PriVi: Towards a General-Purpose Video Model for Primate Behavior in the Wild

Supplementary Material

Table 5. **Overview of the composition of our dataset.** The species distribution of YT-Filtered was estimated by labeling a random subset of 900 frames, for R&O the species distribution is known.

	Subsets		PriVi
	YT-Filt.	R&O	
Unique Hours	250 h	174 h	424 h
Genus/Family [%]			
Macaques	63.1	14.1	43.0
Chimpanzees	7.8	35.7	19.3
Orangutans	4.1	0	2.4
Baboons	1.3	16.2	7.4
True Lemurs	< 1	22.7	9.8
Marmosets	< 1	5.7	2.1
Squirrel monkeys	< 1	5.4	2.0
Others	8.1	0	4.8
Not identifiable	9.4	0	5.5
No primate	6.0	0	3.5
Setting [%]			
In the wild	59.6	62.7	60.9
Wild-like	27.8	22.4	25.6
Indoors	4.1	14.6	8.4
Not identifiable	8.6	0	5.1

A. PriVi Dataset

See Tab. 5 for a breakdown of species and settings in PriVi, Tab. 6 for details about all subsets of PriVi, and Fig. 6 for example frames from the video snippets.

In the following paragraphs, we will give a brief description of each of the source datasets in R&O.

eulemur. Videos are collected from a social learning experiment with red-fronted lemurs which involved feeding boxes with different opening techniques. Lemurs wear collars for identification. Observed behaviors are various types of box interactions, scrounging, and looking at other lemurs. The experiments were filmed with fixed cameras from various camera angles (close-up, bird's eye view, and tree-mounted).

baboon_w. This dataset contains videos of a foraging experiment with wild Guinea baboons (*Papio papio*) in Niokolo Koba National Park in Senegal. The experiment consists of baboons learning to operate a food box that rewards them with peanuts for pulling a lever on its side. Once a sufficient number of baboons were trained to operate the lever one was chosen to be the sole operator for the remainder of the experiment and his social interactions were monitored. Video cameras were placed on fixed tripods a set distance from the box (around 8 m) as part of the set up for both training and experimental sessions with the box and remained on until the box was empty of food. During sessions the primary behavior of the baboons is feeding, and resting, with occasional bouts of aggression and allogrooming.

baboon_a. Miscellaneous, diverse footage of baboons at Niokolo Koba National Park in Senegal. Recordings are not from a single experiment, but focus on social behaviors like grooming and sitting together. Recordings contain some fights.

lemur. This dataset was collected during short focal follows of semi–free-ranging ring-tailed lemurs (*Lemur catta*). Video recordings were obtained using hand-held cameras, and subjects wore tracking collars for identification and monitoring.

assamese. Miscellaneous footage of Assamese macaques (*Macaca assamensis*) expressing natural behavior in the wild at Phu Khieo Wildlife Sanctuary, Thailand. Video data was collected only when visibility was good, i.e. often when the individuals were on the ground or at low canopy levels. Common behaviors include feeding, resting and social interactions. The camera was either hand-held or head-mounted.

barbary_a. Miscellaneous footage of semi-free ranging Barbary macaques (*Macaca sylvanus*) at La Forêt des Singes, Rocamadour in France. Most videos are of macaques walking through a tunnel made of wire mesh. These recordings were done to capture standardized videos of individuals in the population for individual identification and for determining group structure. The dataset also contains recordings not containing the wire mesh tunnel, most of which are of social interactions.

saimiri. Squirrel monkeys were tested at Zoo Basel and Zoo Zurich. At Basel, they lived in three connected indoor and two outdoor enclosures enriched with climbing structures. At Zurich, they were housed in two connected indoor and two outdoor enclosures, with indoor vegetation and large outdoor trees. Both groups were tested in the morning after the first feeding and filmed during an extractive foraging task in one of their home enclosures, with all group members able to access the apparatus mounted on a wooden table in the enclosure center.

barbary_t. This is footage of a risk assessment experiment with barbary macaques. A wooden box contains a rubber snake (high risk) or a cube with snake-like visual appearance

Table 6. **Detailed composition of our dataset.** The proportion of the different sub-datasets corresponds to the number of 3-s snippets in each sub-dataset. Note that we vary the sampling stride between datasets and sample overlapping snippets for small datasets. We also report the duration of unique video material of each dataset. We publish all data except for the chimpanzee subset.

