We Will Track You: A Multi-Criterion Evaluation of Modern Object Detectors for Lizard Behavioral Analysis

Luis Gustavo de Oliveira Miranda[1], **Pedro Henrique Rabelo Leão de Oliveira**[1], **Pedro Henrique Lopes Silva**[2], **Vander Luis de Souza Freitas**[2], **Eduardo José da Silva Luz**[2], **Gladston Juliano Prates Moreira**[2], **Conrado Aleksander Barbosa Galdino**[3], **Jadson Castro Gertrudes**[2]

[1] Postgraduate Program in Computer Science – Federal University of Ouro Preto (UFOP)
Campus Morro do Cruzeiro – Ouro Preto – MG – Brazil

[2]Computing Department – Federal University of Ouro Preto
Campus Morro do Cruzeiro – Ouro Preto – MG – Brazil

[3]Programa de Pós-graduação em Biologia de Vertebrados da Pontifícia Universidade Católica de Minas Gerais – Coração Eucarístico, Belo Horizonte-MG, Brasil

luis.miranda@ufop.edu.br

Abstract. *Behavioral analysis of Tropidurus lizards historically relies on manual inspection of video recordings, which is time-consuming and labor-intensive. In this scenario, machine learning algorithms can help improve inspection efficiency and reduce time spent during evaluations. The present work investigates five recent object detectors for extracting ethological signals. We evaluate YOLOv5 as a baseline, YOLOv12, and three RT-DETR variants. The evaluation protocol integrates classical detection metrics and two ethological measures. Experimental results indicate that YOLOv12 presented the best result, balancing classical and temporal evaluation criteria.*

1. Introduction

Animal communication constitutes a process of information transfer between a signaling individual and a receiver that processes the incoming signal into meaningful information [Alcock 2009]. This process serves a variety of important functions, including conveying fighting ability and physical condition during resource-defense contests [Kaplan 2014], advertising reproductive status, and warning or deterring predators.

Among the many groups in which these phenomena have been investigated, lizards constitute particularly well-established models for the study of visual signal communication. The communicative process operates through the repetition of stereotyped gestures, among which push-ups and headbobs are especially prominent [Carpenter and Ferguson 1977].

Figure 1 presents the three main steps of analysis of visual signal types in lizards. In general, the process relies on a supervised approach, in which behavioral events are manually annotated by domain experts from video recordings (step 2 of Figure 1), including the identification of individual headbobs, a task that imposes a considerable time burden on researchers, given the large volumes of footage. After signal extraction, researchers attempt to understand the emitted signals by annotating patterns within them.

Understanding lizard communication requires turning observable behavior into structured temporal representations (step 2 of Figure 1), and in this context

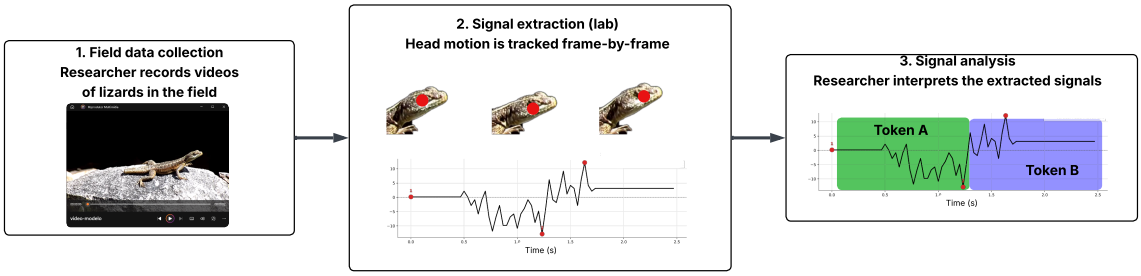


Figure 1. Complete framework for lizard communication analysis. Step 1 involves data collection by researchers; step 2 involves extracting head motion (headbobs) from the videos; and finally, step 3 involves analyzing the extracted signals.

computational approaches become indispensable. Visual Display Action Patterns (DAPs), such as headbobs, have been studied as candidates for structured, multi-component signaling [DeCourcy and Jenssen 1994, Labra et al. 2007, Martins et al. 2004, Vicente and Halloy 2015], and the ReptileRecon framework [Zenóbio et al. 2023] uses YOLO-based detectors to localize head and body centroids in lizards, thereby extracting reliable signals from field recordings. However, signal extraction remains challenging due to occlusion, illumination variability, and detector instability [Zenóbio et al. 2023, Ferrante et al. 2021].

