International Journal of Computer Vision (2025) 133:5062–5082
https://doi.org/10.1007/s11263-025-02408-4

Animal-CLIP: A Dual-Prompt Enhanced Vision-Language Model for Animal Action Recognition

Yinuo Jing[1] · Kongming Liang[1] · Ruxu Zhang[1] · Hao Sun[2] · Yongxiang Li[2] · Zhongjiang He[2] · Zhanyu Ma[1]

Received: 18 September 2024 / Accepted: 23 February 2025 / Published online: 4 June 2025

Abstract

Animal action recognition has a wide range of applications. With the rise of visual-language pretraining models (VLMs), new possibilities have emerged for action recognition. However, while current VLMs perform well on human-centric videos, they still struggle with animal videos. This is primarily due to the lack of domain-specific knowledge during model training and more pronounced intra-class variations compared to humans. To address these issues, we introduce Animal-CLIP, a specialized and efficient animal action recognition framework built upon existing VLMs. To address the lack of domain-specific knowledge in animal actions, we leverage the extensive expertise of large language models (LLMs) to automatically generate external prompts, thereby expanding the semantic scope of labels and enhancing the model's generalization capability. To effectively integrate external knowledge into the model, we propose a knowledge-enhanced internal prompt fine-tuning approach. We design a text feature refinement module to reduce potential recognition inconsistencies. Furthermore, to address the high intra-class variation in animal actions, a novel category-specific prompting method is introduced to generate adaptive prompts to optimize the alignment between text and video features, facilitating more precise partitioning of the action space. Experimental results demonstrate that our method outperforms six previous action recognition methods across three large-scale multi-species, multi-action datasets and exhibits strong generalization capability on unseen animals.

Keywords Animal action recognition · Vision-language pre-training · Prompt learning · External knowledge

1 Introduction

Biodiversity, ecosystems, and the essential services they provide are fundamental for all forms of life on earth, including humans (Romanelli et al., 2015). Animals are integral components of natural ecosystems, driving crucial ecological processes such as seed dispersal, nutrient cycling, and predator–prey dynamics through their actions (Chen et al., 2023). Consequently, monitoring and understanding animal actions are vital for comprehending the complexity of natural ecosystems and promoting biodiversity. Automated animal action recognition through computer vision facilitates extensive monitoring and analysis, not only mitigating the labor and time expenditures associated with management but also enabling the prompt identification of anomalous actions or health issues, thereby advancing the efficacy of targeted conservation and wildlife protection initiatives. Currently, automated animal action recognition is applied in and plays a significant role across various domains, including ethology research (Anderson & Perona, 2014; Graving et al., 2019; von Ziegler et al., 2021), animal disease manage-

Communicated by Anna Zamansky.

✉ Kongming Liang
liangkongming@bupt.edu.cn

Yinuo Jing
jingyinuo@bupt.edu.cn

Ruxu Zhang
zhangruxu@bupt.edu.cn

Hao Sun
sun.010@163.com

Yongxiang Li
liyx25@chinatelecom.cn

Zhongjiang He
hezj@chinatelecom.cn

Zhanyu Ma
mazhanyu@bupt.edu.cn

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China

[2] China Telecom Artificial Intelligence Technology Co. Ltd.,, Beijing 100036, China

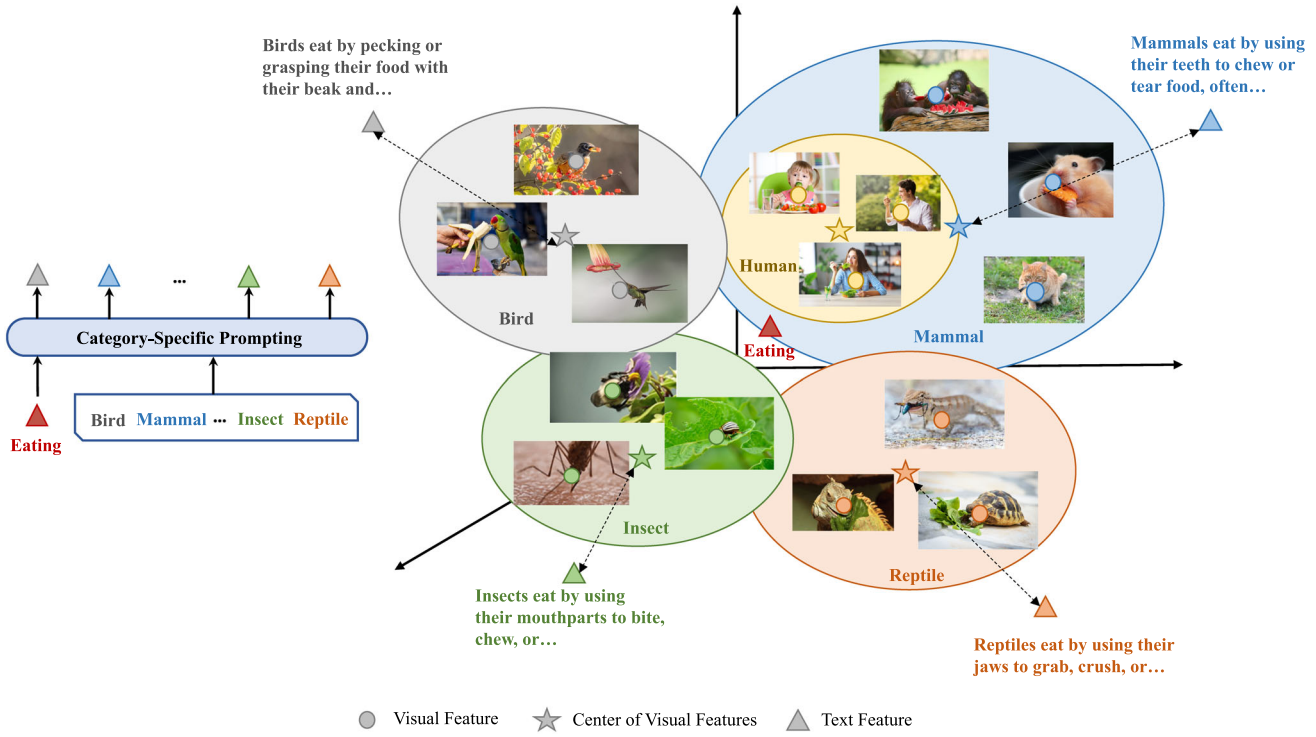


Fig. 1 Action feature space: In contrast to previous work that dealt only with human actions, the animal kingdom includes diverse categories with complex visual features, making it challenging for rigid textual features to align with this visual diversity. By creating tailored textual prompts for different animals, we enhance the granularity of text and improve the alignment between textual and visual features

ment (Feng et al., 2021; Singh et al., 2020), endangered species protection (Nguyen et al., 2017), and human diseases understanding (Anderson & Perona, 2014; Graving et al., 2019; Karashchuk et al., 2021).

In the early stages of research on this field (Nguyen et al., 2021; Yang et al., 2018; Segalin et al., 2021; Geuther et al., 2021; Graving et al., 2019; Ravbar et al., 2019), due to the fact that most animal action data came from experiments in biological science, most action recognition models were focused on a single species within a single laboratory setting. The limitations of both the species and the scene made it difficult to apply the model in broader, more diverse real natural environments.

Currently, animal action recognition has gained increasing attention. The emergence of animal action video data of multiple species in the wild (Ng et al., 2022; Chen et al., 2023; Liu et al., 2023) has made it possible to employ computer vision algorithms for automated animal action recognition in real-world scenarios. Ng et al. (2022) proposed a large-scale dataset called Animal Kingdom for animal action recognition and presented the results of three Convolutional Neural Network-based baselines: I3D (Carreira and Zisserman, 2017), SlowFast (Feichtenhofer et al., 2019), and X3D (Feichtenhofer, 2020). Mondal et al. (2023) adopted a transformer-based framework to establish a non-actor-specific action recognition framework. While these traditional methods achieve decent performance, they still have some drawbacks. These models are all heavily rely on large amounts of training data, which poses significant challenges in generalizing to actions with limited samples and to previously unseen species. It is worth mentioning that in order to not disturb the normal living environment of animals, collecting animal action data can be challenging, which often require complicated design (Pascoe et al., 2000). As shown in Fig. 3c, we conducted a statistical analysis of the distribution of action labels in the Animal Kingdom datasets. The action distribution exhibits a pronounced long-tail pattern, with the majority of actions being represented by a very limited number of samples. Traditional Models struggle to perform effectively on these underrepresented tail actions.

Recently, with the advent of vision-language large models(VLM) (Radford et al., 2021; Chen et al., 2023; Li et al., 2023; Lin et al., 2023; Su et al., 2023; Wang et al., 2022, 2023; Zhang et al., 2023), the extensive pre-training data and large parameter scales of these models have endowed them with robust open-set recognition capabilities, facilitating their application across a range of domains. However, the application potential of these models in animal action recognition tasks remains unknown. We tested the models on two animal action recognition datasets, designing

Table 1 Accuracy (%) comparison of VLMs. They perform well on human-centric datasets but show poor results on animal-centric datasets

Dataset	Agent	VLM	
		Video-LLaVA	Video-Chat2
UCF101	Human	94.55	98.63
MammalNet	Animal	48.31	70.00
LoTE-Animal	Animal	30.63	47.25

question-option pairs where one option is derived from the ground-truth label, and the remaining three options are randomly selected from other labels within the dataset. As shown in Table 1, existing multi-modal large models achieve high accuracy in human-centered action recognition but exhibit low accuracy in animal-centered action recognition.

We analyzed and summarized the potential reasons for the poor performance of VLMs. From the perspective of model training, the models lack domain-specific expertise in animal actions, which may lead to two issues. First, the ability of the models to generalize to new species is limited. For instance, humans leverage their existing knowledge to recognize the actions of unfamiliar species. Once a child learns the *predatory* action of a cat, they can also recognize the same action in a lion. As their knowledge expands, they become increasingly capable of identifying the actions of new species through experience. However, current models lack this domain-specific knowledge, resulting in weak generalization capabilities. Second, predictions are susceptible to label noise. Label noise refers to action labels that are inconsistent with the natural capabilities of specific animals, defined based on common understanding. As shown in Fig. 2, when evaluating animal actions, humans first identify the animal, then narrow down the possible actions, selecting the correct one from a limited set. For example, when a gorilla appears in a video, humans instinctively expect the gorilla to *eat* or *climb*, rather than *dive* or *fly*. Thus, action labels such as *dive* and *fly* are considered noise for gorillas. However, current models treat each action label as equally probable for all animals, ignoring the presence of label noise. As shown in Figs. 3a and b, we analyzed the relationships between each action and animal. It is evident that most actions are associated with only a small subset of animals, and over 50% of animals exhibit no more than five action categories. The relationship between animals and actions is semantically asymmetric, a phenomenon that linguists often observe in subject-verb relations (Tapanainen et al., 1998). This semantic asymmetry entails strong priors, which can significantly assist models in making accurate judgments.

Additionally, from the perspective of the task itself, animal actions present greater intra-class variability compared to human actions. As illustrated in Fig. 1, human action videos are confined to a relatively narrow visual space within the

Fig. 2 Existing models cannot effectively narrow down the selection of actions and are easily confused by a large number of labels due to their lack of specialized knowledge

mammalian category. In contrast, the broader animal kingdom encompasses diverse categories such as birds, insects, and reptiles, each with complex and varied morphological characteristics (Ng et al., 2022). The complexity and dispersion of the visual space make it challenging to cover the large range of corresponding visual features with rigid and inflexible textual features when training vision-language models. Additionally, individuals of the same category may exhibit actions in diverse natural environments, which contain dynamic elements such as swaying tree branches and flowing streams. These environmental factors can distract models from focusing on the animals themselves in the videos, resulting in reduced model performance.

In this paper, we introduce a dual-prompt enhanced vision-language model for animal action recognition, named Animal-CLIP, aiming to address the challenges faced by existing models in applying to this domain.

