ViMoCLIP: Augmenting Static CLIP Representations with Video Motion Cues for Animal Action Recognition

Marcos Rodrigo, Enmin Zhong, Carlos R. del-Blanco, Carlos Cuevas,
Fernando Jaureguizar, Narciso García
Grupo de Tratamiento de Imágenes (GTI), Information Processing and Telecommunications Center,
ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
{marcos.rodrigo, enmin.zhong, carlosrob.delblanco, carlos.cuevas,
fernando.jaureguizar, narciso.garcia}@upm.es

Abstract

We present a novel approach to animal action recognition that augments CLIP's image embeddings with explicit motion and temporal cues, addressing CLIP's limitation of processing only static frames, which prevents it from capturing nuanced dynamics needed to distinguish actions such as "stalking" versus "walking" or "flying" versus "jumping". Our method employs a teacher-student framework: a frozen CLIP model (teacher) provides image features, while a student model learns motion embeddings from optical flow, softly aligning them with the teacher's output. This alignment enables the motion embeddings to encode temporal dynamics while preserving general visual knowledge from the teacher's embeddings. Finally, a Transformer-based network with cross-attention fuses the resulting image and motion embeddings, effectively modeling temporal dependencies. Experiments on the Animal Kingdom dataset show substantial performance gains over prior CLIP-based methods, underscoring the value of integrating motion into pre-trained vision-language embeddings. The complete implementation of our method is publicly available at https://github.com/MarcosRodrigoT/VIMO-CLIP.

1. Introduction

Understanding animal behavior is essential in fields such as wildlife conservation, ecology, and animal welfare. However, automatically recognizing animal behavior poses a significant challenge due to the wide variability in size, shape, and appearance across different species and even among individuals of the same species. Additionally, environmental factors, including diverse backgrounds and habitats, further increase the complexity of this task.

Recent advances in large pretrained vision-language models (VLMs), such as CLIP (Contrastive Language-Image Pre-training) [11] or ALIGN [6], have demonstrated remarkable capabilities in learning general representations based on paired web-scale text-image datasets. These models have shown great generalizability and transferability across a variety of downstream tasks, including zero-shot rare animal classification, animal re-identification and animal pose estimation. Inspired by this success, researchers have attempted to adapt CLIP for animal action recognition by processing videos frame-by-frame, leveraging its pretraining on static image-text pairs. However, directly applying pretrained CLIP models to video tasks results in suboptimal performance since these models are not trained on video-text data and thus struggle to capture the temporal dynamics essential for understanding actions.

Efforts to extend CLIP to video-based tasks, including animal action recognition, have followed several paths. For instance, methods like Category-CLIP [7] process individual frames with CLIP and enhance recognition by introducing category-specific textual prompts (e.g., "a bird is [action]"). Although this approach achieves reasonable results with minimal fine-tuning, it treats videos as a collection of independent frames, ignoring motion continuity. Similarly, Transformer-based models such as MSQNet [8] and Mamba-MSQNet [2] incorporate CLIP-derived text embeddings of action labels, but focus primarily on semantic alignment rather than temporal modeling. Animal-MotionCLIP [15] interleaves optical flow and video frames within the XCLIP [10] framework (a video extension of CLIP), slightly outperforming other state-of-the-art methods on the Animal Kingdom dataset [9]. In addition, it uses multiple classifiers and score aggregation, which increases computational demands. Other strategies include Animal-CLIP [14], which leverages large language models (LLMs) to generate rich action descriptions; and Dual-phase MSQNet [13], which employs a two-stage

