

Article

Towards Sustainable Wildlife Conservation: Automatic Recognition of Endangered Animal Behavior Using a Multimodal Contrastive Learning Framework

Shuyi Liu ¹, Ao Xu ² and Zhenjie Hou ^{2,*} 

¹ School of Overseas Education, Changzhou University, Changzhou 213164, China; 2300400210@smail.cczu.edu.cn

² School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213164, China; xuao@js.chinamobile.com

* Correspondence: houzj@cczu.edu.cn

Abstract

Automatic recognition of endangered animal behavior is crucial for biodiversity conservation and improving animal welfare, yet traditional manual observation remains inefficient and invasive. This work contributes directly to sustainable wildlife management by enabling non-invasive, scalable, and efficient monitoring, which supports long-term ecological balance and aligns with several United Nations Sustainable Development Goals (SDGs), particularly SDG 15 (Life on Land) and SDG 12 (Responsible Consumption and Production). The current deep learning approaches often struggle with the scarcity of behavioral data and complex environments, leading to poor model generalization. To address these challenges, this study focuses on endangered animal behavior monitoring and proposes a multimodal learning framework termed ABCLIP. This model leverages multimodal contrastive learning between video-and-text pairs, utilizing natural language supervision to enhance representation ability. The framework integrates pre-training, prompt learning, and fine-tuning to optimize performance specifically for small-scale animal behavior datasets, with a focus on the specific social and ecological behaviors of giant pandas. The experimental results demonstrate that ABCLIP achieves remarkable accuracy and robustness in recognizing endangered animal behaviors, attaining Top-1 and Top-5 accuracy of 82.50% and 99.25%, respectively, on the LoTE-Animal dataset, which outperforms strong baseline methods such as SlowFast (78.54%/97.55%). Furthermore, in zero-shot recognition scenarios for unseen behaviors, ABCLIP achieves an accuracy of 58.00%. This study highlights the potential of multimodal contrastive learning in wildlife monitoring and provides efficient technical support for precise protection measures and scientific management of endangered species.

Keywords: sustainable wildlife conservation; endangered species monitoring; behavior recognition; multimodal learning; non-invasive monitoring; biodiversity sustainability; animal welfare



Academic Editor: Andreas Kanavos

Received: 4 January 2026

Revised: 26 January 2026

Accepted: 2 February 2026

Published: 5 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Behavioral studies of endangered animals are crucial for improving animal welfare and guiding species recovery efforts. An in-depth understanding of their behavioral patterns is essential for developing and implementing effective conservation measures that enhance their well-being [1]. For example, monitoring the unique inverted marking behavior of giant pandas, the defecation of marking territories, and smelling not only helps

researchers to understand their ecological needs but also provides a basis for developing more effective conservation strategies [2]. Traditional methods of animal behavior analysis rely on manual observation, which is not only time-consuming and labor-intensive but also susceptible to human interference [3]. Therefore, the development of automated non-invasive monitoring technologies represents a critical step toward sustainable conservation practices, minimizing anthropogenic disturbance while enabling continuous evidence-based stewardship of vulnerable species and their ecosystems. This is because animals are highly wary and it is difficult to implant sensors to collect behavioral data [4]. Infrared trap cameras can be an effective solution to these problems by collecting animal behavioral image data while ensuring animal welfare [5,6]. However, image data is often tedious, and manual classification and analysis are inefficient and imprecise [7–9].

With the rapid development of computer vision and deep learning technologies, accurate and efficient results can be obtained by relying on robust deep learning models [10,11]. Previous studies using deep learning have focused on recognizing individual animals by facial features and recognizing the behavior of captive farm animals [12–15]. In the field of behavioral monitoring of endangered protected animals, the application of deep learning methods is still relatively rare [16]. Most current research on video action recognition in animals uses a classical framework that treats the task as a standard 1-n majority voting problem [17–19]. In this framework, the input video is mapped to a predefined set of labels, and the training model translates the action categories in the video into a number or vector by predicting conditional probabilities. In the inference phase, the model is used as the corresponding action category by selecting the index with the highest prediction score. This approach has been successful in many applications, but there are obvious limitations [20–22]. Firstly, it completely ignores the rich semantic information contained in the label text, which is crucial for understanding and distinguishing complex behavioral patterns. Secondly, traditional unimodal video modeling approaches have limited migration capabilities when confronted with data from different or unknown species.

To overcome the above limitations, we propose a novel action recognition model based on a multimodal learning framework called ABCLIP. Our model consists of two unimodal encoders, a feature similarity computation module and a prompt module, where the two unimodal encoders are a text encoder and a video encoder, respectively. Inspired by the related work on human behavior recognition in the field of computer vision [23], we not only enhance the model's representational capability but also achieve flexible zero-sample migration by modeling the task as a video-text multimodal learning problem using supervised signals in natural language.

Our work is inspired by and builds upon the paradigm of adapting large-scale vision-language models like CLIP for video understanding. Pioneering frameworks such as ActionCLIP [24], X-CLIP [25], and VideoCLIP [26] have demonstrated the power of video-text contrastive learning for tasks like action recognition and retrieval. While ABCLIP shares this core multimodal learning philosophy, it is specifically tailored for the unique challenges of endangered animal behavior monitoring: extreme data scarcity, long-tailed class distribution, and complex cluttered natural backgrounds. Architecturally, our novel Multi-Frame Integration Prompt module (especially the Similar-Transf variant) is designed to efficiently distill key spatiotemporal cues from animal videos while being robust to irrelevant environmental noise—a consideration that is less prominent in human-centric video datasets. Furthermore, we systematically explore and validate the “pre-training + prompt learning + fine-tuning” pipeline specifically for few-shot learning in ecological contexts, addressing the critical practical constraint of limited annotated data in conservation biology. Thus, ABCLIP's contribution lies not in proposing a new foundational framework but in

the targeted adaptation and innovation within the established multimodal paradigm for a specialized high-impact application domain.

Most existing animal action recognition models require large amounts of data for training, which not only requires significant computational resources but also long training cycles [27,28]. This is particularly detrimental to the task of studying endangered animal behaviors as we often have difficulty obtaining large amounts of labeled data. In contrast, a framework that uses a combination of pre-training, prompt learning, and fine-tuning has significant advantages. Pre-trained models can be trained with large amounts of general-purpose data to obtain a strong feature representation, while prompt learning can be used to design appropriate prompts so that the model can still perform well in the face of a small amount of task-specific data. Fine-tuning further optimizes the model so that it can be better adapted to specific tasks [29,30]. Therefore, we use a combination of pre-training, prompt learning, and fine-tuning in our multimodal framework. Prompt learning is a new approach that utilizes pre-trained models by designing suitable prompts that allow the model to perform a specific downstream task better with fine-tuning. The prompts can be natural language phrases or sentences, etc., that guide the model to produce the desired output [31]. The prompts can be adapted to the task requirements and are highly flexible. For example, for action recognition of giant pandas, we can generate species-specific prompts through hints to allow the model to learn under specific semantic supervision and achieve better recognition results. In the case of limited resources, prompt learning can be deployed quickly and achieve better results. This approach has achieved good results in the field of natural language processing, but it has not been widely used in video action recognition tasks and especially has not been exploited in animal action recognition monitoring.

We experimentally evaluated the proposed multimodal framework ABCLIP model. The experimental results show that the ABCLIP model proposed in this paper achieves significant performance in the behavioral monitoring of endangered and protected animals. The ABCLIP model can also achieve good performance in zero-sample action recognition, which provides some help in realizing the field of automatic monitoring of animal behavior. Overall, our work has the following contribution points:

- In this paper, the behaviors of endangered protected animals are monitored, including some unique behaviors of giant pandas, which helps researchers to formulate scientific conservation and management measures to effectively protect endangered species.
- To address the data scarcity issue in this domain, we explore and adapt the prompt learning paradigm, which, to our knowledge, has been under-explored in the context of endangered animal behavior recognition from videos.
- We propose the ABCLIP model for multimodal architectures, which achieves significant performance on action recognition tasks by modeling the action recognition task as a video–text multimodal learning problem using supervised signals in natural language.