First, to address the weak generalization ability caused by the lack of expert knowledge, we introduce external prompts. With the rise of large language models (LLMs) (Touvron et al., 2023; Taori et al., 2023), some studies (Liang et al., 2023a, b) have leveraged the rich external knowledge embedded in these models to inject new intelligence into them. To further enrich the semantic knowledge of the model and improve its generalization ability, we use LLMs to expand abstract animal and action labels into more detailed descriptions, incorporating external knowledge. In addition, we combine LL Ms with statistical data to generate relation-

ships between animals and actions, which are subsequently modeled as a graph structure and integrated into the model.

Furthermore, to address the challenges posed by severe label noise and the high intra-class variation in animal actions, we introduce knowledge-enhanced internal prompt fine-tuning. Initially, we design a "Text Feature Refinement Module", which models the relationships between animals and actions in the external prompts, enabling the model to process animal action labels in a manner that reflects their inherent characteristics. This approach mitigates the impact of label noise. Additionally, we propose a category-specific prompting method for generating text and video prompts. Specifically, we leverage the animal categories present in the videos and employ a category-text specific prompting to generate tailored textual prompts based on the features of the respective animal categories. This refinement improves the granularity of the text representations, facilitating a more precise alignment with distinct subspaces in the visual feature space, thereby enhancing text-video feature alignment. Moreover, we introduce a category-visual specific prompting module to strengthen the model's attention to individual animal instances within the video, reducing the influence of background noise on model predictions. To address potential challenges arising from the lack of animal labels in practical scenarios, we further propose a category feature extraction module, thereby expanding the model's applicability.

External and internal prompts complement each other in the model, exhibiting a synergistic effect. The internal prompt focuses on specific animals within the video, generating customized textual and visual prompts aligned with the animal's characteristics, thereby enhancing the precision of visual-text feature alignment. Meanwhile, the external prompt module incorporates domain knowledge from large language models, broadening the semantic scope of the model and facilitating cross-species action recognition, which improves the model's adaptability in handling rare species and actions. The combination of both significantly enhances the model's accuracy and robustness.

Our main contributions are as follows.

- To the best of our knowledge, we are the first to conduct large-scale multi-species animal action recognition research in real-world environments, rather than in controlled laboratory settings, and to propose innovative approaches tailored to these complex conditions. Our approach has surpassed the performance of previous action recognition methods on the Animal Kingdom, MammalNet, and LoTE-Animal datasets, and exhibits robust generalization capabilities on unseen species.
- We introduce an innovative framework for animal action recognition that leverages vision-language models pretrained on extensive data and fine-tuned to align with specific task characteristics.
- We introduce external prompts by leveraging large language models (LLMs) to generate descriptive labels for animals and actions, thereby expanding the semantic scope of limited labels and enhancing the model's generalization capability. Additionally, we utilize LLMs and statistical data to establish relationships between animals and actions.
- We propose knowledge-enhanced internal prompts. First, a text feature refinement module is designed to establish these animal-action relationships in the form of graphs, mitigating the noise in action labels. We further propose a category-specific prompting module that generates tailored text and video prompts for distinct animal categories, addressing the significant intra-class variations in animal actions. Moreover, we introduce a category feature extraction module to tackle the issue of missing ground truth labels for animals.

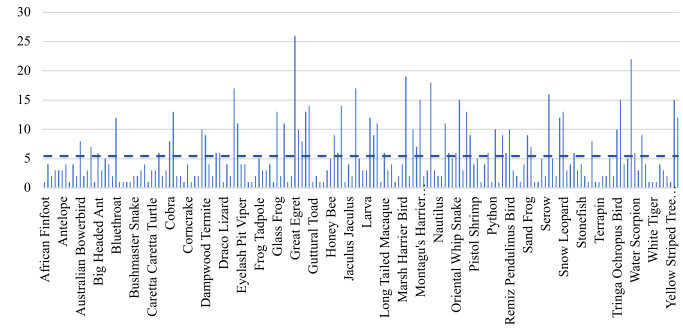
2 Related Work

In this section, we will review relevant research work related to the topic of this paper, including animal action recognition, visual-language pretraining models, prompt learning methods and knowledge-enhanced prompt learning.

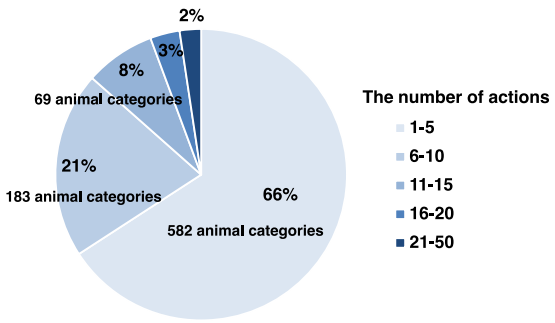
2.1 Animal Action Recognition

Action recognition is about classifying activities performed by one or more subjects in a trimmed video. Although there is an increasing amount of recent work focusing on human action recognition, animal action recognition is still a relatively unexplored field. In this field, most work is limited to a small subset of animals and actions in specific environments. These studies include research on the actions of animals in farming environments, such as cows (Nguyen et al., 2021) and pigs (Yang et al., 2018), to optimize management practices and improve production efficiency. Additionally, there is extensive research on the actions of common laboratory animals to understand the effects of experimental interventions, such as studies on mice (Segalin et al., 2021; Geuther et al., 2021) and drosophilas (Graving et al., 2019; Ravbar et al., 2019). Common methods of these studies are to transfer human action recognition models to specific animals. Since these methods are designed with fine-tuning for specific species, they are often difficult to adapt to changes in species in practical applications.

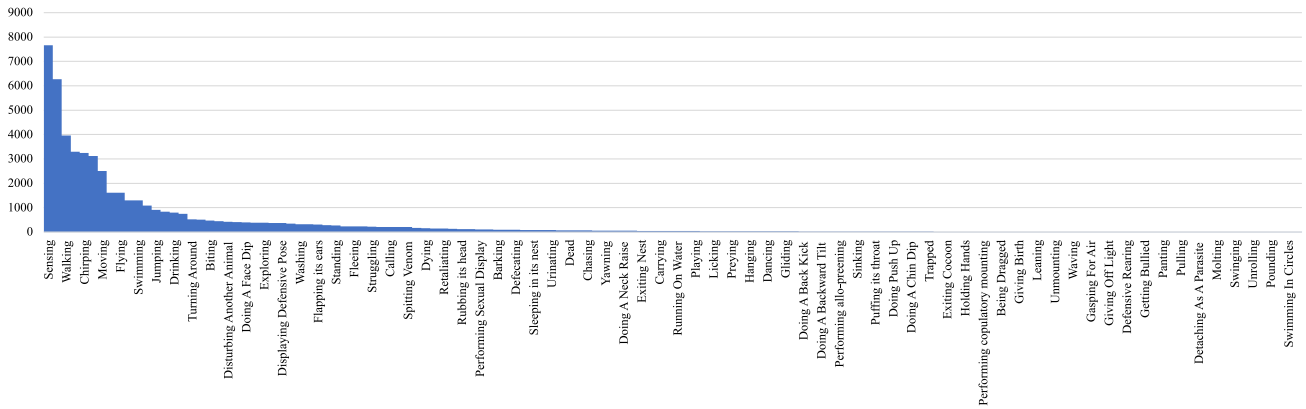
In recent years, with the increasing attention on animal action recognition, various video datasets for multi-species multi-action recognition have emerged. As shown in Table 2, we now have three large-scale animal action datasets available for use. Xun Long Ng et al.'s open source dataset Animal Kingdom (Ng et al., 2022) which contains 30K video



(a) The number of animal labels covered by each action in the Animal Kingdom dataset, with the dashed line indicating the average.



(b) Distribution of action label frequencies across different animal species in the Animal Kingdom dataset.



(c) The frequency of each action label in the Animal Kingdom dataset.

Fig. 3 Animal Kingdom dataset data statistics

Table 2 The comparison of existing open-source animal action datasets covering multiple species, where *Videos*, *Actions*, and *Animals* respectively indicate the number of videos, action labels, and animal labels

Datasets	Videos	Actions	Animals	Source	TD	Labels	LT	Notes
Animal kingdom	30.1K	140	850	web	50h	M	✓	Wide range of animal categories and action types
MammalNet	18K	12	173	web	539h	S	✓	a. Contains only mammals b. Labels are advanced animal behaviors c.The average video duration is long
LoTE-Animal	10K	21	11	wild	21h	S	✓	Contains only endangered species

TD represents Total Duration, and the *Labels* column indicates whether it is Multi-label (M) or Single-label (S). *LT* stands for Long-Tail distribution

sequences involving a diverse range of animals with more than 850 animals across 6 major animal classes. MammalNet (Chen et al., 2023) is a dataset comprising over 150 mammal species and 12 high-level behaviors. Its average video length is longer, requiring more temporal information for animal behavior inference. LoTE-Animal (Liu et al., 2023) is a dataset collected from real environments, containing 11 endangered species. These datasets provide the foundation for our research.

In these works, considering the lack of action recognition models specifically for multiple categories of animals, the authors use human action recognition methods for experiments. Those methods include classic 3D CNN algorithms and two stream algorithms, doing well in extracting representative spatial-temporal features in videos. I3D (Carreira and Zisserman, 2017) is based on 2D ConvNet inflation, expanding the 2D convolution kernels and pooling kernels to 3D form. X3D (Feichtenhofer, 2020) also expect to expand 2D CNN into 3D, but the expanding process was done progres-

sively. It expand a tiny 2D image classification architecture along multiple network axes, in space, time, width and depth. SlowFast (Feichtenhofer et al., 2019) is a two-stream model, which consists of two parts: Slow pathway to capture spatial semantics (objects) and Fast path way to capture shot-time motion. While these methods have shown promising results in human action recognition, there exists a significant domain gap between human and animal action recognition, leading to suboptimal performance of these methods in animal action recognition.

2.2 VLMs and Prompt Learning

Vision-language pretrained models(VLMs) (Miech et al., 2019; Sun et al., 2019a, b; Zhu et al., 2020) that demonstrate great zero-shot generalization ability in downstream tasks, has become the trend in recent years. Such VLMs like ALIGN (Jia et al., 2021) and CLIP (Radford et al., 2021) are trained on large dataset with billions of image-text pairs. In the training phase, vision-language models (VLMs) jointly optimize the image encoder and text encoder to ensure that both modalities are effectively aligned in a shared embedding space. Specifically, the image encoder processes image inputs, and the text encoder processes textual inputs to generate respective embedding representations. The model then learns the correct alignment between images and texts by minimizing the distance between image-text pairs, typically utilizing contrastive loss functions. During the inference phase, VLMs leverage the pretrained text encoder to process category names or descriptions, embedding this information into a fixed linear classifier, thereby forming a zero-shot classifier. This enables the model to make predictions even for categories it has not seen during training. For instance, if the model has been trained with data containing the "cat" and "dog" categories, it can predict a novel category, such as "rabbit", by embedding its description into the classifier and using the zero-shot classifier for prediction, without the need for additional training on specific examples of that category.