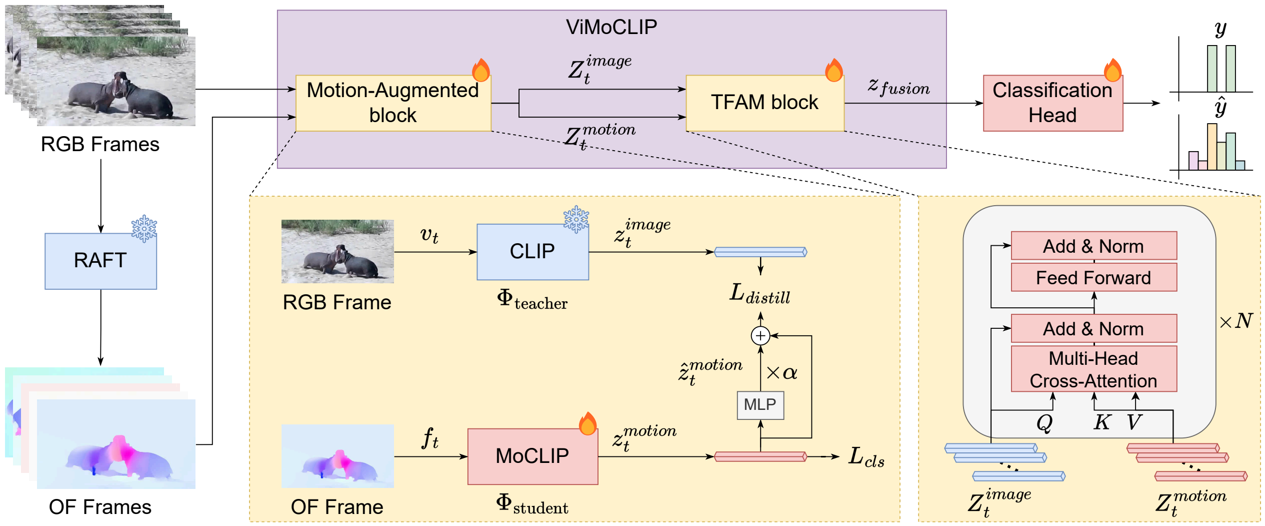

Figure 1. Overview of the proposed ViMoCLIP framework, composed of three blocks: *Motion-Augmented* extends static CLIP embeddings to MoCLIP embeddings by explicitly incorporating motion information from optical flow (OF); *Temporal Transformer-based Fusion of Appearance and Motion (TFAM)* fuses the sequence of CLIP and MoCLIP embeddings into a video-level ViMoCLIP embedding through a temporal cross-attention Transformer; and *Classification Head* that uses ViMoCLIP embedding to predict the animal action categories. Blocks marked with a fire symbol are trainable, whereas those marked with a snowflake symbol are frozen (non-trainable).

pipeline: first identifying the animal species with EfficientNet, and then using a Transformer-based MSQNet to recognize species-specific actions. This second approach tailors recognition to species diversity but introduces scalability challenges due to its dual-phase design.

Despite their strengths, these approaches share an important limitation: they rely heavily on static image-level features and textual priors, without explicitly addressing the motion information essential for action recognition. The suboptimal performance of these adaptations is due to CLIP's original design, which excels at aligning static images with text but lacks mechanisms to process temporal sequences or motion cues, such as those provided by optical flow. This gap is especially pronounced in the animal domain, where actions like ”stalking” versus ”walking” or ”flying” versus ”jumping” depend not only on appearance, but also on subtle movement patterns. Therefore, current methods do not fully exploit the dynamic nature of video, limiting their ability to differentiate such actions effectively. Unlike Animal-MotionCLIP [15], which interleaves RGB and flow tokens inside XCLIP and requires an ensemble of classifiers with score aggregation, ViMoCLIP adopts a completely different teacher–student architecture that (i) aggregates motion information directly into its embeddings, and (ii) removes the need for multiple classifiers.