A fundamental source of the difficulty discussed before lies in how modern detectors operate. Designed to process each frame independently and to optimize spatial accuracy rather than temporal coherence [Zou et al. 2023, Raza et al. 2025], these models are prone to introducing small but consequential fluctuations in bounding box placement. These fluctuations constitute non-biological noise, which can dominate subtle behavioral signals and compromise downstream analyses that assume smooth trajectories. Despite the variety of architectures available, most animal-behavior pipelines adopt a single CNN model without systematic comparison [Rajagukguk et al. 2025]. Within ethology, the YOLO family has become a near-default choice [Zenóbio et al. 2023, Chan et al. 2025].

Classical evaluation metrics such as mean Average Precision (mAP), while widely used, capture spatial localization performance but were developed in the context of still-image protocols and are largely insensitive to temporal artifacts [Padilla et al. 2021, Zhu et al. 2020]. A detector may achieve high mAP while producing trajectories dominated by jitter, rendering standard benchmarks insufficient for ethological applications.

The data acquisition pipeline poses an additional challenge, since generating the frame-by-frame ground-truth annotations required to train such models is both labor-intensive and error-prone. To mitigate this operational bottleneck, knowledge transfer via teacher-student pseudo-labeling has been proposed to leverage large volumes of unlabeled footage [Gou et al. 2021, Li et al. 2023]. This paradigm, however, tends to amplify the teacher's weaknesses through confirmation bias [Li et al. 2022], and its benefit for ethological signal quality has not been empirically established.

These observations motivate our central research question: **RQ.:** *Which detector architecture (YOLOv5, YOLOv12, or RT-DETRs) produces the most reliable head and body trajectories for lizard visual signal extraction?*

We address this question by systematically comparing five detectors on lizard video

recordings: YOLOv5 and YOLOv12 trained on manual labels, and three RT-DETR variants that span fully supervised training, pseudo-label distillation from a YOLOv12 teacher, and subsequent fine-tuning on ground-truth annotations. The evaluation protocol combines classical metrics with two purpose-designed ethological measures: anatomical jitter and spectral noise analysis. We monitor high-frequency content above 8 Hz as an indicator of tracking error, since signal components in this frequency range are highly unlikely to arise from natural lizard movement [DeCourcy and Jenssen 1994, Labra et al. 2007, Martins et al. 2004, Vicente and Halloy 2015]. Our findings suggest adopting YOLOv12 for the signal extraction task.

2. Background

2.1. Object detection and tracking

Object detection localizes and classifies each object instance in an image, typically by predicting a bounding box and a class label. Detectors are trained and benchmarked on still-image datasets such as COCO and PASCAL VOC, and their performance is conventionally summarized by frame-level metrics [Padilla et al. 2021, Zhu et al. 2020].

A defining property of this formulation is that it operates on each frame independently, with no notion of identity or continuity across frames. Tracking, in contrast, links per-frame detections over time to recover an individual's trajectory. In behavioral pipelines, this distinction is consequential: even when frame-level detection is accurate, small frame-to-frame variations in bounding box placement accumulate into temporal artifacts along the trajectory. These frame-level metrics, therefore, do not target temporal fidelity, which is precisely the property that downstream signal analysis depends on.

Two families of detectors are relevant to this work. The YOLO family has evolved through successive incorporations of anchor boxes, multi-scale aggregation, and attention modules [Rajagukguk et al. 2025]. Two waypoints matter here: YOLOv5, the PyTorch-based detector adopted in ReptileRecon [Zenóbio et al. 2023], and YOLOv12, an attention-centric variant that introduces area attention to improve long-range feature aggregation. On the other hand, transformer-based detectors began with DETR [Carion et al. 2020], which formulated detection as set prediction but suffered from slow training. RT-DETR [Zhao et al. 2024] addressed this by combining an efficient hybrid encoder with IoU-aware query selection, becoming the first Transformer-based detector to compete with YOLO in real-time settings. This convergence in operational viability makes a controlled comparison meaningful. However, cross-family comparisons remain rare in animal-behavior research [Rajagukguk et al. 2025].

Knowledge distillation, originally a model-compression technique [Gou et al. 2021], generalizes to semi-supervised learning when a teacher trained on limited labels generates pseudo-labels for unlabeled data used to train a student. Applied to detection, the paradigm is more fragile than in classification: pseudo-labels carry uncertainty in both class and localization, and bounding boxes that are close but misaligned introduce systematic noise [Li et al. 2023]. Pseudo-labeling tends to amplify the teacher's strengths and weaknesses, biasing the student toward objects that the teacher already detects well [Li et al. 2022]. The benefit of distillation is therefore not universal and, in low-data ethological settings, has not been empirically established. The present

work directly tests whether a teacher-student pipeline based on YOLOv12 and RT-DETR improves over the teacher used alone.

