Prompting-MammAlps: Fine-Grained Text-to-Video Retrieval for Camera-Trap Data

Valentin Gabeff[1], Baptiste Maquignaz[1], Jennifer Shan[1], Sepideh Mamooler[1], Gencer Sumbul[1], Blair Costelloe[2,3], Devis Tuia[1], and Alexander Mathis[1,4]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2] Max Planck Institute of Animal Behavior, Konstanz, Germany
[3] University of Konstanz, Konstanz, Germany
[4] alexander.mathis@epfl.ch

Abstract. Automatically retrieving videos from large camera-trap datasets remains challenging. Text-to-Video retrieval (TVR) methods based on large video-language models (VLMs) have potential to retrieve events of interest by describing them with simple text queries. However, current methods often lack spatiotemporal understanding and do not generalize well to ecological data. In this work, we introduce Prompting-MammAlps, the first camera-trap TVR benchmark, and propose a fine-grained and interpretable TVR method. Specifically, we trained a vision transformer to perform spatiotemporal action localization, and convert its output to structured text, describing each video. Independently, ethology-inspired queries are processed by a Large-Language Model (LLM) based coding agent to parse the structured text per video and retrieve videos accordingly. We harnessed the LLM to use functions from a custom parsing library to minimize the risk of LLM hallucinations and to improve method interpretability. This retrieval approach applied on the Prompting-MammAlps benchmark achieved a set-based F1-score of 34% on a test set of 135 ecologically-relevant queries and 775 candidate videos. In comparison the best zero-shot VLM achieved a F1-score of 18%, while also lacking interpretability.
Project page: eceo-epfl.github.io/prompting-mammalps/

Keywords: Text-to-Video Retrieval · Camera-Traps · Benchmark

1 Introduction

Wildlife monitoring is crucial to design accurate ecosystem conservation strategies, to measure their impact over time [21, 39, 67, 70], and more generally to improve our fundamental understanding of how animals behave in complex, real-world environments [2, 8, 20, 41]. To achieve this, ecologists deploy networks of camera-traps –among other remote sensors and other techniques– that capture fine-grained visual information of ecological events with minimal disturbance [11, 13, 55, 64, 65]. However, processing this data remains time-consuming,

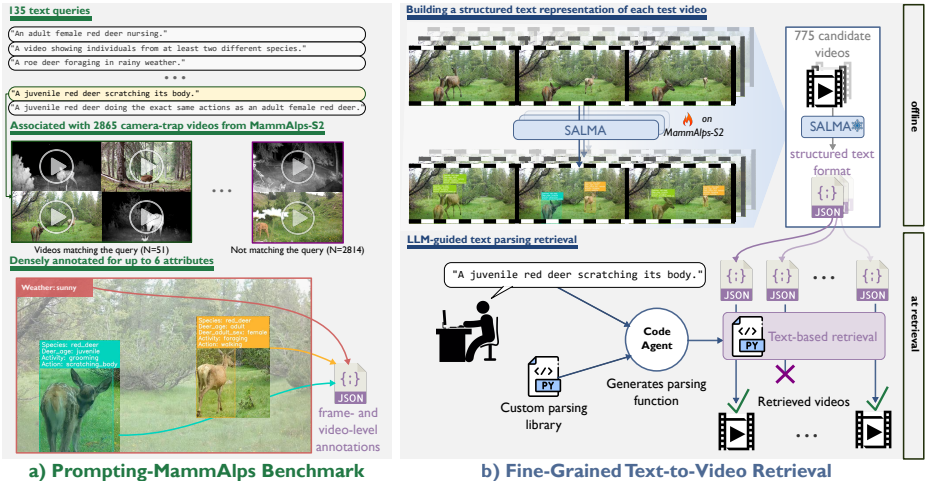

Fig. 1: Overview of Prompting-MammAlps and our TVR method. (a) We associate a densely annotated pool of camera-trap videos with 135 queries spanning six fine-grained visual attributes and diverse spatiotemporal relationships. (b) We perform spatiotemporal action localization on the candidate videos and store each result in a structured `.json` format. For each query, we task a LLM to write a parsing function that assesses if the text representation of a candidate video matches the query or not.

far exceeding the time spent in data analysis [5, 70, 85]. While deep learning advances have been improving camera-trap data processing, for example to filter out false positive images [6, 32] or to recognize species identity for some taxonomic groups [59]; fine-grained visual attributes (*e.g.* animal behavior) are still typically annotated manually by experts [34, 64, 85] or through citizen-science initiatives [54] with notable exceptions [28, 47, 60].

When recording videos, camera-traps can capture rich information on animal behaviors and on social and environmental interactions [7, 9, 10, 28, 68, 74]. While video captures richer behavioral data, the overhead of managing and annotating it, (still) makes images the practical choice for most applications. To lift the constraints on the research questions that can be answered using camera-trap videos, the field requires better methods to automatically process and manipulate collected data. To address this, text-to-video retrieval (TVR) methods are particularly relevant. They retrieve the videos needed to address an ecological research question of interest [72, 73] based on user text queries.

The recent development of video-language models (VLMs) [4, 18, 45, 77, 83, 91] from large corpora of video-text pairs [3, 16, 31, 38, 52, 82, 87] drove major advances in TVR [86, 93]. However, these models still fail to generalize to domain-specific data such as camera-trap imagery [26, 29]. This performance gap is in part due to: (1) the scarcity of publicly available datasets describing ecological events beyond species taxonomy; and (2) to the lack of benchmarks to properly evaluate and guide the development of large VLMs for fine-grained ecological tasks. (3)

Additionally, current TVR methods often rely on retrieving videos based on the latent similarity of few sub-sampled frames with a text prompt [86, 93], which hinders interpretability and under-performs on complex prompts involving spatiotemporal reasoning [27]. While there has been a recent interest in using VLMs to describe long videos as text followed by retrieval from LLMs, the risk of LLM hallucination when describing video content or when retrieving videos given a user prompt remains an open challenge.