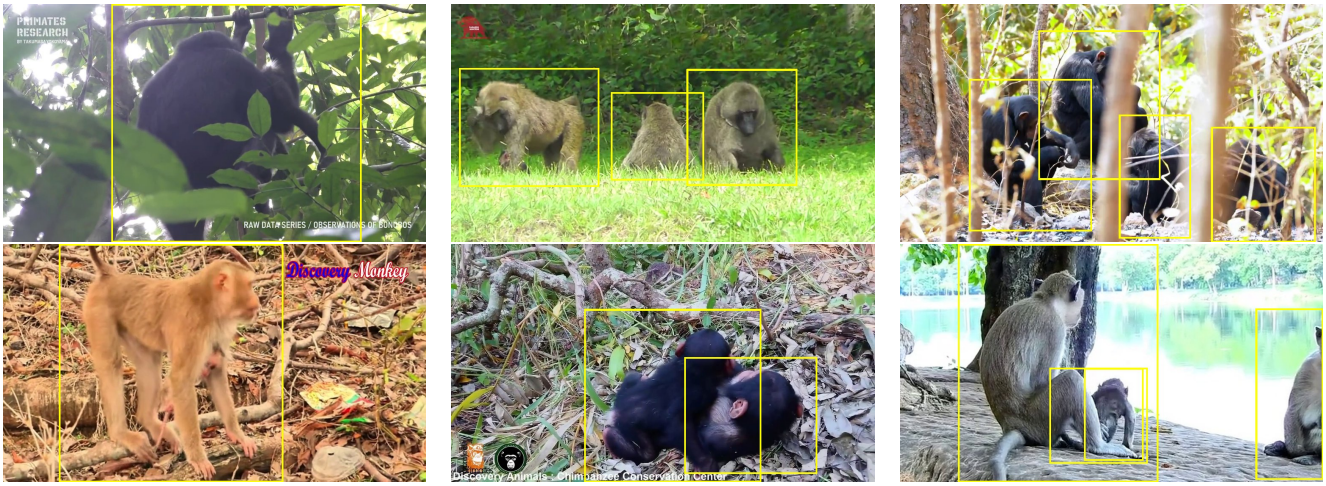
Name	Weight [%]	#Snippets	Unique Hours	Location	Setting	Species	Raw [h]
YT-Filtered	**63.0**	**454k**	**250.0**	diverse	diverse	diverse	**458**
R&O	**37.0**	**267k**	**173.7**				**721**
eulemur [31]	3.9	28k	23.2	Kirindy Forest, Madagascar	Wild, Experiment, Multi-Cam	Red-fronted lemurs (*Eulemur rufifrons*)	311
baboon_w [46, 47]	4.4	31k	26.1	CRP Simenti, Senegal	Wild, Experiment	Guinea baboons (*Papio papio*)	80
baboon_a	1.1	8k	2.2	CRP Simenti, Senegal	Wild	Guinea baboons (*Papio papio*)	2
lemur	4.5	32k	26.9	Affenwald Straußberg, Germany	Semi-free ranging	Ring-tailed lemurs (*Lemur catta*)	38
assamese [49]	0.1	1k	0.3	Phu Khieo Wildlife Sanctuary, Thailand	Wild	Assamese macaques (*Macaca assamensis*)	8
barbary_a	2.2	16k	9.0	La Forêt des Singes, Rocamadour, France	Semi-free ranging	Barbary macaques (*Macaca sylvanus*)	21
saimiri [60]	2.0	15k	12	Basel and Zurich Zoo, Switzerland	Captive	Black-capped squirrel monkeys (*Saimiri boliviensis*)	30
barbary_t [6]	1.6	12k	6.3	La Forêt des Singes, Rocamadour, France	Semi-free ranging, Experiment	Barbary macaques (*Macaca sylvanus*)	6
marmoset [60]	2.1	15k	12.5	University of Zurich, Switzerland	Captive, Multi-Cam	Common marmosets (*Callithrix jacchus*)	100
rhesus [61]	1.3	9k	7.7	German Primate Center, Germany	Captive, Multi-Cam	Rhesus macaques (*Macaca mulatta*)	62
chimpanzee (private)	13.7	99k	47.3	Moyen-Bafing National Park, Guinea	Wild	96.5 % Chimpanzees (*Pan troglodytes*), 3.5 % Guinea baboons (*Papio papio*)	63
PriVi	**100.0**	**721k**	**423.7**				**1179**