2.2. Related work

Computer-vision pipelines for behavioral quantification fall into two complementary traditions [Dell et al. 2014]. Markerless pose estimation, established by DeepLabCut [Mathis et al. 2018] and extended by LEAP [Pereira et al. 2019], represents the animal as a set of keypoints but requires per-frame landmark annotation. Detection-based pipelines, in contrast, represent the animal as a sequence of bounding boxes from which behavioral signals are derived. YOLO-Behaviour [Chan et al. 2025] applied YOLO across five behavioral case studies, achieving F1-scores between 0.62 and 0.94. At the same time, the ReptileRecon framework [Zenóbio et al. 2023] used YOLO to localize head and body centroids in lizards and recovered latent behavioral clusters via unsupervised autoencoders.

Across these works, the detector is typically chosen once and evaluated with frame-level metrics, on the implicit assumption that it provides a stable substrate for temporal analysis. Whether different detectors yield signals with comparable temporal fidelity has not been examined, motivating the comparison presented here.

3. Materials and Methods

3.1. Dataset and Annotation

The corpus consists of field recordings of *Tropidurus* lizards filmed at approximately 30 fps under natural variation in illumination, vegetation occlusion, and camouflage. It extends the corpus first collected by [Zenóbio et al. 2023] by adding further recordings, totaling 47 videos. Of these, 11 were discarded due to processing failures during inference.

Five videos were set aside as the manually annotated detection benchmark; these same videos serve as the held-out evaluation set for pseudo-label distillation (Section 3.3), being the exact source of the evaluation frames for all other models, so the student detector was never exposed to test footage. To avoid placing near-duplicate consecutive frames in different partitions, a subset of frames sampled from within these five videos was split 70/15/15 (train/validation/test) in YOLO format. The remaining 42 videos were processed by every detector for temporal analysis; after removing the 11 failures, $n = 31$ yielded usable temporal signal under the criterion of Section 3.4 and form the temporal evaluation set of Section 4. Classical metrics (mAP, precision, recall) are thus computed on the five annotated videos, and temporal metrics are computed on the 31, identically for every model. As the classical partitions share the same recordings, these metrics reflect *in-domain* detection quality, while generalization is assessed on the 31 disjoint videos.

Annotation was performed in Roboflow with two classes: *body*, excluding the tail, to prevent caudal motion from contaminating the angular signal, and *head*; each frame thus yields a head–body label pair. Data augmentation was applied only to the training partition, expanding it from 329 to 988 label pairs, while keeping validation and test at their original 76 pairs each.

3.2. Detector Configurations

Five configurations were evaluated under a unified protocol, summarized in Table 1. The **YOLOv5** baseline was trained from scratch on our manually annotated dataset, as

described in Section (3.1). Implemented via the Ultralytics native API, this model serves as a baseline reference for all comparisons. To isolate the network architecture as the primary variable, the training used an optimized set of hyperparameters obtained via the Optuna framework [Akiba et al. 2019]. **YOLOv12**, the most recent attention-centric YOLO variant, incorporates area attention to improve localization of small objects under clutter and was trained with the exact same optimized configuration as the baseline to guarantee a fair comparison. The three **RT-DETR** configurations share the RT-DETR-Large architecture and provide the Transformer-based arm of the comparison. The *base* model serves as a fully supervised Transformer baseline, trained on the same ground truth as YOLOv12. The *student* is trained exclusively on pseudo-labels generated by the YOLOv12 teacher, testing whether pseudo-label transfer can match supervised training on ground truth. The *final* model fine-tunes the student on ground truth to test whether pseudo-label pretraining followed by fine-tuning improves over direct supervision alone. Full training settings for all three are listed in Table 1.

Table 1. Summary of the five detector configurations evaluated under the unified protocol. All models use 640×640 input. Hyperparameters marked "Optuna" refer to the optimized set shared by both YOLO models.

Configuration	Architecture	Training data	Epochs	Init. LR ($\times10^{-4}$)	Key settings
YOLOv5 (baseline)	CNN (Ultralytics)	Ground truth	100	6.2	AdamW; Optuna HPs; patience 20
YOLOv12	CNN + area attn.	Ground truth	100	6.2	AdamW; same Optuna HPs as YOLOv5
RT-DETR (base)	Transformer (RT-DETR-L)	Ground truth	100	1	AdamW; batch 16; AMP; COCO-pretrained
RT-DETR (student)	Transformer (RT-DETR-L)	Pseudo-labels (YOLOv12)	100	1	AdamW; batch 8; AMP; no human labels
RT-DETR (final)	Transformer (RT-DETR-L)	Student + GT fine-tune	30	0.1	Attenuated aug.; final-LR factor 0.01