Since prompt engineering can have huge impact on the performance of VLMs, many have focused on improving the prompt module. Based on prompt learning methods in NLP (Shin et al., 2020; Jiang et al., 2020; Zhong et al., 2021), CoOp (Zhou et al., 2022a) and Co-CoOp (Zhou et al., 2022b) improve text prompt by modeling the prompt's context words with learnable vectors while fixing the entire pretrained parameters of CLIP. Bahng et al. (2022) performs visual prompt tuning by creating prompts in the form of pixels. VPT (Jia et al., 2022) modifies parameters prepended into the input sequence of each Transformer layer, fulfilling Visual-Prompt Tuning. Those methods focus on learning prompts on either textual or visual way. There are also methods that adapt both text and vision prompt. For example, MaPLe (Khattak et al., 2022) adapts both text and vision branches simultaneously, promoting strong coupling between the vision-language prompts. UPT (Zang et al., 2022) learns a tiny neural network to jointly optimize prompts across vision and language modalities. There are other methods that focus on optimizing the process of relationship-building between visual and textual domains. For example, BIKE (Wu et al., 2022) utilize bidirectional cross-modal knowledge by implementing a pre-defined lexicon as video attributes in Video-to-Text direction and computing temporal saliency in the Text-to-Video direction. Referring to the above approach, it can be concluded that using the prompt learning approaches can help with generalizing VLMs to downstream tasks. When it comes to level of sophistication of prompting mechanism, the existing methods for generating prompts can be classified into three categories, unified prompts (Shin et al., 2020; Jiang et al., 2020; Zhong et al., 2021), class-level prompts (Zhou et al., 2022a), and instance-level prompts (Zhou et al., 2022b). Unified prompts utilize a fixed template to generate input prompts that remain consistent across tasks and video content, making them applicable to a wide range of scenarios. Class-level prompts generate specific prompts for each class, aimed at differentiating categories during inference, typically used in classification tasks where the same prompt is applied to all instances within a given class. Instance-level prompts, in contrast, create tailored prompts for each individual video or instance, allowing for customization based on the unique characteristics of each input. However, instance-level prompts, while tailored to individual videos, suffer from limited generalization capacity. On the other hand, class-level and unified prompts lack the ability to generate differentiated prompts for different animals, whereas animal action often exhibits significant variability even for the same action.

2.3 External Knowledge-Enhanced VLMs

The method of enhancing model performance using external knowledge firstly appeared in natural language processing tasks, where models learned more semantic information by integrating encyclopedic knowledge and common sense knowledge. With the proliferation of multi-modal vision-language pre-trained model (Gpt4, 2023), more and more multi-modal work draws inspiration from methods in natural language processing, injecting new wisdom into models by introducing external knowledge on the text side.

Some approaches build knowledge graphs, which are effective tools for representing real-world entities and their relationships, and embed them into text embedding to provide richer structured knowledge for text encoding. Currently, knowledge graphs have been applied to various tasks in computer vision, including image classification (Kampffmeyer et al., 2019), panoptic segmentation (Wu et al., 2020), image captioning (Zhao and Wu, 2023), visual question answer-

ing (Hudson & Manning, 2019; Shah et al., 2019), and more. Gu et al. (2023) proposed a text with knowledge graph augmented transformer (TextKG) for video captioning, aiming to integrate external knowledge from knowledge graphs and exploit multi-modality information in video to address the challenge of long-tail words. Naeem et al. (2021) utilized the interdependence among states, objects, and their compositions within a graph structure to facilitate the transfer of relevant knowledge from seen to unseen compositions for zero-shot composition recognition. Inspired by the aforementioned works, we apply knowledge graphs to the field of animal action recognition, representing the inherent relationships between animals and actions.

Other approaches adopt natural language descriptions to expand labels into descriptions, enriching their semantics. Early works relied on external natural language knowledge bases, where K-LITE (Shen et al., 2022) enriches entities in text with Word-Net and Wiktionary knowledge, resulting in an efficient and scalable approach to learning image representations that leverages knowledge about visual concepts, thereby enhancing the model's zero-shot and few-shot transfer capabilities. Some recent works (Huang et al., 2023) utilize the powerful knowledge compression ability of large language models (LLMs) (Chatgpt, 2022; Touvron et al., 2023; Taori et al., 2023) to generate tag descriptions and incorporate them into the training process, which is a natural and effective approach to enhance the open-set capability of tagging models. In this work, we leverage the abundant factual knowledge in LLMs to generate descriptions of animals and actions, obtaining richer semantics beyond labels. This allows us to maximize the potential of the visual encoder, strengthen the alignment between visual and textual modalities, and expect the model to achieve better generalization capabilities.

3 Method

In this section, we first present the overall framework. Then, we explain the method for generating external prompts. Finally, we introduce knowledge-enhanced internal prompt fine-tuning, including detailed descriptions of the text feature refinement module, category-specific prompting, and category feature extraction method.

3.1 Overview

Our proposed Animal-CLIP consists of two key components: external prompts generation and knowledge-enhanced internal prompt fine-tuning. In the external prompts generation stage, we leverage large language models (LLMs) to generate animal descriptions, action descriptions, and animal-action relationships. This approach addresses the challenges of weak generalization and significant label noise in large-scale models caused by their lack of domain-specific knowledge. In the internal prompt fine-tuning stage, our framework aligns with most vision-language pretrained model architectures (Radford et al., 2021; Zhou et al., 2022a, b), comprising vision and language branches. To adapt image-language pretrained models for video understanding, we first encode video frames as image patches using the CLIP pretrained image encoder. To effectively capture spatio-temporal dependencies in the visual data, inspired by X-CLIP (Ni et al., 2022), we incorporate two specialized transformer modules: the cross-frame communication transformer and the multi-frame integration transformer. Video clips, along with animal and action descriptions, are independently processed through dedicated encoders to obtain their respective feature representations.

Specifically, given a video clip v, action labels t, and animal labels c, we employ a video encoder $e_v(\cdot|\theta, \phi)$ and two text encoders $e_t(\cdot|\epsilon)$ to derive the video embedding $\mathbf{v}$, text embedding $\mathbf{t}$ and animal category embedding $\mathbf{c}$, respectively. Here, θ, ϕ and ϵ represent the parameters of the image patch embedding, the newly added transformer blocks, and the text encoder, respectively. Therefore,

$$\mathbf{v} = e_v(v|\theta, \phi), \quad \mathbf{t} = e_t(t|\epsilon), \quad \mathbf{c} = e_t(c|\epsilon). \tag{1}$$

We denote $\mathcal{G}(\cdot|\varphi)$ to represent the text feature refinement module(TFRM), where φ is the parameters of this module. The union of the action label features and animal category features outputted by encoders is denoted as $\mathcal{V} = \mathbf{c} \cup \mathbf{t}$. Finally, $\mathcal{E}$ represents the animal-action relation matrix. The enhanced action features, which are fused with external knowledge, can be represented as

$$\hat{\mathbf{t}} = \mathcal{G}(\mathcal{V}, \mathcal{E}|\varphi). \tag{2}$$

We use $g_p(\cdot|\zeta)$ to represent the category-specific prompting, where ζ denotes its parameters. The video and text representations generated through category-specific prompting, with the input of $\mathbf{v}$, $\hat{\mathbf{t}}$ and $\mathbf{c}$ can be calculated as follows

$$\tilde{\mathbf{v}} = g_p(\mathbf{v}, \mathbf{c}|\zeta), \quad \tilde{\mathbf{t}} = g_p(\hat{\mathbf{t}}, \mathbf{c}|\zeta). \tag{3}$$

Then, we can get the cosine similarity score $\text{sim}(\tilde{\mathbf{v}}, \tilde{\mathbf{t}})$ as follows:

$$\text{sim}(\tilde{\mathbf{v}}, \tilde{\mathbf{t}}) = \frac{\tilde{\mathbf{v}} \cdot \tilde{\mathbf{t}}}{\|\tilde{\mathbf{v}}\| \|\tilde{\mathbf{t}}\|}. \tag{4}$$

During the training process, the parameters θ and ϵ of the image patch embedding and text encoder are initialized with a pretrained image-language model such as CLIP and kept fixed. The parameters ϕ, φ and ζ are responsible for capturing spatio-temporal features, refining action features

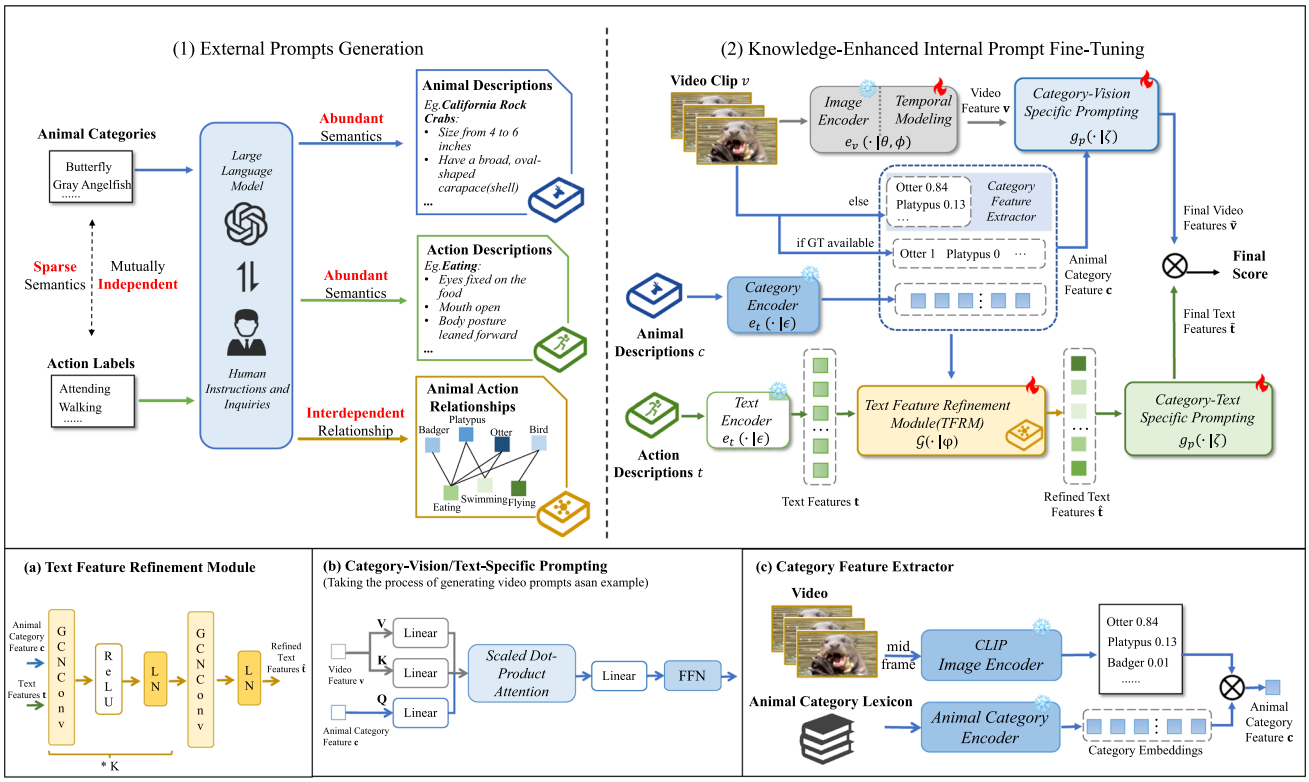


Fig. 4 Our proposed Animal-CLIP consists of two parts: (1) external prompts generation and (2) knowledge-enhanced internal prompt fine-tuning. In the external prompts generation part, we use LLMs to generate three required external prompts: animal, action description, and animal-action relationships. In the internal prompt fine-tuning part, features are extracted for video, animal, and action descriptions using three encoders. Text features are optimized by **a** a text feature refinement module along with animal category features and animal-action relationships to mitigate label noise. Then, video and text features are enhanced via **b** category-specific prompting with animal category features. We also provide a **c** category feature extractor to obtain animal category features in the absence of animal GT. The final prediction score is obtained by computing the similarity between the two representations and applying softmax

and category-specific prompts, respectively, and are the only parameters trained in the network. This approach ensures that the pretrained model's knowledge is utilized while enabling the network to learn features specifically for video-related tasks and for downstream task datasets.