In this work, we tackle the above-mentioned challenges (1)-(3). First, we extend the MammAlps dataset [28] with another season of data, improve the original annotation scheme to better comply with existing ethograms, and add additional individual-level attributes for deer species (MammAlps-S2). Second, we introduce Prompting-MammAlps, the first TVR benchmark for camera-trap videos (Fig. 1a). To this end, we defined 135 open-vocabulary text queries of ecological significance, to which we exhaustively match 2865 videos from MammAlps-S2. And third, we propose a new TVR method (Fig. 1b) to improve interpretability and reasoning over spatiotemporal relationships detected in camera-trap events. In particular, we process candidate videos offline with a fine-tuned video transformer (SALMA), and represent each video's output as a structured `.json` file with frame- and video-level information. At retrieval time, we task a LLM code agent to process user queries and to output a text parsing function following a custom library of common parsing operations. The generated function is then applied to each candidate video text representation and indicates that representation's binary correspondence with the input prompt. As such, the code agent never interacts with the candidate videos' textual representations but only needs to translate the input prompt into a logical sequence of predefined parsing function calls, improving parsing capabilities and limiting the risk of hallucination. The interpretability of our retrieval approach lies in the intermediary frame-level description from SALMA that can be visualized on the retrieved videos, and in the generated parsing function that can be investigated by the end-user. We evaluate our TVR method on the proposed benchmark and compare its retrieval performance with other similarity-based retrieval methods.

2 Related Works

Text-to-video retrieval (TVR) benchmarks have flourished in recent years given the growing need for retrieving data from large video datasets [86, 93]. While first benchmarks used associated captions [3,42,76,79], subtitles [44,52] or coarse metadata [62] as a convenient source of queries, later works progressively created more realistic textual queries for retrieval, with an increased reasoning complexity [43]. Since the associated videos are primarily sourced from popular multimedia platforms such as YouTube, the topic is either too broad [52], or focused on domains not relevant for ecology [92]. Moreover, current benchmarks do not include queries for which no video is associated with them. Together, these reasons strongly limit their use to accurately evaluate recent VLMs for TVR on camera-trap data. Recently, Inquire [73] has become the first text-to-image

retrieval benchmark applied to ecological data: a list of 250 expert-level queries was semi-automatically matched with five million images from the iNaturalist platform [35, 71]. While there is no such benchmark for video data, several video datasets contain fine-grained visual information, which could power VLMs suited for ecological retrieval tasks [10, 17, 25, 40, 46, 53, 60, 61]. Our benchmark is the first TVR benchmark for camera-trap videos and closely represents the context in which TVR could be used in practice: they span various ecological themes of interest and some queries do not have any videos associated with them.

TVR methods using deep learning commonly retrieve videos of high latent similarity with text queries in a joint embedding space shared between both modalities [3, 22–24, 30, 37, 48, 75, 78, 80]. This paradigm received increased attention with the development of large pretrained vision-language models, first for images [58] and later for videos [18, 19, 45, 83], which can generalize well enough on some domains for zero-shot retrieval, while having a significant computational cost for the models employing cross-modal attention [45,66,77]. While methods based on latent similarity or cross-modal integration apply well to images, their adaptation to videos often fails to capture long-term spatiotemporal relationships between entities and lacks explainability [57].

Textual representation of videos appears as a promising solution to these issues by learning to summarize the relevant visual information of a video into a textual form [49, 50, 57, 84]. The powerful reasoning capabilities of modern LLMs can then be used to interpret the textual representation of videos [49, 84], or to reason on the retrieved results [57]. While not directly applied to TVR, RAVU [49] is a question-answering framework based on a VLM and a LLM, which jointly operate to create and parse a video spatiotemporal graph. Then, compositional reasoning on the graph is performed through Chain-of-Thought (CoT) and frames of interest to the question are retrieved by measuring the similarity between the embedded textual representations of a graph node and the grounding query. In X-CoT [57], authors propose an interpretable reranking method based on LLM CoT reasoning. The LLM processes the similarity measures between videos, their structured frame-level captions generated by a VLM, and the input query. The method requires to exhaustively compute a binary preference decision for all video pairs from the top-k list, which does not scale well with k, and associated explanations from the LLM are prone to hallucinations hindering retrieval interpretability.

Our proposed Prompting-MammAlps benchmark is closest to Vendrow et al. [73]. The smaller scale of our benchmark (135 queries matched with 18 hours of video) in comparison to its internet-based counterparts is explained by the difficulty in matching query-video sets representative of fieldwork data, but comes with the advantage of high-quality annotations and domain-specific queries. Our retrieval strategy is inspired by Malik et al. [49] when applied to TVR. It differs by constraining the LLM to perform solely prompt decomposition and parsing code generation using a predefined set of parsing functions without direct examination of the video textual representations, therefore limiting hallucination and improving interpretability.

3 Prompting-MammAlps

3.1 Extending MammAlps: MammAlps-S2

We extended the initial MammAlps dataset [28] (i) by adding another season of densely annotated data (acquired in 2024), (ii) by annotating two new attributes describing deer individuals status for both years, and (iii) by improving the ecological validity of the annotation scheme. Specifically, we

(i) replicated the original data collection protocol by deploying three cameras in the same three sites as in 2023. While the position of the cameras in Site 2 were closely matched, the position of the Site 1&3 cameras required adjustment and did not share the same field-of-view as the ones placed in the first season. We therefore refer to these cameras as C4-C6 (instead of C1-C3). Cameras were deployed in the field for 10 weeks (Jun-Aug 2024) recording at day and night (IR illumination). This period does not overlap with the previous data acquisition period of MammAlps (Aug-Oct 2023; see Supp. Mat. S1.1), adding behavioral and morphological diversity to the dataset, as well as increasing the number of samples for rare species (*i.e.* mountain hare).

(ii) improved the annotation scheme for ungulate species by collecting and integrating behavior descriptions from existing ethograms [12, 56].

(iii) increased label diversity by manually annotating deer age as a binary category (*juvenile* or *adult*) based on the apparition of sexual dimorphism traits, along with sex for adult individuals (*male* or *female*). We focused on deer as they are in the majority of events. Furthermore, we manually refined the dense annotations of the videos from the 2023 dataset as an opportunity to improve the labeling quality and to use a common label space for both seasons.