2. Materials and Methods

2.1. Animal Recording Environment

In this paper, endangered and protected animals are used as research subjects, and the video data used are from the LoTE-Animal dataset [32], which is a long-time-span dataset used for behavioral studies of endangered animals and records detailed behavioral information. The dataset is not only large and diverse but also suitable for a variety of research tasks. The main research task in this paper is animal movement recognition. The animals studied in this paper were distributed in two areas in Sichuan Province, southwestern China: the Wolong National Nature Reserve (102°52′–103°25′ E, 30°45′–31°25′ N) and the Mabian Dafengding Nature Reserve (103°14′–103°24′ E, 28°25′–28°44′ N). The main study area

is located in the Wolong Nature Reserve, which is situated in the northern part of the Hengduan Mountains and is the natural habitat of a wide variety of animals, as well as the reserve with the largest number of giant pandas in Sichuan Province. As shown in Figure 1, video data are collected from infrared-triggered camera traps deployed at multiple locations in the elevation range of the areas where the data are collected, which is the predominant altitude of the distribution of endangered and protected animals. In addition, the video data were collected over a long period, recording seasonal, weather, and diurnal variations in the area where they were located, providing additional contextual information for the animal behavior identification task.

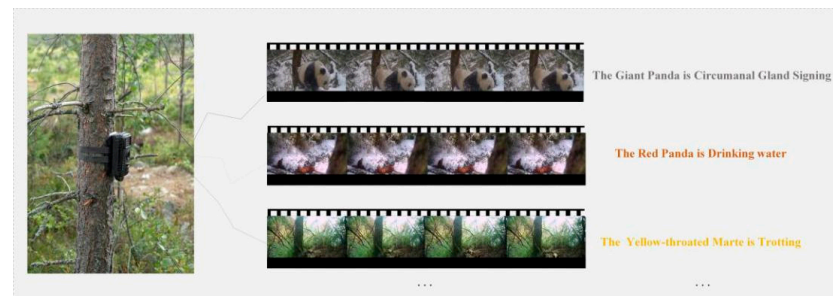


Figure 1. Video data collection example.

2.2. Dataset Statistics

First, based on the criteria of the LoTE-Animal dataset, Table 1 lists the conservation status of the three species studied in this paper in China. These include important conservation species among endangered animals, such as the giant panda and the red panda. Among them, the giant panda is a flagship species in biodiversity conservation with high conservation research value [2].

Table 1. The endangered animal species list. I, II, and III are protection levels and endangered levels in China. Italics are Latin names.

Protection Level	Species	Species	Genus	Family	Order
I	Giant panda	<i>Ailuropoda melanoleuca</i>	<i>Ailuropoda</i>	<i>Ailuropodidae</i>	<i>Carnivora</i>
II	Red panda	<i>Ailurus fulgens</i>	<i>Ailurus</i>	<i>Ailuridae</i>	<i>Carnivora</i>
II	Yellow-throated marten	<i>Martes flavigula</i>	<i>Martes</i>	<i>Mustelidae</i>	<i>Carnivora</i>

Based on the animal subjects studied, we used the typical behaviors of each animal for the action recognition task. We specifically selected 11 behavior types, including walking, smelling, drink water, resting, circumanal gland signing, exploratory, feeding, trotting, urine signing, defecating, and playing. We further divided the behaviors into 6 major categories based on their different manifestations and functional purposes, following standard ethological classification principles [33,34]. Among them, walking and trotting belong to locomotor behaviors, drinking and feeding belong to feeding behavior, smelling, circumanal gland signing, and urine signing belong to communication behavior, exploratory and playing belong to miscellaneous behavior, resting belongs to resting behavior, and defecating belongs to defecating behavior. The six broad categories of behavior are defined as shown in Table 2 and detailed statistics of the dataset are presented in Table 3.

Table 2. The definitions of the six broad behavioral categories.

Behavior Category	Description
Locomotor Behavior	The animal moves through the limbs in a specific sequence of alternating movements, generally staggered from side to side and back to front.
Feeding Behavior	The behavior of an animal in acquiring and ingesting food through actions such as chewing, gnawing, or swallowing. For example, giant pandas feed on bamboo leaves, drink water, and other activities.
Communication Behavior	Behavior in which an animal communicates information by sight, sound, smell, or physical contact. For example, giant pandas engage in activities such as smelling and urine signing.
Resting Behavior	Behavior of animals at rest for recovery and energy replenishment, usually involving lying down or inactivity.
Eliminative Behavior	The behavior of animals excreting feces or urine.
Miscellaneous Behavior	Other behaviors that do not fall into specific classifications include activities such as exploring, playing, etc.

Table 3. Detailed statistics of the LoTE-Animal dataset (illustrated with the giant panda species).

Species	Behavior	Videos (Train)	Videos (Test)	Avg. Duration (s)	Med. Duration (s)	Resolution	Frame Rate (fps)
Giant panda	Walking	85	22	4.2	3.8	1920 × 1080	30
Giant panda	Smelling	92	24	5.1	4.5	1920 × 1080	30
Giant panda	Drink water	18	5	6.3	5.9	1920 × 1080	30
Giant panda	Resting	95	24	8.5	7.2	1920 × 1080	30
Giant panda	Circumanal gland signing	45	12	3.8	3.5	1920 × 1080	30
Giant panda	Exploratory	70	18	7.2	6.8	1920 × 1080	30
Giant panda	Feeding	105	27	9.1	8.5	1920 × 1080	30
Giant panda	Trotting	60	15	2.5	2.3	1920 × 1080	30
Giant panda	Urine signing	50	13	4.0	3.7	1920 × 1080	30
Giant panda	Defecating	25	7	5.5	5.0	1920 × 1080	30
Giant panda	Playing	30	8	6.8	6.2	1920 × 1080	30
Other species (red panda; yellow-throated marten)	All behaviors	≈782	≈196	≈5.0	≈4.5	1920 × 1080	30
Total/average	-	≈1595	≈399	≈5.5	≈5.0	1920 × 1080	30

Note: the dataset is publicly available at <https://lote-animal.github.io/> (accessed on 5 August 2024). The total video count matches the description in Section 2.2 (1994 videos). The ‘other species’ row provides aggregate estimates; their per-behavior distribution mirrors the long-tailed pattern shown in Figure 2. Videos (Train/Test): number of video clips in the training and test sets. Avg. Duration: average clip length in seconds. Med. Duration: median clip length in seconds. Data for red panda and yellow-throated marten follow a similar distribution but are summarized in the ‘other species’ row for conciseness.

The experimental dataset contains 1994 video clips (22 GB). To ensure a fair evaluation and prevent performance inflation from spatial or temporal leakage, we adopted a strict camera-aware splitting strategy. The dataset was split into training (80%) and test (20%) sets stratified by camera trap location (Camera ID). Specifically, all camera locations were first partitioned into two disjoint sets. This guarantees that videos from the same camera location do not appear in both the training and test sets, effectively mitigating potential bias from similar backgrounds, lighting, or recurring individual animals across splits. Within each camera set, we maintained the proportional distribution of species and action categories. This stratified random split was applied to ensure the generalizability of our model to novel locations.