The primary objective is to enhance the cosine similarity between $\tilde{\mathbf{v}}$ and its respective $\tilde{\mathbf{t}}$, prioritizing this alignment while minimizing deviations elsewhere. The loss function during the training process can be represented as follows, where N and C represent the number of samples per batch and the number of action classes, respectively. Then,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp(\sin(\tilde{\mathbf{v}}_i,\tilde{\mathbf{t}}_i)/\tau)}{\sum_{c=1}^{C}\exp(\sin(\tilde{\mathbf{v}}_i,\tilde{\mathbf{t}}_c)/\tau)}\right]. \tag{5}$$

In the inference phase, scores are computed between the video embedding and each action text embedding. These scores are then transformed using the softmax function to derive probabilities for all action classes. The computation of the softmax function is shown in Eq. 6.

$$p\left(j|\tilde{\mathbf{v}}\right) = \frac{\exp(\sin\left(\tilde{\mathbf{v}},\tilde{\mathbf{t}}_j\right)/\tau)}{\sum_{c=1}^{C}\exp(\sin\left(\tilde{\mathbf{v}},\tilde{\mathbf{t}}_c\right)/\tau)}. \tag{6}$$

3.2 External Prompts Generation

Our external prompts primarily consist of three parts: descriptions of action and animal labels, and the potential relationships between animals and actions. Firstly, to deepen the model's semantic understanding, we leverage the insights of large language models (LLMs) to convert semantically constrained action label supervision into extensive semantic descriptions. The prompt design for LLMs is crucial for obtaining descriptions of each category within the label system. We anticipate that the descriptions generated by LLMs will primarily exhibit two characteristics: (i) being as relevant as possible to visual representations, as our model determines animal actions based on visual cues; and (ii) being as diverse as possible to capture richer semantic information. Follow-

ing (Pratt et al., 2022; Huang et al., 2023), we have designed five system prompts specifically for animals and actions, as outlined below:

a. Describe concisely what a/(animals doing) {} looks like;
b. How can you identify a/(animals doing) {} concisely?;
c. What does a/(animal doing) {} look like concisely?;
d. What are the identifying characteristics of a/(animals doing) {}?;
e. Please provide a concise description of the visual characteristics of {}/animals doing {}.

We employ the GPT−3.5-turbo-0125 model for generating responses. Setting the maximum token count to 77 is in line with the tokenization length of the pre-trained text encoder, while adjusting the model's temperature to 0.99 ensures maximum response diversity. We collect 10 model responses for each prompt. The 10 model responses are generated through multiple inferences with the same input. With a temperature setting of 0.99, the model increases randomness and diversity, producing varied responses. This methodology facilitates the exploration of a broader spectrum of possible outcomes. For each label, we generate 50 descriptions and process them through a text encoder to obtain 50 feature tensors. The maximum value for each feature dimension is selected, resulting in a final feature tensor that preserves the most prominent features.

Secondly, we capture the intrinsic relationship between animals and actions. For seen animals and actions, we record the occurrences of animals and actions appearing together in the dataset as their intrinsic relationship. For unseen animals and actions, we query LLM about *Whether a certain animal can perform certain actions* and automatically convert the answers into a binary adjacency matrix.

3.3 Knowledge-Enhanced Internal Prompt Fine-Tuning

In this part, we fully leverage the three types of prompts generated in the external prompts generation stage for knowledge-enhanced internal prompt fine-tuning.

3.3.1 Text Feature Refinement Module

We firstly design a text feature refinement module to embed the animal-action relationships generated in 3.2 into the existing model framework, with the goal of mitigating label noise. We model these relationships as a graph structure and utilize graph convolutional layers to refine action features.

We organize these relationships into a (0,1) matrix format, serving as the adjacency matrix $\mathcal{E} \in \mathbb{R}^{N_E \times N_E}$ for the relationships between actions classes set C and animal classes set D, where $N_E = N_C + N_D$. N_C represents the number of action classes, and N_D represents the number of animal classes. Then the adjacency matrix can be represented as

$$\mathcal{E}_{i,i+j} = \mathcal{E}_{i+j,i} = \begin{cases} 1, & \text{if } D_j \text{ can perform action } C_i. \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Afterward, we introduced a multi-layer graph convolutional network, taking the textual features of actions $\mathbf{t}$ and the weighted animal category features $\mathbf{c}$ as inputs.

The output of the k-th graph convolutional layer can be represented as follows.

$$f^{(k)} = \begin{cases} \mathbf{t} \parallel \mathbf{c}, & k = 0 \\ \mathrm{LN}(\sigma(\hat{\mathcal{E}} f^{(k-1)} W^{(k)})), & k > 0, \end{cases} \tag{8}$$

where LN represents layer normalization computation. $\hat{\mathcal{E}} = \mathcal{E} + \mathcal{I}$. $\mathcal{E}$ and $\mathcal{I}$ represent the adjacency matrix and the identity matrix, respectively. $W^{(k)}$ is the weight matrix of the k-th layer. σ is the ReLU function.

3.3.2 Category-Specific Prompting

Then, we propose category-specific prompting, a novel approach specifically designed for animal action recognition, to enrich the text space and guide the model's attention towards the animals themselves, facilitating better alignment between visual and textual features. In more detail, we expand the pretrained text encoder and video encoder by our prompting scheme to enhance both textual and visual representations automatically. The proposed category-specific prompting module takes the animal category feature $\mathbf{c}$ and text/video features $\mathbf{t}$/$\mathbf{v}$ as input. Taking the process of generating video prompts as an example. We first model the dependency of the two kinds of features using a multi-head attention (Vaswani et al., 2017). The output matrix of the attention mechanism is calculated as:

$$\mathrm{Atention}\,(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V \tag{9}$$

, where d_k represents the dimension of the key.

Multi-head attention is a highly effective mechanism that allows a model to simultaneously attend to information from diverse representation subspaces and positions. This technique involves conducting independent linear transformations of the queries, keys, and values, which are then utilized in parallel attention computations. Finally, the resulting output values are concatenated and projected to produce the final values. Here, we use the animal category feature as query Q, and use video feature as key K and value V. The

multi-head attention is computed as follows.

$$\text{MultiHead}\,(\mathbf{c}, \mathbf{v}) = \text{Concat}\,(\text{head}_1, \ldots, \text{head}_h)\, W^0, \quad (10)$$

$$\text{head}_i = \text{Atention}\left(\mathbf{c}W_i^Q, \mathbf{v}W_i^K, \mathbf{v}W_i^V\right), \quad (11)$$

where the weight matrices W_i^Q, W_i^K, and W_i^V serve to realize projections, while the weight matrix W^0 is employed in the linear output function. Then,

$$\mathbf{v}' = \mathbf{v} + \text{MultiHead}\,(\mathbf{c}, \mathbf{v}). \quad (12)$$

Following the computation of the multi-head attention mechanism, a feed-forward network is employed, which consists of two fully connected layers and one non-linear activation layer. Here is the specific calculation formula:

$$\text{FFN}(\mathbf{v}') = \text{ReLU}\left(W^1\mathbf{v}' + b^1\right)W^2 + b^2, \quad (13)$$

where W^1, W^2, b^1 and b^2 are learnable linear transformation matrices and biases. Afterward,

$$\tilde{\mathbf{v}} = \mathbf{v}' + \text{FFN}(\mathbf{v}'). \quad (14)$$

The final video representation enhanced by category-specific prompting can be represented as:

$$\hat{\mathbf{v}} = \mathbf{v} + \lambda\tilde{\mathbf{v}}, \quad (15)$$

where λ is a learnable parameter to balance the original feature and prompts.

3.3.3 Category Feature Extraction

Although most of the animal action datasets we surveyed include ground truth labels for animals (Liu et al., 2023; Ng et al., 2022; Chen et al., 2023; Khosla et al., 2011; Bourdev, 2012; Del Pero et al., 2015; Del Pero et al., 2017; Mathis et al., 2021), it is possible that some datasets may lack such labels, or that obtaining these labels in practical applications can be challenging (Miao et al., 2019; Van Horn et al., 2018). To broaden the applicability of our model, we propose a method for extracting semantic representations of animals from video segments, aiming to address the issue of the absence of direct animal labels. To elaborate on the details, we first use the CLIP model to obtain probabilities for all animal categories present in the video. Next, we use a pretrained text encoder to extract the text features of each categories. Finally, we fuse these features by weighting them according to their respective probabilities as the category feature of a video clip. More detail of the pipeline is as follows.

Animal Category Recognition To begin, we aim to identify the animal category present in a given video. To achieve this, we utilize the CLIP model, which possesses robust zero-shot recognition capabilities, to acquire animal categories. Specifically, we use the middle frame as the input to the image encoder of CLIP, as we believe it contains the most relevant information for animal recognition among all frames, and use this to obtain image features. Text embeddings are obtained from the frozen CLIP text encoder. We compute the cosine similarities between the image feature $\mathbf{i}$ and text embeddings $\mathbf{c^a}$ and normalize them using the softmax function, and obtain the result as the probabilities of all the animal categories as given by Eq. 16.

$$p\,(k|\mathbf{i}) = \frac{\exp(\text{sim}\,(\mathbf{i}, \mathbf{c^a}_k)/\tau)}{\sum_{n=1}^{N}\exp(\text{sim}\,(\mathbf{i}, \mathbf{c^a}_n)/\tau)}. \quad (16)$$

where N represents the number of animal categories in the lexicon, while τ represents temperature parameter.

Animal Category Encoder Next, the category names in the pre-defined category lexicon are fed into CLIP's pre-trained text encoder to generate text features, as described in Eq. 17. When processing a video clip, the category feature of it is generated by aggregating all text features, weighted by the probabilities of each category. Specifically, to generate category features for a given list of animal categories $A = \{a_1, a_2, ..., a_N\}$, we first tokenize them using CLIP's pre-defined tokenize function. The resulting tokens $T = \{t_1, t_2, ..., t_N\} \in \mathbb{R}^{N\times M}$ are then fed into the pre-trained text encoder $e_t(\cdot|\epsilon)$, which outputs a tensor $C = \{c_1, c_2, ..., c_N\} \in \mathbb{R}^{N\times M\times F}$, where M and F represent the number of category tokens and features, respectively.

$$\mathbf{c_i} = e_t(c_i|\epsilon). \quad (17)$$

The final output category feature $\mathbf{c}$ of the animal category encoder will be calculated as follows,

$$\mathbf{c} = \sum_{t=1}^{N}\mathbf{c_i}\cdot p_i. \quad (18)$$

4 Experiments

In this section, we firstly introduce our experimental setup. Next, we conduct extensive experiments from the perspectives of effectiveness and generalization to validate our method. Additionally, we explore the effectiveness of each component through a series of ablation studies. Finally, we demonstrate the model's performance from a visualization perspective.

4.1 Experimental Setup

4.1.1 Datasets

Compared to previous studies, our research specifically targets the recognition of animal actions within natural environments. To achieve this, we validate our method using three datasets of videos, which feature various species and actions recorded in their natural surroundings.

Animal Kingdom (Ng et al., 2022) This is a diverse dataset for animal action understanding that includes multiple animals and actions. It includes 50 h of video clips featuring 6 major animal categories, including mammals, insects, birds, sea animals, and so on, with more than 850 kind of animals that correspond to 140 different action classes selected from a list used by ethologists. The actions cover a wide range, from life events like molting, to daily activities like feeding. Each video clip includes fine-grained, multi-labeled actions. In addition, the dataset encompasses a diverse range of recording environments, such as grasslands, forests, rivers, and oceans, as well as various weather conditions, including sunny, rainy, and snowy weather. Our experiments consist of two types of settings to evaluate the efficiency and generalization of the proposed method. *Setting 1* we randomly divided the entire dataset into training(24004 videos) and testing(6096 videos) sets in a 4:1 ratio. *Setting 2* we conducted action recognition on unseen animals. Specifically, we selected a subset of 15 common mammals with 14 common action categories, which were then divided into a training set and a testing set. The training set included 10 common mammals, while the testing set comprised the remaining 5 common mammals, with all actions included in the training set.

MammalNet (Chen et al., 2023) This is a large-scale dataset focusing on advanced behaviors of mammals, comprising 18k videos with a total duration of 539 h. It encompasses 12 behaviors exhibited by 173 mammal species. Unlike typical actions, MammalNet emphasizes higher-level behaviors such as *hunting* and *nursing or breastfeeding its baby*. Here, *behavior* denotes the main activity within a period, encompassing multiple atomic actions, and describing video activity bouts in this way is essential for ecological research. Following the partitioning of the original paper, we randomly divided the examples from each category into two sets: 80%(14,553 videos) for training and 20%(3,840 videos) for testing.