We refer to this extended dataset as *MammAlps-S2*. MammAlps-S2 comprises of 2865 videos totaling 18 hours of high-resolution recording acquired from 15 different camera views. The videos were densely annotated such that every animal contains frame-level annotations for its behavior (up to two low-level actions and one high-level activity), individual-level attributes (species, age for deer species, sex for adult deer species), and weather at the video-level. The annotation process yielded 3227 individual tracks (avg. duration of 15.6s), 1.5M bounding boxes, and an average of 1.13 individuals per video. We included a small proportion of false positive (*i.e.* absence of visible vertebrates) videos since these are common in camera-trap datasets.

We randomly split camera-trap acquisition days in two groups to define a train and a test video set while maintaining a similar label distribution across sets. The train and test sets contain 2090 and 775 videos, respectively. We further detail the final MammAlps-S2 dataset in Supp. Mat. S1.1 and illustrate the distribution of attributes across video splits in S1.2.

3.2 Text-to-Video Retrieval Benchmark

To create the TVR benchmark, we identified 135 text queries in collaboration with a behavioral ecologist (B.C.). These queries span four ecological categories: *courtship* related behaviors, *other social behaviors*, *camera reactions*,

and any other *rare* events (Fig. 2A). We additionally included a *common* category designed to retrieve (*i.e.* filter out) the most common events (*e.g.* `"A red deer foraging only."`). We also categorized queries based on their reasoning complexity. *Single* and *multi-attributes* queries require processing information from one, and two or more different attributes (*e.g.* `"A juvenile red deer suckling."`). *Multiple individuals* queries ask information about two or more individuals (*e.g.* `"A video of two or more wolves."`). *Complex* queries require more reasoning than checking for the presence of some attributes (*e.g.* `"An animal reacting to a camera and then foraging."`). We also propose *video comparison* queries which require integrating information from a reference video with the other candidate videos (*e.g.* `"An individual doing the same sequence of actions as the individual of S2_C1_E323_V0074."`).

We used the ground-truth annotations to semi-automatically match the camera-trap videos of MammAlps-S2 to each query. Only videos containing at least a segment fully matching the elements from a query were associated with it, while incomplete matches and non-relevant videos were excluded from this list. We considered the training set of Prompting-MammAlps as the aforementioned list of queries associated with the train videos. Similarly, the test videos, referred to as candidate videos, were associated with the same list of queries to constitute the test retrieval set.

On average, queries are matched with 78.6 videos (min.: 0; max.: 1439). Seven queries (13 when considering the test set) do not have any video associated with them but are still kept in the dataset as a user may look for events that were never captured by any camera (Fig. 2B). Single videos can also be matched to multiple queries, on average to 3.7 of them (min.: 0; max.: 26). Similarly, some videos are not associated with any query (Fig. 2C). Additional benchmark statistics can be found in Supp. Mat. S1.3.

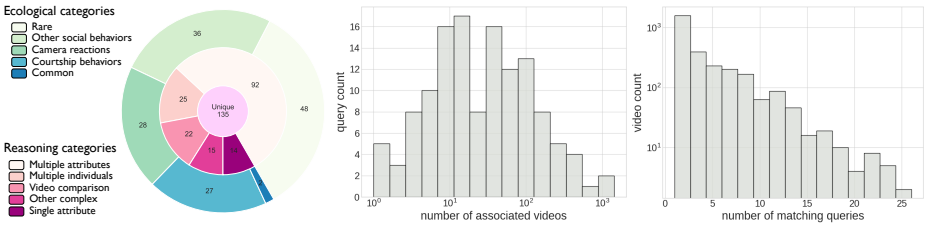

Fig. 2: Prompting-MammAlps Benchmark. (a) The 135 queries are categorized in five ecological and five reasoning categories. One query can belong to multiple categories. (b-c) Number of matching videos per query and inversely (excluding zero entries).

4 Fine-grained text-to-video retrieval strategy

Our end-to-end TVR strategy processes the textual user query and the candidate videos separately. This enables to process candidate videos offline prior to retrieval, improving computational efficiency and speed at retrieval.

4.1 SALMA: Spatiotemporal Action Localization for MammAlps

To process candidate videos, we aim to exhaustively extract the visual information that can be useful to retrieve videos matching the queries. This involves tracking individuals across frames, performing individual-level attributes recognition and action segmentation on each detected animal track. For this, we used a monolithic architecture adapted from DETR [15] and MOTR [88] (Fig. 3). We trained an encoder-decoder transformer model to progressively decode object queries over a sequence of frames $\mathbf{X}_{i,t}, i \in [0, N_v], t \in [0, N_{f,i}]$ where N_v is the number of videos and $N_{f,i}$ the number of frames for video i, and used them to predict both frame- and individual-level attributes $\mathbf{y}_{i,t}$. The encoder (parametrized by θ_{en}) follows the vision transformer architecture from VideoMAE [28, 69, 77] and outputs a sequence of video tokens $\mathbf{f}_{i,t}$. The transformer-based decoder (parametrized by θ_{dec}) progressively decodes learnable object queries $\mathbf{q}_{i,t}$ (initialized at the beginning of every frame sequence to $\mathbf{q}_0$) through cross-attention with $\mathbf{f}_{i,t}$. Each object query j is given as input to six multi-layered perceptions (MLP) heads, predicting bounding-box coordinates and an activation score $s_{i,t,j}$, species identity, actions, activity, deer age, and adult deer sex categories, respectively. Differently from Carion et al. [15], we did not add a null-object class to indicate if an object query $q_{i,t,j}$ is not associated with any detection. Instead, we applied a fixed threshold σ_s on $s_{i,t,j}$ to indicate whether $q_{i,t,j}$ is associated with an animal track ($s_{i,t,j} \geq \sigma_s$) or if it is an inactive query ($s_{i,t,j} < \sigma_s$).

During training, we learned the parameters of SALMA in two curriculum steps. Firstly, we learned θ_{en}, θ_{dec}, q_0, and the MLP head parameters used to predict bounding box coordinates and the query activation scores $s_{i,t,j}$. The matching between ground-truth animal detections and each object queries is done through Hungarian matching based on association costs [15]. We applied an L_1 loss and a $gIoU$ loss on the bounding box coordinates, and used binary-cross entropy to learn $s_{i,t,j}$. To maintain the temporal consistency of object queries, we computed an InfoNCE contrastive loss where pairs of object queries $q_{i,t,j}$ and $q_{i,t+1,j}$ are considered as positives, while any other pair at this time step in the video and in the batch are considered as negatives [51]. Secondly, we learned the remaining MLP heads parameters using categorical cross-entropy, while still optimizing for the parameters of the first curriculum step.