To overcome these challenges, we propose a novel animal action recognition framework called ViMoCLIP, which consists of three main blocks. In the first block, named the Motion-Augmented Block, CLIP embeddings are enriched with explicit motion information, forming motion-aware embeddings that capture subtle dynamics of animal behaviors. Specifically, we adopt a teacher-student architecture, where a frozen CLIP teacher model extracts robust static features from RGB frames, while a Transformer-based student model, MoCLIP, captures motion dynamics by processing optical flow derived from consecutive RGB frames. The student's motion embeddings are softly aligned with the teacher's image embeddings via a distillation process, effectively integrating appearance and motion representations through contrastive learning involving both image and optical flow modalities. The second block, the Temporal Transformer-based Fusion of Appearance and Motion (TFAM) Block, processes the resulting sequence of CLIP and MoCLIP embeddings—one of each type per video frame—using a Transformer architecture with cross-attention. This approach effectively captures temporal dependencies and multimodal interactions, producing a comprehensive, video-level representation that surpasses existing methods relying solely on static image embeddings combined frame-by-frame. Finally, the Classification Head leverages this video-level ViMoCLIP embedding to predict one or more (non-exclusive) animal behavior categories.

2. Methodology

Overall Approach. Figure 1 provides a schematic overview of our proposed ViMoCLIP framework for animal action recognition. First, the Motion-Augmented Block enhances standard CLIP embeddings with explicit motion-aware representations (MoCLIP embeddings) derived from optical flow. These motion embeddings are softly aligned with CLIP's image embeddings using a teacher-student distillation process. Next, the Temporal Transformer-based Fusion of Appearance and Motion (TFAM) Block integrates these embeddings through a Transformer architecture with cross-attention, effectively modeling temporal and multi-modal interactions. Finally, the Classification Head utilizes the resulting video-level ViMoCLIP embedding to predict multiple animal behavior categories.

Motion-Augmented Block. This component augments CLIP embeddings with explicit motion information via a teacher-student architecture. Let $\mathbf{v} = \{v_t\}_{t=0}^{T}$ represent the T input video frames, from which we generate optical-flow frames $\mathbf{f} = \{f_t\}_{t=1}^{T}$ using RAFT [12]. A frozen CLIP model, Φ_{teacher}, serves as the teacher and extracts per-frame CLIP image embeddings:

$$\mathbf{z}_t^{\text{image}} = \Phi_{\text{teacher}}(v_t). \tag{1}$$

Meanwhile, a Transformer-based student network, Φ_{student}, processes the corresponding optical-flow frames to yield motion-aware (MoCLIP) embeddings:

$$\mathbf{z}_t^{\text{motion}} = \Phi_{\text{student}}(f_t). \tag{2}$$

To align $\mathbf{z}_t^{\text{motion}}$ with $\mathbf{z}_t^{\text{image}}$, we adopt a residual scheme: the motion embedding is passed through an MLP head $h(\cdot)$, scaled by a factor $\alpha \in [0, 1]$, and then added to the original motion embedding:

$$\hat{\mathbf{z}}_t^{\text{motion}} = \mathbf{z}_t^{\text{motion}} + \alpha \cdot h\big(\mathbf{z}_t^{\text{motion}}\big). \tag{3}$$

A cosine-similarity-based distillation loss is applied to encourage $\hat{\mathbf{z}}_t^{\text{motion}}$ to remain close to $\mathbf{z}_t^{\text{image}}$, thereby transferring pretrained appearance knowledge to the motion pathway without suppressing the distinct motion features:

$$\mathcal{L}_{\text{distill}} = \frac{1}{T} \sum_{t=1}^{T} \left[1 - \frac{\mathbf{z}_t^{\text{image}} \cdot \hat{\mathbf{z}}_t^{\text{motion}}}{\|\mathbf{z}_t^{\text{image}}\| \, \|\hat{\mathbf{z}}_t^{\text{motion}}\|} \right]. \tag{4}$$