LoTE-Animal (Liu et al., 2023) This dataset focuses on the actions of endangered species, comprising a total of 10k video clips involving 21 actions of 11 endangered species. Unlike the previous two datasets, videos in this dataset are collected from real wild environments, featuring lower clarity and more realistic camera noise. We randomly partitioned the dataset into 70% for training (6,994 videos) and 30% for testing (2,997 videos).

Table 3 Category frequency distribution across different segments of the long-tail distribution

Datasets	Animal Kingdom	MammalNet	LoTE-Animal
Head	(500, +∞)	(2000, +∞)	(1000, +∞)
Middle	(100, 500]	(1100, 2000]	(140, 1000]
Tail	(0, 100]	(0, 1100]	(0, 140]

To verify the effectiveness of the model across different segments of the long-tail distribution, we partitioned the data from the three datasets based on the frequency of category occurrences. The specific partitioning rules are shown in Table 3.

4.1.2 Compared Methods

We evaluate our proposed method by comparing it to seven state-of-the-art action recognition techniques. *Traditional methods* Approaches that use convolutional neural networks (CNNs) and transformers, such as I3D, X3D, and MViT. *Vision-language pretrained model-based methods* These methods are similar to ours in that they also leverage CLIP's pretrained model as a foundation, including VideoPrompt, EVL, Text4Vis, and the baseline model. *Baseline model* The visual encoding component utilizes a frozen CLIP visual encoder, augmented with the fine-tuned two transformer blocks for temporal modeling as detailed in Sect. 3.1. For the textual encoding component, a frozen CLIP text encoder is employed. The features extracted from both components are subsequently processed with a softmax function to compute prediction probabilities, resulting in the output of the baseline model. As the Animal Kingdom dataset is multi-labeled, we retrained the above methods to accommodate multi-label learning. We presented the experimental results of two settings: *Ours*, in which animal category features were generated directly using ground truth labels, represents the potential of our model. $Ours^{\dagger}$, in which animal category features were extracted from videos using a category feature extraction module. It is worth noting that the comparison methods we used all ignore the introduction of animal ground truth labels. The setup of $Ours^{\dagger}$ also facilitates a clear and fair comparison with other methods that do not incorporate animal ground truth labels.

4.1.3 Implementation Details

Our model is implemented in PyTorch and trained on a single NVIDIA RTX 3090. For training, we use the Adam optimizer with an initial learning rate of 8e-6. We set the batch size to 16 and accumulation steps to 8, training the model for 30

Table 4 Performance comparison with the state-of-the-art action recognition models on the Animal Kingdom dataset

Methods	6 major animal classes						Imbalance			Overall
	Amphibian	Bird	Sea animal	Insect	Mammal	Reptile	Head	Medium	Tail	
I3D Carreira and Zisserman (2017)	15.22	48.12	43.34	16.91	30.32	25.23	45.12	40.61	25.47	32.07
X3D Feichtenhofer (2020)	23.78	53.31	45.44	26.30	38.46	31.01	51.29	46.25	33.46	39.21
MViT Fan et al. (2021)	19.25	50.19	40.61	26.51	36.78	28.08	50.96	45.20	32.67	38.42
VideoPrompt Ju et al. (2022)	31.00	50.13	40.57	33.65	41.07	47.68	58.27	52.49	40.12	45.81
EVL Lin et al. (2022)	29.75	52.07	55.00	40.57	48.20	50.36	60.34	56.21	44.35	49.61
Text4Vis Wenhao et al. (2024)	33.91	**56.70**	48.04	37.18	47.44	47.52	60.04	56.22	44.99	49.95
Baseline	28.53	52.31	49.57	39.52	48.59	51.33	60.11	55.21	44.51	49.42
Ours	**37.85**	54.21	**61.87**	**50.42**	**55.72**	**62.70**	**61.89**	**59.17**	**56.59**	**57.99**
Ours†	32.01	53.01	53.95	38.26	52.16	54.34	60.44	56.67	45.44	50.40

epochs. We utilize CLIP ViT-B/16 as the backbone architecture. In the Ours† model, we employ the Bio-CLIP text and visual encoders to extract animal category features. Videos are cropped to 224x224 size, with 8 frames extracted per video through uniform sampling. During testing, to alleviate memory burden, we validate two videos at a time.

4.1.4 Evaluation Metrics

Regarding the multi-label Animal Kingdom dataset, in line with other studies on multi-label learning (Wang et al., 2016, 2017; Huynh et al., 2020), we adopted the mean Average Precision (mAP) (Veit et al., 2017) as our primary evaluation metrics.

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}(R(k) - R(k-1)) \cdot P(k) \tag{19}$$

where $R(k)$ and $P(k)$ are the Recall and Precision values at the k-th sample, respectively. N is the total number of classes, and n is the total number of samples. For the single-label MammalNet and LoTE-Animal datasets, we employ top-k accuracy as the evaluation metric.

To assess the overall performance of our method on both seen and unseen animals, we also introduced the Harmonic Mean (Hm) as an evaluation metric.

$$\text{Hm} = \frac{2 \cdot \text{mAP}_\text{S} \cdot \text{mAP}_\text{U}}{\text{mAP}_\text{S} + \text{mAP}_\text{U}} \tag{20}$$

4.2 Results

4.2.1 The Effectiveness of the Method

To evaluate the effectiveness of our method, we conducted experiments under *Setting 1*. Table 4 presents the performance of our proposed Animal-CLIP model with state-of-the-art traditional and vision-language pretrained methods on the six major animal categories in the Animal Kingdom, as well as the overall performance. Our model achieves a new state-of-the-art performance in terms of all evaluation metrics. Notably, when provided with ground truth animal labels, our proposed model outperforms the best vision-language pretraining method, Text4Vis, by 8.04% in mAP metric, respectively. Furthermore, it demonstrates a significant improvement over the best traditional method, X3D, in the mAP metric by 18.78%. Additionally, our method outperforms the baseline by 8.57%, indicating that our proposed prompt fine-tuning approach effectively leverages the potential of vision-language models and demonstrates strong adaptability in animal action recognition. We also tested the model's ability to handle long-tail distribution data, and our method achieves best results across various segments, outperforming the second-best by 12.08% in the tail categories. This suggests that our approach effectively addresses the challenge of insufficient data in tail categories, offering advantages for practical applications. Furthermore, we also present the performance of the method that utilizes animal category features extracted by the category feature extractor, rather than relying on ground truth animal labels. As demonstrated, our model still outperforms the existing methods.

Figure 5 illustrates a comparison between our model and other models across six major animal categories and various segments of the long-tail distribution. It is evident that our approach performs well across all animal categories and segments. Our model demonstrates outstanding performance in each animal category, highlighting its comprehensiveness and generalization capability. Moreover, our method exceeds the baseline model by 12.08% in the tail categories, indicating its effectiveness in addressing the issue of insufficient data for these categories. These results all underscore the advantages of our approach in practical applications.

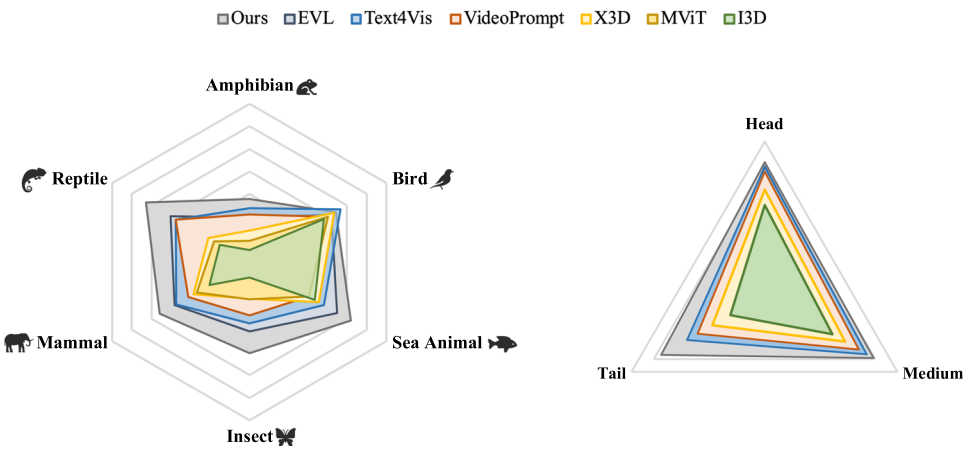


Fig. 5 Radar map of action recognition accuracy for six major categories of animals and radar map of recognition accuracy for different segments of the long-tail distribution

Table 5 Performance comparison with the state-of-the-art action recognition models on the MammalNet and LoTE-Animal datasets

Methods	MammalNet				LoTE-Animal			
	Head	Medium	Tail	Overall	Head	Medium	Tail	Overall
I3D	70.31	54.10	44.26	57.06	84.00	83.76	47.56	81.94
X3D	77.35	60.47	47.48	62.68	86.01	84.37	47.56	83.58
MViT	72.53	56.58	46.89	59.49	75.11	74.75	25.77	72.33
VideoPrompt	77.98	57.62	47.73	62.11	71.67	45.56	11.59	61.52
EVL	77.35	60.47	47.48	61.09	72.47	67.34	25.00	68.52
Text4Vis	79.23	72.61	62.10	71.77	77.51	60.15	20.73	69.84
Baseline	81.18	73.70	63.90	73.15	81.41	87.44	46.95	81.11
Ours	83.90	75.46	65.73	**75.55**	86.20	85.66	43.90	**83.74**
Ours†	82.23	75.88	62.10	73.91	82.14	88.58	41.46	81.61

Table 5 presents the results of our method on the MammalNet and LoTE-Animal datasets, showing that our approach achieves the best overall accuracy metrics. On MammalNet, Animal-CLIP surpasses the second-best baseline method by 2.40% and achieves the best performance across all segments of the long-tail distribution. On LoTE-Animal, vision-language model-based methods generally perform poorly, likely due to the limited presence of endangered species in the training data of large models. Nevertheless, Animal-CLIP still performs well, exceeding Text4Vis by 13.9%. We also report the accuracy using our animal feature extraction module to obtain animal features, rather than directly using ground truth, to simulate scenarios where animal labels may be missing in practical applications. It is evident that our method still outperforms previous methods on MammalNet and also surpasses the previous vision-language pretraining model-based methods on LoTE-Animal, demonstrating strong competitiveness compared to traditional methods. In practical applications, the utilization of more advanced animal recognition models has the potential to further enhance performance.

Table 6 Generalization ability comparison on unseen animals with the state-of-the-art action recognition models on the Animal Kingdom dataset, where CSP represents Internal Category-Specific Prompting, and EP represents External Prompting

Methods	mAP_S	mAP_U	Hm
I3D	43.47	18.17	25.63
MViT	38.31	17.01	23.56
X3D	46.80	24.98	32.57
VideoPrompt	60.73	18.15	27.95
EVL	58.61	20.14	29.98
Text4Vis	42.76	30.47	35.58
Baseline	59.85	27.53	37.71
CSP	63.42	29.34	40.12
CSP+EP	61.12	31.00	41.14
Ours	60.67	34.74	**44.18**
Ours†	61.33	33.41	43.26

4.2.2 The Generalization Ability of the Method

To test the generalization ability of our model, we conducted experiments using *Setting 2*. As shown in Table 6, the results indicate that our model outperforms the other six methods

Table 7 Ablation study on different parts of our model

Methods			Animal Kingdom				MammalNet			
CSP	EP	TFRM	Head	Middle	Tail	Overall	Head	Middle	Tail	Overall
			60.11	55.21	44.51	49.42	81.18	73.70	63.09	73.15
✓			61.69	59.16	52.17	55.28	83.34	73.70	65.40	74.69
✓		✓	62.43	59.83	52.82	55.95	83.90	75.63	63.91	75.03
✓	✓		60.70	58.44	56.00	57.28	83.55	75.46	64.99	75.18
✓	✓	✓	61.89	59.17	56.59	**57.99**	83.90	75.46	65.73	**75.55**

Table 8 Ablation study on category-specific prompts in both the text and video branches

Methods		Animal Kingdom				MammalNet
		Setting 1	Setting 2			
text prompt	video prompt	mAP	mAP_S	mAP_U	Hm	mAP
		49.42	63.15	24.85	35.67	73.15
✓		55.16	61.03	28.13	38.51	74.45
	✓	53.42	61.09	28.52	38.89	74.40
✓	✓	**55.28**	**63.42**	**29.34**	**40.12**	**74.69**

on seen animal categories. However, there is a decline in performance with unseen animals. Nevertheless, even for unseen animals, our model still performs exceptionally well and remains the best compared to the other six methods, demonstrating its strong generalization capability. Furthermore, our approach shows significant improvement on the comprehensive evaluation metric Hm. The introduction of internal category-specific prompting exceeds the second-best score by 2.41%, while the external prompts further enhance performance by 1.02%. After the introduction of TFRM, our overall model's Hm metric surpasses the baseline model by 6.47%. This suggests that our model is particularly well-suited to address the challenges posed by the rich diversity of animals in natural environments, where data collection remains both challenging and constrained.

