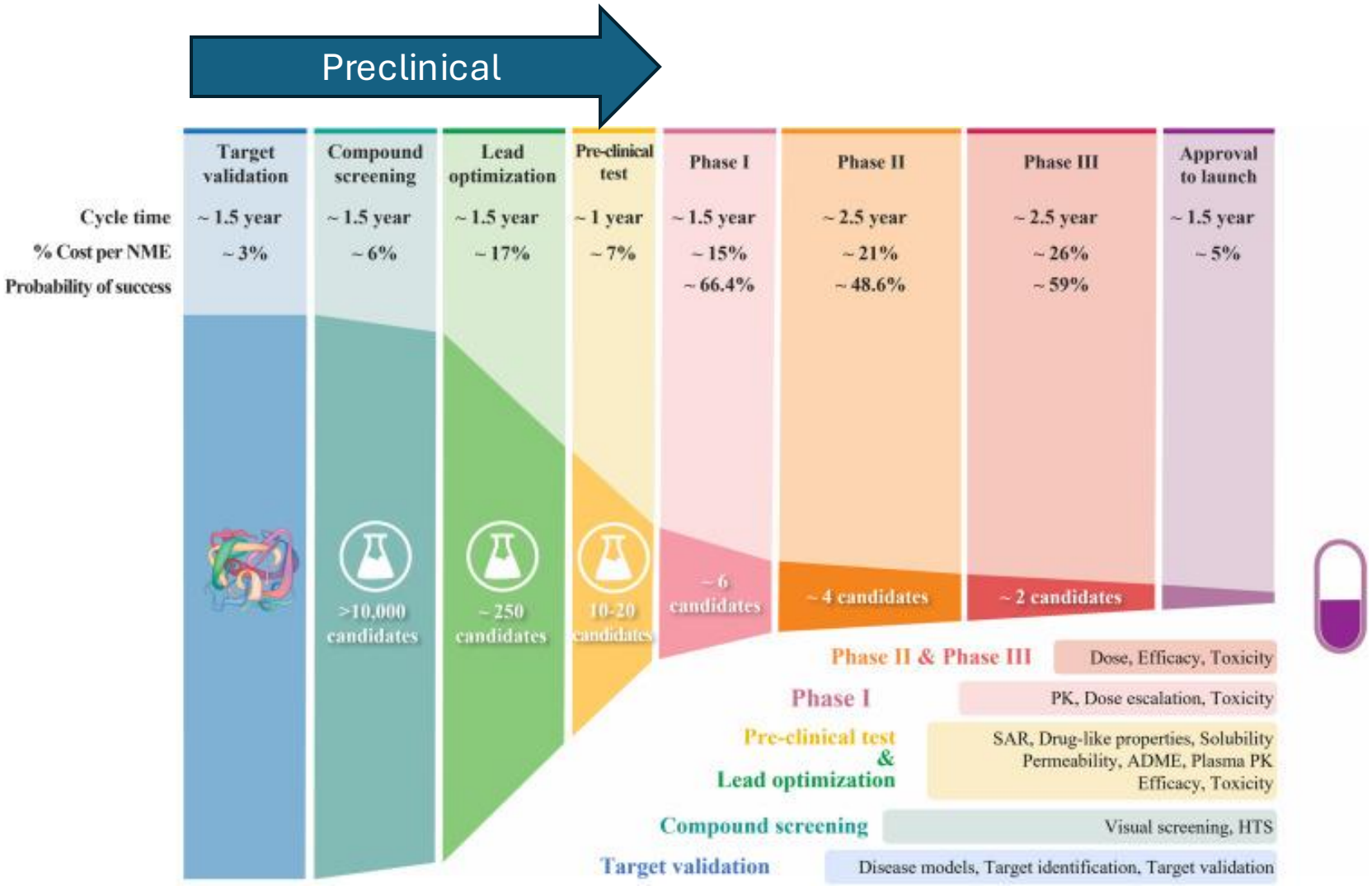
Robust Multi-Object Tracking for In Vivo Behavioral Phenotyping