Additionally, a cross-entropy classification loss is applied to improve the discriminative capability of the motion features for multi-label animal action recognition:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{K} \sum_{k=1}^{K} y_k \log\big(\tilde{y}_k\big), \tag{5}$$

where $y \in \{0, 1\}^K$ represents the ground-truth labels for K possible actions (with y_k denoting the label for the k-th action) and $\tilde{y} \in [0, 1]^K$ are the corresponding predicted probabilities from an accessory classification head that takes the average of all the $\mathbf{z}_t^{\text{motion}}$ in a video as input. This simple yet effective video-level embedding enables efficient adaptation of features to a specific domain without incurring significant memory overhead during training. The overall teacher-student objective, $\mathcal{L}_{\text{TS}}$, is then given by:

$$\mathcal{L}_{\text{TS}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{distill}}, \tag{6}$$

ensuring that the motion representations achieve both semantic alignment and discriminative strength.

TFAM Block. The Temporal Transformer-Based Fusion of Appearance and Motion block aggregates the per-frame image and motion embeddings (CLIP and MoCLIP embeddings, respectively) into a unified video-level representation, denoted as ViMoCLIP. Let

$$\mathbf{Z}^{\text{image}} = \{\mathbf{z}_t^{\text{image}}\}_{t=1}^{T} \quad \text{and} \quad \mathbf{Z}^{\text{motion}} = \{\mathbf{z}_t^{\text{motion}}\}_{t=1}^{T},$$

be the sequences of image and motion embeddings obtained from the Motion-Augmented Block. In the TFAM block, these sequences are fused via multi-head cross-attention. In particular, for each head $i \in \{1, \ldots, H\}$, we compute:

$$Q_i = \mathbf{Z}^{\text{image}} W_i^Q, \quad K_i = \mathbf{Z}^{\text{motion}} W_i^K, \quad V_i = \mathbf{Z}^{\text{motion}} W_i^V,$$

where W_i^Q, W_i^K, and W_i^V are learnable projection matrices. The attention output for head i is then given by:

$$\text{head}_i = \text{softmax}\Big(\frac{Q_i K_i^\top}{\sqrt{d_k}}\Big) V_i,$$

with d_k being the key dimension. The outputs of all heads are concatenated and projected via:

$$\mathbf{z}_{\text{fusion}} = \text{Concat}(\text{head}_1, \ldots, \text{head}_H) W^O,$$

where W^O is a learnable output matrix. The fused embedding $\mathbf{z}_{\text{fusion}}$ (i.e., ViMoCLIP) encapsulates both temporal dependencies within each modality and inter-modal interactions, providing a comprehensive representation for robust animal action recognition.

Implementation Details. We train and validate our model on the standard splits of the Animal Kingdom dataset [9]. Both RGB and optical-flow videos originally have a resolution of 640 x 360 pixels and are resized to 224 x 224 during preprocessing. Optical-flow videos consist of one frame less than their RGB counterparts, as flow is computed between consecutive RGB frames using RAFT [12]. RAFT is chosen for accuracy, but ViMoCLIP is agnostic to the flow backbone. The Motion-Augmented Block is trained independently with a cosine similarity distillation loss. After training, it is frozen and used in inference mode to generate CLIP and MoCLIP embeddings for the TFAM Block. The TFAM and Classification Head Blocks are trained jointly. All training and evaluations were conducted on two Nvidia RTX 4090 GPUs.

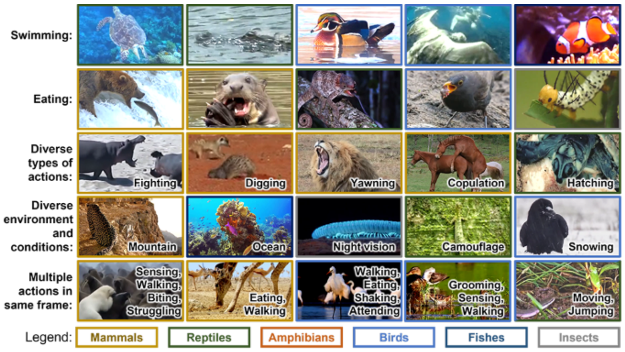
Swimming:
Eating:
Diverse types of actions:
Fighting
Digging
Yawning
Copulation
Hatching
Diverse environment and conditions:
Mountain
Ocean
Night vision
Camouflage
Snowing
Multiple actions in same frame:
Sensing, Walking, Biting, Struggling
Eating, Walking
Walking, Eating, Shaking, Attending
Grooming, Sensing, Walking
Moving, Jumping
Legend:
Mammals
Reptiles
Amphibians
Birds
Fishes
Insects

Figure 2. Examples of the Animal Kingdom dataset with different animal species in different environments and exhibiting diverse behaviors.