4.2.3 Ablation Study

In this part, we conduct comprehensive ablation studies to analyze the effectiveness of our proposed Animal-CLIP method. Specifically, we firstly perform a main ablation study to investigate the contribution of different components of the model. Next, we examine the impact of various prompting methods and different lexicons. We then test the effect of the number of multi-head attention layers in category-specific prompting and the influence of graph convolutional layers on model performance.

Ablation Study of Different Components As shown in Table 7, we tested the contribution of External Prompting (EP), Internal Category-Specific Prompting (CSP), and the Text Feature Refinement Module (TFRM) to the results on the Animal Kingdom and MammalNet datasets. It can be seen that adding CSP to the baseline provides a 5.86% and 1.54% improvement, respectively. Further adding EP results in an additional 2.00% and 0.49% gain, while adding TFRM contributes a 0.67% and 0.34% improvement. This indicates that each component of our model is effective, and their combined use leads to further performance enhancement. Ultimately, our model achieves improvements of 8.57% in mAP and 2.40% in accuracy compared to the baseline.

Table 9 The effect of different lexicons

Lexicon	Baseline ours
6 major parent classes	49.42 $\xrightarrow{+2.39\%}$ 51.81
897 species classes	49.42 $\xrightarrow{+5.86\%}$ 55.28

Ablation Study on Prompting Methods We first conducted ablation experiments to explore the contribution of adding Category-Specific Prompts in the visual and text branches. Next, we compared the hand-crafted prompting and instance prompting mentioned in 2.2, and further validated that our method, explicitly tailored for the animal action recognition task, achieved notable improvements. Specifically, we present the results of our ablation experiments on prompting methods in Table 8. On the Animal Kingdom dataset (Setting 1) and the MammalNet dataset, text prompts resulted in gains of 5.74% and 1.3%, while video prompts contributed gains of 4.00% and 1.25%, respectively. Moreover, when combining text and video prompts, our model's performance saw further improvement, achieving gains of 5.86% and 1.54% over the baseline. On the Animal Kingdom dataset (Setting 2), text prompts led to an overall gain of 2.84%, video prompts resulted in a gain of 3.22%, and the combination of both achieved a gain of 4.45%. These results highlight the impor-

Table 10 Performance comparison with the hand-crafted prompt on the Animal Kingdom dataset

Methods	Head	Middle	Tail	Overall
Hand-craft	60.13	55.27	40.45	46.99
Ours	61.89	59.17	56.59	**57.99**

Table 11 Generalization ability comparison on unseen animals with the instance-level prompt on the Animal Kingdom dataset

Methods	mAP_S	mAP_U	Hm
Instance-level	**64.87**	25.47	36.58
Ours	63.62	**31.67**	**42.29**

tance of using both text and video prompts in our proposed method to achieve superior performance.

We also explored the impact of different prompting methods on model performance. As shown in Table 10, we first used unified hand-crafted prompts. We designed five prompts, including: "a video of action {}", "Animal action of {}", "Playing action of {}", "Doing a kind of action, {}", and "Video classification of {}". During training, we randomly selected one of these prompts, and during testing, we tested all five prompts and averaged the results to obtain the overall performance. It can be observed that compared to our method, the performance of the hand-crafted prompts decreased by 11.00%. This further confirms that unified prompts cannot generate different text and video representations for different animals and scenarios, making them unsuitable for animal action recognition tasks. Secondly, as shown in Table 11, we tested the performance of instance-level prompts on unseen animals. Instance-level prompts, while achieving higher accuracy by tailoring prompts to individual videos and capturing fine-grained features, suffer from limited generalization ability due to their reliance on specific input characteristics and lack of adaptability to unseen categories. The category-specific prompts outperformed instance-level prompts by 6.20% in mAP on unseen animals, demonstrating the stronger generalization ability of our proposed method. Our approach also showed a 5.71% improvement in the Hm metric.

*Exploration of Different Lexicons*We explored the impact of obtaining animal category features from different lexicons. Specifically, we selected 6 major parent animal category names and 897 more fine-grained species names as different lexicons. As shown in Table 9, compared to the baseline method that did not incorporate animal category information, our approach improved the mAP metric by 2.39%. Additionally, our results demonstrate that refining the animal categories in the lexicon led to further enhancements in experimental performance, resulting in a 5.86% improvement. This indicates that using more fine-grained species names allows for a more detailed division of the text subspace, leading to further enhancement in model performance.

The Impact of the Number of Layers of Multi-head Attention As shown in Table 12, we investigated the impact of varying numbers of multi-head attention layers on the experimental results for category-specific prompting. Our findings indicate that as the number of multi-head attention layers increases, the model performance consistently improves. Specifically, for the Animal Kingdom and LoTE-Animal datasets, we observed that the model achieved the best performance with three layers of attention, and performance began to decline when the number of layers was further increased. However, for the MammalNet dataset, the model performance continued to improve with additional layers. The reason why increasing the number of layers does not further improve accuracy is twofold: on one hand, the model has already sufficiently extracted the features from the video data after a certain number of layers, and additional layers do not contribute more effective information. On the other hand, an excessive number of layers may lead to overfitting, particularly when the training data is limited, causing the model to overfit specific scenes or patterns, thereby restricting its generalization ability and preventing further accuracy improvements. The reason why the highest accuracy for the MammalNet dataset corresponds to a higher number of layers than the other two datasets is likely due to the fact that MammalNet involves more complex, high-level behaviors and consists of longer videos, which contain more intricate action patterns or long-term dependencies. These require more layers of multi-head attention to capture the information effectively. In contrast, the actions in the other two datasets are relatively simpler or involve shorter dependencies, resulting in diminishing returns from increasing the number of layers.

The Impact of the Number of Layers of Graph Convolution We also conducted ablation experiments on the number of graph convolutional layers. As shown in Fig. 13, we found that for the Animal Kingdom and LoTE-Animal datasets, the model performed best with 2 layers. Increasing the number of layers initially led to a reduction in accuracy. However, for the MammalNet dataset, performance improved with the addition of layers, reaching optimal results with three layers. Further increases in the number of layers resulted in a decline in performance.

4.2.4 Visualization Results

In this part, we will present our visualization results. First, we will show the generated external prompts to illustrate the extension of label semantics. Next, we will display the T-SNE visualization of the feature space. Then, we will present the attention maps in our visual feature space. Finally, we will showcase the recognition results of our method.

	Action
Drinks Water	*When animals drink water, they typically **lean down to reach the water source**, often **using their mouths to lap up the water**. This action may vary depending on the species, but common characteristics include **the bending of the neck downwards, tongue extension, and sometimes the submerging of the mouth into the water**.*
	*When identifying animals drinking water, observe for the following key behaviors: 1. Approach: Animals generally **approach the water source** cautiously, especially if they are in a vulnerable position. 2. Posture: When drinking, many animals **lower their heads towards the water source while keeping their bodies relatively still**.*
	*Animals drinking water typically display a characteristic posture involving **lowering their heads towards the water source, often with a dipping or lapping motion**. They may **use their tongues or mouths to scoop up water, sometimes creating ripples or splashes in the process**. Common visual cues include **the bending of the neck, rhythmic movements, and the occasional lifting of the head to swallow**.*
Sleep	*During sleep, animals may adopt various positions depending on their species and habitat. They typically **close their eyes, relax their muscles, and may change positions throughout the night**. Some animals may **curl up, stretch out, or nestle in a safe and comfortable spot to rest**. **Their breathing may slow down**, and their body temperature may drop slightly.*
	*You can identify animals in various sleep positions by observing their body posture and behaviors. For example, animals that **curl up into a ball, with their head tucked under their body**, are likely in a deep sleep. Similarly, animals that are **lying down with their eyes closed and breathing slowly** are also likely asleep. Some animals may also **exhibit twitching or movement** during sleep,*
	*When animals are sleeping, they usually **exhibit relaxed body postures, closed eyes, and slowed or rhythmic breathing**. Many animals may also **curl up or tuck their head under their wing or leg** while sleeping. Some animals, like cats and dogs, may **twitch or move their paws as if they are dreaming** during sleep. Additionally, some animals may have specific sleep behaviors.*
Fights against Other Animals	*When identifying animals engaged in fights against other animals, you can look for the following signs: 1. Aggressive behavior: **Watch for aggressive postures** such as raised **hackles, bared teeth, growling, or hissing**. 2. Physical confrontation: **Look for physical interactions** such as **biting, scratching, or wrestling between the animals**.*
	*Animal fights can vary greatly depending on the species involved. Common behaviors in animal fights include displaying of intimidation through vocalizations, postures, or displays of size or strength, aggressive physical contact such as **biting, kicking, or swatting, and attempts to establish dominance or defend territory or resources**. Fights may end with one animal retreating or submitting to the other.*
	*Animal fights can involve physical aggression such as **biting, clawing, or kicking, as well as vocalizations and displays of dominance**. These battles often aim to establish territory, dominance hierarchy, or access to mates, and can be intense and sometimes violent. The combatants may use their natural weapons, such as **horns, antlers, or tusks,** to overpower their opponent.*
	Animal
Panthera	*Panthera is a genus of big cats that includes tigers, lions, leopards, and jaguars. They are characterized by their **powerful build, muscular bodies, sharp retractable claws, and typically have sleek fur with distinctive spots or stripes**. They are apex predators known for their stealth, agility, and hunting skills.*
	*Here are some key characteristics to help identify members of the genus Panthera concisely: 1. They have **a robust and muscular build**. 2. They have **a large head with powerful jaws and prominent canine teeth**. 3. They have **a distinctive roset**…*
	*Panthera is a genus of large cats that includes species such as lions, tigers, leopards, and jaguars. To identify a member of the Panthera genus concisely, look for the following key characteristics: **1. Powerful build with a compact and muscular body. 2. Sharp retractable claws. 3. Cat-like pupils (oval-shaped).***
Vulpes	*Vulpes, commonly known as foxes, are **medium-sized** carnivorous mammals with a distinctive pointed snout, erect triangular ears, and bushy tail. They typically have **a reddish-orange coat**, although coloration can vary among species. Foxes have **a slender build and agile movements**, with keen eyesight and excellent hearing.*
	*Vulpes is a genus that includes various species of foxes. To identify a fox belonging to the Vulpes genus, you can look for the following key characteristics: **1. Pointed muzzle: Foxes typically have a long, pointed muzzle compared to other canids. 2. Large, bushy tail: Foxes have a distinctive long and bushy tail that helps them**...*
	*Vulpes is the genus that includes foxes, which are **small to medium-sized canids**. Foxes have several identifying characteristics, including: **1. Bushy tail: Foxes have a long, bushy tail that often has a white tip. 2. Pointed ears: Foxes have large, pointed ears that help them to hear well.***
Gulo	*The gulo, commonly known as a wolverine, can be identified by **its stocky build, dark fur with light-colored markings on its face, sharp claws, and bushy tail**. Wolverines are known for their aggressive and tenacious behavior, as well as their ability to thrive in **snowy and rugged environments**.*
	*Gulo is the scientific genus name for wolverines, which are large carnivorous mammals known for their strength and aggressive behavior. Wolverines have several identifying characteristics, including: **1. Stocky build: Wolverines have a robust body with strong legs, well-suited for their rugged habitat. 2. Dense fur: Their fur is thick, oily, and often dark brown in color**...*
	*Gulo is the scientific genus for wolverines, which are the largest members of the weasel family. Wolverines have several identifying characteristics, including: **1. Size: Wolverines are stocky and muscular animals, typically measuring about 65 to 107 cm (26 to 42 inches) in length, not including the bushy tail which is about 17 to 26**...*

