Supplementary Materials for “MammalNet: A Large-scale Video Benchmark for Mammal Recognition and Behavior Understanding”

In the supplementary material, we provide: 1) more details for training the classification and detection models, 2) more analysis for the comparison of separately recognizing animal and behavior vs. their joint recognition, 3) the complete version of our mammal taxonomy, 4) the demonstration of compositional low-shot animal and behavior recognition, 5) the demonstration of the long-tail distribution of animals, behaviors and their compositions, 6) the confusion matrix for animal and behavior prediction, 7) the visualization of our behavior localization interface.

1. Training Details

1.1. Training Details for Animal and Behavior Classification

We train all the recognition models with their officially released code. Specifically, for the configuration of training I3D [1], C3D [8], SlowFast [2] models, we use the base learning rate 0.1, cosine decay learning rate scheduler, 196 training epochs, 34 warmup epochs and the batch size 256. We sample 16 frames per clip with the sampling rate of 24. For the configuration of training MViT v2 [5] model, we apply the base learning 0.0001, cosine decay learning rate scheduler, 200 training epochs, 30 warmup epochs, and the batch size 256. We sample 16 frames per clip with the sampling rate of 16.

1.2. Training Details for Behavior Detection

Feature extraction. We firstly extract the frames from each video with 25 FPS and also extract the optical flow with TV-L1 [3, 6] algorithm. After that, we finetune an I3D [1] model, that has been pretrained on Kinetics 400 [4] dataset, on our MammalNet, and then use it to generate the features for each RGB and optical flow frame. Since each video has variable duration, we perform the uniform interpolation to generate 100 fixed-length features for each video. Finally, we concatenate the RGB and optical flow features into a 2048-dimensional embedding as the model input.

Model training. We train all the detection models with their officially released code and the default configurations. For training ActionFormer [10] model, we apply the base learning rate 0.001, cosine decay learning rate scheduler, 30 training

	Animal Classification				Behavior Classification			
Baselines	Many	Medium	Few	All	Many	Medium	Few	All
	12	28	133	173	4	4	4	12
SlowFast [2] (joint loss)	**58.3**	**43.1**	16.6	24.5	**45.1**	**32.7**	14.8	**30.9**
SlowFast (separate losses)	57.0	42.4	**20.2**	**26.9**	44.3	32.0	**15.2**	30.4
C3D [8] (joint loss)	**58.3**	45.4	19.1	26.8	**44.6**	**36.0**	**15.9**	**32.2**
C3D (separate losses)	58.1	**46.7**	**21.6**	**28.8**	42.4	34.0	15.3	30.6
I3D [1] (joint loss)	**58.6**	**42.9**	**16.9**	**24.8**	**46.3**	**35.0**	**14.8**	**32.1**
I3D (separate losses)	57.3	42.4	16.6	24.3	43.8	29.3	12.3	28.4
MViT V2 [5] (joint loss)	66.7	56.0	23.4	32.5	**50.9**	**42.4**	**20.0**	**37.8**
MViT V2 (separate losses)	**67.1**	**56.2**	**24.2**	**33.2**	50.3	38.7	16.0	35.0

Table 1. Per-class Top-1 accuracy for animal, behavior and their compositional prediction. All the models are initialized with the weights pretrained on Kinetics 400 dataset [4].

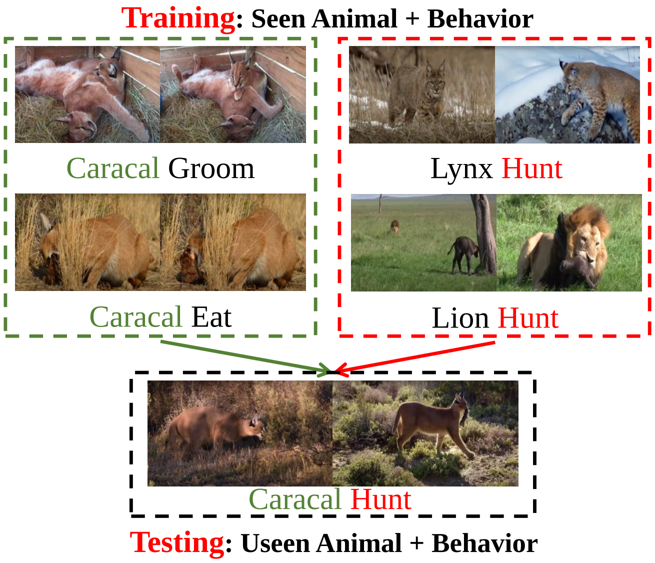

Training: Seen Animal + Behavior
Caracal Groom
Caracal Eat
Lynx Hunt
Lion Hunt
Caracal Hunt
Testing: Useen Animal + Behavior

Figure 1. Demonstration of compositional zeroshot animal and behavior recognition.

epochs, 5 warmup epochs, and the batch size 16. For training TAGS [7] model, we apply the base learning rate of 0.0004, step decay learning rate scheduler, 20 training epochs, and the batch size 200. For training CoLA [9] model, we apply the base learning rate of 0.0001, 50 training epochs, and the batch size 256.