(low risk) and a peanut (high reward) or a popcorn (low reward). It was studied whether macaques take the food or not. The camera is fixed in the wooden box, recording outwards.

marmoset. Marmosets were housed in heated indoor enclosures with ad libitum access to outdoor spaces. Both areas were enriched with natural branches, climbing structures, wood chips, or soil with vegetation. Animals were filmed during an extractive foraging task in their home enclosure, with all group members able to access the apparatus mounted on a wooden table placed in the enclosure center.

rhesus. Recordings are of a foraging experiment with one or two rhesus macaques in a white experiment room. The floor in the experiment room is covered with piles of wood chips, some of which contain varying amounts of food. It was observed whether macaques learn in which piles they can expect food. Recordings were captured with several fixed cameras on walls and ceiling.

chimpanzee. This dataset was collected by the Moyen Bafing Chimpanzee Project and consists of camera-trap footage of chimpanzees in a savanna mosaic in Guinea. Bushnell 4K no-glow cameras were deployed to document a wide range of behaviors, including tool-use. This dataset cannot be published yet as it is part of an ongoing research project.

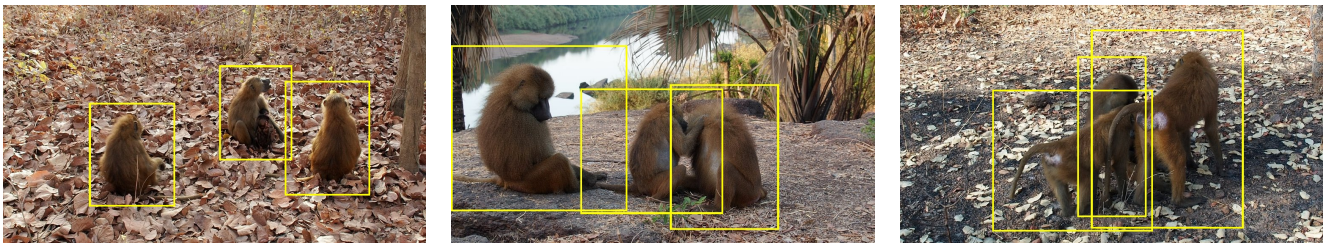


(a) YT-Filtered

(b) eulemur

(c) baboon_w

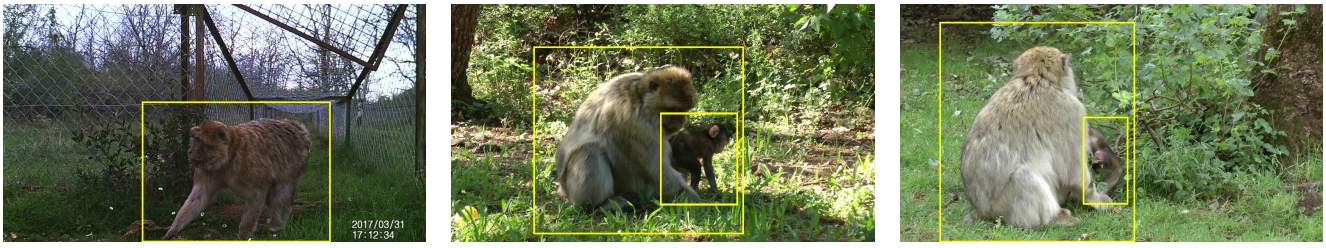
Figure 6. **Examples of PriVi dataset samples.** Examples sorted by sub-dataset. We only show center frames of the video snippets, please see the supplementary material for videos. Bounding boxes of detected primates in yellow. (continued on next page)