Reuben Retnam, Pietro Artoni,
Tamas Kiss, Zsigmond Benko

Takeda Discovery Statistics

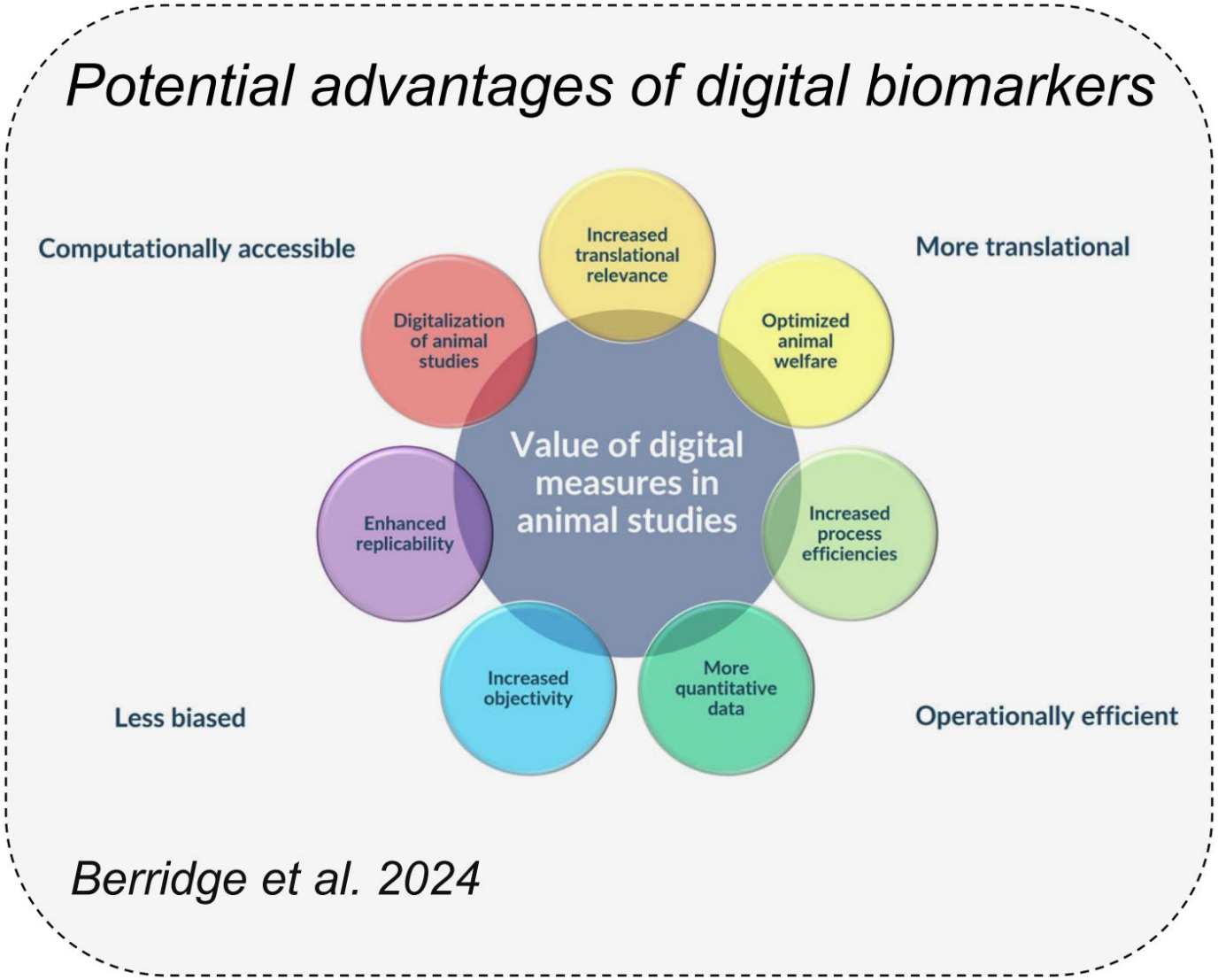
How do we efficiently sort compounds in the preclinical phase?



How do we obtain actionable data from in vivo studies?