At inference, we ran SALMA on every frame of a candidate video, propagating object queries over frame sequences. To account for the input resolution ratio difference between training (1:1) and inference (16:9), we ran inference on two square-cropped views of the image, and used Hungarian matching to match predictions across both views. For individual- and video-level attributes, we took the most commonly predicted value over predicted tracks and the entire video,

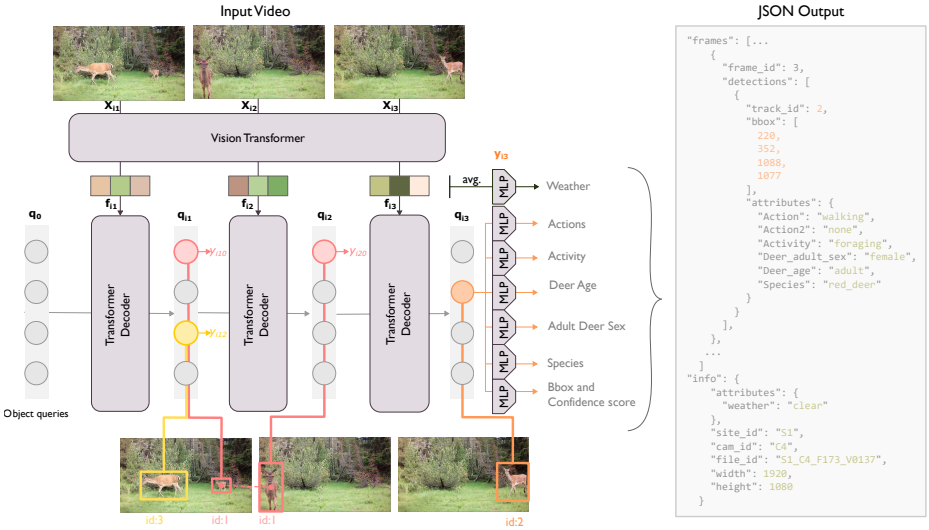

Fig. 3: From raw videos to their structured text representations. SALMA progressively decodes object queries over a sequence of frames and outputs a track activation score for each object query $q_{i,t,j}$ (shown in color if above the activation threshold σ_s) and its associated list of predictions $y_{i,t,j}$. We created a `.json` file containing video- and frame-level information given the predicted attributes of the activated object queries.

respectively. The output of each decoding step was collected in a `.json` text file: for every frame t, we wrote a list of detection objects, each with an associated track id j corresponding to the activated object query $q_{i,t,j}$, and containing a list of attributes corresponding to the predicted labels $y_{i,t,j}$. The weather attribute was added in the "info" section of the file.

4.2 LLM-guided text parsing

To process text queries, we translated each query into an executable function that operates on a video's `.json` representation and returns a boolean value indicating whether the annotation satisfies the query's semantic constraints. We leveraged a LLM to synthesize these functions from natural language queries. Specifically, we designed a library of 23 functions that operate over structured video annotations. These primitive functions can be combined to form executable predicate functions without having to input the content of the `.json` file to the LLM. They can be used to check for the presence of some specific attributes (*e.g.* `check_tracks_contains_species(tracks, species_name)`, or to extract general information (*e.g.* `get_unique_activities(tracks)`) about a given track. We list them in Supp. Mat. S1.7. The library specification was provided to the LLM as a prompt together with the query, the possible labels, and three in-context reasoning examples from unseen queries, enabling the LLM to generate

a predicate function composed of the available primitives. The resulting predicate functions were then executed on every candidate video's text representation. If the conditions stated in the function were all met for a given `.json` file, we considered it a match to the query and returned the associated video to the user.

5 Experiments

To evaluate our approach on Prompting-MammAlps, we first trained SALMA on the 2090 training videos for 500 epochs in each curriculum learning step. We used a large pretrained vision transformer encoder backbone [77]. At inference, we processed the 775 candidate test videos using a temporal stride of four frames and applied forward-filling to derive missing entries. Additional architecture and training details can be found in Supp. Mat. S1.4.

We used smolagent [63] for the LLM-based coding framework and evaluated the prompt interpretation and parsing code generation capabilities of four different LLMs: Qwen3-8B [81]; Llama-3.1-8B-Instruct [1]; Mistral-7B-Instruct [36], and Apertus-8B-Instruct [33]. Our framework allows for efficient multi-step reasoning, since any deviation from the list of provided parsing functions or from the possible attribute labels results in an error that is forwarded to the code agent, initiating a new trial. The agent finished its execution once it successfully calls the generated function on a default `.json` file without any error, or if it failed to generate an executable function after more than ten trials.

We compared our approach against zero-shot evaluation of CLIP4Clip [48], GRAM [19] and, InternVideo2.0 1B [77]. Since Vendrow et al. [73] found SigLIP SO400m-14 [89] to be the most effective retrieval model, we also applied this backbone to CLIP4Clip, to which we refer as SigLIP4Clip. To better compare with our supervised approach, we fine-tuned CLIP4Clip on query-video pairs from the train set of query-video associations. We provide further details for these experiments in Supp. Mat. S1.5.

Since the methods used as comparison provide a ranking based on a distance [19] or a similarity score [48,77] between the textual and the visual modality, they are not directly comparable to ours. Our approach does not compute a relevance score but only a binary matching decision for every candidate video and query pair and we cannot compute standard ranking-based metrics such as mAP or Recall@k. Instead, we computed the F1-score between the retrieved set of candidate videos and the ground-truth set for every query. For queries not associated with any ground-truth video, we computed the F1-score based on the filtered set (*i.e.* what is left after retrieval). We then averaged each F1-score over the list of queries. For the comparison baselines, we computed the optimal similarity decision threshold for each individual query on the training set, which contains the same queries but different videos, and then computed the F1-score of the binarized output using this optimized decision threshold. Additionally, we cannot compute the performance of these methods on queries requiring a comparison with a reference video as it requires a two-steps reasoning process, and we therefore exclude these queries from the comparative analysis.