3.3. Pseudo-label Distillation Pipeline

The teacher–student pipeline applies the standard formulation of pseudo-label distillation [Gou et al. 2021, Li et al. 2023] to a low-data ethological setting; its stages are summarized in Table 2. The YOLOv12 model (Section 3.2) acts as the *teacher*, generating pseudo-bounding-boxes on frames absent from the manually annotated set. To build the synthetic dataset, every third frame was sampled to reduce redundancy, and only detections with a confidence score above 0.65 were retained and saved in YOLO format. To prevent information leakage, five videos were strictly excluded from this process and reserved for evaluation, maintaining consistency with the YOLOv5 and YOLOv12 test sets. The RT-DETR-L *student* is trained exclusively on this synthetic dataset. Choosing a Transformer-based student, distinct from the CNN-based teacher, was motivated by cross-architecture distillation, in which the student inherits a supervisory signal with a different inductive bias. The *final* stage fine-tunes the student on ground truth, completing the three-stage pipeline: ground-truth supervision of the teacher, pseudo-label transfer to the student, and ground-truth fine-tuning. This is the configuration most commonly advocated in semi-supervised detection [Li et al. 2023] and the most favorable condition for measuring the benefit of distillation here.

Table 2. Three-stage pseudo-label distillation pipeline. The teacher supervises on ground truth, transfers knowledge to the student via pseudo-labels, and the student is finally fine-tuned on ground truth.

Stage	Model	Role / supervision
1. Teacher	YOLOv12 (CNN)	Trained on ground truth; generates pseudo-bounding-boxes on held-out frames
2. Transfer	—	Frames sampled every third frame; detections with confidence > 0.65 retained as pseudo-labels (YOLO format)
3. Student	RT-DETR-L (Transformer)	Trained exclusively on the synthetic (pseudo-labeled) dataset; cross-architecture distillation
4. Fine-tune	RT-DETR-L (Transformer)	Student fine-tuned on ground truth; final configuration of the pipeline

3.4. Evaluation Protocol

We combine classical detection metrics with two ethological measures of temporal fidelity and a tracking-robustness diagnostic, summarized in Table 3. Classical performance was assessed on the held-out test split using precision, recall, mAP@50, and mAP@50–95.

Table 3. Evaluation protocol: classical detection metrics, two ethological measures of temporal fidelity, and a tracking-robustness diagnostic.

Category	Metric	Definition / configuration	Better
Classical	Precision, Recall	Computed on the held-out test split via each framework's native validation routine	higher
	mAP@50, mAP@50–95	Standard COCO-style mean average precision	higher
Temporal	Anatomical jitter J	$J = \mathrm{std}(\Delta\tilde{d}_t)$, $\Delta\tilde{d}_t = (d_{t+1} - d_t)/\bar{d}$ (normalized head–body distance); raw consecutive pairs with both detections present	lower
	Spectral noise > 8 Hz	Summed single-sided amplitude ($\frac{2}{N}\lvert\mathrm{FFT}\rvert$) of the centered vertical signal, normalized by mean head–body distance $\bar{d}$, resampled onto a common $[0, 15]$ Hz grid and summed over bins above 8 Hz; Welch PSD (128-sample segments) shown separately for inspection	lower
Robustness	Frame loss rate	Per-video % of frames missing head or body; videos $> 15\%$ excluded from spectral analysis	lower
Statistics	Paired Wilcoxon	Signed-rank test, casewise deletion; effective $n = 27$ to $n = 30$ per pair. Boxplots/medians on common $n = 31$ subset	—

The first temporal metric quantifies frame-to-frame variability of the head–body distance. Under a rigid-body approximation, this distance should be approximately constant across consecutive frames. High variability indicates centroid drift inconsistent with anatomy. The anatomical jitter:

$$J = \mathrm{std}\big(\{\Delta\tilde{d}_t\}_{t\in\mathcal{T}}\big), \qquad \Delta\tilde{d}_t = \frac{d_{t+1} - d_t}{\bar{d}}, \qquad \bar{d} = \frac{1}{|\mathcal{R}|}\sum_{t\in\mathcal{R}} d_t. \tag{1}$$

is computed only on consecutive pairs in which both detections are present in the raw data, so that J reflects detector instability rather than interpolation artifacts. Lower J indicates greater anatomical consistency.

The second temporal metric measures the amount of vertical signal amplitude that lies outside the biologically plausible frequency range of headbob displays. Because frequency analyses specific to our target species are currently absent from the literature, we established a biological threshold from the median values of related taxa, in

which stereotyped headbob units typically last more than $\sim 100ms$, constraining meaningful frequency content to roughly $8Hz$ [DeCourcy and Jenssen 1994, Labra et al. 2007, Martins et al. 2004, Vicente and Halloy 2015]. Amplitude above this threshold is interpreted as detector-induced noise. The processing configuration (FFT, interpolation range, and Welch PSD settings) is given in Table 3.