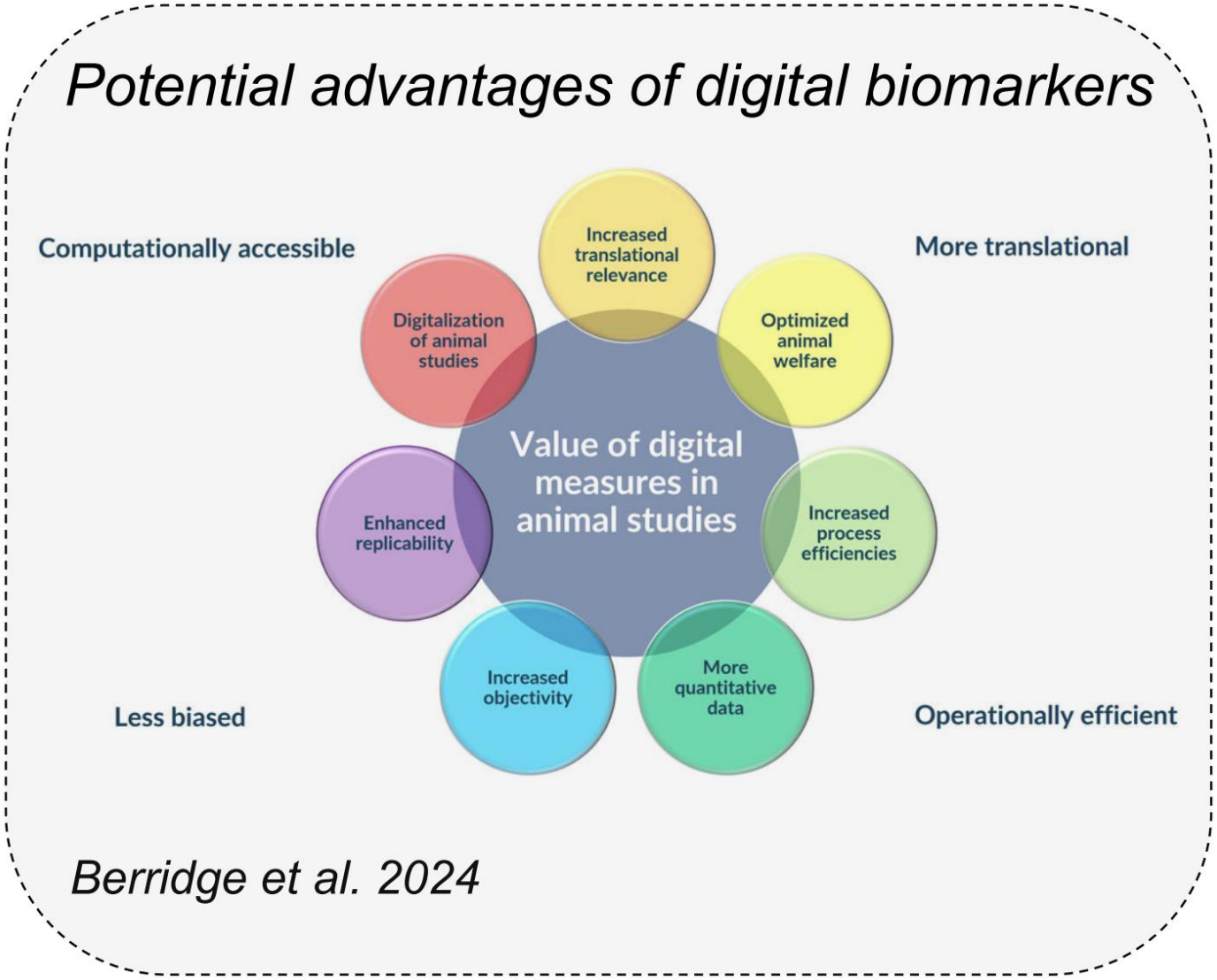
Video: A Rotarod test

What's the Utility of Digital Biomarkers In Vivo?



- Bring the experiment to the animal, not the animal to the experiment

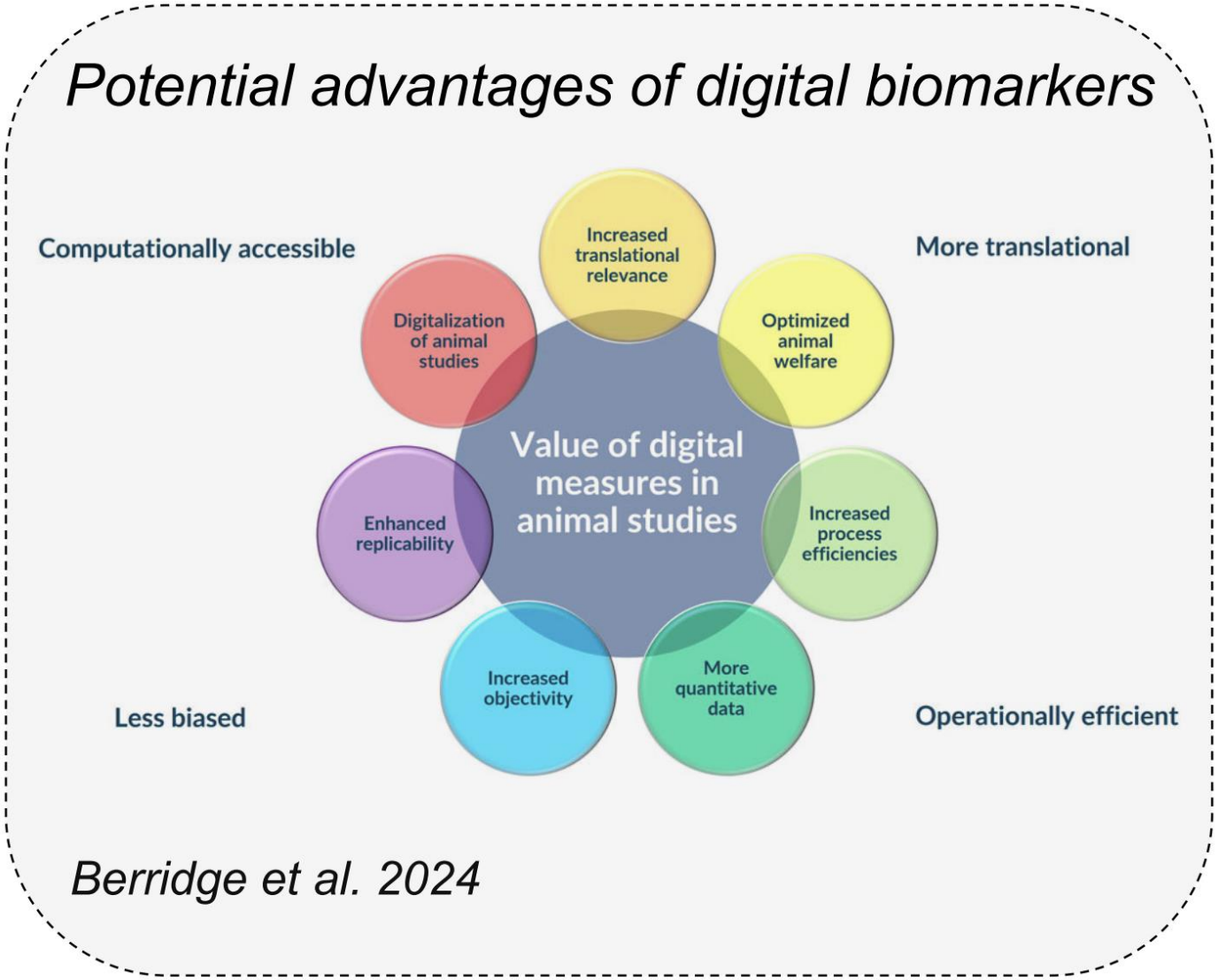
What's the Utility of Digital Biomarkers In Vivo?



- Bring the experiment to the animal, not the animal to the experiment

Figure: A Home Cage Monitoring System

What's the Utility of Digital Biomarkers In Vivo?



- Bring the experiment to the animal, not the animal to the experiment

> Cancers (Basel). 2023 Sep 29;15(19):4798. doi: 10.3390/cancers15194798.