6 Results

6.1 End-to-end retrieval

We report the end-to-end retrieval performance of our approach and of the zero-shot performance of comparative VLMs (Tab. 1). Since our method is bounded by the capabilities of the code agent to generate parsing functions, we report the best possible retrieval performance obtained when the generated parsing functions are directly applied to the ground-truth annotations (row Agent + Oracle). When applying the generated functions to the output from SALMA instead of the ground-truth (row Agent+SALMA), we obtained an averaged F1-score of 0.34, suggesting that SALMA does not reliably extract the visual information necessary for accurate retrieval. For the reasoning categories, our approach was most effective on simple queries referring to a single visual attribute, and performed comparatively lower on queries involving multiple individuals.

VLMs zero-shot retrieval performances were all lower than our proposed method, highlighting poor generalization to this ecological domain. To bridge this domain gap, we fine-tuned CLIP4Clip on MammAlps-S2 by using the training query-video pairs, which led to a consistent gain in performance across categories. However, this naïve supervised approach still underperformed in comparison to ours, suggesting that more complex retrieval strategies are required to perform well on the proposed benchmark (Tab. 1). The comparison remains limited since the coding agent has access to common reasoning operations through the parsing library.

When comparing the four LLM performances on this task (Tab. 2), we observed that Qwen3-8B yielded the best generative capabilities with a F1-score of 0.87, and we therefore used this model for the remainder of the study.

Table 1: F1-Score retrieval performance on Prompting-MammAlps. †: fine-tuned on MammAlps-S2, zero-shot evaluation otherwise; n/a: VLMs performance cannot be evaluated on queries referencing an additional video (Vid. Comp.).

Ecological Cat. (Support)	**All (135)**	**w/o Vid. Comp. (113)**	**Rare (48)**	**Courtship (27)**	**Other Social (36)**	**Cam. Reaction (28)**	**Common (2)**
Agent + Oracle	0.87	0.89	0.94	0.79	0.90	0.82	0.98
CLIP4Clip [48]	n/a	0.17	0.16	0.13	0.26	0.14	0.40
SigLIP4Clip	n/a	0.17	0.17	0.12	0.25	0.12	0.37
GRAM [19]	n/a	0.16	0.16	0.08	0.25	0.12	0.35
InternVideo2.0 [77]	n/a	0.18	0.16	0.15	0.28	0.12	0.40
CLIP4Clip† [48]	n/a	0.26	0.29	0.22	0.30	0.18	0.70
Agent + SALMA†	**0.34**	**0.37**	**0.43**	**0.34**	**0.35**	**0.21**	**0.87**
Reasoning Cat. (Support)	**All (135)**	**w/o Vid. Comp. (113)**	**1 attr. (14)**	**2+ attr. (92)**	**2+ indiv. (25)**	**Vid. Comp. (22)**	**Complex (15)**
Agent + Oracle	0.87	0.89	1.00	0.89	0.80	0.75	0.69
CLIP4Clip [48]	n/a	0.17	0.12	0.18	0.13	n/a	0.13
SigLIP4Clip	n/a	0.17	0.13	0.18	0.10	n/a	0.14
GRAM [19]	n/a	0.16	0.11	0.17	0.11	n/a	0.13
InternVideo2.0 [77]	n/a	0.18	0.10	0.19	0.15	n/a	0.15
CLIP4Clip† [48]	n/a	0.26	0.29	0.25	0.20	n/a	0.25
Agent + SALMA†	**0.34**	**0.37**	**0.44**	**0.37**	**0.23**	**0.19**	**0.26**

Table 2: Oracle performance comparison between LLM code agents. Qwen3-8B is the model yielding the best retrieval performance when applying generated functions to the ground-truth `.json` representation of candidate videos.

LLM	F1-score
Mistral-7B-Instruct [36] + Oracle	0.38
Apertus-8B-Instruct [33] + Oracle	0.41
LLama-3.1-8B-Instruct [1] + Oracle	0.42
Qwen3-8B [81]+ Oracle	**0.87**

6.2 SALMA

We analyzed the performance of SALMA by first computing its multi-object tracking capabilities after the first (SALMA-MOT) and the second (SALMA) curriculum steps (Tab. 3a). As a comparison, we report the tracking performance obtained on the candidate videos when using MegaDetector v5a [32] for detection, followed by ByteTrack [90] for tracking. While this latter approach gave the highest tracking accuracy (HOTA) and the least number of identity swaps, the detection performance of SALMA-MOT was superior and led to the best IDF1 score, while being end-to-end and with minimal post-processing steps. The higher detection accuracy might be explained by the fact that our approach is directly fine-tuned on MammAlps-S2, while MegaDetector [32] has been trained on a vast set of camera-trap datasets, but excluding the alps. The tracking performance of SALMA-MOT slightly degraded after the second curriculum step (SALMA) as we add more learning objectives. Removing the contrastive objective on the object queries $\mathbf{q}_{i,t}$ also decreased the tracking performance of SALMA-MOT while yielding a relatively similar detection accuracy. We therefore argue that this loss is useful to maintain temporal consistency in the context of multi-object tracking, as it was previously demonstrated for object-centric learning [51]. Yet SALMA does not have an explicit object tracking learning objective based on the track ID as a memory-based model would have [14, 88]. Both objectives are compatible and could be explored to improve the performance of SALMA.

Second, we report the behavior segmentation and multi-attributes recognition performance of SALMA after the second curriculum step (Tab. 3b). The averaged F1-scores per attribute were within the same range of previously reported results from the first benchmark of MammAlps [28], while the task was considerably simpler (*i.e.* multi-attribute classification from pre-processed clips).

6.3 Agent-based prompt interpretation

We report ablation results on the context provided to the Qwen3-8B agent (Tab. 4). The vanilla agent did not have access to any parsing function. We still provided it a template of the `.json` structure and a description of the possible labels, as the task would have been unnecessarily complicated otherwise.