2. Separate Animal and Behavior Prediction vs. Joint Prediction

We also train all the baseline models with just the animal classification loss for animal recognition and just the behavior classification loss for behavior recognition. We compute the per-class top-1 accuracy and summarize all the results in Table 1. Comparing joint vs. separate training, we find that joint training is able to improve the behavior recognition performance by improving $\sim$2.2 per-class accuracy in average of all the baselines, indicating that recognizing the animal category can help improve the behavior understanding. On the other hand, performing the animal recognition alone is better than the joint training together, improving $\sim$1.5 top-1 accuracy on average for all the baselines.

3. Full animal taxonomy

We demonstrate the full animal taxonomy in the Fig. 11. The full taxonomy consists of different layers such as *order*, *family*, *genus*, *sub-family* and *tribe*. In our MammalNet, we treat the *genus*, *sub-family* and *tribe* from the lowest level as the animal categories during the classification. For some genera that are very hard to distinguish and also hard to find sufficient number of videos from YouTube, such as *African elephant* and *Asian elephant*, we group them together as an union of genera. We also expand the searching keywords with more general names such as *elephant* in order to find more relevant videos. Additionally, *bovidae* family has the sub-layers of *tribe* and *sub-family* instead of *genus* in scientific mammal taxonomy, hence we classify the animals in the *tribe* and *sub-family* level. Finally, we have:
17 orders: e.g. artiodactyla, primates, etc.
69 families: e.g. bovidae, cervidae, etc.
162 genera: e.g. antilocapra, felis, etc.
5 tribes: e.g. aepycerotinae, bovini, etc.
6 sub-families: e.g. alcelaphinae, caprini, etc.

4. Illustration of compositional low-shot animal and behavior recognition

To better demonstrate the compositional low-shot animal and behavior recognition, we visualize one zero-shot example in Fig. 1. During the training, the model has seen *Caracal* performing behaviors such as *groom* and *eat*, but it has never seen the *hunt* behavior. However, it has seen the *hunt* behavior performed by other animals (e.g. *lynx* and *lion*), and the target is to evaluate if the model can also successfully predict the *Caracal hunt*.

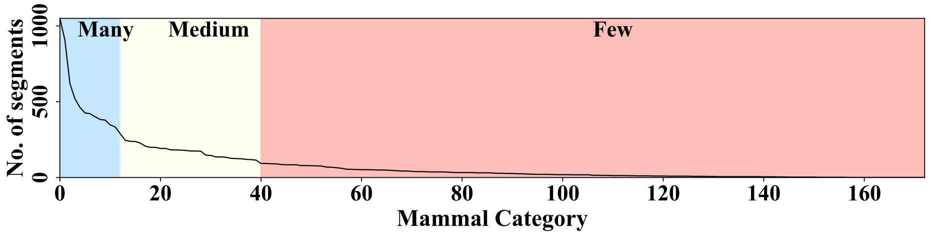


Figure 2. **The number of trimmed videos per each mammal category.** The animal with the frequency $>$ 300 is grouped into **many**. The animal with the frequency $\leq$ 300 and $>$ 100 is grouped into **medium**. The animal with the frequency $\leq$ 100 is grouped into **few**. We rank the animals based on their frequency.

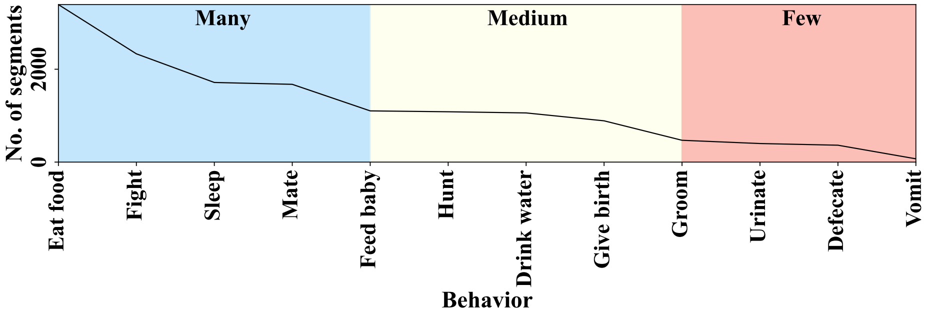


Figure 3. **The number of trimmed videos per each behavior.** The behavior with the frequency $>$ 1,500 is grouped into **many**. The behavior with the frequency $\leq$ 1,500 and $>$ 500 is grouped into **medium**. The behavior with the frequency $\leq$ 500 is grouped into **few**. We rank the behavior based on their frequency.