Activity in Group-Housed Home Cages of Mice as a Novel Preclinical Biomarker in Oncology Studies

> Genes Brain Behav. 2017 Jun;16(5):564-573. doi: 10.1111/gbb.12374. Epub 2017 Mar 29.

Decreased home cage movement and oromotor impairments in adult Fmr1-KO mice

S J Bonasera [1], T R Chaudoin [1], E H Goulding [2], M Mittek [3], A Dunaevsky [4]

> Proc Natl Acad Sci U S A. 2007 Feb 6;104(6):1983-8. doi: 10.1073/pnas.0610779104. Epub 2007 Jan 29.

The power of automated high-resolution behavior analysis revealed by its application to mouse models of Huntington's and prion diseases

Other reasons to use digital biomarkers

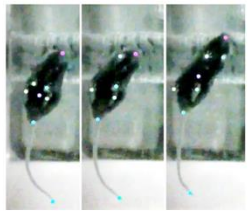
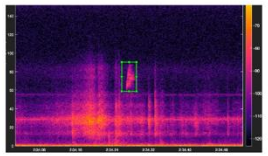
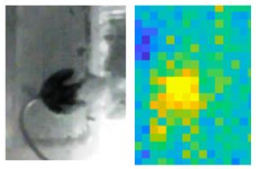
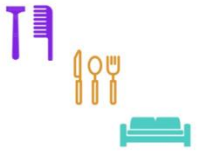
- Digital biomarkers enable longitudinal measurements for low researcher effort.

- Adaptive designs. We can use cues to know when to enroll an animal in a study, for example when its behavioral phenotype is sufficiently altered.
  - Magnify the phenotype (model window)
  - Greater chance to detect a noticeable treatment effect

- …and reverse digital translation

What digital metrics are useful in humans?

Reverse digital translation

Preclinical		Clinical	
Quadruped gait		Biped gait	
Ultrasonic Vocalizations		Speech biomarkers	
Temperature monitoring		Temperature monitoring	
Core behaviors		Quality of life	

Challenges in Digital Biomarker Development

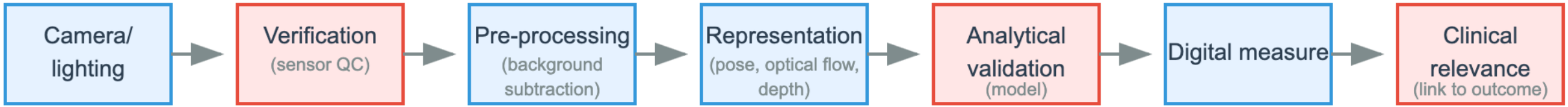
- Biomarker Choice
  - Does the biomarker describe an important correlate of the pathology we are exploring?
  - What evidence does that correlate provide?
- Value Demonstration
  - Does the biomarker add new evidence, complement current evidence, or neither?
  - How do we gain consensus from preclinical and clinical teams?
- Data Engineering and Analysis
  - Can we turn additional data into better evidence?
  - How do we follow good statistical practices in our design and analysis?

Targeted Biomarkers through Supervised ML

- Activity <> metabolic expenditure & frailty indices.
- Respiration <> pulmonary toxicity & ALS progression.
- Temperature <> inflammation
- Sleep architecture <> neurodegeneration & sleep
- Supports DIVA goal of biomarkers with clear mechanism ties

- Use annotated frames or multimodal ground truth
- NNs for frame-wise detection; Transformers for temporal context; pose-based classifiers using key-point kinematics
- Analytical validation: AUROC vs. expert labels, distance from expert label

From Video to Digital Biomarker



Video Acquisition & Verification

- Many home-cage/other systems out there
- Top-, side-, multi-view cameras remove occlusions
- IR illumination keeps circadian cycle undisturbed
- Automated QC: focus drift, occlusion detection
- V5 “Verification” requirement: prove raw signal fidelity over study duration