As a diagnostic, the per-video frame loss rate was recorded as the percentage of frames in which either the head or the body was missing. Systematic detection failure is a distinct mode of inadequacy from per-frame jitter, so the loss rate was retained as an independent metric. The exclusion criterion applied before spectral analysis is noted in Table 3.

Paired Wilcoxon signed-rank tests were applied with case-wise deletion: each pair of detectors was compared on the videos where both produced valid temporal measurements. This uses all available paired observations rather than discarding videos in which any single detector failed, thereby maximizing statistical power for each comparison. Boxplot visualizations and reported median values reference a common subset of videos in which the majority of detectors produced consecutive head–body frame pairs, ensuring a shared reference set for visual comparison and descriptive statistics. Effective sample sizes are listed in Table 3. All the implementation is available in our GitHub repository[1]

4. Results

On the held-out test split (76 images, 152 instances), all detectors achieved high mAP@50 (≥ 0.815), confirming that lizard heads and bodies are well-localized at the lenient IoU threshold. The more demanding mAP@50–95 reveals clearer separation, with YOLOv12 (0.595), RT-DETR Base (0.500), RT-DETR Final (0.493), RT-DETR Student (0.474), and YOLOv5 (0.418). These results can be seen in Table 4.

Table 4. Classical detection metrics on the held-out test split broken down by class.

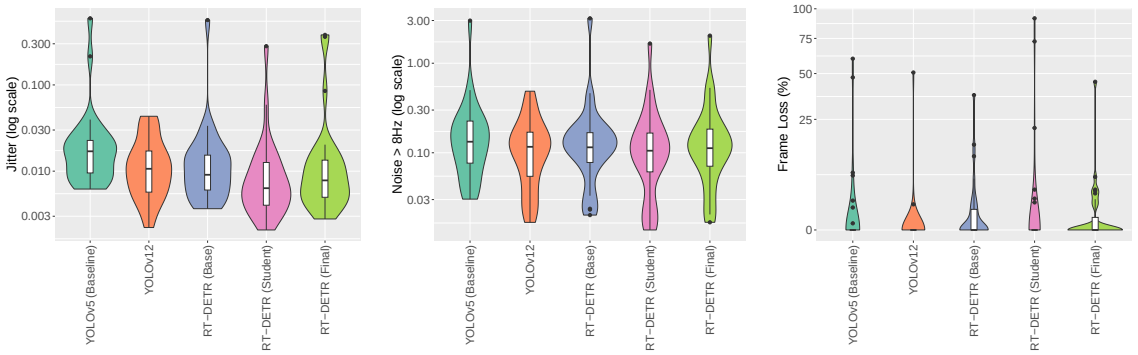
Configuration	Class	Precision	Recall	mAP@50	mAP@50–95
YOLOv5 (Baseline)	All	0.829	0.883	0.933	0.418
	Body	0.838	0.818	0.900	0.312
	Head	0.820	0.947	0.966	0.525
YOLOv12	All	0.908	0.932	0.967	**0.595**
	Body	0.817	0.882	0.940	0.459
	Head	1.000	0.983	0.995	0.731
RT-DETR (Base)	All	**0.987**	0.885	**0.975**	0.500
	Body	1.000	0.789	0.963	0.388
	Head	0.974	0.981	0.986	0.612
RT-DETR (Student)	All	0.896	0.763	0.815	0.474
	Body	0.863	0.553	0.647	0.287
	Head	0.928	0.974	0.984	0.661
RT-DETR (Final)	All	0.947	**0.952**	0.950	0.493
	Body	0.932	0.903	0.911	0.376
	Head	0.962	1.000	0.989	0.610

First, the head class is consistently easier than the body class, with an average mAP@50–95 gap of ~ 0.26, consistent with the head's fewer degrees of postural freedom

[1]Source code: `https://github.com/Persapius/Multi-Criterion-Eval`

and more rigid morphology. Second, RT-DETR (Student) shows asymmetric degradation on the body class (recall 0.553, mAP@50–95 0.287) while preserving high head detection performance (recall 0.974, mAP@50–95 0.661). The phenomenon reflects a known property of pseudo-label distillation, where the Student inherits the Teacher's limitations and cannot correct them with pseudo-labels alone [Li et al. 2022]. Classical metrics suggest that YOLOv12 is the strongest single-model configuration but do not distinguish detectors by temporal fidelity, which we examine next.

