# A Whisker of Truth: A Multimodal Interdisciplinary Machine Learning Approach to Vocal, Visual, and Tactile Signals in the Domestic Cat

van Toor, Astrid; Schötz, Susanne; Hirsch, Elin

2025

*Document Version:*
Publisher's PDF, also known as Version of record

Link to publication

*Citation for published version (APA):*
van Toor, A., Schötz, S., & Hirsch, E. (2025). *A Whisker of Truth: A Multimodal Interdisciplinary Machine Learning Approach to Vocal, Visual, and Tactile Signals in the Domestic Cat*. 1-8.
https://openreview.net/attachment?id=HIkzPhD0Zo&name=pdf

*Total number of authors:*
3



LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

# A Whisker of Truth: A Multimodal Interdisciplinary Machine Learning Approach to Vocal, Visual, and Tactile Signals in the Domestic Cat

Astrid van Toor | Susanne Schötz | Elin Hirsch
ML eng., MSc AI | PhD Phonetics | PhD Ethology
LUND UNIVERSITY

## Why Cat Communication Matters

Over 600 million domestic cats live with humans worldwide, yet human interpretation of feline behavioural signals remains poor — even experienced owners achieve only modest accuracy classifying vocalisations by context [1] and recognition of subtle negative behavioural cues during play interactions often approximates chance levels [2].

### The Problem:

- Subtle welfare changes often go undetected until clinical presentation
- Existing AI approaches treat vocalisations in isolation, ignoring visual and tactile signals
- Prior datasets lack expert annotation or suffer from subject leakage in validation [3, 4]

### Our Approach:

A **multimodal framework** combining vocal, visual, and tactile signals to:

1. Classify behavioural state (validated via leave-one-cat-out CV)
2. Enable personalised deviation detection (flag changes from individual baseline)

We train a multimodal encoder on expert-annotated cat-human interactions to classify behavioural state, context, and valence. This pushes the model to learn embeddings that capture behaviourally meaningful distinctions. For welfare monitoring, we don't need labeled pathology data; we model per-individual distributions over these embeddings during healthy enrolment, then flag statistical departures. Classification validates the representation; deviation detection uses the representation embedding.

## Data: Expert-Annotated Multimodal Datasets

### Meowsic (Acoustic Primary) [5]

- **Vocalisations**: 1,799 segmented
- **Cats**: 63 Adults
- **Features**: 100-point F0 contours, MFCCs, duration
- **Types**: Meows, Trills, Growls, Howls, Hiss + combinations
- **Contexts**: 15+ (food, door, cuddle, tbox, etc.)

### CHC (Cat-Human Communication) [6]

- **Interactions**: 81 multimodal sessions
- **Pairs**: 19 cat-human pairs
- **Visual**: Frame level BORIS annotations
- **Valence**: Owner & expert judged (ICC=0.95)

Within-cat repeated observations: same cats recorded in both "normal" (food, cuddle) and "stressed" (tbox, vet) contexts.

## Data Annotation

### Acoustic Labels

labelling protocol [7]:

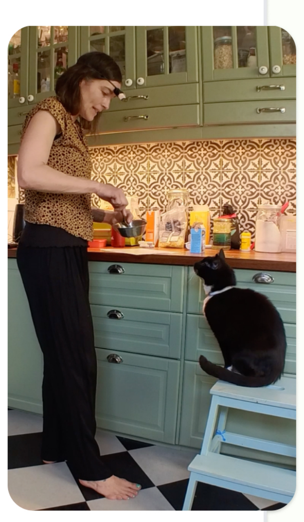
```
Type        | Context | Mental State | Example
-----------|--------|--------------|-----------------
Me (Meow)   | food    | con2          | Food soliciting, content
Tr (Trill)  | gree    | att1          | Greeting, attention-seeking
Gr (Growl)  | hunt    | dis2          | Hunting, medium discontent
Ho (Howl)   | terr    | aro3          | Territorial, high arousal
```