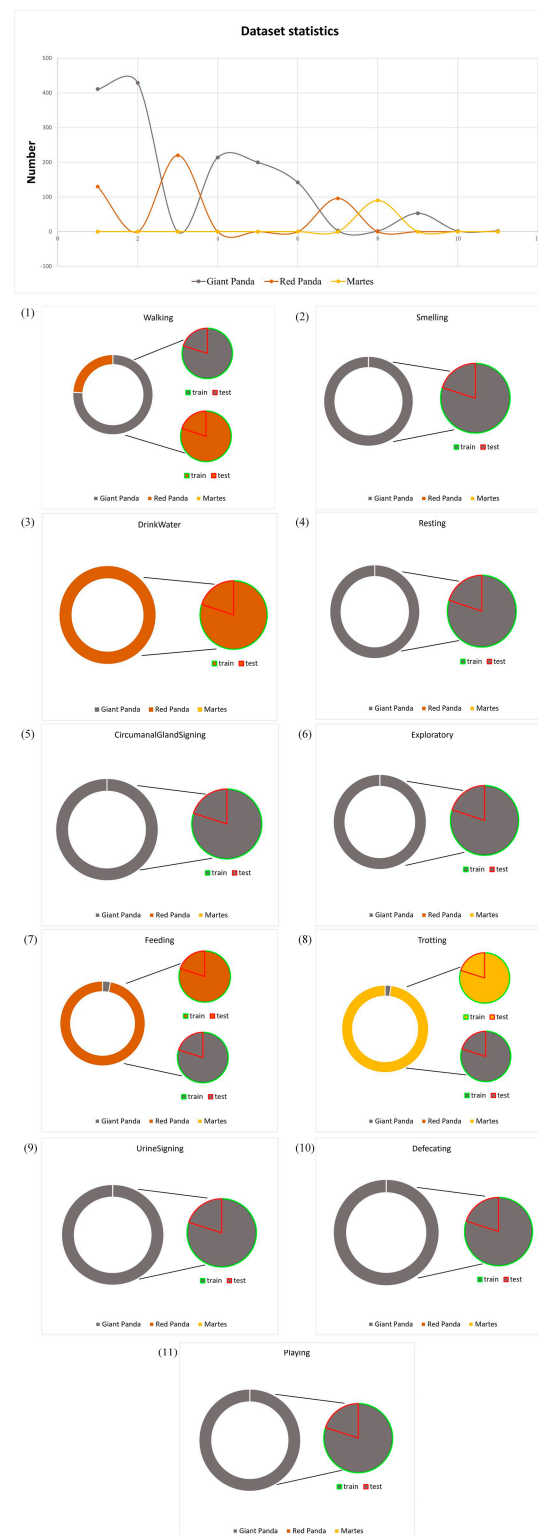


Figure 2. Statistical distribution of experimental data. (1–11) corresponds to the 11 actions on the x-axis in the main figure. Each circular pie chart is the proportion of three animals in each action. The corresponding pie chart shows the distribution of the number of videos for each species in the training set and test set.

2.3. Method

ABCLIP is a new model for action recognition based on a multimodal learning framework. It consists of two unimodal encoders, a feature similarity calculation module, and all

the prompt modules. The two unimodal encoders of the model are the text encoder and the video encoder. The overall framework of ABCLIP is shown in Figure 3a.

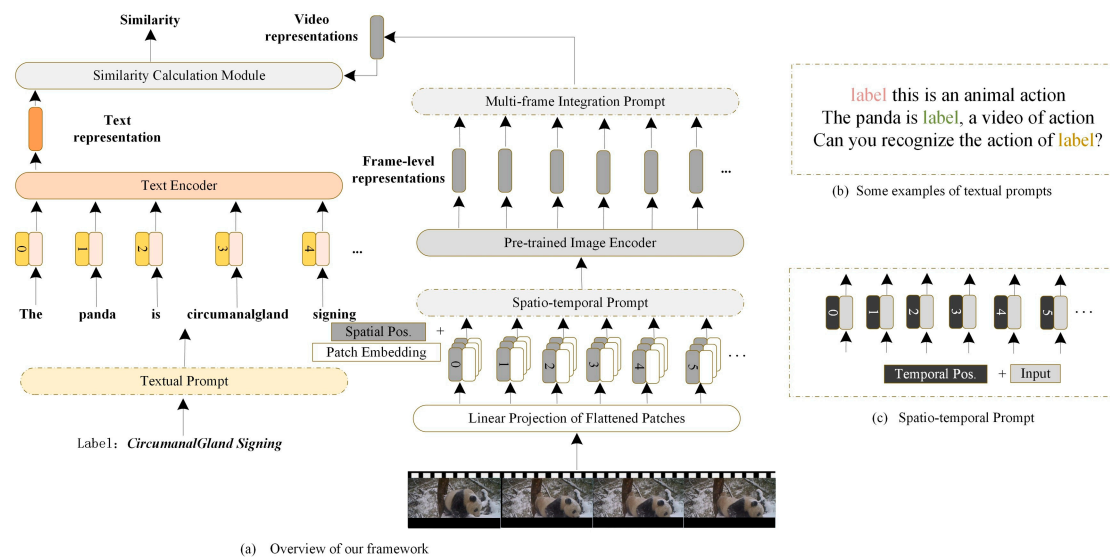


Figure 3. Overview of our proposed ABCLIP.

Specifically, as shown in the left branch of Figure 3a, the text modality involves generating text prompt strings from action labels via the text prompt module, which are then input into the text encoder to obtain the feature representations of the text modality. As illustrated in the right branch of Figure 3a, the video modality involves processing video data through the visual prompt module and image encoder to obtain the feature representations of the video modality. Finally, the similarity calculation module computes the similarity between the text features and video features. The ABCLIP model optimizes the feature representations by employing contrastive learning, enabling the model to better distinguish between action features of different categories while maintaining the compactness of features within the same category. Specifically, by using positive and negative sample pairs, the model calculates similarity and applies a contrastive loss function to optimize the entire framework.

The main modules of the proposed framework are described in detail below:

2.3.1. Textual Prompt Module

In natural language processing (NLP), prompt learning is a learning approach that exploits the flexibility and power of pre-trained language models by designing appropriate prompts to directly guide the model to perform a specific task. Prompt learning allows tasks to be performed with few samples (few-shot learning) or even without exemplar samples (zero-shot learning) in the absence of large amounts of labeled data. This makes prompt learning particularly useful in data-scarce scenarios. Textual prompts are usually natural language phrases or sentences that contain task information or requirements in some form that guide the model to generate the desired output. A prompt is a piece of text input to a pre-trained language model that contains the context and goal of the task. For example, for the task of recognizing the movements of endangered animals, the textual prompt could be something like “This giant panda is ‘walking’”, which contains an animal-specific prompt, where ‘walking’ is the ‘label’.

We used templates in the text prompts to modify the original input action labels into text string prompts. As shown in Figure 3b, we specifically used three prompt templates, where some unfilled slots need to be filled with predefined action labels. We formally

define a set of three reproducible prompt templates $P = \{p_1, p_2, p_3\}$, where [ACTION] and [SPECIES] are slots to be filled:

p_1 : “A video of a [SPECIES] [ACTION].”

p_2 : “Footage showing an animal engaged in [ACTION].”

p_3 : “This clip contains the behavior [ACTION].”

Application across species/actions:

(1) For regular training/evaluation

All three templates are used. For each video–text pair, the action label (e.g., “walking”) is filled into the [ACTION] slot of each template. The [SPECIES] slot in p_1 is filled with the actual species (e.g., “giant panda”). The three resulting text strings are encoded separately, and their similarity scores with the video features are averaged.

(2) For zero-shot evaluation

Prompts for unseen actions are generated using the same templates and filling rules, leveraging the model’s learned cross-modal alignment.