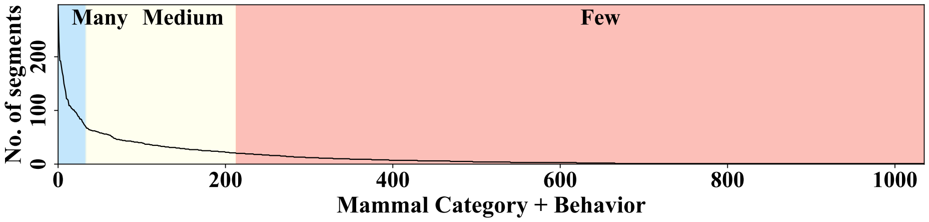


Figure 4. **The number of trimmed videos per each composition of animal category and behavior.** The composition with the frequency $>$ 70 is grouped into **many**. The composition with the frequency $\leq$ 70 and $>$ 20 is grouped into **medium**. The composition with the frequency $\leq$ 20 is grouped into **few**. We rank the composition based on their frequency.

5. The long-tail distribution of animal, behavior and their composition

Our data has the skewed distribution in terms of the animal category, behavior and also their composition. To get better insights, we split the categories into *many*, *medium* and *few* groups based on the their frequency. We show the animal distribution in Fig. 2, the behavior distribution in Fig. 3 and their compositional distribution in Fig. 4.

Animal split: the animals with the frequency $>$ 300 are grouped into *many*. The animals with the frequency $\leq$ 300 and $>$ 100 are grouped into *medium*. The animals with the frequency $\leq$ 100 are grouped into *few*.

Behavior split: The behaviors with the frequency $>$ 1,500 are grouped into *many*. The behaviors with the frequency $\leq$ 1,500 and $>$ 500 are grouped into *medium*. The behaviors with the frequency $\leq$ 500 are grouped into *few*.

Composition split: The compositions with the frequency $>$ 70 are grouped into many. The compositions with the frequency $\leq$ 70 and $>$ 20 are grouped into medium. The compositions with the frequency $\leq$ 20 are grouped into few.

6. Confusion Matrix for Behavior Prediction

Confusion matrices for predicting animal and behavior. We show the confusion matrix for behavior recognition in Fig. 5. We observe that more frequent behaviors, such as *eat food* and *fight*, have less ambiguity, while less frequent behaviors, such