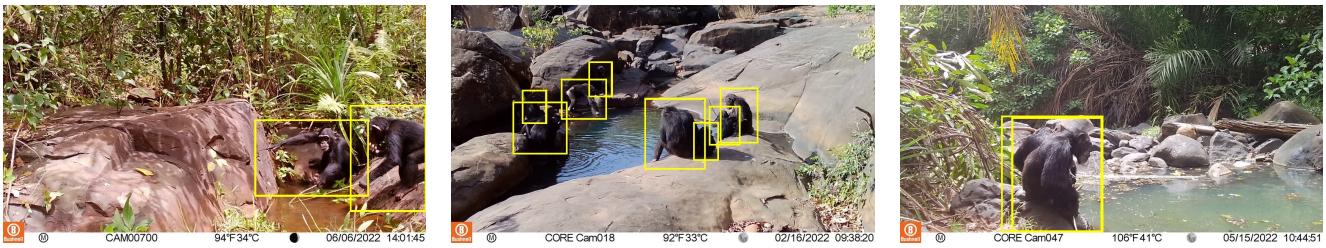
(d) baboon_a

(e) lemur

(f) assamese



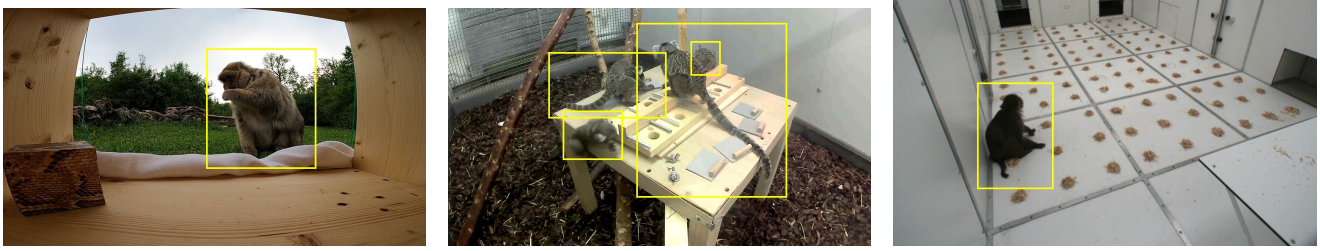
(g) barbary_a



(h) chimpanzee

Figure 6. **Examples of PriVi dataset samples.** Examples sorted by sub-dataset. We only show center frames of the video snippets, please see the supplementary material for videos. Bounding boxes of detected primates in yellow. (continued on next page)

(i) saimiri



(j) barbary_t (k) marmoset (l) rhesus

Figure 6. **Examples of PriVi dataset samples.** Examples sorted by sub-dataset. We only show center frames of the video snippets, please see the supplementary material for videos. Bounding boxes of detected primates in yellow.

B. Data Pipeline

Relevance Labeling. Labeling was performed on randomly selected center frames from YouTube snippets after cut detection and discarding videos shorter than 3 seconds.

Labeling criteria: *Label a video as relevant if the video visibly contains a primate (drawings, dolls, reflections, or computer generated content do not count) or if it is a recording of the natural habitat of primates (e.g. forest, savannah) and might contain a primate. In addition: Humans in a video are acceptable, if primate(s) and not the humans are the main focus of the video. If humans are the main focus of the video, label as irrelevant. If it is hard to recognize anything on an image (blurry, bad lighting), label as irrelevant. If a video contains text overlaying large parts of the video, label as irrelevant (Logos or text in the corners is fine).*

CLIP embeddings and relevance filter. We use `laion/CLIP-ViT-H-14-laion2B-s32B-b79K` from HuggingFace to extract CLIP embeddings [59]. We utilize a 2-layer MLP with input dimension 1024, the CLIP embeddings' dimension, and hidden dimension 256. A dropout layer with probability 0.7 is also applied. We train for 18 epochs using Adam with a learning rate of 10^{-4}. We use a relevance threshold of 3.5.

Cut Detection. We utilize PySceneDetect [14] for detecting cuts in videos. PySceneDetect calculates pixel value differences between frames in HSV colorspace and splits videos at locations with high change. We are using the adaptive detector with a threshold of 3.0.

Data Contamination Screening. We compare all samples of the target datasets and YT-Filtered with a CLIP similarity > 0.9 and found none of ChimpACT, ChimpBehave or BaboonLand within YT-Filtered. Only 4 s of video material from the training set of PanAf500 (without any labels) overlap with YT-Filtered. For another train video, we found few YT-Filtered snippets recorded from the same camera trap at different times.

Primate Detection. We are using `IDEA-Research/grounding-dino-base` from HuggingFace as zero-shot object detector [38]. We use the prompt "monkey.primate.ape." and a box confidence threshold of 0.35.

C. Evaluation Protocol

See Tab. 7 for details about our four target datasets for evaluation.