(3) For species-specific prompt ablation (Please refer to the Section 4 for details).

We compare generic prompts (p_2, p_3) with species-specific prompts (e.g., fixing p_1 as “A video of a giant panda [ACTION]”).

All prompt texts are tokenized and embedded using CLIP’s original tokenizer before being fed into the text encoder. The prompt templates themselves are frozen (non-learnable) in our main experiments, focusing on leveraging the pre-trained model’s prior knowledge. The model learns to better align the generated text features with video features during training.

2.3.2. Visual Prompt Module

For the video encoder module in the video modal branch, we use CLIP [30] pre-trained image encoders, which are pre-trained on large-scale image data and have strong image understanding capabilities. The ABCLIP model exploits this property to extract rich image features for each video frame. However, CLIP is a framework for processing image–text data, where the image encoder does not efficiently extract temporal information from the video. Image and video are not the same; the video modality is composed of multiple frames (images), and the video itself has additional temporal information for the image, and the temporal information is crucial for the action recognition task. Then, it is necessary to design appropriate prompt modules to be added to the pre-trained model to extract video features. We have designed the following two types of visual prompt modules in the video modal branch.

1. Spatiotemporal Prompt

As shown in Figure 3c, this prompt operates on the input before it is fed to the image encoder. Given a video, first, the video is decomposed into a series of consecutive frames. Each frame is treated as a separate image and undergoes pre-processing steps, such as resizing and normalization, to ensure that they match the input requirements of the model. Next, each frame is fed into a pre-trained image encoder to extract visual features. In this process, the spatial position embedding of each frame is used to capture the layout and spatial relationships of the objects within the frame, while we label the order of the frames by incorporating additional learnable temporal position embeddings to ensure that the model can perceive the temporal information of the frames (between each image) in the video. This allows the model to extract visual features of the video content by learning the spatial and temporal dimensions of the video.

2. Multi-Frame Integration Prompt

Given an animal behavior video with F-frames, the video data is processed by the pre-module of the video modality, which is responsible for modeling the interaction between the tokens extracted from the same temporal index (the same image frame). The final model has to obtain the feature information of the whole video and then compare it with the text feature information for learning, so how to aggregate the extracted frame-level features is a key issue. For this reason, we designed a Multi-Frame Integration Prompt module in the model to interact between different frame-level features and aggregate them into a single video-level feature. The Multi-Frame Integration Prompt module is divided into three different integration prompting methods. The specific integration prompt methods are shown in Figure 4.

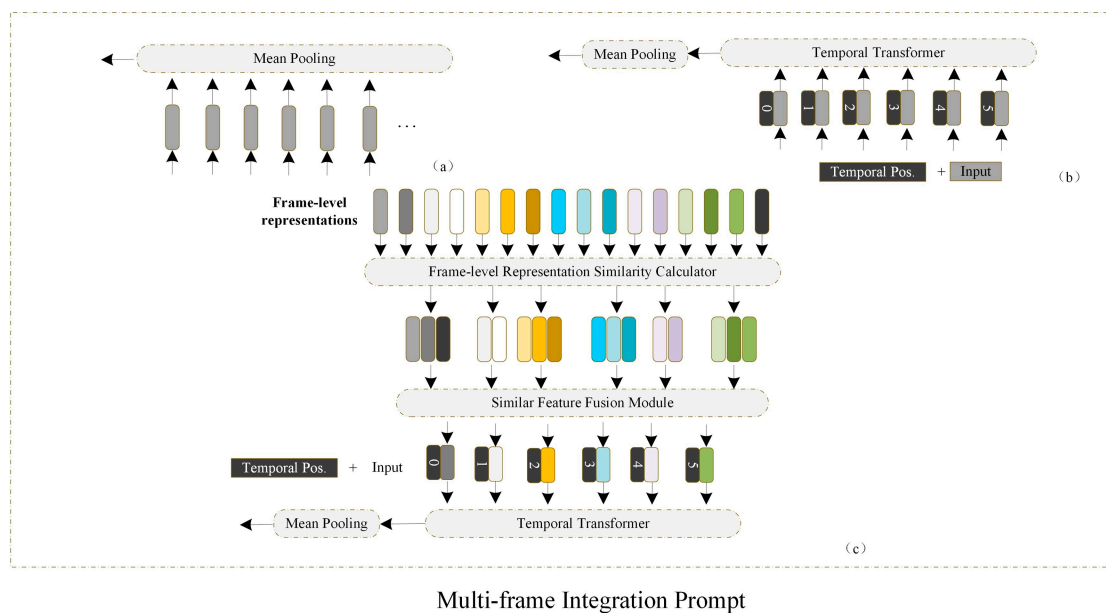


Figure 4. Multi-Frame Integration Prompt module. (a) represents the method using average pooling directly; (b) represents feeding frame-level features into a Temporal Transformer to capture temporal dependencies among different frame features; (c) represents an efficient aggregation prompting method.

Figure 4a represents a direct approach to average pooling, where video features are obtained by averaging features at different frame levels. Average pooling is one of the simple but effective methods. By taking the average of multiple frame-level features, average pooling does not focus only on the extreme values in the features as maximum pooling does, but rather it integrates the information of all frames, which can retain the semantic information of the video more comprehensively. The implementation of average pooling is very simple and does not require complex operations or parameter tuning.

Figure 4b represents the further input of frame-level features into the temporal Transformer to capture the temporal dependencies of different frame features. A temporal Transformer is a model based on the Transformer architecture, which is specifically designed for processing and modeling time-series data or temporal information. The temporal Transformer is capable of capturing temporal dependencies and dynamics in data in tasks such as video processing, action recognition, sequence prediction, etc. Temporal Transformer adopts an architecture similar to that of traditional Transformer, including a multi-head self-attention mechanism, feed-forward neural network, position embedding, and residual connectivity. We first add learnable temporal positional embeddings on the frame-level features to help the model understand the temporal structure of the data and then input them

into the temporal Transformer to efficiently integrate the temporal information between the input frame features through the self-attention mechanism. The subsequent operations are the same as average pooling since, for the frame-level task, feature representations with the same number of input frame features are output.

Figure 4c represents an efficient aggregation prompting method. The difference between this method and the previous two methods is that it does not use all the frame-level features. The aggregated prompting method finally uses a small and discriminative amount of feature information, which not only can better extract the timing information of the video but also can make the whole model less complex and computational. The whole model can achieve a certain slimming effect and reduce the computational overhead required by the model, which makes it better to be applied in animal behavior monitoring and management. This is done as follows. Given a video consisting of 16 frames, the model will get 16 frame-level features through the output of the video pre-modal image encoder. As shown in Figure 4c, different-colored features represent different frame-level features. After that, we employ a decision-making mechanism for feature grouping and fusion. First, features are chronologically split into odd and even frames. Then, the cosine similarity between each odd-frame feature and all even-frame features is calculated. Grouping is based on predefined thresholds: if the similarity is above an acceptance threshold (α , e.g., 0.8), the odd-frame feature is assigned to the corresponding even-frame's group; if it is below a rejection threshold (β , e.g., 0.2), it forms no group at this stage; if the similarity is between α and β , the decision is delayed.

Implementation of delayed decision: During training, for gradient stability, undecided frames are temporarily treated as independent single-frame groups. During inference, for efficiency and determinism, an undecided frame is assigned to the group of the even-frame with which it has the highest similarity. If its similarity to all even-frames is below α , it remains a standalone group.

This process yields multiple groups of similar features. Features within each group are fused (e.g., by averaging) to obtain intra-group fused features. These group features are then fed into a temporal Transformer to capture inter-group temporal dynamics efficiently. Finally, average pooling aggregates the output into the final video-level feature representation.