Table 3: Tracking and attribute performance. (a) SALMA-MOT was trained for multi-object tracking only. Removing the contrastive loss between the object queries **q** decreases tracking quality. (b) Attributes were predicted after the MOT step. False positive tracks that were not matched to any ground-truth predictions also counted as false positive attribute predictions, and conversely for false negative tracks.

Model	HOTA↑	IDF1↑	DetA↑	ID_{swp} ↓
SALMA-MOT w/o InfoNCE loss	67.5	75.4	64.9	3485
SALMA-MOT	72.9	**84.5**	**67.7**	192
SALMA	69.0	81.1	62.6	202
MegaDetector [32] → ByteTrack [90]	**76.2**	81.4	67.2	**93**

(a) Multi-object tracking performance

Model	Species	Activity	Actions	Deer Age	Adult Deer Sex	Weather
SALMA	0.48	0.40	0.29	0.70	0.75	0.76

(b) Per-frame F1-score (macro-avg.) per attribute category

Adding the custom parsing library had the strongest effect on retrieval performance (+0.34 F1-score), which is expected as the logic to extract complex information such as action sequences is error-prone for the LLM. We added explicit constraints on the label space in the last row, by raising an error and returning it to the agent when it calls elements not belonging to the available label list. While it provided minimal performance gain (+0.03), it proved to be useful to disambiguate between "Actions" and "Activities".

Table 4: Ablation of the LLM framework components using ground-truth annotations. The last row corresponds to the Agent + Oracle row in Tab. 1.

Qwen3-8B	**F1-score (gain)**
vanilla	0.44 (−)
+ parsing library	0.78 (+0.34)
+ three in-context examples	0.84 (+0.06)
+ explicit constrain on possible labels	0.87 (+0.03)

6.4 Retrieval process decomposition

To illustrate our approach, we show a step-by-step retrieval process for the following user query: `"An individual sharing at least one activity with any individual from S1_C1_E66_V0152 but of a different species."` (Fig. 4 and Supp. Mat. S1.8). The generated code from the selected LLM agent correctly

extracted the activities and the species from both the reference video and each candidate video, but failed to compare them at the individual level, which is a first source of retrieval error. We then visualized predictions from SALMA on the reference video. The extracted list of activities contained *grooming*, which is a false prediction. Videos A and B are correctly returned, despite the inexact LLM's logic and some attribute misclassifications. Video C is a false positive because the *roe deer* is not *foraging*. Videos D and E are correctly filtered out, although if the parsing function logic was correct, video D would have been returned to the user as it contains a false positive *roe deer* detection *foraging*. The model failed to return Video F as it detected a *red deer* instead of a *chamois*.

This illustrative process shows the interpretability of our TVR method. Not only can the user visualize intermediary predictions from SALMA, but they can also understand how the LLM interpreted their query through examination of the generated function, and potentially refine their request accordingly.

7 Discussion

The retrieval performance of baselines demonstrates that our benchmark poses difficulties to current open-source VLMs that process a relatively small number of frames per video. Using an agent-based strategy, we show that with an oracle tracking, segmentation and classification model, one can achieve high retrieval performance. The lower performance obtained when using SALMA's predictions therefore suggests that extracting complex visual information and linking it to natural language remains the performance bottleneck more than domain-specific language understanding in this context. The benchmark design also meets some unaddressed needs such as the ability to evaluate performance on queries not matched to any video, or to integrate information from an additional reference video. Overall, Prompting-MammAlps is a unique TVR benchmark, which translates needs from ecologists into a benchmark to better evaluate advances in domain-specific video-language understanding.

Unlike most common TVR benchmarks, Prompting-MammAlps is not ranking-based but operates on a set of associated videos per query. Our methodology could be extended to a ranking-based retrieval strategy by returning a similarity score between a user query and a given `.json` file instead of the binary association. This would allow to answer more complex queries containing continuous information such as `"A juvenile following a hind."`, and would constitute a natural extension. The benchmark design could also be extended by leveraging the available audio modality, as it is done in other related TVR works [19, 43].

Beyond our proposed benchmark, the dense tracking annotations from the MammAlps-S2 dataset can be further exploited to create other multi-modal benchmarks. For example, one could use them to automatically generate video captions or visual question-answer pairs without requiring to explicitly process the videos (see Supp. Mat. S1.9). The dense tracking annotations can also benefit method development and evaluation for animal detection and tracking in complex environments with frequent occlusions.

Fig. 4: Retrieval process decomposition. For illustrative purposes, we chose the "no label constrain" version of Qwen3-8B (third row from Tab. 4) which produced a concise code but with a logic error that we discuss. Example frames are shown; numbers in parentheses indicate video counts.

8 Conclusion

We propose Prompting-MammAlps, a TVR benchmark containing 18h of high-resolution wildlife camera-trap recordings densely annotated at the frame-level matched to 135 expert-level queries. This specific visual and semantic domain brings unique challenges that are not well addressed by current VLMs. Our analysis suggests that perception rather than reasoning in VLMs is the biggest bottleneck to solve the benchmark. We hope that the present study will serve as inspiration to keep fostering development in video-language modeling and understanding, with the ultimate objective of supporting biologists in data management, processing and analysis.

Acknowledgments

We thank members of the Mathis Group for Computational Neuroscience & AI (EPFL) and of the Environmental Computational Science and Earth Observation Laboratory (EPFL) for their feedback and fieldwork efforts. We also thank members of the Swiss National Park monitoring team for their valuable support and feedback. The project was approved by the Research Commission of the National Park. This project was partially funded by EPFL's SV-ENAC I-PhD program (G.V.), Swiss SNF grant (320030-227871), and the Horizon Europe grant 101213369 (DVPS).