Figure 2 summarizes the per-video distributions of the two temporal measures and the robustness diagnostic across detectors. On the evaluation subset, median jitter values were 0.0162 (YOLOv5), 0.0106 (YOLOv12), 0.0093 (RT-DETR Base), 0.0061 (RT-DETR Student), and 0.0077 (RT-DETR Final). All four post-baseline detectors reduced median jitter relative to YOLOv5, ranging from 37% (Student) to 65% (YOLOv12) of the baseline value. Within the post-baseline group, RT-DETR (Student) was lowest, followed by RT-DETR (Final), while YOLOv12 and RT-DETR (Base) were nearly identical, as shown in Figure 2a, which displays the log-scale distribution of this metric.



(a) Stability: jitter (↓). **(b) Noise >8 Hz (↓).** **(c) Frame loss (↓).**

Figure 2. Performance comparison between the models across three distinct metrics. For all panels, ↓ indicates that lower values are better.

With respect to the spectral amplitude above 8 Hz, the pattern was qualitatively similar, with median values from 0.126 (YOLOv5) to 0.109 (RT-DETR Student). YOLOv5 accumulated the highest median supra-threshold amplitude, though absolute differences across detectors were smaller than for jitter and the statistical separation weaker.

All five detectors share the same low-frequency peak corresponding to the dominant biological motion, indicating that the underlying behavioral signal is recovered consistently across architectures. Differences are concentrated in the mid-band ($\sim$4–8 Hz), where YOLOv5 and RT-DETR (Base) show elevated amplitude, and in the supra-threshold band (>8 Hz), where they again carry more amplitude than YOLOv12, RT-DETR (Student), and RT-DETR (Final). Detectors with higher anatomical jitter also distribute more amplitude outside the biologically plausible range. Figure 2b shows the log-scale distribution of this metric and its noise, while Figure 3 presents the spectral density using Welch's method.

Median frame loss was at or near zero for all detectors, but mean loss rates were considerably higher (18.5% for YOLOv5 to 28.9% for RT-DETR Student), driven by a small number of complete-loss events. Across the corpus, these events (loss $\geq 99\%$) occurred in 2 videos for YOLOv5, 4 for YOLOv12, 5 for RT-DETR Base, 6 for RT-DETR

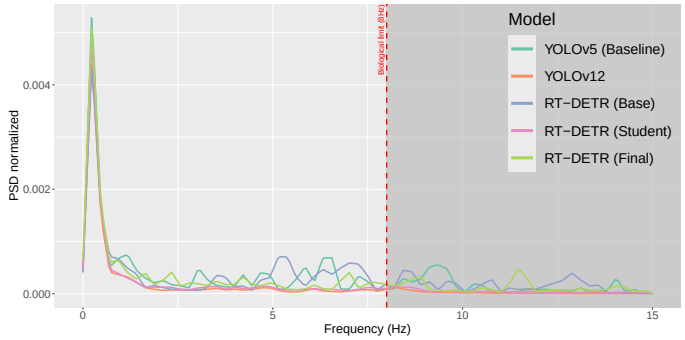


Figure 3. Average power spectral density of all models using Welch's method. The dashed red line marks the 8 Hz biological limit.

Student, and 5 for RT-DETR Final. Loss patterns are only partially correlated across detectors. For example, in one video, where YOLOv12 loses all frames, RT-DETR (Base) recovers the sequence with only 2% loss, whereas in another video, RT-DETR (Base) loses 37% of frames. At the same time, the other detectors operate without loss, consistent with architectural diversity producing distinct visual blind spots. When restricted to the consensus subset (loss $< 15\%$ for all detectors), mean loss rates fell below 5% across the board, matching the distribution shown in Figure 2c.

Paired Wilcoxon signed-rank tests revealed three patterns. YOLOv5 produces significantly higher jitter than each of the other detectors — YOLOv12 ($p = 0.011$) and all three RT-DETR variants ($p < 0.001$) —, so any modern detector evaluated improves anatomical stability over the legacy YOLOv5 used in ReptileRecon [Zenóbio et al. 2023]. For spectral amplitudes above 8 Hz, only RT-DETR (Student) reaches statistical significance relative to the baseline ($p = 0.010$); the remaining comparisons were not significant, indicating that high-frequency noise is harder to suppress than frame-to-frame jitter. Within the post-baseline group, RT-DETR (Student) yields significantly lower jitter than both YOLOv12 ($p = 0.008$) and RT-DETR (Base, $p = 0.005$), while YOLOv12 is statistically equivalent to both RT-DETR (Base) and RT-DETR (Final) in jitter, and RT-DETR (Final) shows lower jitter than RT-DETR (Base) ($p = 0.022$).