### Visual Labels (BORIS & ethograms) [8-12]:

- Tail: up/halfway, parallel, down, vertical, wrapped, fast, slow
- Ears: forward, back/angled, flattened
- Body: sitting, standing, crouching, locomotion
- Contact: rub, sniff/lick, touch/knead, soft gaze

### Mental State Labels:

- con (content) / dis (discontent) / str (stressed) / att (attention) / aro (aroused) / foc (focused)
- Intensity: 1 (mild) → 2 (medium) → 3 (strong)



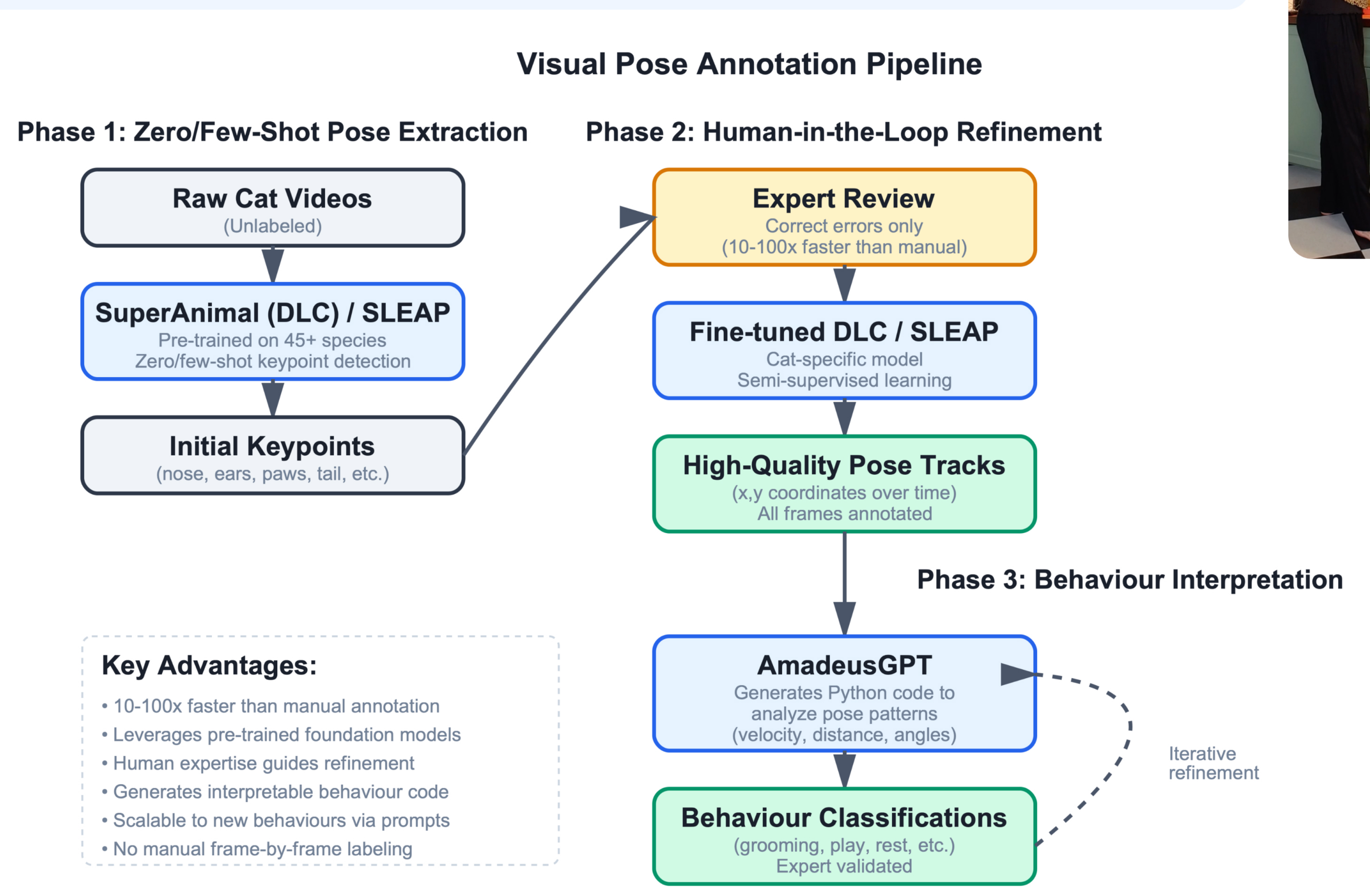



*Fig 1:* *Semi-supervised labelling pipeline to extract keypoints for pose estimation using SuperAnimal [13] (DeepLabCut / DLC) and/or SLEAP [14]. Human-in-the-loop refinement ensures robust pose labels, reducing annotation time by 10-100x. Semi-supervised tactile identification is obtained by feeding time-series pose data to AmadeusGPT [15] to generate hypotheses like: "Cat ear angle changed immediately after contact onset → tactile influence on behaviour", which are then validated and refined by our expert team.*

