The SA-FARI Dataset: Segment Anything in Footage of Animals for Recognition and Identification

Supplementary Material

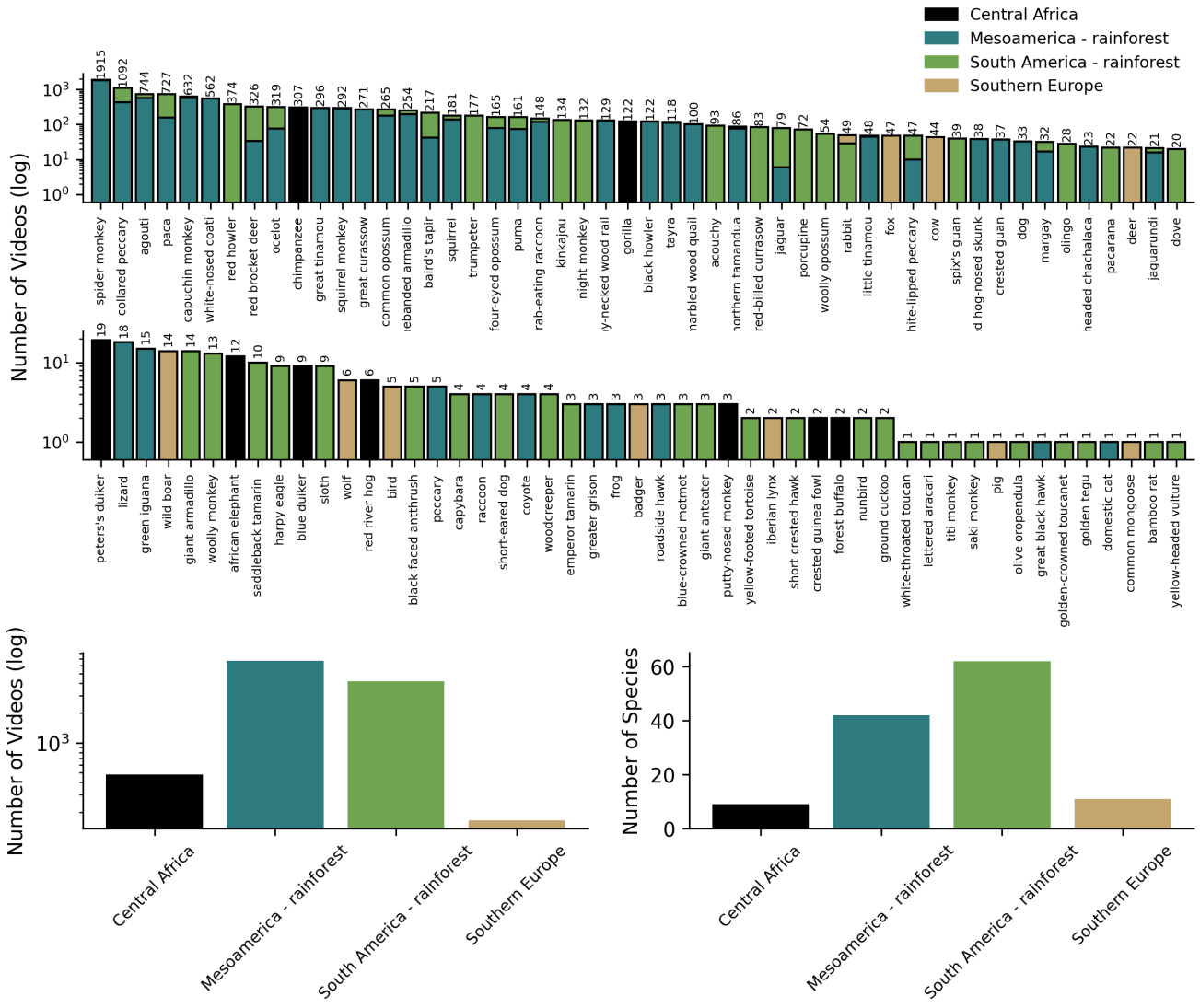
A. Additional Dataset Statistics

Fig. S1 shows the distribution of species annotations per continent, and Fig. S2 shows the distribution of the number of videos per location.