Figure: Video from YouTube of a Home Cage Monitoring System

A Brief History of Home-Cage Analytics

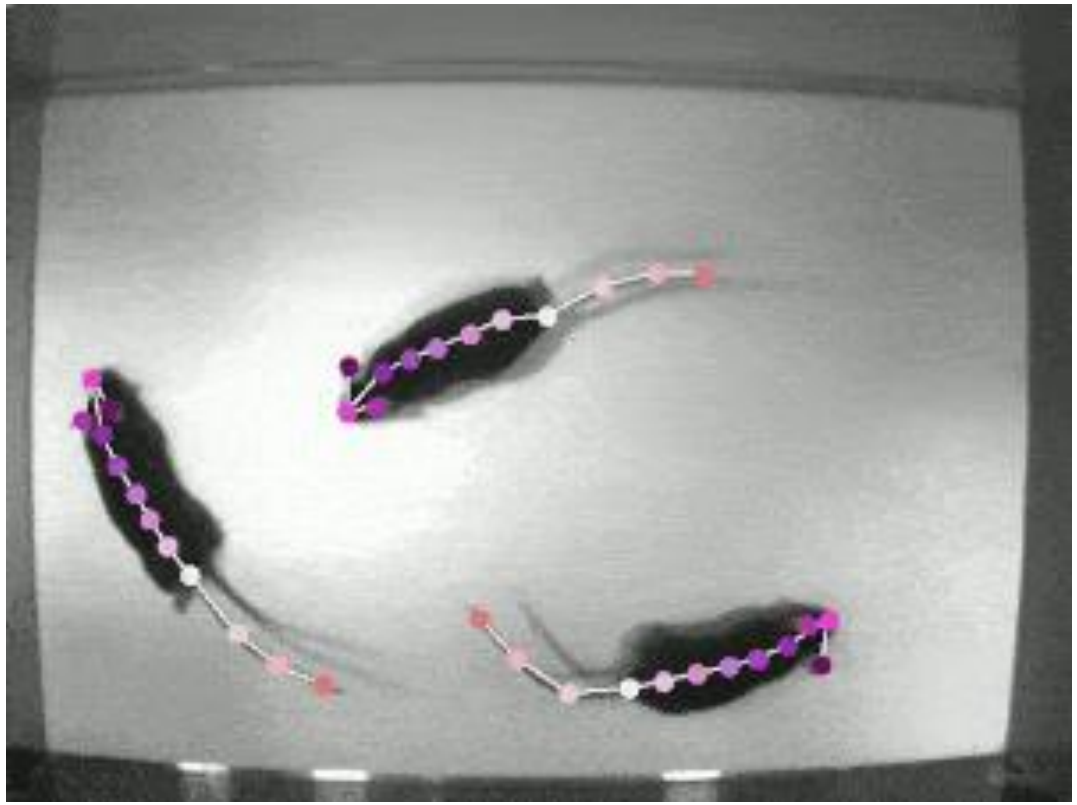
- Early home-cage systems (e.g., HomeCageScan, EthoVision unsupervised modules) relied on background subtraction + blob tracking
- Breakthroughs: LEAP (2018), DeepLabCut (2018) made key-point tracking feasible with ~hundreds of labelled frames
- These pose coordinates enabled engineered kinematic features (gait speed, limb angle) and density-based clustering (UMAP + HDBSCAN) to separate behaviors.
- Unsupervised:
  - MoSeq (Datta 2017) used depth + a HMM to identify sub-second motifs in behavior
  - VAME (Luxem 2022) trained VAEs on pose to discover motifs in freely moving mice
  - B-SOID clusters pose data then trains a supervised model to predict cluster membership on new data
- Supervised:
  - DeepEthogram (Bohnslav 2021) uses CNNs previously trained for human action recognition to predict specific behaviors

A Review of Multi-Object Tracking Tools

- Tracking by detection: SORT/DeepSORT (2016-2017) use Kalman filter + Hungarian matching to match tracks to detected objects; DeepSORT added appearance features
- ByteTrack (2021) added a second association step for low confidence detections, assessing their similarities with tracklets
- Joint detection and embedding trackers were popular in the early 2020s
- Transformer based trackers, such as those based on Segment Anything 2 like SAM2MOT (2025), are now popular
- Other trackers such as ConsistencyTrack (2024) frame tracking as a denoising problem

DeepLabCut & Markerless Pose Estimation

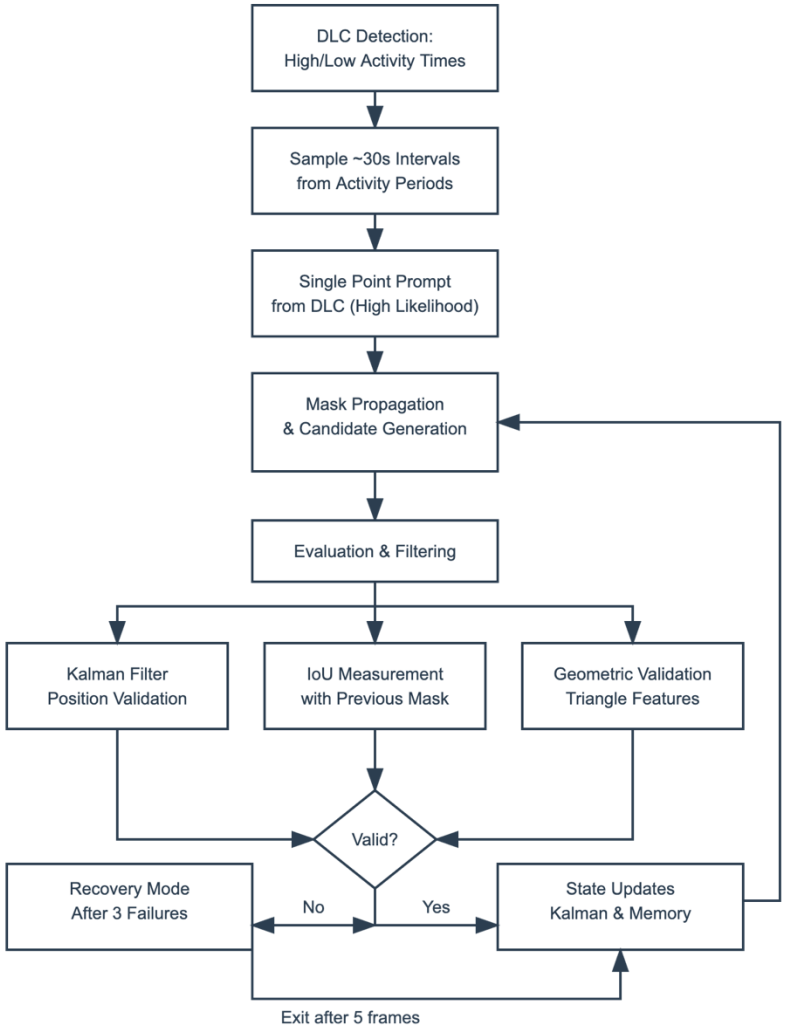
- Released in 2018, DeepLabCut performs pose estimation with ~hundreds of labelled frames
- Uses pre-trained ResNet models to predict keypoints in animals
- Easy to train, but quality is sometimes insufficient for more complex markers