## Architecture: Multimodal Pipeline

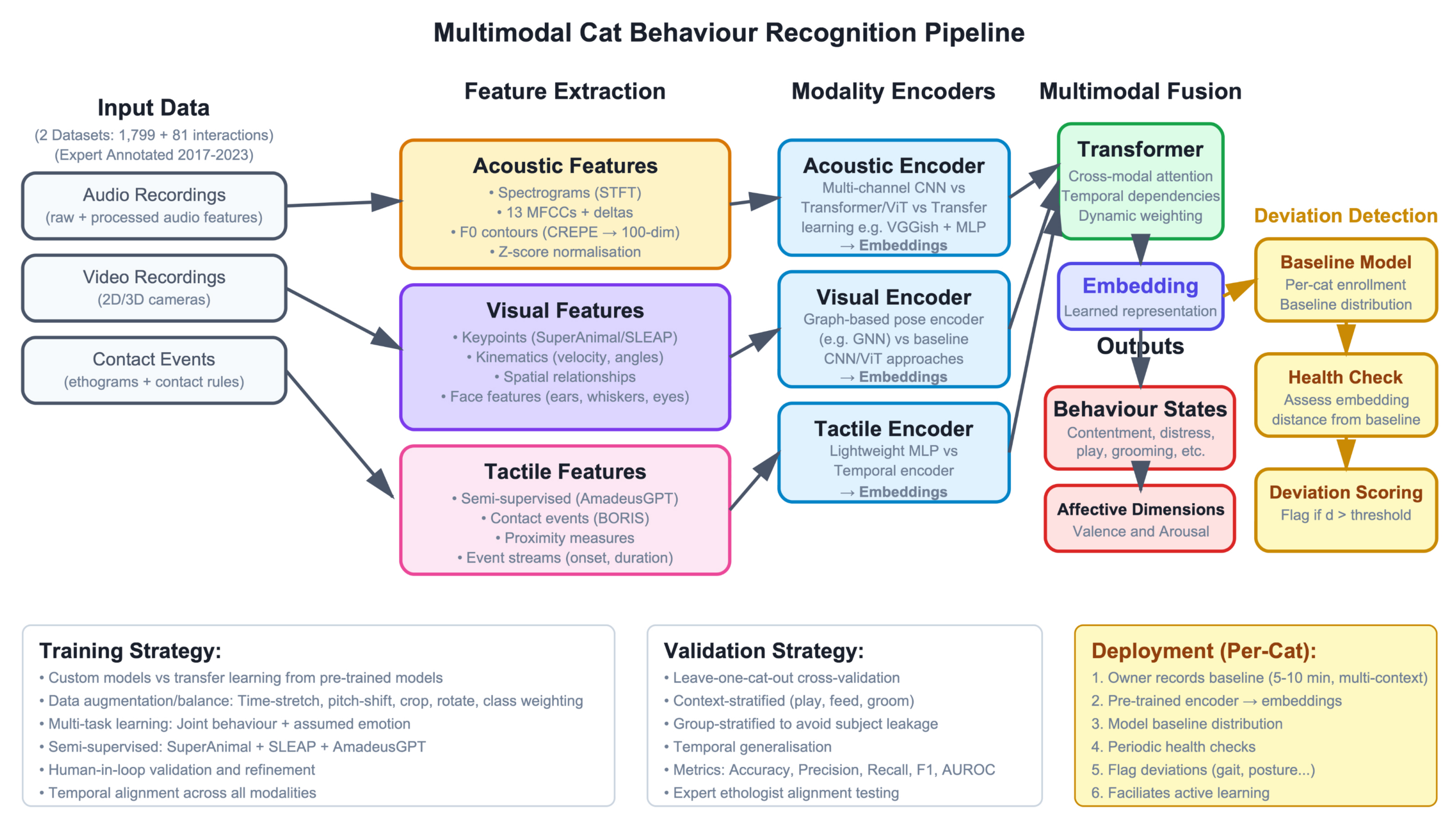



*Fig 2:* *Per-modality encoders extract acoustic (F0, MFCCs), visual (SuperAnimal/SLEAP downstream keypoints), and tactile (derived contact events) features. Transformer fusion produces embeddings used for (A) classification during training and (B) deviation scoring during deployment.*

## Feature Extraction: Per-Modality Processing

### ACOUSTIC

- Input: 44.1kHz WAV, audio segmentation
- Features: STFT spectrograms, 13 MFCCs, F0 contours