The computational complexity of the grouping step is $O((F/2)^2 \times d)$, where F is the total number of frames and d is the feature dimension. This is typically lower than the $O(F^2 \times d)$ complexity of applying a full-sequence Transformer directly to all frames, making our method efficient for the modest F (e.g., 16) used in our experiments.

2.3.3. Encoder Module

Both the text encoder and the image encoder in the model use the CLIP [30] pre-trained model, and the specific architecture follows the CLIP encoder architecture. The encoder uses the Transformer architecture, which consists of multiple layers stacked on top of each other. This architecture performs well in natural language processing tasks, where the text encoder first transforms the input text sequence into a dense vector representation. Each vocabulary word is mapped to a fixed dimensional vector by an embedding matrix. Since the Transformer has no inbuilt sequence processing capability, positional encoding is added to represent the position of the word in the sequence. The positional encoding is added to the input embedding, and a multi-head self-attention mechanism is used in the encoder, which captures the relationship between different positions in the sequence through multiple attention heads. Each attention head computes the attention independently, and then the results are stitched together to get the final representation by linear transformation. Specifically, the text encoder in the ABCLIP model employs a 12-layer Transformer structure

containing 512 dimensions of hidden layers and using 8 attention heads per layer. It extracts text feature representations from the activation of the highest layer at the end-of-sequence symbol (EOS). The image encoder in the ABCLIP model uses a Vision Transformer. Vision Transformer (ViT) is a deep learning model based on a self-attention mechanism in which a multi-head self-attention mechanism is also employed, dedicated to processing visual tasks, especially image classification. Unlike traditional Convolutional Neural Networks (CNNs), ViT splits the input image into small pieces (patches) and then processes these image pieces through a multi-layer attention mechanism. This approach allows ViT to efficiently capture global and local features in an image without using convolutional layers, thus achieving performance that matches or even exceeds that of traditional CNNs. The image encoder in the ABCLIP model uses a Vision Transformer (ViT). We employ the standard ViT architecture as used in CLIP. Both ViT-B/32 and ViT-B/16 share the same ‘base’ (B) configuration in terms of Transformer depth and width: they have 12 Transformer encoder layers, a hidden size of 768, and 12 attention heads. The key difference lies in the patch size: ViT-B/32 uses a patch size of 32×32 pixels, while ViT-B/16 uses 16×16 pixels, leading to finer-grained input representations at the cost of increased computational load. In our experiments, we primarily used ViT-B/16 as the image encoder backbone.

2.3.4. Feature Similarity Calculation Module

As shown in Figure 3 of the overall model framework, the whole model is divided into two modal branches, separate branches to obtain the overall feature representations of the different modalities. The text modality is used to extract the feature representations of the text of the input labels, and the video modality is used to extract the spatiotemporal features of the input video. To bring the paired video and text representations closer together, the model measures the direct symmetric similarity of the two modalities in terms of cosine distance in the later similarity computation module, and the cosine distances of the features of the two modalities are obtained by the following equation:

$$s(x, y) = \frac{v \cdot t^T}{\|v\| \|t\|}, s(y, x) = \frac{t \cdot v^T}{\|t\| \|v\|} \quad (1)$$

In this context, x and y represent the video and text modalities, respectively, while v and t denote the encoded features of the video and text modalities, respectively. Then, the similarity scores between video-to-text and text-to-video are obtained using the following formula:

$$Z_i^{x2y}(X) = \frac{\exp(s(x, y_i)/\tau)}{\sum_{j=1}^n \exp(s(x, y_j)/\tau)} \quad (2)$$

$$Z_i^{y2x}(y) = \frac{\exp(s(y, x_i)/\tau)}{\sum_{j=1}^n \exp(s(y, x_j)/\tau)} \quad (3)$$

where τ is a learnable temperature parameter and n is the number of training pairs. Given n as the number of training pairs, the ground-truth similarity scores are set such that the probability for positive pairs is 1 and for negative pairs is 0. In our task, multiple videos within a batch may correspond to the same text label, making it suboptimal to treat the similarity scores as a strict ‘1-of-N’ classification problem (as with standard cross-entropy loss) as it would incorrectly treat videos sharing the same label as negative pairs. The symmetric KL divergence (Equation (4)) measures the distance between the predicted and ground-truth similarity distributions in a softer manner, which is more suitable for this many-to-many matching scenario. It encourages the model to learn a smoother similarity distribution between a video and all relevant text labels.

Consequently, the model employs Kullback–Leibler (KL) divergence, defined as the text-to-video contrastive loss, to globally optimize the ABCLIP model as

$$\mathcal{L} = \frac{1}{2} \mathbb{E}_{(x,y) \sim A} [KL(z^{x2y}(x), l^{x2y}(x)) + KL(z^{y2x}(y), l^{y2x}(y))] \quad (4)$$

In this context, A represents the entire animal behavior video training set, $z^{x2y}(x)$ and $z^{y2x}(y)$ denote the video-to-text and text-to-video similarity scores, respectively, while $l^{x2y}(x)$ and $l^{y2x}(y)$ represent the corresponding ground-truth similarity scores. The entire model is optimized using this contrastive loss.

3. Results

3.1. Experimental Procedures

Due to the extremely imbalanced animal behavior data, we conducted experiments with both regular behavior recognition and zero-shot behavior recognition settings. Specifically, we isolated the actions of defecating and playing for the zero-shot behavior recognition experiments. In the following descriptions, any reference not mentioning the zero-shot setting refers to the regular experimental setup. During training, we used the AdamW optimizer to train the ABCLIP model. The base learning rate for pre-trained parameters was set at 5×10^{-7} , while the base learning rate for new modules with learnable parameters was 5×10^{-6} . The model was trained for 40 epochs, with a learning rate warm-up over the first 10% of the total training period, followed by a gradual decay of the learning rate from the target value to zero according to a cosine function for the remainder of the training. We used a video segment-based input frame sampling strategy, with a specific sampling rate of 16 frames per segment. The training was conducted on a Tesla V100-PCIE-32GB GPU, taking approximately 0.717 h. Compared to I3D and SlowFast, our training speed is significantly faster. During the inference phase, to achieve optimal performance, we employed multi-view inference, applying three spatial crops and 10 temporal clips for each video using the best-performing model. The final prediction results are the average of the similarity scores across all views. Figure 5 provides a detailed overview of the changes in the loss function, model accuracy, and learning rate throughout the experimental stages.

3.2. Comparison with Existing Models

In this section, we evaluate the performance of different action recognition methods on our adopted dataset. Table 4 presents the recognition accuracy achieved by various methods, covering the currently popular action recognition models, which include 3D-CNN-based methods, 2D-CNN-based methods, Transformer-based methods, and the multimodal method used in this paper. As shown in Table 4, the first section using 3D-CNN achieves better recognition performance compared to the second section using 2D-CNN methods. This is because 3D models can directly process spatiotemporal information, capturing temporal changes and multi-scale features of actions, thereby enhancing the model's understanding of complex action patterns. Although the third section employs Transformer-based methods, the recognition performance is not significantly better and even slightly lower than that of the first section. This is due to the relatively small scale of the dataset and resource constraints. The fourth section represents our adopted method, showing that ABCLIP achieves superior recognition performance compared to the first three sections. Our model, through video–text multimodal learning, utilizes the supervisory signals of natural language to strengthen the model's representation capabilities, enhancing its generalization ability. Overall, by combining pre-training, prompt learning, and fine-tuning strategies, the model achieves competitive performance (e.g., 82.50% Top-1 accuracy)

even with the limited scale of animal behavior data available for training, demonstrating the effectiveness of this paradigm in data-scarce scenarios.