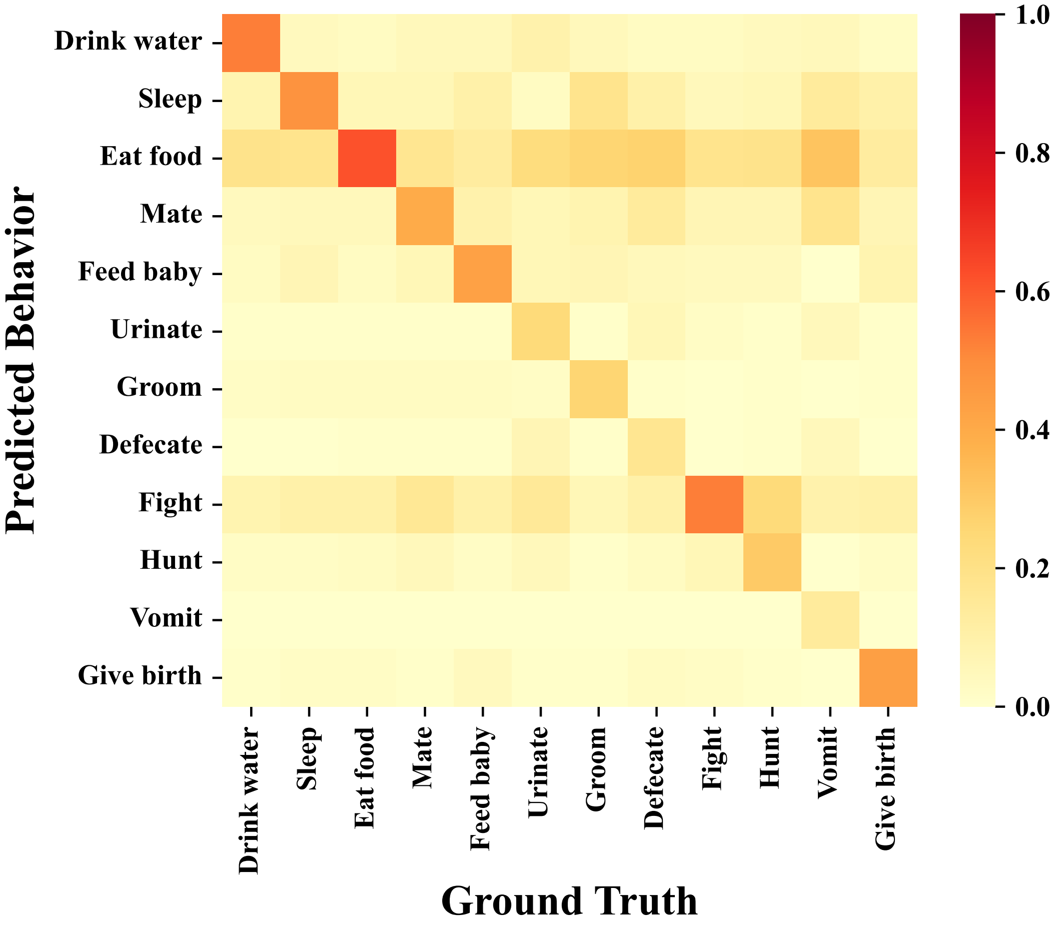


Figure 5. Confusion matrix visualization for behavior recognition.

as *vomit* and *defecate*, are often misclassified with, e.g., *eat food*. Also, some behaviors can mistakenly be predicted as other similar behavior: for example, the *hunt* behavior is often mistakenly predicted as *fight*. We show the confusion matrix for animal recognition in the supplement.

7. Confusion Matrix for Animal Prediction

We compute the confusion matrix for the animal prediction. We demonstrate its visualization in the Fig. 6. Through this confusion matrix, we found that some animals are easily mis-classified into other similar-appearance animals. For example, *lynx* is often mis-classified into *panthera*. the *caracal* is often mis-classified into *panthera* as well. *muntiacus* is often mis-classififed into *caprini* and *deer*. However, for some animals such as *giraffe* and *elephant*, they are much less mis-classified due to their unique body shape compared to other animals.

8. Annotation Interface Demonstration

We demonstrate the final animal behavior verification interface in Fig. 8 and also its instruction description in Fig. 7. Additionally, we also demonstrate the behavior localization interface in Fig. 10 and also the instruction description in Fig. 9. We assign each video to 5 different Amazon Mechanical Turk (AMT) workers and ask them to temporally localize the video range which exists the animal behavior. Each worker is required to annotate all the temporal ranges in which appears the target animal behavior. Before each worker started to participate in the annotation, they need to be evaluated if they have clearly understand our task with 20 multi-choice questions, and they can only work on our behavior localization task only when they can correctly answer ≥90% questions.