3. Experiments

To evaluate the effectiveness of the proposed ViMoCLIP, we conducted experiments on the Animal Kingdom dataset [9] to address the problem of general recognition of animal behavior. It contains more than 50 hours of video and 30K annotated video sequences with multi-label, fine-grained actions. It comprises 850 species of mammals, fishes, amphibians, reptiles, birds, and insects, with 140 action classes spanning life stages, daily activities, and social interactions (e.g., molting, feeding, playing, etc.). Some examples are displayed in Fig. 2.

Several state-of-the-art approaches have been used in the result comparison, which can be categorized into vision-language models (Category-CLIP [7], Animal-MotionCLIP [15], MSQNet [8], MSQNet-VideoMAE [8], MSQNet-TimeSformer [8], and Mamba-MSQNet [3]) and video models (I3D[1], Slowfast [5], and X3D [4]). Table 1 shows the recognition results. Our proposed ViMoCLIP method achieves the best recognition results using RGB images and the derived optical flow, without using other modalities such as text. This can be justified by the fact that including motion cues can be crucial for fine-grained animal actions in videos. The ViMoCLIP version that only uses RGB images is the second best, surpassing works that use vision-language models and explicitly text features. The reason is that other works proposed complex neural network architectures that are trained end-to-end including the image and video extraction of features. This requires huge memory resources for training that commonly restricts the temporal analysis to a maximum of 16 frames per video. In comparison, the ViMoCLIP framework is focused on the temporal analysis of compact and efficient (image and motion) embeddings, which are obtained from the independently trained models CLIP and MoCLIP. This strategy requires much less memory for the training, allowing to process a much larger quantity of frames per video, and ultimately improving the recognition accuracy performance. Although the absolute mAP gain over the previous best Animal-MotionCLIP baseline is $+2.46$, it occurs in a regime where recent state-of-the-art methods have been separated by less than 3 points (71.20–74.63 mAP in Table 1). In a 140-class multi-label setting, this 3.3% relative improvement corresponds to a considerable number of additional correctly recognized behaviors and is achieved without relying on the text modality required by Animal-MotionCLIP.

Method	Image	Text	Motion	mAP(%)
I3D [1]	✓			16.48
SlowFast [5]	✓			20.46
X3D [4]	✓			25.25
Category-CLIP [7]	✓	✓		55.36
MSQNet [8]	✓	✓		55.59
MSQNet-VideoMAE [8]	✓	✓		71.20
Dual-phase MSQNet [13]	✓	✓		72.50
MSQNet-TimeSformer [8]	✓	✓		73.10
Animal-MotionCLIP [15]	✓	✓		73.9
Mamba-MSQNet [2]	✓	✓		74.60
Animal-MotionCLIP [15]	✓	✓	✓	74.63
ViMoCLIP (ours)	✓			75.83
ViMoCLIP (ours)	✓		✓	**77.09**

Table 1. Performance comparison with the state-of-the-art models on the Animal Kingdom dataset. The best result, obtained with the proposed ViMoCLIP, is highlighted in bold.

Another practical advantage of ViMoCLIP is that text-based strategies, although effective, require generating careful text annotations or captions for the videos (usually databases do not include these text annotations, as is the case of Animal Kingdom), which is costly and raises concerns about how reliably text can capture complex or overlapping behaviors.