- Encoder: CNN or fine-tuned wav2vec/animal2vec → multi-dimensional vector

***E.g.*** Falling F0 contours in stressed contexts (cat carrier); rising-falling in positive contexts (food, greeting); duration shorter in positive valence (Schötz et al., 2024).

### VISUAL

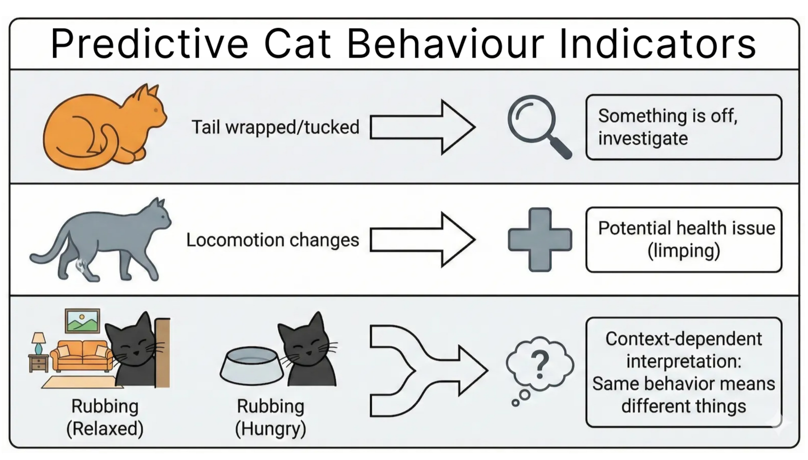
- Input: Video → Fine-tuned SuperAnimal [13] / SLEAP [14]
- Behaviours: Rule-based from joint angles/positions
- Encoder: Temporal C(R)NN on keypoint sequences → multi-dimensional vector