Fig. 6 External prompts generated by the LLM are displayed, with important visual cues highlighted in **bold**

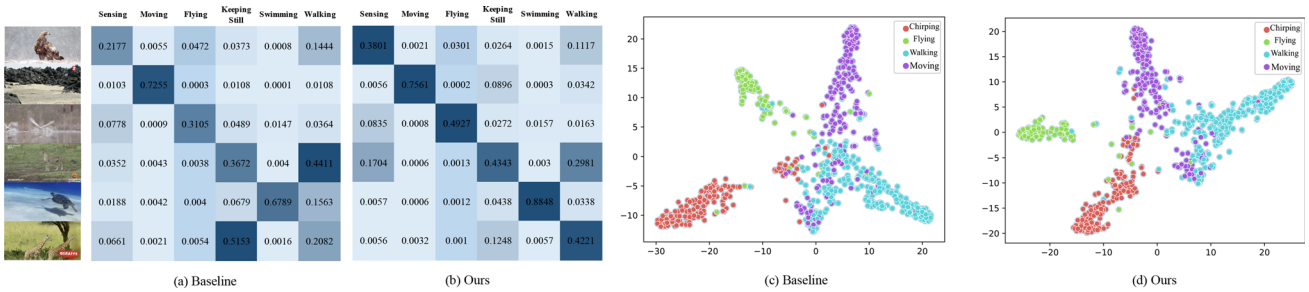


Fig. 7 The final similarity scores from both the baseline model and ours, as well as the T-SNE visualization results on PACS

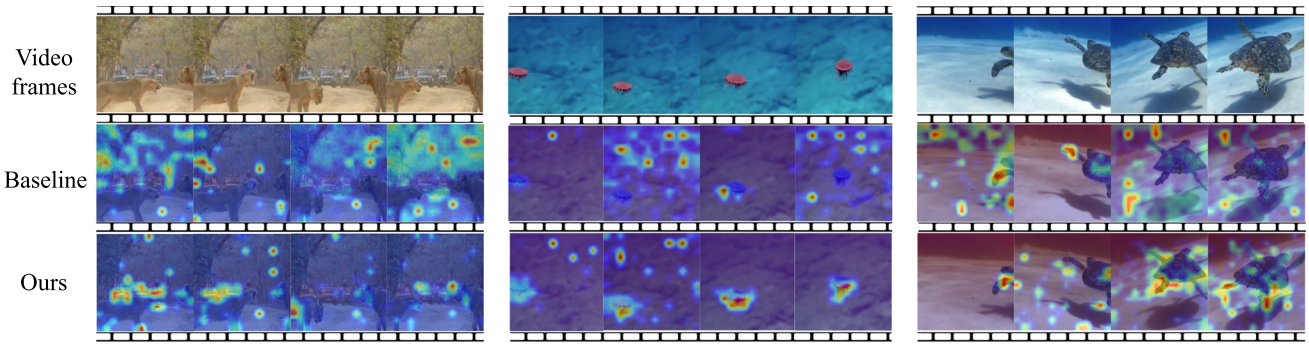


Fig. 8 Visual attention maps. Our method focuses more on the animals themselves, demonstrating better robustness

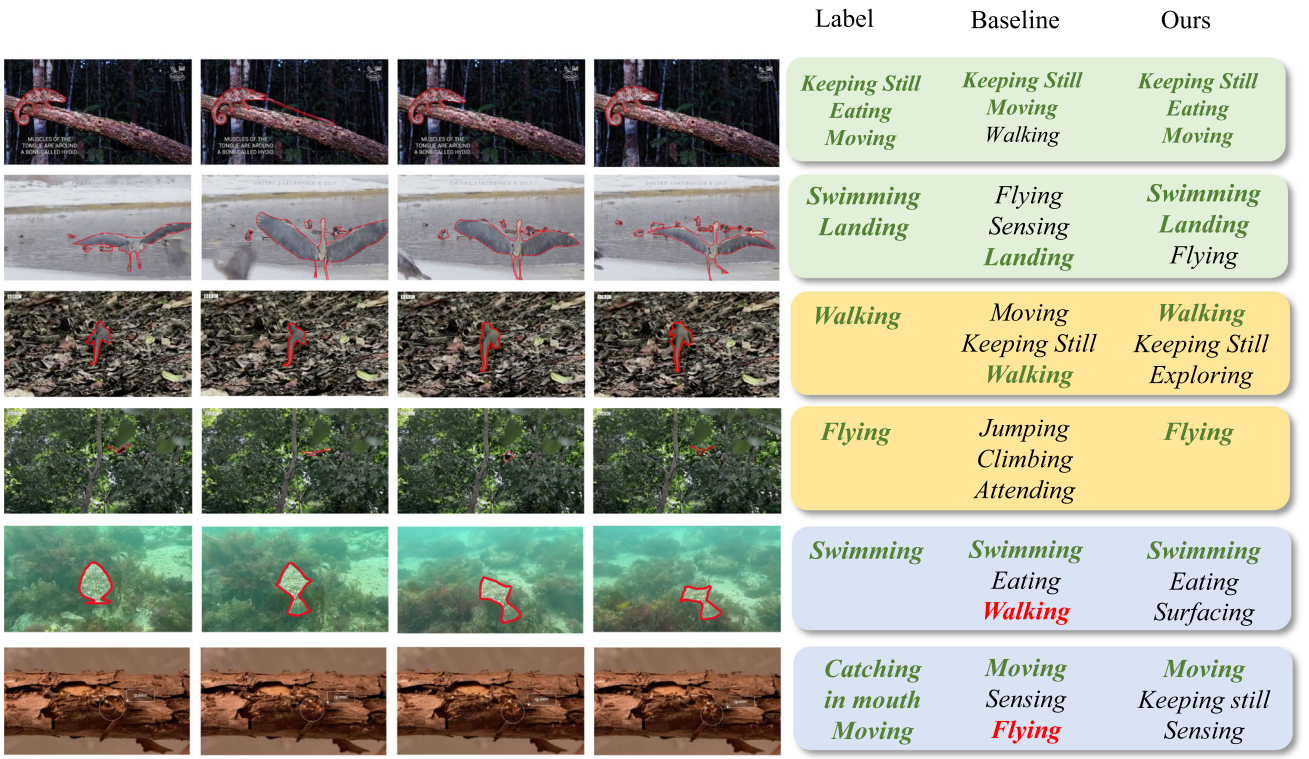


Fig. 9 Some prediction results of the baseline and our method (the animals are highlighted with red edge lines for ease of observation)

Table 12 Effects of the number of layers of multi-head attention in category-specific prompting

Layers	Animal Kingdom				MammalNet				LoTE-Animal			
	Head	Middle	Tail	Overall	Head	Middle	Tail	Overall	Head	Middle	Tail	Overall
1	61.39	56.96	47.43	51.81	83.07	74.96	65.57	75.03	85.77	86.28	51.83	84.05
2	61.28	57.64	48.42	52.57	82.58	74.79	64.66	74.51	85.33	86.66	49.39	83.71
3	61.69	59.16	52.17	**55.28**	83.34	73.70	65.40	74.69	86.11	87.94	44.51	**84.31**
4	61.10	58.05	51.23	54.37	84.39	75.13	64.74	**75.31**	85.72	86.40	53.66	84.15

Table 13 Effects of the number of layers of graph convolution

Layers	Animal Kingdom				MammalNet				LoTE-Animal			
	Head	Middle	Tail	Overall	Head	Middle	Tail	Overall	Head	Middle	Tail	Overall
2	60.45	59.10	56.55	**57.75**	84.25	74.62	64.99	75.18	86.20	85.66	43.90	**83.74**
3	58.99	57.06	54.86	56.00	83.90	75.46	65.73	**75.55**	72.11	85.83	37.80	73.77
4	58.46	53.36	50.88	52.58	81.95	75.54	64.66	74.51	54.75	87.31	32.93	62.12

External Prompts To provide a detailed view of our external prompts, we present examples generated by large language models in Fig. 6. We adopted the LLM prompts introduced in Sect. 3.2 to generate 50 responses for each category, with varying contents. For both animals and actions, we selected three categories, each with three examples. It can be observed that the external prompts include several visual cues (highlighted in bold), which provide the model with substantial domain knowledge and expand the originally limited label semantics.

T-SNE Visualization We present the final similarity scores of the video and text modalities, as shown in Fig. 7a and b. The diagonal represents the similarity scores with ground truth labels. It can be observed that our method demonstrates improved similarity scores between videos and ground truth labels, while exhibiting reduced similarity scores with non-ground truth labels. To further explore the effectiveness of our proposed method, we conducted T-SNE visualization of the similarity scores for category-specific prompts and the baseline model. We selected four common actions in the Animal Kingdom dataset: "Chirping," "Flying," "Walking," and "Moving." As shown in Fig. 7 c and d, our method outperforms the baseline in clustering, particularly for action labels with significant intra-class variation, such as "Moving" and "Walking." The similarity scores computed by our method show better separation, with clusters being more distinct from each other. This further validates the efficacy of our method, which integrates animal category information to mitigate intra-class variation for identical actions.

Attention Maps We provide attention map visualizations for video frames to elucidate the effectiveness of our approach. These maps reveal that, compared to the baseline model, our method enables the model to concentrate more on the animal itself rather than on the surrounding environment, thereby reducing the impact of dynamic environmental elements. For instance, as shown in the second column, the baseline model focuses primarily on the sea water, while our model is more attuned to the jellyfish itself. This difference highlights the superior performance of our method.

Table 14 The comparison of noise index(%) between our method and the baseline on different datasets

Methods	Animal Kingdom	MammalNet	LoTE-Animal
Baseline	14.44	11.18	10.26
Ours	10.50	6.47	7.61

Recognition Results We present the recognition results of our proposed method and the baseline model, as shown in Fig. 9. The first two video clips contain dynamic elements, such as swaying tree branches and falling snow, and include occlusions between animals or between animals and the environment. However, even under these challenging conditions, our method achieved better prediction performance compared to the baseline model. The next two video clips feature animal instances with backgrounds of similar colors, making it difficult to distinguish the animals from the environment. Despite this, our method was still able to accurately recognize the animals' actions. These results demonstrate the effectiveness of our model in predicting animal actions across various species and environments.

Additionally, leveraging external knowledge, our model effectively addresses the issue of label noise. As shown in Fig. 9, we present the top three predicted actions output by the models. In the fifth row, a fish is depicted, with the baseline model predicting the action as *walking*. In the sixth row, an ant colony is shown, with the baseline model predicting the action as *Flying*. These predictions are inconsistent with reality, as these animals cannot perform the corresponding actions in nature. Our model is capable of correcting these "impossible actions" predictions, making the recognition more accurate. Furthermore, we quantitatively calculated the

noise in the predicted results. We defined a noise index as the proportion of actions within the top five predicted categories that are not feasible for the animal to perform. As shown in Table 14, we compared the noise index of our approach with that of the baseline, and the results demonstrate that our method effectively reduces model noise across all three datasets.

5 Limitations

Although our method yields promising results, there are still some limitations. First, the introduction of additional parameters increases the risk of overfitting, particularly when applied to small datasets. For instance, in the Setting 2 experiment, as the model components increased, the performance on the seen test set declined. Furthermore, the inclusion of predicted animal categories reduces the accuracy of action recognition. Additionally, the external prompts generated by the model are sometimes inaccurate, especially in the case of certain animals, where the prompts do not align with the true characteristics of the animals, thereby affecting the accuracy of action classification.