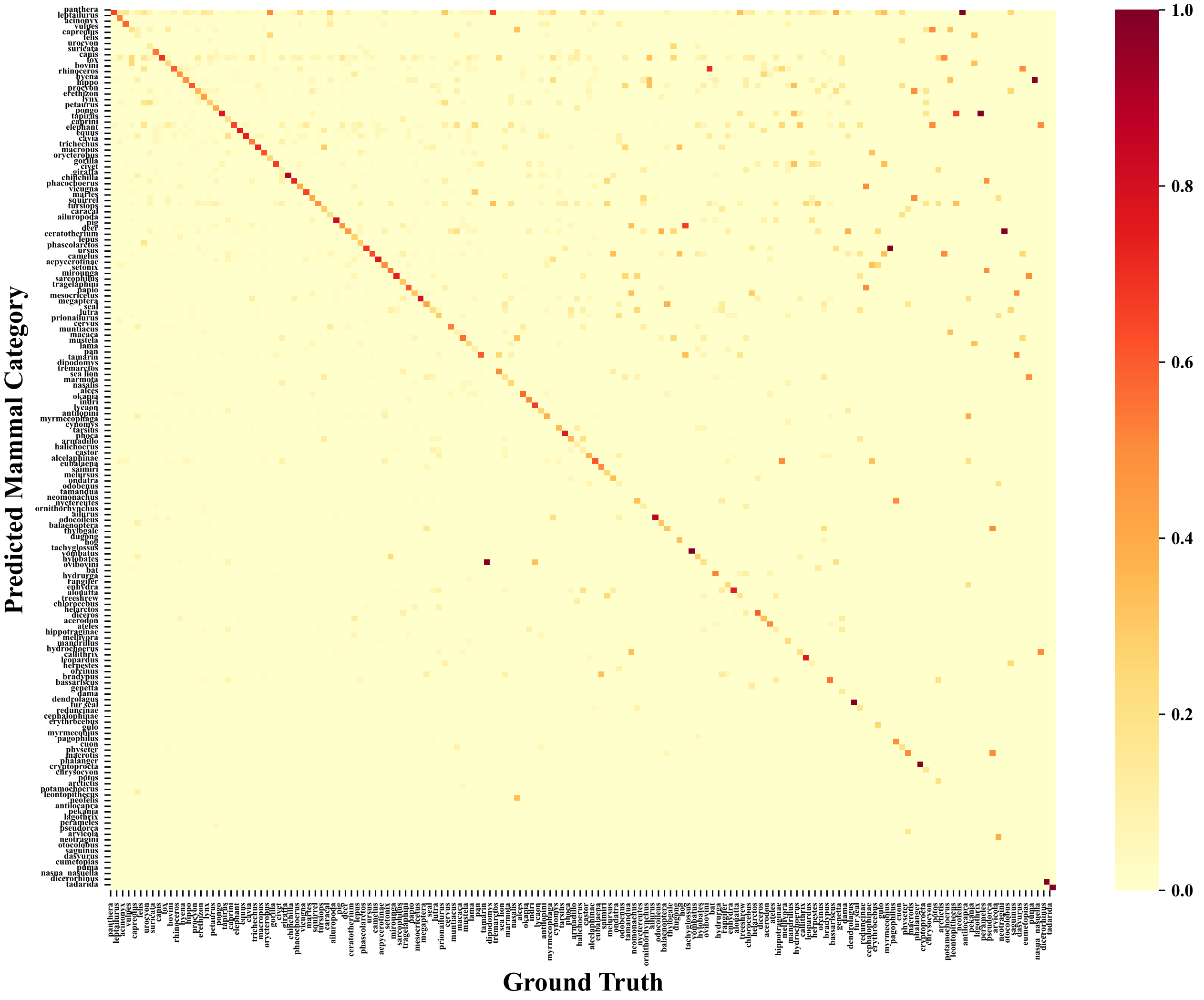
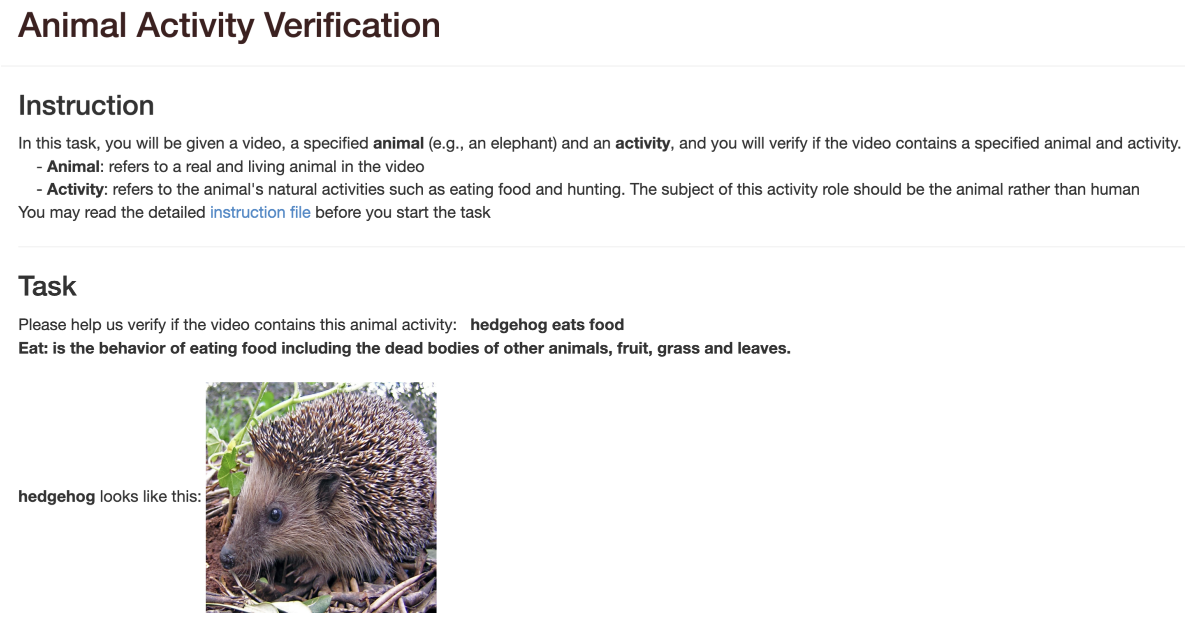


Figure 6. Confusion matrix for animal prediction. We visualize the confusion matrix for predicting the animals in our testing data.



Animal Activity Verification

Instruction

In this task, you will be given a video, a specified **animal** (e.g., an elephant) and an **activity**, and you will verify if the video contains a specified animal and activity.