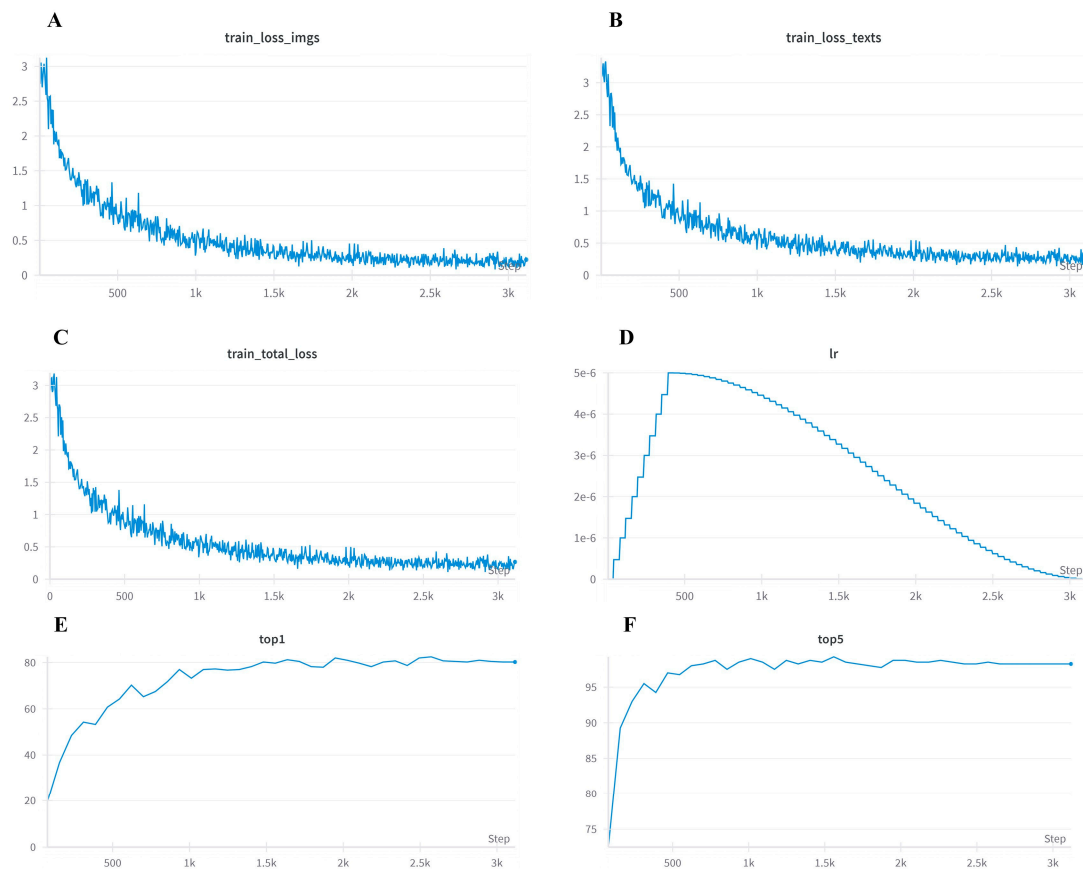


Figure 5. The variations in various data throughout the experimental phases. (A–C) depict the changes in the loss functions for the visual modality, text modality, and overall model during the training phase, respectively. (D) illustrates the variation in the learning rate during training. (E,F) display the changes in Top-1 and Top-5 accuracy, respectively.

Table 4. Comparison with existing models (Mean $\pm$ Std for ABCLIP over 3 runs).

Model	Source	Top-1 Acc (%)	Top-5 Acc (%)
I3D	CVPR2018	77.15	98.34
SlowOnly	ICCV2019	74.13	97.03
SlowFast	ICCV2019	78.54	97.55
TPN	CVPR2020	73.34	94.32
TimeSformer (spaceOnly)	ICML2021	71.70	96.67
TimeSformer (jointST)	ICML2021	75.64	97.86
TimeSformer (divST)	ICML2021	77.24	97.99
ABCLIP (ours)	Ours	82.50 $\pm$ 0.35	99.25 $\pm$ 0.15

3.3. Zero-Shot Behavior Recognition

To evaluate the generalization capability of ABCLIP to entirely unseen behaviors, we conducted a zero-shot recognition experiment. We completely removed all the videos of two behaviors, ‘playing’ and ‘defecating’, from the training set. During testing, the model was required to recognize videos containing these two unseen behaviors along with the nine seen ones.

- **Baselines:** We compared against two baselines, (1) Random Guess (chance level: $2/11 \approx 18.18\%$) and (2) CLIP Zero-Shot, using the frozen pre-trained CLIP image and text

encoders (without our task-specific fine-tuning) with the same prompt templates for nearest-neighbor matching.

- Results: As shown in Table 5, ABCLIP achieves a zero-shot recognition accuracy of 58.00%, significantly outperforming both baselines. This demonstrates that our multimodal fine-tuning enables effective transfer of learned visual–textual associations to novel behavior concepts.
- Analysis: The per-class breakdown indicates better performance on ‘playing’ (~65%) than on ‘defecating’ (~51%), likely due to the more distinctive and dynamic visual features of playful actions.

Table 5. Zero-shot recognition performance (Mean $\pm$ Std over 3 runs).

Method	Zero-Shot Accuracy (%)
Random Guess	18.18
CLIP zero-shot (pre-trained)	42.15 $\pm$ 1.20
ABCLIP (ours)	58.00 $\pm$ 1.05

3.4. Ablation Study

In this section, we validate the effectiveness of the proposed model in animal behavior recognition. We conducted detailed ablation experiments to compare the effectiveness of different prompt modules and modalities in ABCLIP. Table 6 presents the recognition performance when using different prompt methods within the model. The Spatiotemporal Prompt yielded suboptimal performance (Top-1: 77.37%). We hypothesize that directly prepending learnable temporal position embeddings to the input of the pre-trained image encoder may alter the input distribution that the encoder was optimized for during large-scale image pre-training. This shift could destabilize the extraction of robust visual features, whereas our Multi-Frame Integration Prompt module operates on the encoder’s output features, better preserving pre-trained representations while modeling temporal relations.

Table 6. Ablation of the different prompt modules (Mean $\pm$ Std over 3 runs).

Video Prompt	Top-1 Acc (%)	Top-5 Acc (%)
Spatiotemporal Prompt	77.37 $\pm$ 0.52	97.02 $\pm$ 0.30
Multi-Frame Integration (Mean Pooling)	80.88 $\pm$ 0.41	97.80 $\pm$ 0.25
Multi-Frame Integration (Transf)	82.20 $\pm$ 0.38	98.89 $\pm$ 0.19
Multi-Frame Integration (Similar-Transf)	82.50 $\pm$ 0.35	99.25 $\pm$ 0.15

Table 7 demonstrates the benefits of the Multimodality used in this study. Under the supervision of the semantic information from label texts, the multimodal learning framework significantly improved performance, increasing accuracy by 3.10% compared to the Unimodality. This indicates that the multimodal framework is highly beneficial for learning representations for animal action recognition. Overall, ABCLIP provides a powerful tool for monitoring animals’ behavior.

Table 7. Ablation of the multimodal framework (Mean $\pm$ Std over 3 runs).

Accuracy	Unimodality	Multimodality
Top-1	79.40 $\pm$ 0.45	82.50 $\pm$ 0.35
Top-5	97.20 $\pm$ 0.28	99.25 $\pm$ 0.15

3.5. Ablation on Loss Functions

We further conducted an ablation study to compare the symmetric KL divergence loss with two common alternatives: InfoNCE loss and cross-entropy (CE) loss (applied on the

video–text similarity matrix). Results averaged over 3 random seeds are shown in Table 8. While all losses are effective, the symmetric KL divergence achieves slightly better and more stable performance, justifying our choice.