Video: Example of DeepLabCut Pose Estimation (from the DLC website)

Our Analysis Pipeline

- Utilize DeepLabCut predictions to prompt a modified version of SAM 2 to track specific parts of the animal more accurately
  - Predict bouts of interest (e.g. movement); stratified sample equally from these bouts and other times
  - Use stable keypoint predictions to initialize tracking with a larger model
- Uses Segment Anything 2 as a base model, but adds various problem-specific features
  - Distractor-resolving, high-quality, and dynamic memory for multiple objects
  - Geometric constraints
  - Unique implementation of Kalman filtering



SAM2-based Workflow for Object Tracking

Algorithm Overview: Mask Generation

- Mask Propagation & Candidate Generation:
  - The SAM 2 predictor takes previous-frame masks as prompts → propagates to next frame, yielding 3 candidate masks per object
- We have $a_i = f\left(x_i | m(a_k, p_1)\right)$, where
  - a_i is the mask for frame i
  - x_i are the features from frame i
  - $m(a_k, p_1)$ are memory embeddings based on the prompts from frame 1 and a certain amount of previous masks
- Memory embeddings are generated from
  - Spatial feature maps (image embeddings from past frames)
  - Object Pointers (mask features from mask decoder)
- Model implements cross-attention between current frame features and memory bank

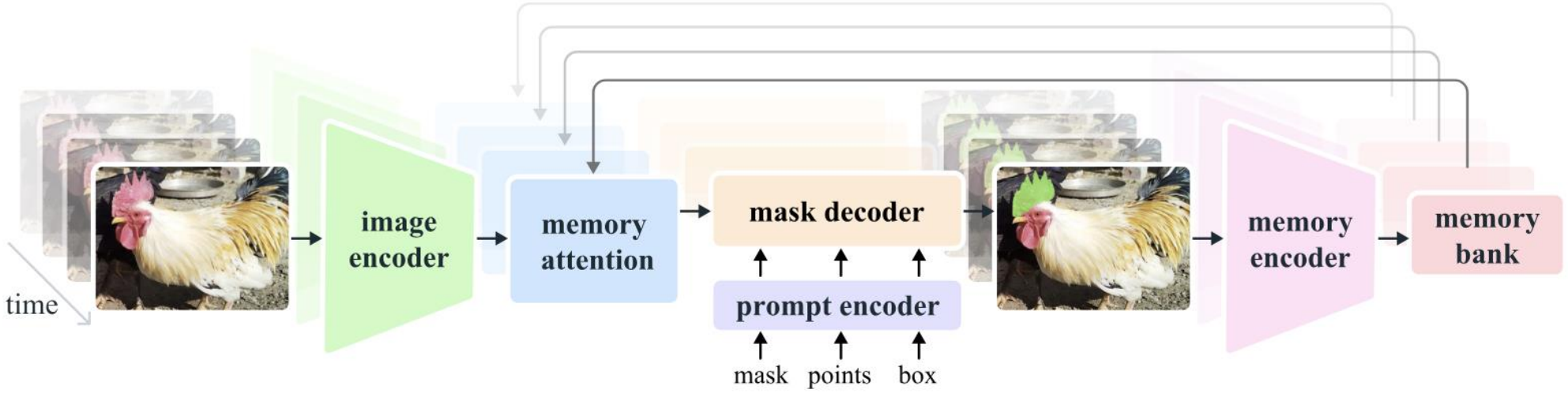


Figure: SAM 2 Architecture (from the SAM 2 paper, 2024)

Algorithm Overview: Mask Validation

- Kalman Filter Validation:
  - For each object, run a constant-velocity Kalman filter on bounding box center
  - Use high process noise and initial uncertainty to reflect animal movement
  - Compare candidate bounding boxes against the Kalman filter's predicted position using IoUs
  - Object could *either* be overlapping with the previous frame *or* overlapping with KF prediction; if no valid masks generated, discard track
- Geometric Validation:
  - Compute closest points between object masks (using FAISS on CPU or GPU)
  - Construct shape features (distance, angles) based on these points
  - Validate current geometric relationships against historical statistics
    - Criteria can be mean + 3SD of the last ~10 seconds or based off of known cutoffs

Distractor-Aware Memory for SAM 2 (Videnovic 2024)

- Concept: 'distractors' come into view before segmentation
  - Objects that are similar to the one of interest
- Use largest connected component of alternative masks from SAM 2; if IoU < threshold, add frame to memory
  - Separation suggests a mask of a different thing vs. a low quality mask for the object of interest