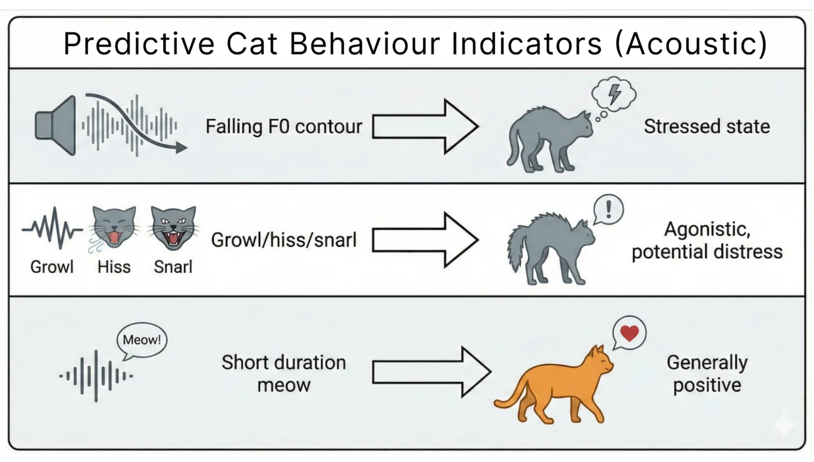
***E.g.*** Tail up = affiliative; tail vertical/wrapped = fear/stress/pain indicator. Visual behaviours associate with valence more clearly than acoustic alone [6].

### TACTILE (derived)

- Input: Cat + human keypoint proximity
- Detection: Distance thresholds + motion patterns
- Events: Rub, stroke, hold, knead → multi-dimensional vector

***E.g.:*** Rub behaviour strongly associated with positive valence [6]. Contact patterns encode relationship quality.




## Training Objectives & Validation

We use a classification head for behavioural context to steer embeddings into encoding behaviourally meaningful distinctions for behavioural deviations analysis.
Validation Protocol:

- **Leave-one-cat-out CV:** all data from one cat held out per fold
- **Metric:** Macro F1 (handles class imbalance)
- **Ablation:** Audio-only vs. visual-only vs. multimodal
- **Manual expert review**

**Preventing subject leakage:** Unlike prior work, we ensure no individual appears in both train and test splits and experts remain in the loop during initial annotation and development (critical for generalisation).

## Outcomes & Impact

**Research Contributions:**

- First multimodal cat-human interaction benchmark with expert annotation
- Rigorous data-splits and validation preventing subject leakage
- Open-source pipeline adaptable to other species

**Applications:**

Early welfare alerts, Deviation from behavioural baseline, Behaviour interpretation, Classification with explainable features, Human-cat relationship, Longitudinal communication patterns. Single modalities miss the full picture. Cats communicate through integrated vocal, postural, and contact signals [6, 7].

References
[1] Nicastro & Owren (2003). Classification of domestic cat vocalisations by naive and experienced human listeners.
[2] Henning et al. (2025). Recognition of negative behavioural cues in cat-human play interactions.
[3] Ntalampiras et al. (2021). Acoustic classification of individual cat vocalizations in evolving environments.
[4] Ntalampiras et al. (2019). Automatic Classification of Cat Vocalizations Emitted in Different Contexts.
[5] Schötz et al. (2019). Phonetic methods in cat vocalisation studies: A report from the Meowsic project.
[6] Hirsch, Weijer, Schötz (2024). Vocal, visual, and tactile signals in cat–human communication: A pilot study.
[7] Schötz. (2020) Phonetic Variation in Cat–Human Communication.
[8] Gamba & Friard (2016). BORIS.
[9] Stanton et al. (2015) A standardized ethogram for the felidae: A tool for behavioral researchers.
[10] Deputte et al. (2021) Heads and Tails: An Analysis of Visual Signals in Cats, Felis catus.
[11] Finka et al. (2022) Investigation of humans individual differences as predictors of their animal interaction styles, focused on the domestic cat.
[12] Siegford et al. (2023) The quest to develop automated systems for monitoring animal behavior.
[13] Ye et al. (2024) SuperAnimal pretrained pose estimation models for behavioral analysis.
[14] Pereira et al. (2022) SLEAP: A deep learning system for multi-animal pose tracking.
[15] Ye et al. (2023) AmadeusGPT: a natural language interface for interactive animal behavioral analysis.