actions
]
output:List[List[str]]: [[red_deer, walking]]

step 4
return animal_action_pairs
output:Final answer: [[red_deer, walking]]
```

Annotation without code format

```
output:Final answer: [[red_deer, walking]]
```