Supplementary Information
A confidence-convergence decision layer for reliability-aware sequential recognition under limited observations

G. L.

Supplementary tables

Supplementary Table S1: Simulation dataset and scenario settings

Item	Setting	Description
Behavior types	$K = 3$	Stable, search-like, and intermittent correction behaviours
Episodes per trial	$N_{\mathrm{ep}} = 900$	Class-balanced, 300 episodes per behavior type
Complete horizon	$T = 100$	Full observation length of each episode
Window ratio	$\gamma = 0.2$–1.0	Limited-observation settings
Feature dimension	$M = 6$	Motion, stage, and observation-quality indicators
Noise level	$\sigma_o \in \{0.03, 0.06, 0.10\}$	Low, medium, and high observation noise
Missing ratio	$r_m \in \{0, 0.1, 0.2, 0.3\}$	Random missing observations
Maximum delay	$\delta_{\max} \in \{0, 1, 2, 3\}$	Bounded measurement delay
Dataset split	60% / 20% / 20%	540 training, 180 validation, and 180 test episodes per trial
Random trials	$N_{\mathrm{trial}} = 30$	Repeated runs with independent initial states and disturbances

Supplementary Table S2: Simulation assumptions and parameter roles

Component	Setting	Role in the simulation
Behavior abstraction	Stable, search-like, intermittent correction	Produces separable but partially overlapping motion-evolution patterns
State disturbance	Zero-mean process disturbance in position, speed, heading, acceleration, and heading rate	Represents unmodelled manoeuvre variation within each behavior class
Observation noise	Gaussian noise in the main protocol; Laplace and mixed-outlier noise in sensitivity tests	Controls feature uncertainty and tests sensitivity to distributional shape
Missing observations	Independent feature mask with $r_m \in \{0, 0.1, 0.2, 0.3\}$	Represents intermittent loss of feature measurements or sensor reports
Measurement delay	Bounded delay with $\delta_{\max} \in \{0, 1, 2, 3\}$ and safe normalization	Represents delayed arrival of underwater sensing information
Class priors	Balanced priors in the main protocol; 0.6/0.3/0.1 in sensitivity tests	Tests whether non-convergence depends on balanced simulated classes
Validation selection	Thresholds selected on validation splits within each trial	Separates parameter tuning from final test evaluation

Supplementary Table S3: Baseline methods for comparison

Method	Sequential update	Convergence criteria	Output
Static classifier	No	No	Forced label
Bayesian without convergence	Yes	No	Forced label
Threshold-only recognition	Yes	Single confidence threshold	Label or continue
Selective posterior	Yes	Validation-selected reject threshold	Label or abstention
Calibrated selective	Yes	Calibrated reject threshold	Label or abstention
Early-exit GRU	Yes	Calibrated early-exit threshold	Label or abstention
HMM-based recognition	Yes	No	Forced label
GRU-based recognition	Yes	No	Forced label
Proposed method	Yes	Dominance + stability + separation	Label / continue / non-convergence

Supplementary Table S4: Evaluation metrics

Metric	Symbol	Purpose
Accepted-decision accuracy	A_{lw}	Accuracy among accepted recognition decisions
All-episode accuracy	A_{all}	Accuracy over all test episodes, including non-convergence or abstention cases
Mean decision time	T_{dec}	Timeliness of accepted recognition
Confidence convergence time	T_{conv}	Time at which all convergence criteria are first satisfied
Wrong accepted-decision rate	R_{wad}	Incorrect labels among accepted outputs
Non-convergence rate	R_{nc}	Frequency of insufficient-evidence outputs
Coverage	C	Proportion of test episodes with accepted output
Accepted-decision risk	$100 - A_{\mathrm{lw}}$	Error risk among accepted outputs
Expected calibration error	ECE	Calibration quality of accepted confidence outputs
Robustness degradation	ΔQ	Performance change under uncertainty

Supplementary Table S5: Mean performance comparison under the standard limited-observation setting over 30 random trials

Method	A_{lw} (%)	A_{all} (%)	T_{dec}	R_{wad} (%)	R_{nc} (%)
Static classifier	75.8	75.8	N/A	24.2	N/A
Bayesian without convergence	79.0	79.0	40.0	21.0	N/A
Threshold-only	82.0	77.1	22.8	16.9	6.0
Selective posterior	84.9	78.1	25.5	14.1	8.0
Calibrated selective	86.7	79.8	27.0	12.3	8.0
Early-exit GRU	87.9	81.1	30.7	11.1	7.8
HMM	81.5	81.5	40.0	18.5	N/A
GRU	83.5	83.5	40.0	16.5	N/A
Proposed	91.2	82.1	29.4	7.9	10.0

Supplementary Table S6: Mean performance of the proposed method under different observation-window ratios over 30 random trials

Window ratio γ	T_w	A_{lw} (%)	T_{dec}	R_{wad} (%)	R_{nc} (%)
0.2	20	85.0	16.5	9.3	38.0
0.4	40	91.2	29.4	7.9	10.0
0.6	60	93.0	38.7	6.6	5.0
0.8	80	95.2	46.2	4.7	2.0
1.0	100	96.1	50.8	3.9	1.0

Supplementary Table S7: Mean ablation results for confidence convergence criteria over 30 random trials

Criterion combination	A_{lw} (%)	T_{dec}	R_{wad} (%)	R_{nc} (%)	ECE
Dominance only	82.0	22.8	16.9	6.0	0.086
Dominance + stability	88.0	28.6	10.6	12.0	0.061
Dominance + separation	87.4	27.9	11.2	10.8	0.064
Full criteria	91.2	29.4	7.9	10.0	0.044

Supplementary Table S8: Additional sensitivity tests under distribution shift and class imbalance

Scenario	A_{lw} (%)	R_{wad} (%)	R_{nc} (%)	Macro-F1 (%)
Gaussian reference	91.2	7.9	10.0	88.7
Laplace noise, matched variance	89.1	9.8	12.0	86.2
Mixed Gaussian outliers, 5% impulses	87.6	11.5	15.0	84.1
Imbalanced priors, 0.6/0.3/0.1	88.9	9.6	11.8	85.4