References

1. AI@Meta: Llama 3 model card (2024), `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`
2. Anderson, D.J., Perona, P.: Toward a science of computational ethology. Neuron **84**(1), 18–31 (2014)
3. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)
4. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
5. Beery, S., Morris, D., Yang, S.: Efficient pipeline for camera trap image review. arXiv preprint arXiv:1907.06772 (2019)
6. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
7. Berger-Tal, O., Polak, T., Oron, A., Lubin, Y., Kotler, B.P., Saltz, D.: Integrating animal behavior and conservation biology: a conceptual framework. Behavioral Ecology **22**(2), 236–239 (2011)
8. Bonnetto, A., Mamooler, S., Li, C., Mathis, A.: The behavior biopsy: Interpreting animal behavior as embodied, situated, and hierarchical. Current Opinion in Neurobiology **98**, 103201 (2026)
9. Brookes, O., Mirmehdi, M., Kuhl, H., Burghardt, T.: Chimpvlm: Ethogram-enhanced chimpanzee behaviour recognition. arXiv preprint arXiv:2404.08937 (2024)
10. Brookes, O., Mirmehdi, M., Stephens, C., Angedakin, S., Corogenes, K., Dowd, D., Dieguez, P., Hicks, T.C., Jones, S., Lee, K., et al.: Panaf20k: a large video dataset for wild ape detection and behaviour recognition. International Journal of Computer Vision pp. 1–17 (2024)
11. Burton, A.C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J.T., Bayne, E., Boutin, S.: Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. Journal of applied ecology **52**(3), 675–685 (2015)
12. Cap, H., Aulagnier, S., Deleporte, P.: The phylogeny and behaviour of cervidae (ruminantia pecora). Ethology Ecology & Evolution **14**(3), 199–216 (2002)
13. Caravaggi, A., Banks, P.B., Burton, A.C., Finlay, C.M., Haswell, P.M., Hayward, M.W., Rowcliffe, M.J., Wood, M.D.: A review of camera trapping for conservation behaviour research. Remote Sensing in Ecology and Conservation **3**(3), 109–122 (2017)

14. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
15. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
16. Carreira, J., Zisserman, A.: 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)
17. Chen, J., Hu, M., Coker, D.J., Berumen, M.L., Costelloe, B., Beery, S., Rohrbach, A., Elhoseiny, M.: 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)
18. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
19. Cicchetti, G., Grassucci, E., Sigillo, L., Comminiello, D.: Gramian multimodal representation learning and alignment. arXiv preprint arXiv:2412.11959 (2024)
20. Couzin, I.D.: Collective cognition in animal groups. Trends in cognitive sciences **13**(1), 36–43 (2009)
21. Couzin, I.D., Heins, C.: Emerging technologies for behavioral research in changing environments. Trends in Ecology & Evolution **38**(4), 346–354 (2023)
22. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11583–11593 (2021)
23. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9346–9355 (2019)
24. Dong, J., Wang, Y., Chen, X., Qu, X., Li, X., He, Y., Wang, X.: Reading-strategy inspired visual representation learning for text-to-video retrieval. IEEE transactions on circuits and systems for video technology **32**(8), 5680–5694 (2022)
25. Duporge, I., Kholiavchenko, M., Harel, R., Rubenstein, D., Crofoot, M., Berger-Wolf, T., Lee, S., Wolf, S., Barreau, J., Kline, J., et al.: Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos. arXiv preprint arXiv:2405.17698 (2024)
26. Dussert, G., Miele, V., Van Reeth, C., Delestrade, A., Dray, S., Chamaille-Jammes, S.: Zero-shot animal behavior classification with image-text foundation models. bioRxiv pp. 2024–04 (2024)
27. Fragomeni, A., Damen, D., Wray, M.: Leveraging auxiliary information in text-to-video retrieval: A review. arXiv preprint arXiv:2505.23952 (2025)
28. Gabeff, V., Qi, H., Flaherty, B., Sumbul, G., Mathis, A., Tuia, D.: 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)
29. Gabeff, V., Rußwurm, M., Tuia, D., Mathis, A.: Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. International Journal of Computer Vision pp. 1–17 (2024)

30. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision. pp. 214–229. Springer (2020)
31. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
32. Hernandez, A., Miao, Z., Vargas, L., Dodhia, R., Lavista, J.: Pytorch-wildlife: A collaborative deep learning framework for conservation (2024)
33. Hernández-Cano, A., Hägele, A., Huang, A.H., Romanou, A., Solergibert, A.J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E.F., Hakimi, I., Giraldo, J.G., Ismayilzada, M., , et al.: Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. arXiv preprint arXiv:2509.14233 (2025)
34. Hofmeester, T.R., Young, S., Juthberg, S., Singh, N.J., Widemo, F., Andrén, H., Linnell, J.D., Cromsigt, J.P.: Using by-catch data from wildlife surveys to quantify climatic parameters and timing of phenology for plants and animals using camera traps. Remote Sensing in Ecology and Conservation **6**(2), 129–140 (2020)
35. iNaturalist: inaturalist. `https://www.inaturalist.org` (2026), accessed: 2026-02-17
36. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D.L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral-7b. arXiv preprint arXiv:2310.06825 **3** (2023)
37. Jiang, J., Min, S., Kong, W., Wang, H., Li, Z., Liu, W.: Tencent text-video retrieval: hierarchical cross-modal interactions with multi-level representations. IEEE Access (2022)
38. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
39. Kays, R., Crofoot, M.C., Jetz, W., Wikelski, M.: Terrestrial animal tracking as an eye on life and planet. Science **348**(6240), aaa2478 (2015)
40. Kholiavchenko, M., Kline, J., Ramirez, M., Stevens, S., Sheets, A., Babu, R., Banerji, N., Campolongo, E., Thompson, M., Van Tiel, N., 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)
41. Koger, B., Deshpande, A., Kerby, J.T., Graving, J.M., Costelloe, B.R., Couzin, I.D.: Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. Journal of Animal Ecology **92**(7), 1357–1371 (2023)
42. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
43. Kriz, R., Sanders, K., Etter, D., Murray, K., Carpenter, C., Recknor, H., Guallar-Blasco, J., Martin, A., Yang, E., Van Durme, B.: Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24149–24158 (2025)
44. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. In: European Conference on Computer Vision. pp. 447–463. Springer (2020)

45. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
46. Liu, D., Hou, J., Huang, S., Liu, J., He, Y., Zheng, B., Ning, J., Zhang, J.: 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)
47. Liu, J., Hou, J., Liu, D., Zhao, Q., Chen, R., Chen, X., Hull, V., Zhang, J., Ning, J.: A joint time and spatial attention-based transformer approach for recognizing the behaviors of wild giant pandas. Ecological Informatics **83**, 102797 (2024)
48. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
49. Malik, S., Yamada, M., Singh, A., Aggarwal, D.: Ravu: Retrieval augmented video understanding with compositional reasoning over graph. arXiv preprint arXiv:2505.03173 (2025)
50. Mamooler, S., Qi, H., Gabeff, V., Montariol, S., Bosselut, A., Mathis, A.: Fine-tuning vision-language models for animal behavior analysis. In: LLM for Scientific Discovery: Reasoning, Assistance, and Collaboration (2025)
51. Manasyan, A., Seitzer, M., Radovic, F., Martius, G., Zadaianchuk, A.: Temporally consistent object-centric learning by contrasting slots. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5401–5411 (2025)
52. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2630–2640 (2019)
53. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: 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)
54. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences **115**(25), E5716–E5725 (2018)
55. O'Connell, A.F., Nichols, J.D., Karanth, K.U.: Camera traps in animal ecology: methods and analyses, vol. 271. Springer (2011)
56. Prikhod'ko, V., Zvychainaya, E.Y.: Behavior and phylogenetic relations among artiodactyla families (artiodactyla, mammalia). Biology Bulletin Reviews **1**(4), 345–357 (2011)
57. Pulakurthi, P.R., Wang, J., Rabbani, M., Dianat, S., Rao, R., Tao, Z.: X-cot: Explainable text-to-video retrieval via llm-based chain-of-thought reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 31172–31183 (2025)
58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
59. Rigoudy, N., Dussert, G., Benyoub, A., Besnard, A., Birck, C., Boyer, J., Bollet, Y., Bunz, Y., Caussimont, G., Chetouane, E., et al.: The deepfaune initiative:

a collaborative effort towards the automatic identification of the french fauna in camera-trap images. bioRxiv pp. 2022–03 (2022)
60. Rodriguez-Juan, J., Ortiz-Perez, D., Benavent-Lledo, M., Mulero-Pérez, D., Ruiz-Ponce, P., Orihuela-Torres, A., Garcia-Rodriguez, J., Sebastián-González, E.: Visual wetlandbirds dataset: Bird species identification and behavior recognition in videos. Scientific Data **12**(1), 1200 (2025)
61. Rogers, M., Gendron, G., Valdez, D.A.S., Azhar, M., Chen, Y., Heidari, S., Perelini, C., O'Leary, P., Knowles, K., Tait, I., et al.: Meerkat behaviour recognition dataset. arXiv preprint arXiv:2306.11326 (2023)
62. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3c–a research video collection. In: International Conference on Multimedia Modeling. pp. 349–360. Springer (2018)
63. Roucher, A., del Moral, A.V., Wolf, T., von Werra, L., Kaunismäki, E.: 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents` (2025)
64. Rovero, F., Kays, R.: Camera trapping for conservation. Conservation technology **79** (2021)
65. Rowcliffe, J.M., Kays, R., Kranstauber, B., Carbone, C., Jansen, P.A.: Quantifying levels of animal activity using camera trap data. Methods in ecology and evolution **5**(11), 1170–1179 (2014)
66. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024)
67. Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J.T., Burton, C., Townsend, S.E., Carbone, C., Rowcliffe, J.M., Whittington, J., et al.: Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment **15**(1), 26–34 (2017)
68. Tobias, J.A., Pigot, A.L.: Integrating behaviour and ecology into global biodiversity conservation strategies. Philosophical Transactions of the Royal Society B **374**(1781), 20190012 (2019)
69. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
70. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., et al.: Perspectives in machine learning for wildlife conservation. Nature communications **13**(1), 792 (2022)
71. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
72. Vendrow, E., Chae, J., Kurinchi-Vendhan, R., Eckert, I., Hall, J., Jarzyna, M., Miyajima, R., Oliver, R., Pollock, L., Shrack, L., et al.: Inquire-search: A framework for interactive discovery in large-scale biodiversity databases. arXiv preprint arXiv:2511.15656 (2025)
73. Vendrow, E., Pantazis, O., Shepard, A., Brostow, G., Jones, K.E., Mac Aodha, O., Beery, S., Van Horn, G.: Inquire: A natural world text-to-image retrieval benchmark. Advances in Neural Information Processing Systems **37**, 126500–126514 (2024)
74. Vogg, R., Lüddecke, T., Henrich, J., Dey, S., Nuske, M., Hassler, V., Murphy, D., Fischer, J., Ostner, J., Schülke, O., et al.: Computer vision for primate behavior analysis in the wild. Nature Methods pp. 1–13 (2025)

75. Wang, J., Sun, G., Wang, P., Liu, D., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Text is mass: Modeling as stochastic embedding for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16551–16560 (2024)
76. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4581–4591 (2019)
77. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
78. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021)
79. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
80. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language alignment. In: The Eleventh International Conference on Learning Representations (2022)
81. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
82. Yang, A., Nagrani, A., Laptev, I., Sivic, J., Schmid, C.: Vidchapters-7m: Video chapters at scale. Advances in Neural Information Processing Systems **36**, 49428–49444 (2023)
83. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10714–10726 (2023)
84. You, Z., Wen, Z., Chen, Y., Li, X., Zeng, R., Wang, Y., Tan, M.: Toward long video understanding via fine-detailed video story generation. IEEE Transactions on Circuits and Systems for Video Technology **35**(5), 4592–4607 (2024)
85. Young, S., Rode-Margono, J., Amin, R.: Software to facilitate and streamline camera trap data management: A review. Ecology and Evolution **8**(19), 9947–9957 (2018)
86. Younis, A.N., Ramo, F.M.: Ai-based video retrieval: A literature review. In: 2025 3rd International Conference on Cyber Resilience (ICCR). pp. 1–7. IEEE (2025)
87. Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., Choi, Y.: Merlot reserve: Neural script knowledge through vision and language and sound. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16375–16387 (2022)
88. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: European conference on computer vision. pp. 659–675. Springer (2022)
89. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)

90. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European conference on computer vision. pp. 1–21. Springer (2022)
91. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024)
92. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
93. Zhu, C., Jia, Q., Chen, W., Guo, Y., Liu, Y.: Deep learning for video-text retrieval: a review. International Journal of Multimedia Information Retrieval **12**(1), 3 (2023)