- **Animal**: refers to a real and living animal in the video
- **Activity**: refers to the animal's natural activities such as eating food and hunting. The subject of this activity role should be the animal rather than human

You may read the detailed instruction file before you start the task

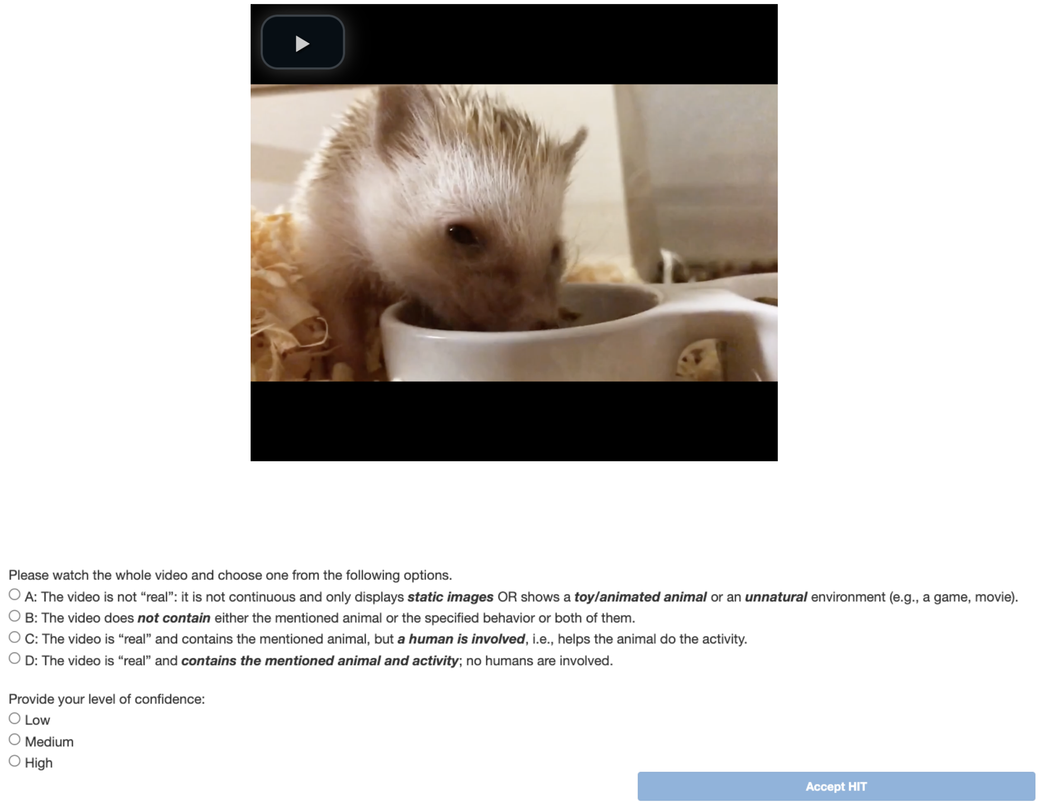
Task

Please help us verify if the video contains this animal activity: **hedgehog eats food**

Eat: is the behavior of eating food including the dead bodies of other animals, fruit, grass and leaves.

hedgehog looks like this:

Figure 7. The instruction for animal and behavior verification annotation



Please watch the whole video and choose one from the following options.

○ A: The video is not “real”: it is not continuous and only displays ***static images*** OR shows a ***toy/animated animal*** or an ***unnatural*** environment (e.g., a game, movie).

○ B: The video does ***not contain*** either the mentioned animal or the specified behavior or both of them.

○ C: The video is “real” and contains the mentioned animal, but ***a human is involved***, i.e., helps the animal do the activity.

○ D: The video is “real” and ***contains the mentioned animal and activity***; no humans are involved.

