A. Supplement

A.1. Experimental Details

A.1.1 Animal Kingdom Experiments

We employ a transformer encoder with 6 layers and 8 attention heads per layer for all feature representations. The network is trained using a cosine annealing schedule with a peak learning rate of 3×10^{-5} over 15 epochs. We use focal loss with $\alpha = 0.25$ and $\gamma = 2.0$, apply weight decay of 1×10^{-4} and mixup augmentation with α value of 0.2.

The distribution of actions and species are shown in S6 and S7

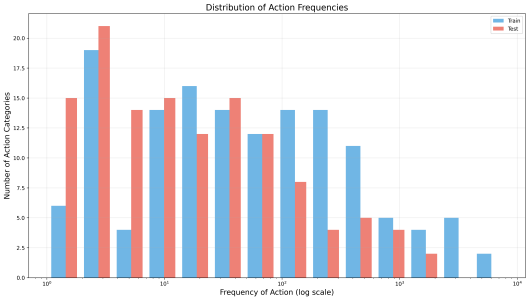

Figure S6. **Animal Kingdom Action Distribution.**

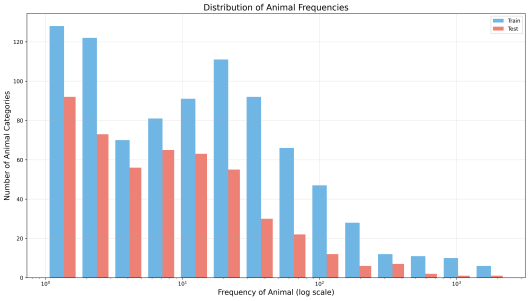

Figure S7. **Animal Kingdom Species Distribution.**

A.1.2 CharadesEgo Experiments

Similarly, transformer encoder with 6 layers and 8 attention heads per layer for all feature representations. The network is trained using a cosine annealing schedule with a peak learning rate of 1×10^{-4} over 15 epochs. We use focal loss with $\alpha = 1.0$ and $\gamma = 2.0$, apply weight decay of 1×10^{-3} and mixup augmentation with α value of 0.3.

The distribution of actions is shown in S8

A.1.3 UCF-HMDB Experiments

We use a transformer encoder with 6 layers and 8 attention heads per layer for all feature representations. The network is trained using a cosine annealing schedule with a peak learning rate of 1×10^{-4} over 15 epochs. We use focal loss with $\alpha = 1.0$ and $\gamma = 2.0$, apply weight decay of 1×10^{-4} and mixup augmentation with α value of 0.0.

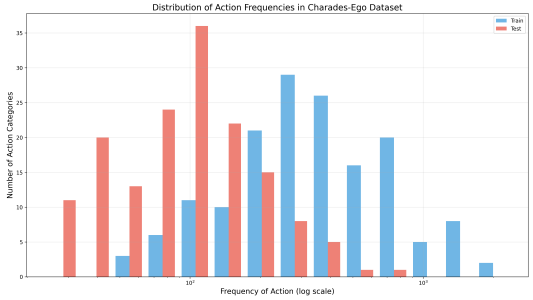

Figure S8. **Charades-Ego Action Distribution.**

A.1.4 Grid-Search

We use a transformer encoder with 6 layers and 8 attention heads per layer for all feature representations. The network is trained using a cosine annealing schedule with a peak learning rate of 1×10^{-4} over 15 epochs. We use focal loss with $\alpha = 1.0$ and $\gamma = 2.0$. For CharadesEgo (Fig. 5 a and b), we apply weight decay of 1×10^{-3} and mixup augmentation with α value of 0.3. For HMDB and HMDB-UCF (Fig. 5 c and d), we apply weight decay of 1×10^{-4} and mixup augmentation with α value of 0.0.