Table 8. Ablation of different loss functions (Mean $\pm$ Std).

Loss Function	Top-1 Accuracy (%)	Top-5 Accuracy (%)
Cross-entropy (CE)	81.23 $\pm$ 0.41	98.05 $\pm$ 0.21
InfoNCE	81.89 $\pm$ 0.38	98.67 $\pm$ 0.18
KL divergence (ours)	82.50 $\pm$ 0.35	99.25 $\pm$ 0.15

4. Discussion

Beyond technical performance, the ABCLIP framework demonstrates a viable pathway for integrating artificial intelligence into sustainable wildlife conservation strategies. By reducing dependence on labor-intensive and invasive monitoring methods, our approach promotes resource-efficient and ethically sound conservation practices. This technological shift is essential for developing scalable evidence-based management policies that ensure the long-term survival of endangered species and the health of their habitats, which are core objectives of sustainability science.

Monitoring the daily behavior of animals is challenging. For example, giant pandas are elusive and highly alert, making it extremely difficult to observe and record their behavior using traditional methods. Therefore, achieving automatic monitoring of animal behavior is essential. With the application of infrared-sensing automatic monitoring cameras in animal monitoring, researchers can collect video data on the daily behavior of animals. This provides strong support for developing models to automatically recognize and monitor animal behavior. However, compared to other current animal action recognition tasks, video data of animal actions remains scarce. The ABCLIP model proposed in this paper addresses this issue by employing pre-training, prompting, and fine-tuning. Specifically, to tackle the challenge of requiring large amounts of data, we used CLIP’s pre-trained models for both the text encoder and video encoder in our network. Then, we incorporated prompt operations and finally fine-tuned the network using a behavior dataset of endangered animals. This framework allows the pre-trained model to be better applied to downstream tasks, leading to effective recognition outcomes.

The training method adopted by the ABCLIP model greatly alleviates the problem of limited data in deep learning tasks. This training paradigm is effective for behavior recognition tasks of endangered animals.

In our experiments, the selected behavior categories are diverse and rare. Regarding the study of giant panda behavior, most current research focuses on common behavior observed in captivity, with comprehensive studies on giant panda behavior being scarce. Captive animals, influenced by factors such as food resources, climate, and habitat environment, exhibit behavior that may not accurately reflect their true characteristics. The model can identify and address threats, formulate effective conservation measures, and aid in protecting endangered species.

The experimental results demonstrate that our proposed ABCLIP model performs well in recognizing the behavior of three endangered species: the giant panda, the red panda, and the yellow-throated marten. Unlike traditional single-modal frameworks, the ABCLIP model approaches the action recognition task as a video–text matching problem within a multimodal learning framework. This framework enhances video representation through additional semantic language supervision and enables the model to perform zero-shot action recognition without any further labeled data. In the zero-shot experimental setting, the ABCLIP model achieved satisfactory recognition results. Through multiple

experiments and averaging the results, ABCLIP achieved a recognition accuracy of 58%, whereas traditional supervised learning methods cannot perform zero-shot recognition. Multimodal methods can simultaneously process and understand different types of data (such as images and text), enabling the model to represent information more richly and comprehensively. Multimodal models can leverage complementary information between images and text to improve the accuracy of understanding and generation. Notably, we also discovered this in additional experiments. In these experiments, we specifically focused on monitoring giant pandas and employed specific animal prompts.

To provide a clear and stable visualization of the effect of species-specific prompts, Figure 6 compares the performance with and without such prompts across ten core behavioral categories, excluding 'DrinkWater'. This exclusion is intentional for this analysis for two reasons: (1) 'DrinkWater' is an extreme long-tail category with very few samples (see Table 3), making its accuracy metric statistically less reliable for fine-grained comparison; (2) the primary aim of this analysis is to demonstrate the consistent benefit of prompting on common well-represented behaviors. As Figure 6 shows, incorporating species-specific prompts leads to performance gains or maintenance across most categories, with notable improvements in 'Communication Behaviors' like smelling and UrineSigning.

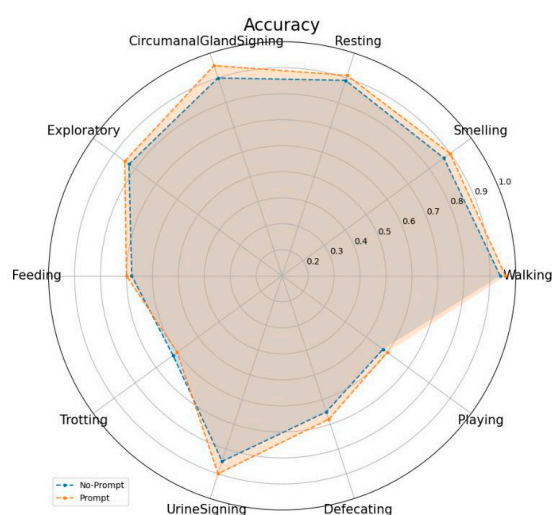


Figure 6. Comparison of recognition performance on ten core behavioral categories (excluding DrinkWater) with and without species-specific prompts.

The experimental results, as shown in Figure 6, indicate that incorporating specific text prompts performed better than using action labels alone. The specific animal prompts added to the text prompt module further enhanced the informational supervision of the text modality. Adding specific animal prompts (such as “giant panda”) helps the model to better understand the context of behavior descriptions. For instance, “giant panda eating bamboo” provides a clearer subject than simply “eating bamboo,” enabling the model to match and recognize the behavior more accurately. Specific animal prompts can help the model to focus on particular categories of images and behavior during training and inference, thus improving recognition accuracy. Specific animal prompts assist the model in better learning the behavior patterns of particular animals during training, which enhances generalization when encountering unseen behavior during inference. For example, the model might not have seen specific images of “giant panda playing,” but, by understanding the specific prompt “giant panda,” it can better infer and recognize related behavior. Overall, the proposed system can assist animal caretakers and managers by providing timely, accurate insights into animal behavior. It enables the early detection of abnormal behavior, helping to identify health issues before they become critical. This technology offers an efficient

scalable solution to improve the overall management and welfare of animals in various settings, including farms, research institutions, and zoos.

5. Conclusions

This study presents a multimodal learning framework, ABCLIP, that advances the toolkit for sustainable wildlife conservation through automated and non-invasive behavior monitoring of endangered species. Although the ABCLIP model demonstrates excellent performance in behavior recognition, it is not without limitations. The prompt-based multimodal approach, while effective for data-scarce and zero-shot scenarios, may not be the most computationally efficient solution for deployment settings with abundant labeled data and a fixed set of behaviors, where simpler task-specific classifiers could be more suitable. Future work could explore hybrid or alternative architectures, such as more specialized transformer-based frameworks [35] or methods that explicitly model relational structures in data [36] tailored for such scenarios.

Although the ABCLIP model proposed in this paper demonstrates excellent performance in behavior recognition, it still has certain limitations. For instance, while our experimental data includes some rare behaviors of endangered animals, data on the feeding behavior of certain protected species is relatively limited. Therefore, future research needs to develop a larger and more comprehensive behavior monitoring model. Overall, this study provides valuable insights for monitoring endangered animal behavior, with the experimental results validating the advantages of the multimodal framework in animal behavior recognition. It not only offers significant implications for the practical application of automated animal behavior recognition but also provides a scientific basis for improving endangered animal welfare and optimizing management practices. This will contribute to enhancing conservation strategies for endangered species and promoting further improvements in animal welfare.