PanAf500 [9] consists of 125 minutes of camera trap videos from 18 field sites in tropical Africa capturing chimpanzees and gorillas. Following the established protocol [9, 56], we train and evaluate on 16-frame miniclips cropped to primates using ground truth bounding boxes. Each miniclip shows one of nine behaviors.

Table 7. **Properties of our target datasets for evaluation. CV**: 5-fold cross-validation instead of a test set. **ML:** multi-label. **SL:** single-label. **Num. Samples:** ChimpACT: Number of annotated (not-interpolated) frames, i.e. every 10th frame. Others: Number of miniclips.

	Num. Samples			Classes		Evaluation Protocol
	train	val	test			
ChimpACT	50k	6.8k	7.6k	23	ML	frames w/ det.
PanAf500	9k	0.8k	1.9k	9	SL	miniclip (16f)
BaboonLand	17.8k	-	5.6k	13	SL	miniclip (90f)
ChimpBehave	9.2k	-	CV	7	SL	miniclip (20f)

ChimpBehave [22] is a dataset of chimpanzees in an indoor enclosure at the Basel zoo. Videos were captured using handheld cameras and behavior recognition is evaluated on 20-frame miniclips cropped to primate bounding boxes and only featuring a single behavior. The dataset features seven behavior classes, which are a subset of the PanAf500 classes. Instead of a dedicated test set, ChimpBehave utilizes five-fold cross-validation.

BaboonLand [18] consists of 18 drone recordings of wild-living olive baboons residing in Mpala, Kenya. From 30 min of densely annotated 5.3 k resolution drone footage, Duporge et al. [18] extracted 20 h of spatio-temporal miniclips, centered on each animal. Each miniclip is annotated into one of twelve behavior classes plus an additional class for occlusions. Annotation is single-label per miniclip with majority vote over per-frame labels.

ChimpACT [40] contains 2 h of video footage of chimpanzees recorded at the Leipzig Zoo. It features frame-wise bounding boxes and multi-label behavior annotations across 23 classes. Two different behavior recognition tasks exist: one with [40] and one without [41] access to ground truth bounding boxes. We predict the labels for a primate i at frame j by sampling a 64-frame miniclip around frame j and producing a crop centered at i's bounding box with padding to incorporate scene context.

Evaluation Metrics. Prior works used differing naming conventions for evaluation metrics. BaboonLand [18] refers to accuracy as Top-1 Micro and to balanced accuracy as Top-1 Macro. ChimpBehave [22] refers to accuracy as Top1 and to balanced accuracy as MCA.

Training Data Size Ablation. For our evaluation of how well our model and X3D perform with less labeled data (Sec. 5.3), we create subsets of the dataset with 50 %, 25 %, and 10 % of the training dataset. We randomly sample three different versions of the dataset per desired size. During

sampling we ensure that there is at least one sample of each class in each subset.

D. Implementation Details and Prior Work

Our method. We always train five attentive classifiers, evaluate each on the evaluation dataset, and report the mean evaluation score.

Origin of prior work results (Sec. 5.2, Tab. 2). On ChimpACT, results for InternVideo-L and VideoPrism-g are from [73] and results for AlphaChimp and X3D are our own reproduction. On PanAf500, results for ChimpVLM are from [8]. All other results are from the respective dataset papers (ChimpACT [40], PanAf500 [9], BaboonLand [18], ChimpBehave [22])

AlphaChimp. There are two different evaluation protocols for ChimpACT: AlphaChimp [41] and PySlowFast-based methods [40] calculate mAP over 18 classes, excluding five tail classes with very low support (begging, being begged from, taking object, losing object, erection), and perform early stopping on the test set, while results reported in VideoPRISM [73] report mAP over 23 classes and do not utilize the test set for model selection. We decided to follow the VideoPRISM protocol and reproduce AlphaChimp results using this protocol for fair comparison. Using the official AlphaChimp [41] implementation and leaving hyperparameters unchanged, we are able to match the reported results in [41] (36.5 mAP in our reproduction; reported results are 34.3 mAP). We then switch to early stopping on validation instead of test and calculating mAP over all 23 classes, leaving everything else constant. As we noticed high variance in the evaluation scores obtained after each epoch, we report the mean over 5 runs. This yields a mAP of 25.35 for AlphaChimp.