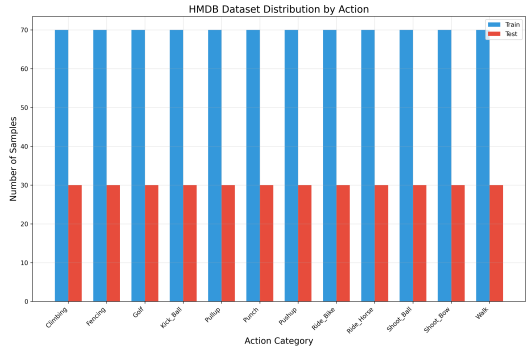

Figure S9. **HMDB51 Distribution.**

A.2. Self-Supervised Frozen Backbones

A.2.1 Stable Diffusion v2 Extraction

Stable Video Diffusion builds upon Stable Diffusion v2 (SDv2) by adding temporal layers in between the layers of a pretrained SDv2, and training only them on video data. The output of these temporal layers is gated by a learned continuous α parameter between 0 and 1 for each layer. When $\alpha = 1.0$, no information from the temporal is added and the

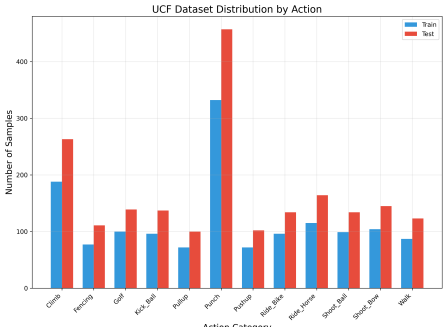

Figure S10. **UCF101 Distribution.**

model is effectively encoding the frames independently using SDv2. Therefore, we set $\alpha = 1.0$ for all layers in order to use the same backbone to extract SDv2 features.

A.2.2 V-JEPA and VideoMAEv2 Extraction

We extract VideoMAEv2 features using the official repository and pretrained encoder provided by the authors. Specifically, we compute spatiotemporal representations for each video segment and then average across the spatial dimension to obtain a single feature vector.

A.3. Action Localization

We can also analyze which parts of a video most contribute to the prediction of a specific action. For a given video, we can crop it into fixed patches and provide as input to the model the same patch over time. Then we can evaluate the prediction of an action over all patches and produce a heatmap like in Figure S11, which shows which parts of a video are most predicted as a prey being eaten, and we can localize the action in space. Resolution is not an issue for our model since we average pool the video features.