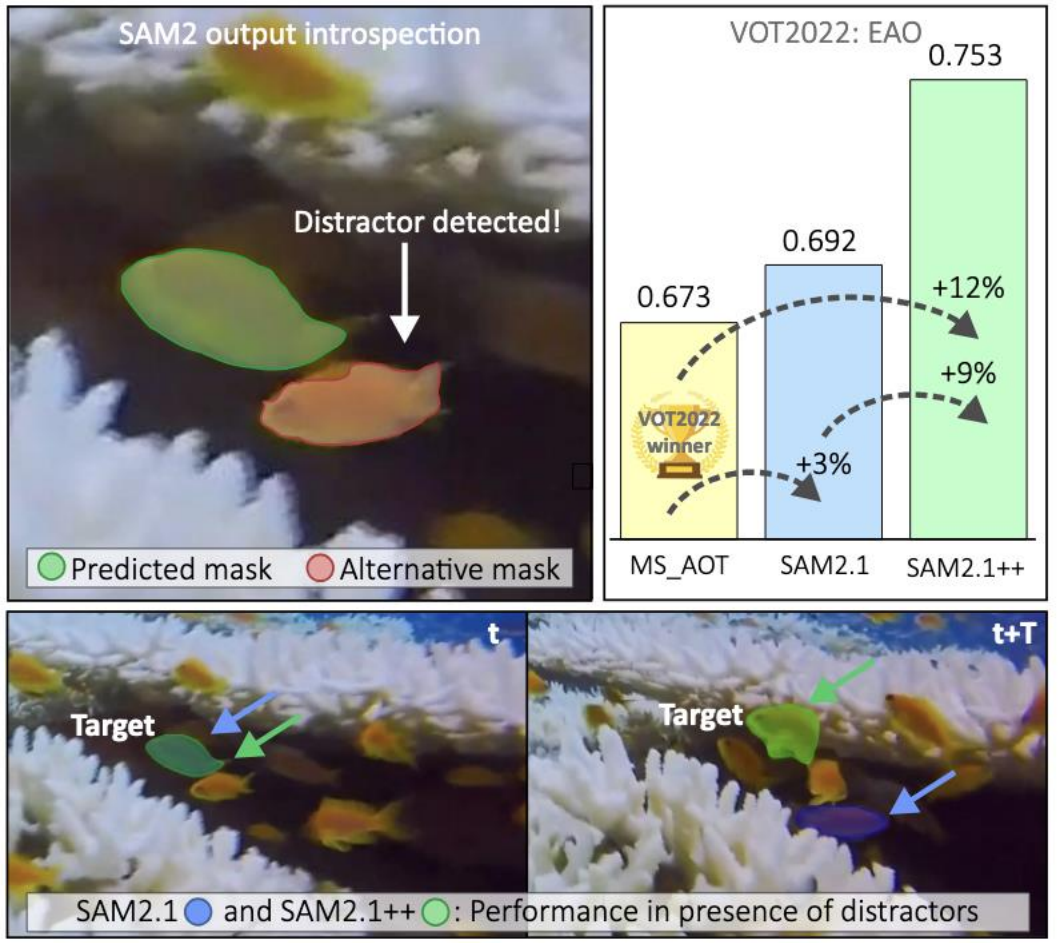


Figure: Example of distractor detection from Videnovic et al 2024

Algorithm Overview: Memory Management

- Update distractor resolving memory (DRM) when valid predictions meet size and consistency thresholds.
  - Current object size is between 90-110% of its average size in the last 10 frames
  - The minimum IoU between the chosen mask and the union of the chosen and alternative masks < 0.9
- Update high quality memory (HQM) when:
  - Current object size is between 90-110% of its average size in the last 10 frames
  - The IoU between frames for the chosen mask is > 0.9
- Purge old DRM or HQM frames (> 10 seconds)

Algorithm Overview: Track Recovery

- After ≥3 missed detections for an object, enter recovery mode
- Purges memory frames in the last ~5 seconds and disables geometric filters for that object
- Exit recovery mode after 5 consecutive valid frames.

Evaluation Dataset: DiDi

I had to cut our internal eval datasets from this talk ☹

Single-Object Performance on the DiDi Dataset

- Our model performs similar to SAM2.1/SAM2.1++ on the single-object case, trading accuracy (mean IoU) for robustness (proportion of frames where the tracker does not completely lose the target).

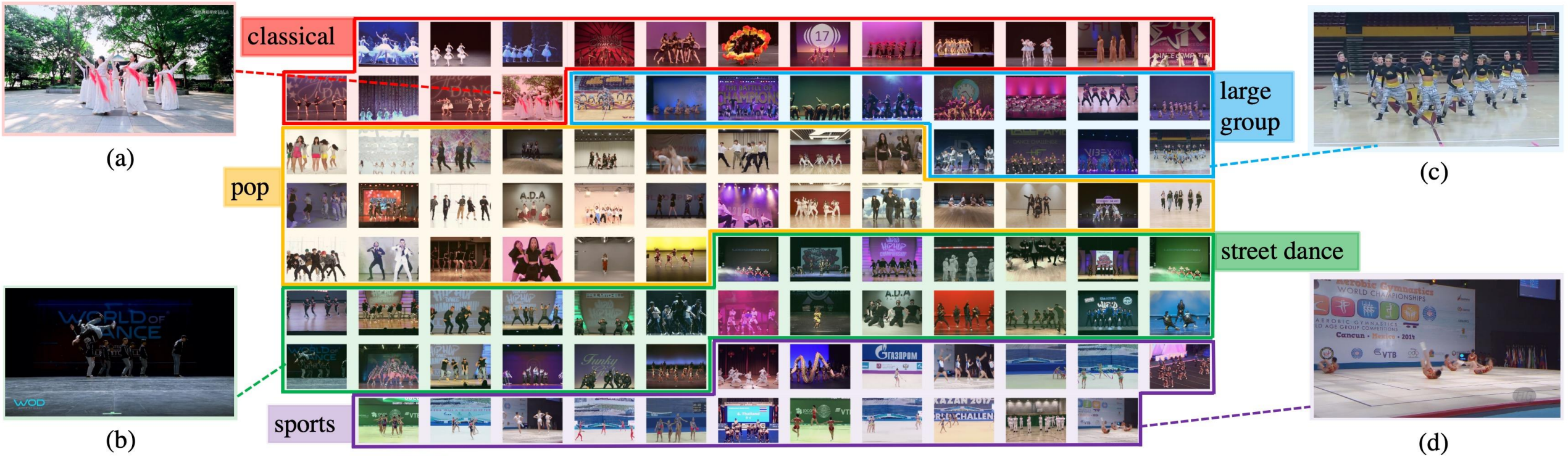
	Accuracy	Robustness
Our Model	0.707	0.96
SAM2.1++	0.727	0.944
SAM2.1	0.72	0.887