The Student attains the lowest classical detection performance among post-baseline detectors (mAP@50–95 $= 0.474$, body recall 0.553) and the highest complete-loss rate (6 of 42 videos), yet on the consensus subset it yields the lowest median jitter and is statistically superior to YOLOv12 and RT-DETR (Base). Because each pairwise comparison is restricted to videos in which both detectors produced valid measurements, this advantage is not an artifact of differing video subsets; it may, however, reflect the Student's more conservative operating point rather than an architectural gain, since it inherits the Teacher's confident detections without extending coverage to harder cases, becoming more selective yet more stable when it fires. This pattern is consistent with reported confirmation bias in semi-supervised detection [Li et al. 2022], though our design cannot separate a distillation-induced effect from the operating-point difference.

The full three-stage pipeline (RT-DETR Final) was designed to combine the Student's temporal stability with supervised detection accuracy. It only partially achieves this: detection accuracy recovers to mAP@50–95 $= 0.493$ and body recall 0.903, but the final fine-tuning stage significantly reverses the Student's jitter advantage ($p = 0.030$,

median rising from 0.0061 to 0.0077), leaving the Final model statistically equivalent to a single-stage YOLOv12 in both temporal metrics. The pipeline thus restores the same temporal regime as YOLOv12 trained directly on ground truth, at the cost of substantially greater engineering effort: a synthetic pseudo-labeled dataset, a training stage on those labels, and a final fine-tuning stage. The Student retains a jitter advantage over YOLOv12 only at the cost of reduced coverage, making it less suitable for pipelines that require continuous behavioral signals. Each stage of the pipeline, therefore, either fails to improve over YOLOv12 or introduces limitations that compromise downstream analysis.

5. Conclusion

This work addressed the problem of extracting headbob trajectories from lizard video recordings, where small frame-to-frame fluctuations in detector output introduce non-biological noise that can obscure the subtle signals carried by headbobs. We compared five detector configurations under a unified protocol, spanning the legacy YOLOv5 baseline used in ReptileRecon, an attention-centric YOLOv12, and three RT-DETR variants that cover fully supervised training, pseudo-label distillation from a YOLOv12 teacher, and subsequent ground-truth fine-tuning. Beyond classical localization metrics, we evaluated each configuration with two purpose-designed ethological measures, anatomical jitter and spectral noise above 8 Hz, together with a frame-loss analysis, so that temporal fidelity rather than per-frame accuracy alone guided the comparison.

Returning to our research question (**RQ**), the evidence points to YOLOv12. All four post-baseline detectors significantly outperform YOLOv5 on anatomical jitter, confirming that the legacy detector currently used in ReptileRecon is suboptimal from a temporal-fidelity standpoint. Within the post-baseline group, the RT-DETR Student achieves the lowest median jitter and is statistically superior to YOLOv12, RT-DETR (Base), and RT-DETR (Final) on this metric, yet it does so with substantially reduced detection coverage, making it less suitable for pipelines requiring continuous signals. The subsequent ground-truth fine-tuning step measurably reverses this temporal advantage ($p = 0.030$) without offering any compensating gain over a single-stage YOLOv12 baseline. Because the pseudo-label distillation pipeline does not yield a configuration that simultaneously improves over direct YOLOv12 supervised training in both detection accuracy and temporal fidelity, the findings argue for adopting YOLOv12 as the operational replacement for YOLOv5 in this domain, and against the additional complexity of multi-stage teacher–student pipelines for low-data ethological signal extraction.

Limitations: Four constraints bound the scope of these findings. First, evaluating each detector at its F1-optimal confidence threshold (ranging from 0.077 for YOLOv12 to 0.752 for RT-DETR Base) conflates architectural effects with precision/recall trade-offs, and fixed-threshold sensitivity analyses remain future work. Second, the operational recommendation for YOLOv12 is conditional on a single genus (*Tropidurus*) and corpus, and thus requires cross-species validation. Third, the $8Hz$ threshold is a proxy derived from related taxa, functioning as a conservative indicator of non-biological noise rather than an exact physiological cutoff. Finally, we deliberately evaluate raw, unfiltered outputs rather than applying standard temporal regularizers such as Kalman filtering or SORT, since heavy smoothing risks attenuating the high-frequency biological signals characteristic of lizard displays.

Future work: Fixed-threshold sensitivity analyses would separate architectural effects from precision/recall trade-offs and clarify how much of each detector's behavior is intrinsic. Extending the evaluation to additional genera and corpora would test whether the recommendation of YOLOv12 generalizes beyond *Tropidurus*, and would help refine the 8 Hz proxy toward a more species-grounded physiological cutoff. Finally, because a temporally stable base detector is a prerequisite for any downstream smoothing, integrating temporal regularizers onto YOLOv12 trajectories is a promising next step, provided that the post-processing preserves the high-frequency components that carry behavioral meaning.