X3D. We are using X3D-L for all experiments, keeping the original input size of $16 \times 312 \times 312$. For scaling X3D with less labels (Fig. 5) on PanAf500, we utilize the PySlowFast codebase, providing the dataset in Kinetics format. For reporting X3D on ChimpACT (Tab. 2) and for scaling X3D with less labels on ChimpACT (Fig. 5), we predict on mini-clips using our ChimpACT dataloader instead of following the AVA protocol, matching the evaluation setting we are using.

E. Additional Results for Our Method

Performance Breakdown. Figure 7 shows a breakdown of model performance by action class on ChimpACT and PanAf500. For chimpanzees in a zoo setting (ChimpACT), we find good performance for classes with visually distinct appearances or characteristic motion patterns, like eating (89 % mAP), carrying (75 %), resting (73 %), or playing (72 %).

Table 8. **Joint domain- and dataset-level pretraining reliably mitigates catastrophic forgetting.** Results are on *val* sets.

	ChimpACT		PanAf500	
	mAP	mAP_w	Acc	B-Acc
PriVi	38.75	54.32	89.65	79.95
+ DaLP: ChimpACT	41.43	57.22	**84.93**	**69.05**
+ DaLP: PanAf500	**32.90**	**48.91**	90.53	87.29
Joint PriVi & ChimpACT	40.68	56.39	**89.91**	**80.89**
Joint PriVi & PanAf500	**37.93**	**55.80**	91.97	85.74

For chimpanzees in the wild (PanAf500, camera traps), the majority classes (sitting, walking, standing, 85-91 % Acc), as well as rare classes with distinct motion (climbing up and down, 87-89 % Acc) perform well. Sitting on back (14 % Acc) and hanging (46 %) underperform, we speculate this is due to PanAf500's evaluation protocol of cropping to bounding boxes, which loses global information.

For ChimpBehave, we see a high performance in general, with 90-98 % Acc for sitting, standing, walking each. For baboons in the wild (BaboonLand, drone footage), we observe a similar pattern as with ChimpACT, with Walking/Running (95 % Acc), Sitting/Standing (87 % Acc) and Drinking (83 %). Behaviors requiring fine-grained information, like Self-Grooming, Being Groomed, Grooming Somebody, Mutual Grooming perform worse (14-38 % Acc).

Joint pretraining instead of DaLP. To mitigate catastrophic forgetting, we experiment with joint pretraining on PriVi and the target dataset instead of first training on PriVi and then performing dataset-level pretraining (DaLP) on the target dataset. We start with the V-JEPA checkpoint and jointly pretrain for 75 k steps with a 50:50 mix of PriVi and the target dataset. We find that joint pretraining improves performance on the selected target dataset without reducing performance on other target datasets, see Tab. 8. While this requires considerably more training compute than DaLP, it might be good option for scenarios where catastrophic forgetting is a concern.

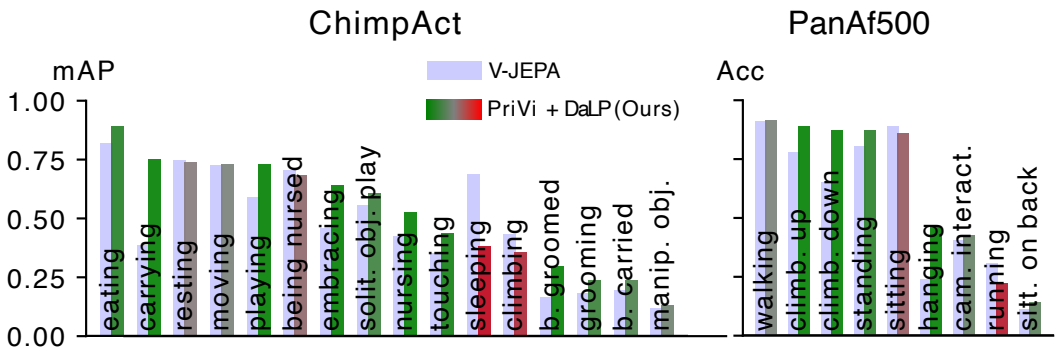


Figure 7. **Performance of our method (PriVi + DaLP) by action class.** Color gradient shows improvement ■, comparable results ■, or deterioration ■ compared to V-JEPA ■.