For additional insights, we conducted an ablation study (Table 2), varying the combination of input information (*Image* and/or *Motion*) and attention-fusion mechanisms. In most configurations, incorporating optical flow alongside RGB frames improves the mAP compared to single-modality baselines (e.g., cross-attention reaches 77.09% vs. 75.83% for RGB-only). However, self-attention with embedding concatenation (74.56%) falls slightly below the RGB-only setting, indicating that the manner of fusing flow can significantly affect outcomes. Notably, self-attention with temporal concatenation achieves 76.99% mAP, underscoring how explicitly encoding temporal information yields stronger multi-label animal action recognition.

4. Conclusion

In this work, we introduced ViMoCLIP, a novel framework for animal action recognition that augments CLIP's robust vision embeddings with explicit motion representations. Through a teacher-student design, the frozen CLIP

Table 2. Ablation study examining the roles of image vs. motion inputs, attention mechanisms (self or cross), and embedding concatenation strategies. We report mAP (%) for each combination of settings. The best result is highlighted in bold.

Image	Motion	Attention	Concat dim	mAP(%)
✓	✗	self	-	75.83
✗	✓	self	-	51.05
✓	✓	self	embedding	74.56
✓	✓	self	temporal	76.99
✓	✓	cross	-	**77.09**

model extracts appearance features while a dedicated student network captures optical flow. A residual fusion module then combines these complementary embeddings for stronger temporal and contextual modeling of behaviors. Experiments on the Animal Kingdom dataset reveal that our method significantly surpasses previous solutions, demonstrating that motion-aware features are crucial for effective animal action recognition. Future research could enhance these results by incorporating additional modalities (including text) and exploring even richer temporal modeling strategies.

5. Acknowledgement

This work was supported in part by Ministerio de Ciencia, Innovación y Universidades (MCIU)/Agencia Estatal de Investigación (AEI)/10.13039/501100011033 of the Spanish Government under projects PID2020-115132RB (SARAOS), PID2023-148922OA-I00 (EEVOCATIONS), and TED2021-130225A-I00 (ANEMONA) and by the European Union NextGenerationEU/PRTR.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 4

[2] Edoardo Fazzari, Donato Romano, Fabrizio Falchi, and Cesare Stefanini. Selective state models are what you need for animal action recognition. *Ecological Informatics*, 85: 102955, 2025. 1, 4

[3] Edoardo Fazzari, Donato Romano, Fabrizio Falchi, and Cesare Stefanini. Artemis: animal recognition through enhanced multimodal integration system. *International Journal of Machine Learning and Cybernetics*, 2025. 4

[4] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. 4

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 4

[6] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 1

[7] Yinuo Jing, Chunyu Wang, Ruxu Zhang, Kongming Liang, and Zhanyu Ma. Category-specific prompts for animal action recognition with pretrained vision-language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 5716–5724, New York, NY, USA, 2023. Association for Computing Machinery. 1, 4

[8] Anindya Mondal, Sauradip Nag, Joaquin M Prada, Xiatian Zhu, and Anjan Dutta. Actor-agnostic multi-label action recognition with multi-modal query. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2023. 1, 4

[9] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding, 2022. 1, 3, 4

[10] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[12] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 3

[13] An Yu, Jeremy Varghese, Ferhat Demirkiran, Peter Buonaiuto, Xin Li, and Ming-Ching Chang. Dual-phase msqnet for species-specific animal activity recognition. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2024. 1, 4

[14] Yifan Zheng, Erhao Li, Lei Zhu, et al. Animal-clip. https://github.com/PRIS-CV/Animal-CLIP, 2023. Accessed on 04/02/2025. 1

[15] Enmin Zhong, Carlos R. del Blanco, Daniel Berjon, Fernando Jaureguizar, and Narciso Garcia. Animalmotionclip: Embedding motion in clip for animal behavior analysis. Paper presented at the CV4Animals Workshop, 2024. Accessed: 04/02/2025. 1, 2, 4