6 Future Work

In the future, our focus will first be on model pruning to eliminate redundant parameters and enhance performance on small datasets. Second, we aim to improve the effectiveness of animal recognition models, maximizing the potential of action recognition approaches. Additionally, we will investigate methods for enhancing the quality of external prompts and developing answer selection strategies to obtain more accurate responses. Ultimately, our goal is to integrate multiple datasets to increase data volume, thereby training a more robust visual-language model to advance the development of a more powerful animal action recognition system with broader practical applicability.

7 Conclusion

To address the challenges faced by existing visual-language pretraining models in animal action recognition tasks, such as significant intra-class variation and severe environmental interference due to task specificity, and the weak generalization ability as well as severe label noise due to the limitations of large models in domain-specific knowledge, we propose Animal-CLIP, a dual-prompt enhanced visual-language pretraining model. To address the lack of domain-specific knowledge, we use large language models to generate external prompts, offering detailed descriptions of animal and action labels, along with their relationships. To tackle intra-class variation and reduce label noise, we introduce knowledge-enhanced internal prompt fine-tuning. We includes an internal category-specific prompting module that generates unique text and video prompts for each video. We also develop a text feature refinement module to reduce inconsistencies and label noise. Our extensive experiments on three multi-species, multi-action datasets show that our method outperforms six prior approaches, demonstrating strong generalization capabilities.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (NSFC) No. 62476029, 62225601, and U23B2052, sponsored by Beijing Nova Program, supported in part by the BUPT Excellent Ph.D. Students Foundation No. CX20241086, and in part by scholarships from China Scholarship Council (CSC) under Grant CSC No. 202406470082.

Data Availability All datasets used in this study are open-access and have been cited in the paper. The code of this work will be available in the Github repository, https://github.com/PRIS-CV/Animal-CLIP.

Declarations

Conflict of interest The authors declared that they have no Conflict of interest to this work.

References

Anderson, D. J., & Perona, P. (2014). Toward a science of computational ethology. *Neuron, 84*(1), 18–31.

Bahng, H., Jahanian, A., Sankaranarayanan, S., & Isola, P. (2022). *Exploring visual prompts for adapting large-scale models, 1*(3), 4. arXiv:2203.17274.

Bourdev, L. (2012) Dataset of keypoints and foreground annotations for all categories of pascal 2011

Carreira, J., & Zisserman, A. (2017) 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).

Chatgpt, 2022.

Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., & Xu, B. (2023) X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160

Chen, J., Hu, M., Coker, D. J., Berumen, M. L., Costelloe, B., Beery, S., Rohrbach, A., & Elhoseiny, M. (2023) 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)

Del Pero, L., Ricco, S., Sukthankar, R., & Ferrari, V. (2017). Behavior discovery and alignment of articulated object classes from unstructured video. *International Journal of Computer Vision, 121*, 303–325.

Feichtenhofer, C., (2020) X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 203–213)

Feichtenhofer, C., Fan, H., Malik, J., & He, K.(2019) Slowfast networks for video recognition. In *Proceedings of the IEEE Int'l Conference on Computer Vision* (pp. 6202–6211)

Feng, L., Zhao, Y., Sun, Y., Zhao, W., & Tang, J. (2021). Action recognition using a spatial-temporal network for wild felines. *Animals, 11*(2), 485.

Geuther, B. Q., Peer, A., He, H., Sabnis, G., Philip, V. M., & Kumar, V. (2021). Action detection using a neural network elucidates the genetics of mouse grooming behavior. *Elife, 10*, Article e63207.

Gpt4, 2023.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., & Couzin, I. D. (2019). Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife, 8*, Article e47994.

Gu, X., Chen, G., Wang, Y., Zhang, L., Luo, T., & Wen, L. (2023) Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18941–18951)

Huang, X., Huang, Y. J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., & Zhang, L. (2023) Open-set image tagging with multi-grained text supervision. arXiv e-prints, pages arXiv–2310

Hudson, D. A., & Manning, C. D. (2019) Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 6700–6709

Huynh, D., & Elhamifar, E. (2020) A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* pp. 8776–8786

Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022) Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conf., Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII* pp. 709–727. Springer

Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Quoc Le, Sung, Y-H., Zhen Li, & Duerig, T. (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In *Int'l Conference on Machine Learning* (pp. 4904–4916)

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics, 8*, 423–438.

Ju, C., Han, T., Zheng, K., Zhang, Y., & Xie, W. (2022) Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conf., Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV* pp. 105–124. Springer

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., & Xing, E. P. (2019) Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11487–11496)

Karashchuk, P., Rupp, K. L., Dickinson, E. S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B. W., & Tuthill, J. C. (2021). Anipose: a toolkit for robust markerless 3d pose estimation. *Cell Reports, 36*(13), Article 109730.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2022) Maple: Multi-modal prompt learning. arXiv preprint arXiv:2210.03117

Khosla, A., Jayadevaprakash, N., Yao, B., & Li, F. F. (2011) Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings CVPR workshop on fine-grained visual categorization (FGVC), volume 2*

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. (2023) Mvbench: A comprehensive multi-modal video understanding benchmark. arXiv preprint arXiv:2311.17005

Li, Y., Wu, C. Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2021) Multiscale vision transformers. In *Proceedings of the IEEE Int'l Conf. on Computer Vision* pp. 6824–6835

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., & Liu, Z. Otter: A multi-modal model with in-context instruction tuning.

Liang, K., Wang, X., Wei, T., Chen, W., Ma, Z., & Guo, J. (2023a) Attribute learning with knowledge enhanced partial annotations. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1715–1719. IEEE

Liang, K., Wang, X., Zhang, H., Ma, Z., Guo, J. (2023b) Hierarchical visual attribute learning in the wild. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3415–3423

Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., Dai, J., Qiao, Y., & Li, H. (2022) Frozen clip models are efficient video learners. In *Computer Vision–ECCV 2022: 17th European Conf., Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV* pages 388–404. Springer

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2023) Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122

Liu, D., Hou, J., Huang, S., Liu, J., He, Y., Zheng, B., & Zhang, J. (2023) 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)

Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., & Mathis, M. W. (2021) Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* pp. 1859–1868

Miao, Z., Gaynor, K. M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M. S., McInturff, A., Bowie, R. C. K., Nathan, R., Yu, S. X., et al. (2019). Insights and approaches using deep learning to classify wildlife. *Scientific Reports, 9*(1), 8137.

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019) Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE Int'l Conference on Computer Vision*, pp. 2630–2640

Mondal, A., Nag, S., Prada, J. M., Zhu, X., & Dutta, A. (2023) Actor-agnostic multi-label action recognition with multi-modal query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (pp. 784–794)

Naeem, M. F., Xian, Y., Tombari, F., & Akata, Z. (2021) Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 953–962)

Ng, X. L., Ong, K. E., Zheng, Q., Ni, Y., Yeo, S. Y., & Liu, J. (2022) Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 19023–19034).

Nguyen, H., Maclagan, S. J., Nguyen, T. D., Nguyen, T., Flemons, P., Andrews, K., & Phung, D. (2017) Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. In *2017 IEEE international conference on Data Science and Advanced Analytics (DSAA)* (pp. 40–49).

Nguyen, C., Wang, D., Von Richter, K., Valencia, P., Alvarenga, F. A., & Bishop–Hurley, G. (2021) Video-based cattle identification and action recognition. In *2021 Digital Image Computing: Techniques and Applications (DICTA)* (pp. 01–05).

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022) Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conf., Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV* (pp. 1–18) Springer

Pascoe, J., Ryan, N., & Morse, D. (2000). Using while moving: Hci issues in fieldwork environments. *ACM Transactions on Computer-Human Interaction (TOCHI), 7*(3), 417–437.

Pero, L. D., Ricco, S., Sukthankar, R., & Ferrari, V. (2015) Articulated motion discovery using pairs of trajectories. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2151–2160).

Pratt, S., Covert, I., Liu, R., & Farhadi, A. (2022) What does a platypus look like? generating customized prompts for zero-shot image classification

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021) Learning transferable visual models from natural language supervision. In *Int'l Conference on Machine Learning* (pp. 8748–8763)

Ravbar, P., Branson, K., & Simpson, J. H. (2019). An automatic behavior recognition system classifies animal behaviors using movements and their temporal context. *Journal of neuroscience methods, 326*, Article 108352.

Romanelli, C., Cooper, D., Campbell-Lendrum, D., Maiero, M., Karesh, W. B., Hunter, D., & Golden, C. D. (2015) Connecting global priorities: biodiversity and human health: a state of knowledge review. World Health Organistion/Secretariat of the UN Convention on Biological.

Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J. J., Perona, P., Anderson, D. J., & Kennedy, A. (2021). The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *Elife, 10*, Article e63720.

Shah, S., Mishra, A., Yadati, N., & Talukdar, P. P. (2019). Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8876-8884)

Shen, S., Li, C., Xiaowei, H., Xie, Y., Yang, J., Zhang, P., Gan, Z., Lijuan Wang, L., Yuan, C. L., et al. (2022). K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems, 35*, 15558–15573.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020) Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980

Singh, A., Pietrasik, M., Natha, G., Ghouaiel, N., Brizel, K., & Ray, N. (2020) Animal detection in man-made environments. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, (pp. 1438–1449).

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., & Cai, D. (2023) Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355

Sun, C., Baradel, F., Murphy, K., & Schmid, C. (2019a) Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019b) Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE Int'l Conferernce on Computer Vision* pp. 7464–7473

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023) Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6):7

Tapanainen, P., Piitulainen, J., & Jarvinen, T. (1998) Idiomatic object usage and support verbs. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018) The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 8769–8778

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017) Attention is all you need. Advances in neural information processing systems

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., & Belongie, S. (2017) Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* pp. 839–847

von Ziegler, L., Sturman, O., & Bohacek, J. (2021). Big behavior: Challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology, 46*(1), 33–44.

Wang, Z., Chen, T., Li, G., Xu, R., & Lin, L. (2017) Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE Int'l Conf. on Computer Vision* (pp. 464–472)

Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., & Jiang, Y. G. (2023) Chatvideo: A tracklet-centric multimodal and versatile video understanding system. arXiv preprint arXiv:2304.14407

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016) Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 2285–2294)

Wang, J., Chen, D., Zuxuan, W., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G., & Yuan, L. (2022). Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems, 35*, 5696–5710.

Wenhao, W., Sun, Z., Song, Y., Wang, J., & Ouyang, W. (2024). Transferring vision-language models for visual recognition: A classifier perspective. *International Journal of Computer Vision, 132*(2), 392–409.

Wu, W., Wang, X., Luo, H., Wang, J., Yang, Y., & Ouyang, W. (2022) Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. arXiv preprint arXiv:2301.00182

Wu, Y., Zhang, G., Gao, Y., Deng, X., Gong, K., Liang, X., & Lin, L. (2020) Bidirectional graph reasoning network for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 9080–9089

Yang, Q., Xiao, D., & Lin, S. (2018). Feeding behavior recognition for group-housed pigs with the faster r-cnn. *Computers and Electronics in Agriculture, 155*, 453–460.

Zang, Y., Li, W., Zhou, K., Huang, C., & Loy, C. C. 2022 Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225

Zhang, H., Li, X., & Bing, L. (2023) Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858

Zhao, W., & Wu, X. (2023) Boosting entity-aware image captioning with multi-modal knowledge graph. IEEE Transactions on Multimedia

Zhong, Z., Friedman, D., & Chen, D. (2021) Factual probing is [mask]: Learning vs. learning to recall. arXiv preprint arXiv:2104.05240

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a) Conditional prompt learning for vision-language models. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* pp. 16816–16825

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b). Learning to prompt for vision-language models. *Int'l Journal of Computer Vision, 130*(9), 2337–2348.

Zhu, L., & Yang, Y. (2020) Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 8746–8755