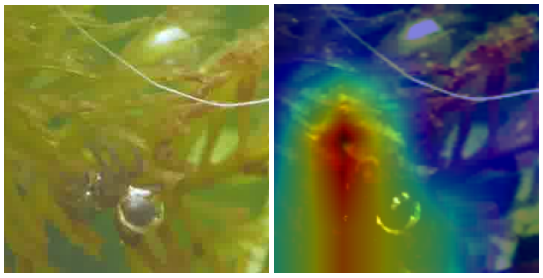

Figure S11. Action localization

A.4. Cross-Species Generalization

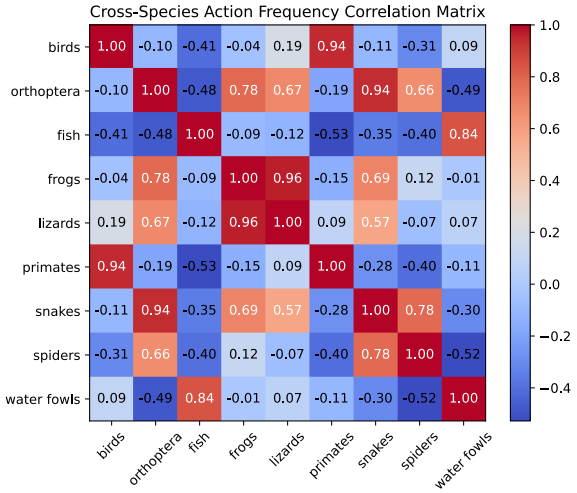

Figure S12. **Action frequencies correlations across species.**

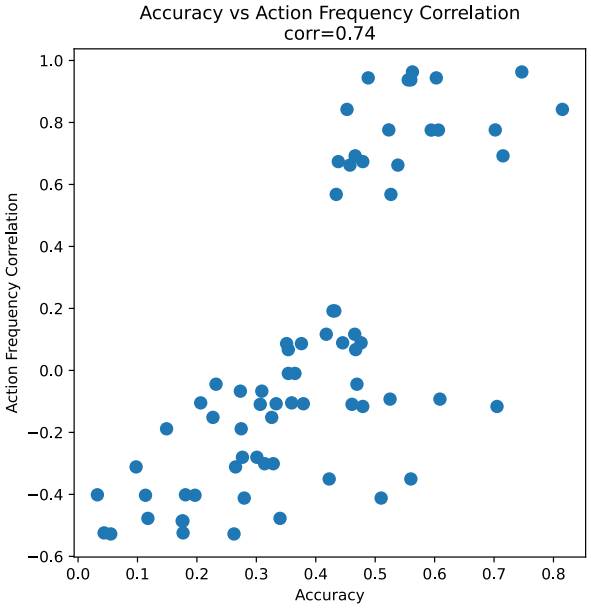

Figure S13. **Accuracy vs action frequency correlation across species.**

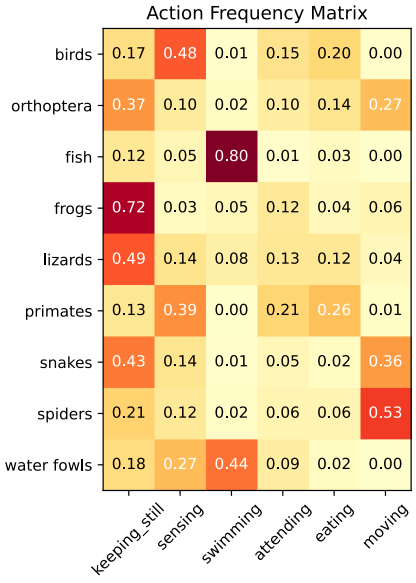

Figure S14. **Action frequencies across species.**

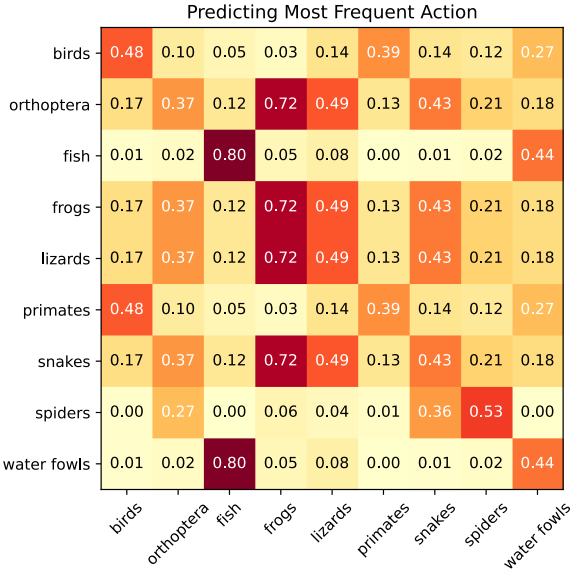

Figure S15. **Results of predicting the most frequent action from the training species in the test species.**

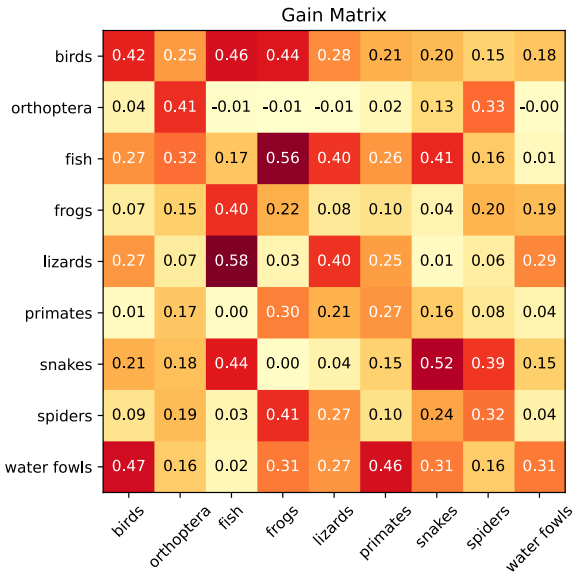

Figure S16. **ActionDiff gains over predicting the most frequent class from the training species in the test species.**

A.5. Choice of UNet Layer

Layer	Animal Kingdom		Charades-Ego		UCF101-HMDB51	
	Full Dataset (mAP)	Unseen Species (acc)	$1^{st} \rightarrow 1^{st}$ mAP	$3^{rd} \rightarrow 1^{st}$ mAP	U $\rightarrow$ H Acc	H $\rightarrow$ U Acc
ActionDiff (ours)						
Layer 3	55.16	35.51	16.98	15.1	17.17	17.57
Layer 6	71.12	40.90	21.07	16.6	33.92	39.67
Layer 9	**80.79**	**51.49**	**36.5**	**30.2**	**75.6**	**81.5**

Table S5. **Effect of UNet layer on all datasets.** Results improve with deeper decoder layers across all metrics. **Bold** denotes best result.