Use of artificial intelligence: During the writing of this document, we used the *Claude Opus 4.8 LLM*, *Gemini 3.5 flash*, and *Grammarly* for grammatical correction and text organization. We also take full responsibility for all content, including checking for plagiarism and correcting all text.

Acknowledgements

This study was financed in part by the Universidade Federal de Ouro Preto (UFOP), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, the Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq, grants 308400/2022-4, 309676/2026-6) and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, grants APQ-02036-24, APQ-01768-24).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proc. ACM SIGKDD*.

Alcock, J. (2009). *Animal behavior: An evolutionary approach*. Sinauer associates.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229.

Carpenter, C. C. and Ferguson, G. W. (1977). Variation and evolution of stereotyped behavior in reptiles. *Biology of the Reptilia*, 7:335–554.

Chan, A. H. H., Putra, P., Schupp, H., Köchling, J., Straßheim, J., Renner, B., Schroeder, J., Pearse, W. D., Nakagawa, S., Burke, T., Griesser, M., Meltzer, A., Lubrano, S., and Kano, F. (2025). Yolo-behaviour: A simple, flexible framework to automatically quantify animal behaviours from videos. *Methods in Ecology and Evolution*, 16(4):760–774.

DeCourcy, K. R. and Jenssen, T. A. (1994). Structure and use of male headbob signals by the lizard Anolis carolinensis. *Animal Behaviour*, 47(2):251–262.

Dell, A. I., Bender, J. A., Branson, K., Couzin, I. D., de Polavieja, G. G., Noldus, L. P., Pérez-Escudero, A., Perona, P., Straw, A. D., Wikelski, M., and Brose, U. (2014). Automated image-based tracking and its application in ecology. *Trends Ecol. Evol.*, 29(7):417–428.

Ferrante, G. S., Rodrigues, F. M., Andrade, F. R. H., Goularte, R., and Meneguette, R. I. (2021). Understanding the state of the art in animal detection and classification using computer vision technologies. In *Proc. IEEE Int. Conf. Big Data*, pages 3056–3065.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Kaplan, G. (2014). Animal communication. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(6):661–677.

Labra, A., Carazo, P., Desfilis, E., and Font, E. (2007). Agonistic interactions in a Liolaemus lizard: structure of head bob displays. *Herpetologica*, 63(1):11–18.

Li, H., Wu, Z., Shrivastava, A., and Davis, L. S. (2022). Rethinking pseudo labels for semi-supervised object detection. *Proc. AAAI Conf. Artif. Intell.*, 36(2):1314–1322.

Li, Z., Xu, P., Chang, X., Yang, L., Zhang, Y., Yao, L., and Chen, X. (2023). When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10555–10579.

Martins, E. P., Labra, A., Halloy, M., and Thompson, J. T. (2004). Large-scale patterns of signal evolution: an interspecific study of Liolaemus lizard headbob displays. *Animal Behaviour*, 68(3):453–463.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, M. W., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289.

Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L., and da Silva, E. A. B. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3).

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., and Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1):117–125.

Rajagukguk, R. A., yeon Lee, S., yeon Park, J., Daniel, K. F., rin Lee, C., Chen, Z., Liu, D., Norton, T., Park, J., and woon Hong, S. (2025). Deep learning for visual animal monitoring (detection, tracking, pose estimation, and behavior classification): A comprehensive review. *Smart Agricultural Technology*, 12:101539.

Raza, A., Hanif, F., and Mohammed, H. A. (2025). Analyzing the enhancement of cnn-yolo and transformer based architectures for real-time animal detection in complex ecological environments. *Scientific Reports*, 15(1):39142.

Vicente, N. and Halloy, M. (2015). Male headbob display structure in a neotropical lizard, Liolaemus pacha: relation to social context. *Herpetological Journal*, 25(1):49–53.

Zenóbio, J. G. F., Silva, P. H. L., Luz, E. J. d. S., Moreira, G. J. P., Galdino, C. A. B., and Gertrudes, J. C. (2023). Reptilerecon: Um arcabouço para extração e análise de sinais de lagartos. *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. (2024). Detrs beat yolos on real-time object detection. In *Proc. IEEE/CVF CVPR*, pages 16965–16974.

Zhu, H., Wei, H., Li, B., Yuan, X., and Kehtarnavaz, N. (2020). A review of video object detection: Datasets, metrics and methods. *Applied Sciences*, 10(21).

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276.