Author Contributions: Conceptualization, Z.H.; methodology and software, S.L.; writing, A.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We thank the anonymous reviewers for their constructive feedback.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Black, S.A. Assessing presence, decline, and extinction for the conservation of difficult-to-observe species. In *Problematic Wildlife II: New Conservation and Management Challenges in the Human-Wildlife Interactions*; Springer: Cham, Switzerland, 2020; pp. 359–392.
2. Songer, M.; Delion, M.; Biggs, A.; Huang, Q. Modeling impacts of climate change on giant panda habitat. *Int. J. Ecol.* **2012**, *2012*, 108752. [CrossRef]
3. Smith, J.A.; Gaynor, K.M.; Suraci, J.P. Mismatch between risk and response may amplify lethal and non-lethal effects of humans on wild animal populations. *Front. Ecol. Evol.* **2021**, *9*, 604973. [CrossRef]
4. Walker, K.A.; Trites, A.W.; Haulena, M.; Weary, D.M. A review of the effects of different marking and tagging techniques on marine mammals. *Wildl. Res.* **2011**, *39*, 15–30. [CrossRef]
5. Jewell, Z. Effect of monitoring technique on quality of conservation science. *Conserv. Biol.* **2013**, *27*, 501–508. [CrossRef]

6. Pimm, S.L.; Alibhai, S.; Bergl, R.; Dehgan, A.; Giri, C.; Jewell, Z.; Joppa, L.; Kays, R.; Loarie, S. Emerging technologies to conserve biodiversity. *Trends Ecol. Evol.* **2015**, *30*, 685–696. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Di Cerbo, A.R.; Biancardi, C.M. Monitoring small and arboreal mammals by camera traps: Effectiveness and applications. *Acta Theriol.* **2013**, *58*, 279–283. [\[CrossRef\]](#)
8. 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. *Front. Ecol. Evol.* **2017**, *15*, 26–34. [\[CrossRef\]](#)
9. Tabak, M.A.; Norouzzadeh, M.S.; Wolfson, D.W.; Sweeney, S.J.; Vercauteren, K.C.; Snow, N.P.; Halseth, J.M.; Di Salvo, P.A.; Lewis, J.S.; White, M.D.; et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods Ecol. Evol.* **2019**, *10*, 585–590. [\[CrossRef\]](#)
10. 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 (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 19023–19034.
11. Von Ziegler, L.; Sturman, O.; Bohacek, J. Big behavior: Challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology* **2021**, *46*, 33–44. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Astill, J.; Dara, R.A.; Fraser, E.D.; Roberts, B.; Sharif, S. Smart poultry management: Smart sensors, big data, and the internet of things. *Comput. Electron. Agric.* **2020**, *170*, 105291. [\[CrossRef\]](#)
13. Hou, J.; He, Y.; Yang, H.; Connor, T.; Gao, J.; Wang, Y.; Zeng, Y.; Zhang, J.; Huang, J.; Zheng, B.; et al. Identification of animal individuals using deep learning: A case study of giant panda. *Biol. Conserv.* **2020**, *242*, 108414. [\[CrossRef\]](#)
14. Li, Q.; Chu, M.; Kang, X.; Liu, G. Temporal aggregation network using micromotion features for early lameness recognition in dairy cows. *Comput. Electron. Agric.* **2023**, *204*, 107562. [\[CrossRef\]](#)
15. Liu, M.S.; Gao, J.Q.; Hu, G.Y.; Hao, G.F.; Jiang, T.Z.; Zhang, C.; Yu, S. MonkeyTrail: A scalable video-based method for tracking macaque movement trajectory in daily living cages. *Zool. Res.* **2022**, *43*, 343. [\[CrossRef\]](#)
16. Zhang, Y.J.; Luo, Z.; Sun, Y.; Liu, J.; Chen, Z. From beasts to bytes: Revolutionizing zoological research with artificial intelligence. *Zool. Res.* **2023**, *44*, 1115. [\[CrossRef\]](#)
17. Qiao, Y.; Guo, Y.; Yu, K.; He, D. C3D-ConvLSTM based cow behaviour classification using video data for precision livestock farming. *Comput. Electron. Agric.* **2022**, *193*, 106650. [\[CrossRef\]](#)
18. Sun, G.; Liu, T.; Zhang, H.; Tan, B.; Li, Y. Basic behavior recognition of yaks based on improved SlowFast network. *Ecol. Inform.* **2023**, *78*, 102313. [\[CrossRef\]](#)
19. Wang, Y.; Li, R.; Wang, Z.; Hua, Z.; Jiao, Y.; Duan, Y.; Song, H. E3D: An efficient 3D CNN for the recognition of dairy cow's basic motion behavior. *Comput. Electron. Agric.* **2023**, *205*, 107607. [\[CrossRef\]](#)
20. Cheng, M.; Yuan, H.; Wang, Q.; Cai, Z.; Liu, Y.; Zhang, Y. Application of deep learning in sheep behaviors recognition and influence analysis of training data characteristics on the recognition effect. *Comput. Electron. Agric.* **2022**, *198*, 107010. [\[CrossRef\]](#)
21. Fuentes, A.; Yoon, S.; Park, J.; Park, D.S. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. *Comput. Electron. Agric.* **2020**, *177*, 105627. [\[CrossRef\]](#)
22. Li, C.; Xiao, Z.; Li, Y.; Chen, Z.; Ji, X.; Liu, Y.; Feng, S.; Zhang, Z.; Zhang, K.; Feng, J.; et al. Deep learning-based activity recognition and fine motor identification using 2D skeletons of cynomolgus monkeys. *Zool. Res.* **2023**, *44*, 967. [\[CrossRef\]](#)
23. Wang, M.; Xing, J.; Liu, Y. Actionclip: A new paradigm for video action recognition. *arXiv* **2021**, arXiv:2109.08472. [\[CrossRef\]](#)
24. Wang, M.; Xing, J.; Mei, J.; Liu, Y.; Jiang, Y. ActionCLIP: Adapting Language-Image Pretrained Models for Video Action Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 625–637. [\[CrossRef\]](#)
25. Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; Ji, R. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval (Version 2). *arXiv* **2022**, arXiv:2207.07285.
26. Schiappa, M.C.; Rawat, Y.S.; Shah, M. Self-Supervised Learning for Videos: A Survey. *ACM Comput. Surv.* **2023**, *55*, 1–37. [\[CrossRef\]](#)
27. Fernandes, A.F.A.; Dórea, J.R.R.; Rosa, G.J.D.M. Image analysis and computer vision applications in animal sciences: An overview. *Front. Vet. Sci.* **2020**, *7*, 551269.
28. Sharma, A.; Jain, A.; Gupta, P.; Chowdary, V. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access* **2020**, *9*, 4843–4873. [\[CrossRef\]](#)
29. Gao, T.; Fisch, A.; Chen, D. Making pre-trained language models better few-shot learners. *arXiv* **2020**, arXiv:2012.15723.
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), Online, 18–24 July 2021; pp. 8748–8763.
31. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

32. 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 (ICCV), Paris, France, 2–6 October 2023; pp. 20064–20075.
33. Gherardi, F.C. Operative versus conceptual classification of animal behaviour. *Hist. Philos. Life Sci.* **1983**, *5*, 87–99.
34. Packard, J.M.; Ribic, C.A. Classification of the behavior of sea otters (*Enhydra lutris*). *Can. J. Zool.* **1982**, *60*, 1362–1373. [[CrossRef](#)]
35. Wang, S. Development of an automated transformer-based text analysis framework for monitoring fire door defects in buildings. *Sci. Rep.* **2025**, *15*, 43910. [[CrossRef](#)] [[PubMed](#)]
36. Wang, S.; Moon, S.; Eum, I.; Hwang, D.; Kim, J. A text dataset of fire door defects for pre-delivery inspections of apartments during the construction stage. *Data Brief* **2025**, *60*, 111536. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.