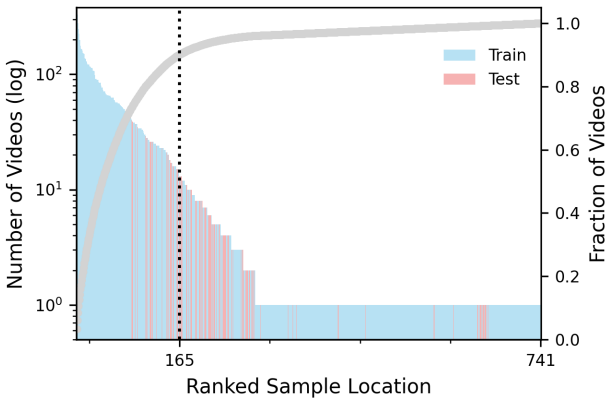
B. Metrics

In this section we provide an overview of the segmentation and tracking metrics we use in § 4. For more details, we refer the reader to the SAM 3 [12] paper.

IDF1. IDF1 measures the accuracy of maintaining object identities in multi-object tracking. It calculates the ratio of correctly matched detections (where both detection



Supplementary Figure S1. **Video and Species Category Distribution per SA-FARI Ecoregion.** Video counts per species category and per continent.



Supplementary Figure S2. **Number of Videos per Camera Trap Location (Ranked).** Vertical line marks the minimum number of locations that account for 90% of the videos. Note, train and test videos come from different locations.

and identity are correct) to the average number of ground-truth and predicted detections, balancing precision and recall while penalizing identity switches.

HOTA. Higher Order Tracking Accuracy (HOTA) jointly evaluates detection and association in tracking. It decomposes performance into detection accuracy (DetA) and association accuracy (AssA), providing a balanced view of how well objects are detected and tracked over time [31].

pHOTA. Phrase-based HOTA (pHOTA) adapts HOTA for open-vocabulary tracking. Each video–noun phrase pair is treated as a unique sample, enabling class-agnostic evaluation. This is ideal for open-world settings, where tracked objects are specified by phrases rather than fixed categories.

TETA. Track Every Thing Accuracy (TETA) builds upon the HOTA metric, while extending it to better deal with multiple categories and incomplete annotations. It consists of three parts: a localization score, an association score, and a classification score [29].

cgF$_1$. Classification-gated F1 (cgF$_1$) is designed for open-vocabulary segmentation and tracking. It combines localization quality (positive micro F1, pmF$_1$) and classification ability (image-level Matthews correlation coefficient, IL_MCC), rewarding models that are both accurate in localizing objects and calibrated in predicting their presence. This metric addresses the limitations of traditional AP in large label spaces.

C. Baseline Selection

This section provides additional analyses that motivate our choice of SAM 3 as the unique baseline. We support this choice in two ways: (1) by showing SAM 3 is the strongest model on public benchmarks that are independent of SA-FARI, and (2) by showing that SA-FARI can also benefit a non-SAM 3 model.

Model Metric	**LVVIS** mAP	**BURST** HOTA	**GMOT40** HOTA
GLEE	9.3	20.2	29.9
LLMDet + SAM3 Tr	15.2	33.3	24.9
SAM3	**36.3**	**44.5**	**60.3**

Supplementary Table S1. **Results on selected public benchmarks.** Performance on LVVIS (mAP) and BURST/GMOT40 (HOTA). This table serves as a non-SA-FARI reference point for comparing models.

Choice of SAM 3 as a baseline. We use SAM 3 as our primary baseline because it is a state-of-the-art model for open-vocabulary video segmentation and tracking, and it provides a strong reference point for evaluating SA-FARI. To justify this choice, we compare SAM 3 against other representative open-vocabulary baselines on selected public video benchmarks that do *not* involve SA-FARI (Tab. S1). SAM 3 performs substantially better in this setting, motivating our use of SAM 3 as the main baseline throughout the paper.