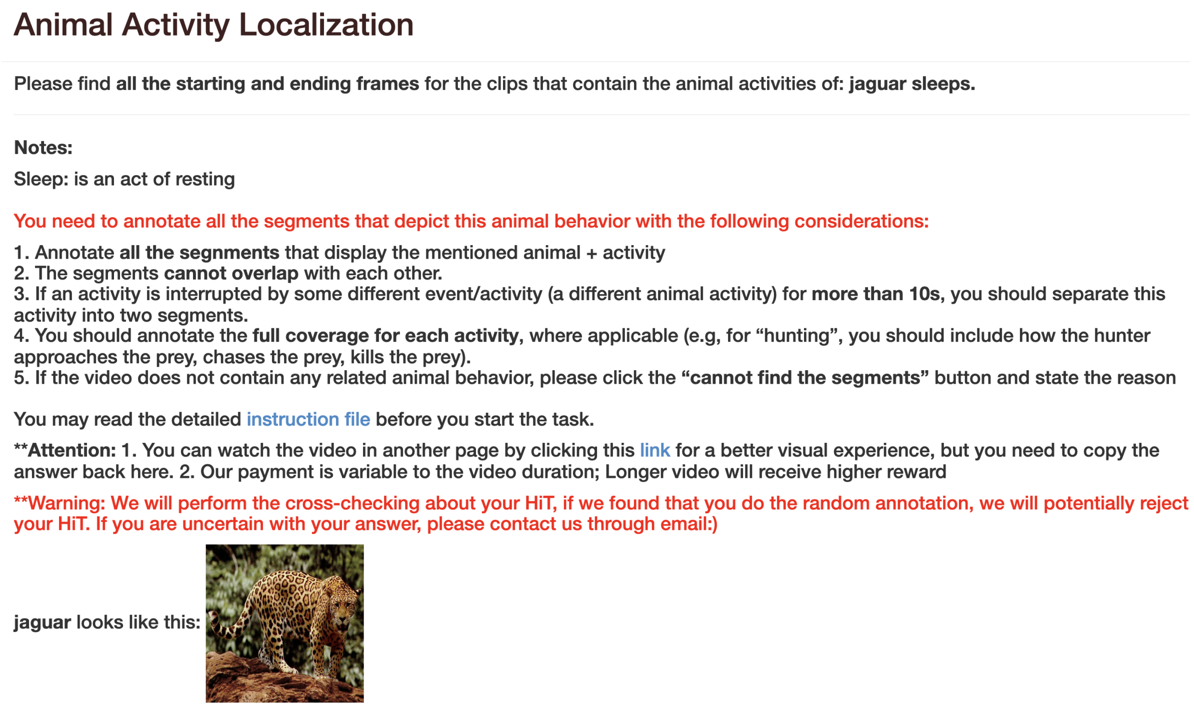
Provide your level of confidence:

○ Low

○ Medium

○ High

Figure 8. The interface for animal behavior verification.



Animal Activity Localization

Please find **all the starting and ending frames** for the clips that contain the animal activities of: **jaguar sleeps.**

Notes:

Sleep: is an act of resting

You need to annotate all the segments that depict this animal behavior with the following considerations:

1. Annotate **all the segnments** that display the mentioned animal + activity
2. The segments **cannot overlap** with each other.
3. If an activity is interrupted by some different event/activity (a different animal activity) for **more than 10s**, you should separate this activity into two segments.
4. You should annotate the **full coverage for each activity**, where applicable (e.g, for "hunting", you should include how the hunter approaches the prey, chases the prey, kills the prey).
5. If the video does not contain any related animal behavior, please click the **"cannot find the segments"** button and state the reason

You may read the detailed instruction file before you start the task.

****Attention:** 1. You can watch the video in another page by clicking this link for a better visual experience, but you need to copy the answer back here. 2. Our payment is variable to the video duration; Longer video will receive higher reward

**Warning: We will perform the cross-checking about your HiT, if we found that you do the random annotation, we will potentially reject your HiT. If you are uncertain with your answer, please contact us through email:)

jaguar looks like this:

Figure 9. The instruction for behavior localization

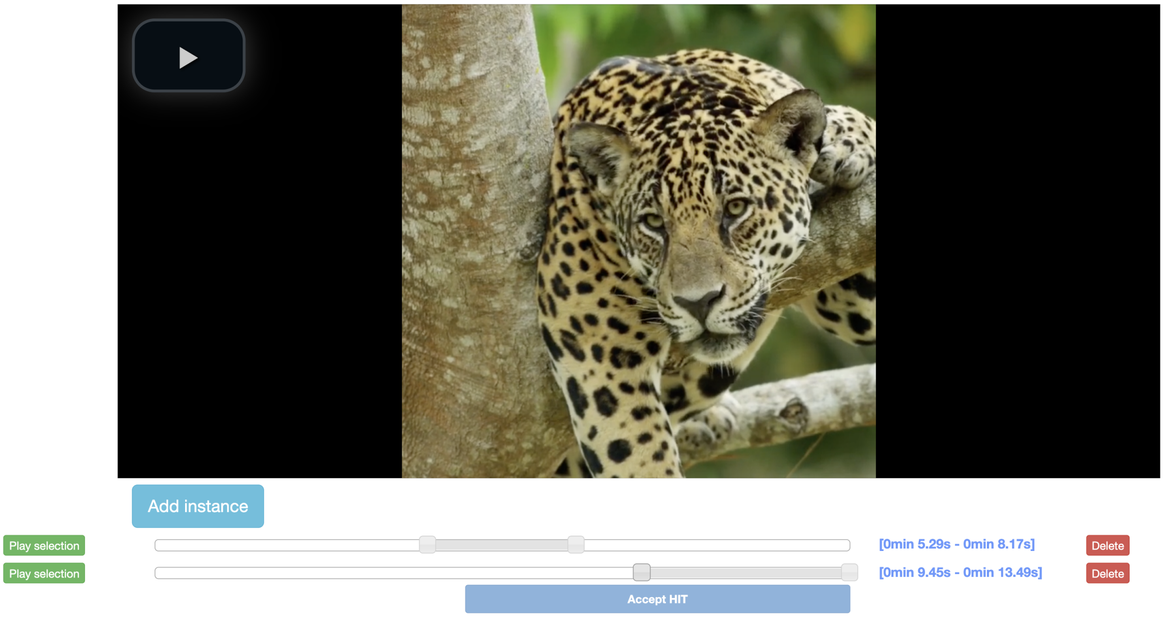


Figure 10. The animal behavior localization interface

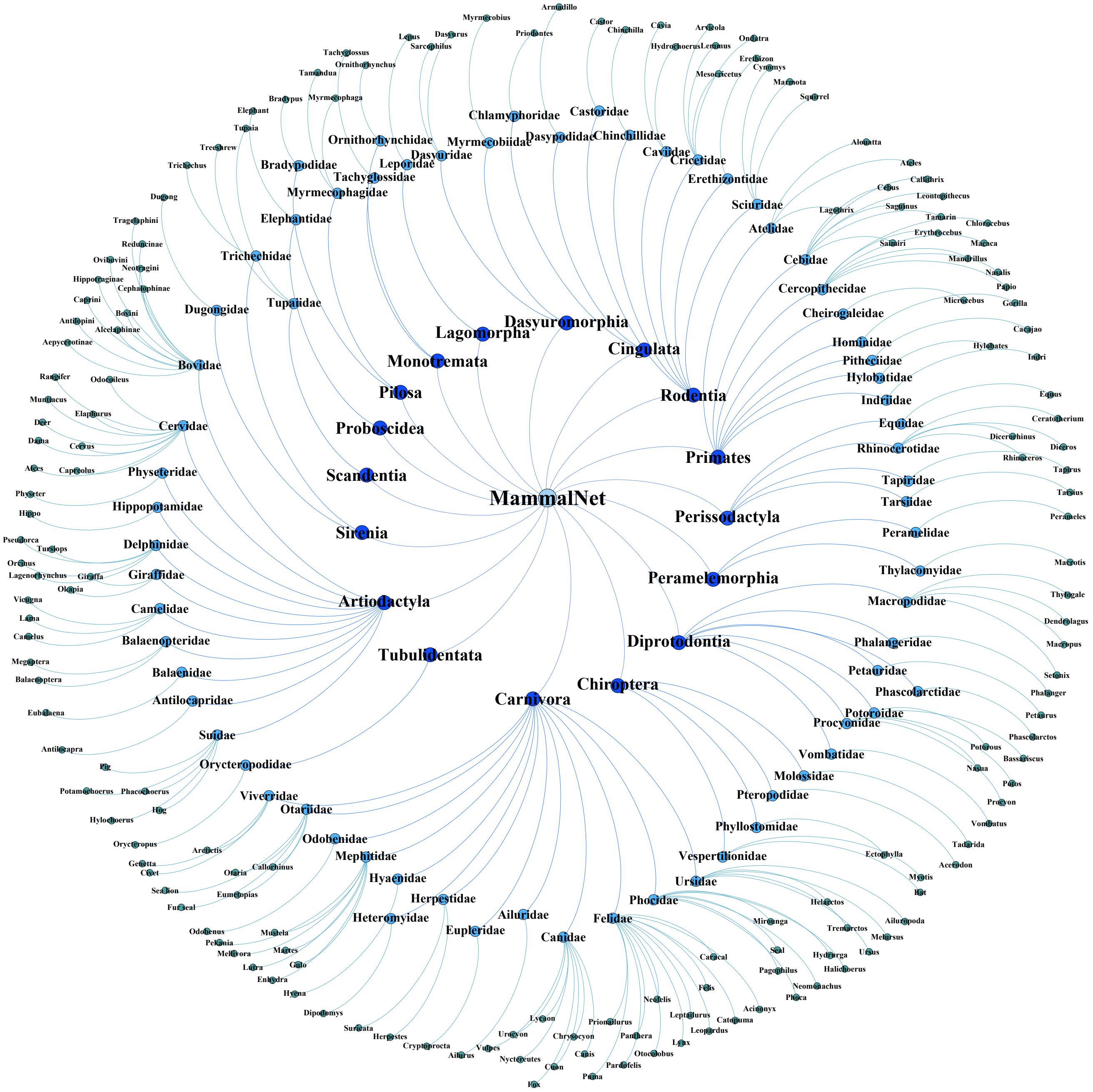


Figure 11. The full taxonomy of MammalNet. We show 17 orders, 69 families, 162 genera, 5 tribes and 6 sub-families in this full animal taxonomy.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[3] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 1

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1

[6] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981. 1

[7] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Temporal action detection with global segmentation mask learning. *European Conference on Computer Vision*, 2022. 2

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[9] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16010–16019, June 2021. 2

[10] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 2022. 1