Figure: A simple video from DiDi

Single-Object Example: Dog

Evaluation Dataset: DanceTrack

- Multi-object tracking dataset emphasizing uniform appearance and diverse motion



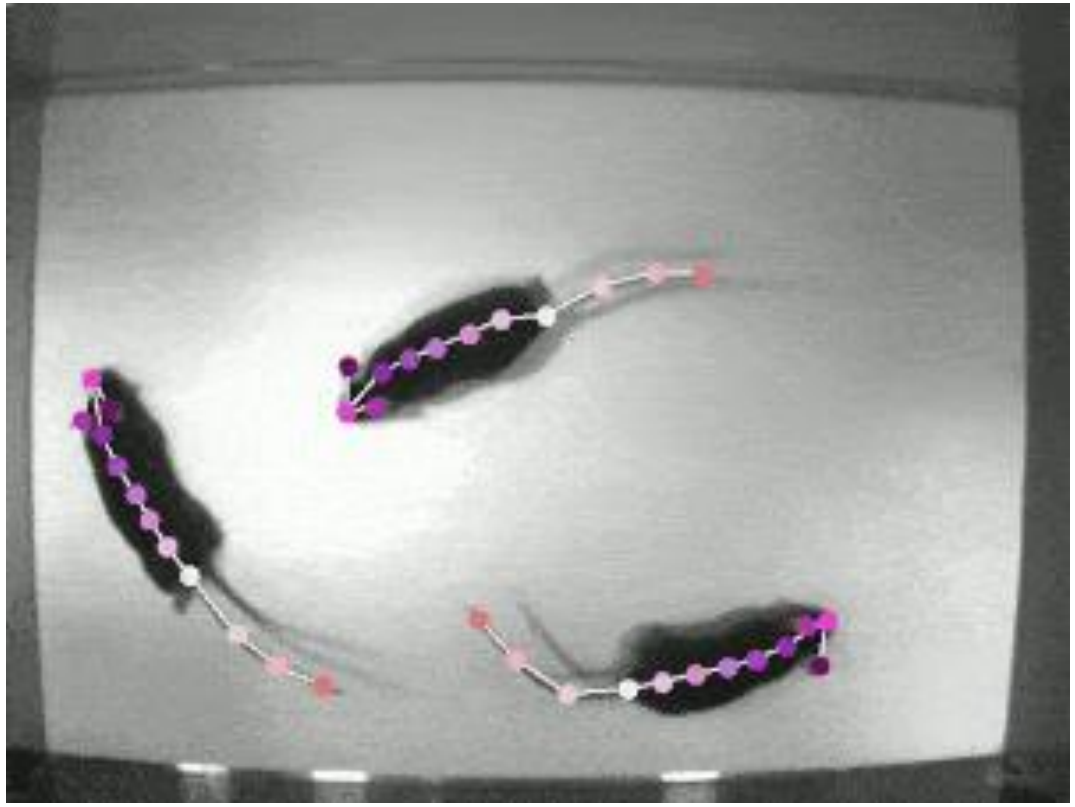
DanceTrack Performance

- HOTA: higher order tracking accuracy, geometric mean of detection accuracy and association accuracy
- IDF1: F1 score of correct identity predictions, harmonic mean of precision and recall

Methods	Publication	HOTA	IDF1
SORT	ICIP2016	47.9	50.8
DeepSORT	ICIP2017	45.6	47.9
FairMOT	IJCV2021	39.7	40.8
CenterTrack	ECCV2020	41.8	35.7
QDTrack	CVPR2021	45.7	44.8
GTR	CVPR2022	48.0	50.3
ByteTrack	ECCV2022	47.3	52.5
MOTR	ECCV2022	54.2	51.5
SUSHI	CVPR2023	63.3	63.4
MOTRv2	CVPR2022	69.9	71.7
ColTrack	ICCV2023	72.6	74.0
FineTrack	CVPR2023	52.7	59.8
OC-SORT	CVPR2023	54.6	54.6
DiffMOT	CVPR2024	62.3	63.0
Hybrid-SORT	AAAI2024	65.7	67.4
AED	arXiv2024	66.6	69.7
MOTIP	arXiv2024	73.7	79.4
SAM2MOT	arXiv2025	**75.8**	83.9
Our method	Ours	65.3	**90.6**

Why do we care about robustness that much?

- Our method is excellent at preserving the identities of objects, but may drop tracks more than other models
- This is by design, as we prefer missing data with a possibility of object reacquisition to incorrect data or identity confusion
  - Consider how identity confusion between paws, ears, etc would affect the DeepLabCut results below
  - A few identity confusions can destroy your metrics and biomarker utility



Video: Example of DeepLabCut Pose Estimation (from the DLC website)

Multi-Object Example: Ballet

Better Health, Brighter Future