SA-FARI helps models beyond SAM 3. To check that SA-FARI is not only beneficial for SAM 3, we ran a small-scale study fine-tuning LLMDet on 10k frames from SA-FARI. Fine-tuning improves AP from 32% to 47% on the test set (@1 fps), suggesting that the value of SA-FARI extends beyond the SAM 3 family of models.

Note, results on publicly available datasets (Tab. S1) also serve as a non-SA-FARI neutrality test, indicating that SAM 3's advantage does not stem from pseudo-labelling with SAM family models. Another key comparison is SAM 3 vs. SAM 3 FT in the main text. As SAM 3 generates the pseudo-labels, any circularity would favor the base model (or all SAM 3 variants equally), not the fine-tuned one specifically.

D. Annotation Details

Annotation quality and exhaustivity auditing. All samples in SA-FARI are audited for annotation quality and exhaustivity. The verification rejection rate is 8.4%. Inter-rater agreement was measured on a subset of 867 samples, yielding a 90.7% pass rate among three annotators (all confirmed exhaustive, high-quality masklets). Rejected jobs were redone until satisfactory, so all SA-FARI annotations passed verification for exhaustivity and mask quality. Annotations were produced by a pool of experienced annotators; we selected the top-performing 20% based on audit results.

Video frame rate (6fps) and release artifacts. All videos (train and test) are uniformly downsampled to 6fps due to annotation cost, prioritizing broader coverage of videos over denser per-video annotation. All results are

reported on the downsampled videos. For reproducibility, we publish both the original videos and the downsampled videos used for annotation and evaluation.

E. SAM 3 Fine-Tuning Details

We describe the training recipes for both fine-tuning variants: **SAM 3 FT** (fine-tuned exclusively on SA-FARI) and **SAM 3 (SA-FARI)** (SAM 3 training mixture with SA-FARI included).

Base Model. Both variants initialize from a SAM 3 checkpoint. The architecture consists of a ViT-L+ vision backbone (d=1024, 32 layers, 16 heads, MLP ratio 4.625, patch size 14) with RoPE positional embeddings and windowed attention (window size 24, global attention at layers 7, 15, 23, 31). The text encoder is a 24-layer transformer (d=1024, 16 heads). The detector uses a 6-layer deformable transformer encoder with cross-modal fusion and a 6-layer deformable transformer decoder with 200 queries and iterative box refinement. Segmentation uses a 3-stage upsampling pixel decoder with a presence head. Input resolution is 1008×1008.

SAM 3 FT. SAM 3 uses a two-stage inference pipeline: a detector that produces per-frame detections, followed by a SAM 2 tracker that propagates masks across the video. We only fine-tune the detector, and the SAM 2 tracker is used with its original pretrained weights. Within the detector, we freeze the vision backbone and fine-tune the remaining components (text encoder, transformer encoder/decoder, and segmentation head) for 8 epochs. We use AdamW with gradient clipping (max norm 0.1) and weight decay of 0.1 (excluding bias and LayerNorm parameters). Learning rates are set to 0.5% of the original SAM 3 training rates: 8×10^{-6} for the transformer encoder/decoder and segmentation head, and 5×10^{-7} for the text encoder. We apply an inverse square root schedule with 1,500 warmup steps and 750 cooldown steps. Training uses bfloat16 mixed precision with activation checkpointing on all major components.

SAM 3 (SA-FARI). This variant uses the standard SAM 3 training recipe with SA-FARI included at 2% of the training mixture. All hyperparameters remain unchanged from SAM 3 training.

Losses. Both variants use the standard SAM 3 losses: mask focal loss (α=0.25, γ=2.0, weight 200), mask dice loss (weight 10), box L1 loss (weight 5), box GIoU loss (weight 2), and focal classification loss (weight 100, γ=2). Object-query matching uses a Hungarian matcher with focal classification cost (2.0), L1 box cost (5.0), and GIoU cost (2.0), supplemented by one-to-many matching (top-k=4, weight 2.0).

Data Augmentation. We apply multi-scale random resize (400/500/600 pixels), random crop (384–600 pixels), final resize to 1008×1008, normalization (mean and std 0.5), and random blur (p=0.2).