EMMA: Extracting Multiple physical parameters from Multimodal Data
Supplementary Material

Farhat Shaikh Ayan Banerjee Sandeep Gupta
IMPACT Lab, School of Computing & Augmented Intelligence (SCAI)
Arizona State University, Tempe, AZ
{fshaik12, abanerj3, Sandeep.Gupta }@asu.edu

S1. Ablation Study

This supplement expands on the architecture behind EMMA's multi-modal, physics-informed estimator by isolating the impact of (i) forcing inputs and audio wavelength, (ii) the LTC network vs. alternative sequence models, (iii) implicit dynamics, and (iv) invariant knowledge. We follow the same dynamical systems introduced in the main paper (pendulum), and the real-world rover cases with hidden inputs. See the architecture and training details in the main paper (Fig. 2; Secs. 3-4). *Where not stated otherwise, the loss and simulator are identical to the main setup.*

S1.1. Case Study with No Forcing Input (Pendulum)

S1.1.1. LTC Architecture: Pendulum

Setup. We ablated the LTC-NN using three alternative recurrent architectures (LSTM, GRU, and Transformer) as they share similar sequential structure. The pendulum example estimates two parameters and has no external force input $u(t)$, making it less complex.

From the results in Table S1, all architectures estimate very similar parameters with comparable accuracy, making our EMMA structure robust and versatile. This confirms that irrespective of architecture, EMMA is efficient on less complex examples with no force input.

S1.2. Case Study with Forcing Input

S1.2.1. LTC Architecture: Rover

Setup. The motivation and justification for using LTC-NN arises when we have a forcing input $u(t)$ and a more complex example. We used the same three alternative architectures as for the pendulum but on the more complex Rover example, which has multiple parameters to estimate along with external force.

The results in Table S2 show that the accuracy of other architectures degrades compared to LTC-NN on the more complex forced dynamics example. LTC-NN is therefore the most suitable architecture choice for EMMA, yielding accurate results with faster convergence.

Architecture	45 cm	90 cm	150 cm
LSTM	0.49±0.20	0.98±0.41	1.35±0.69
GRU	0.49±0.20	0.98±0.41	1.35±0.69
Transformer	0.49±0.20	1.07±0.42	1.35±0.69
LTC	0.49±0.20	0.98±0.41	1.35±0.69
Ground Truth L (m)	0.45	0.90	1.50

(a) Estimated length L (m)

Architecture	45 cm	90 cm	150 cm
LSTM	0.054±0.023	0.054±0.023	0.045±0.023
GRU	0.054±0.023	0.054±0.023	0.045±0.023
Transformer	0.054±0.023	0.053±0.021	0.045±0.023
LTC	0.054±0.023	0.054±0.023	0.045±0.023
Ground Truth	0.05	0.05	0.05

(b) Damping time-constant τ (1/s)

Table S1. Comparison of different architectures using the Pendulum (no forcing) example.

Parameters	GT	**LTC (Ours)**	GRU	LSTM	Transformer
X-arm length m	0.178	**0.173**	0.202	0.168	0.169
Y-arm length m	0.144	**0.133**	0.195	0.174	0.197
Mass kg	26.88	**27.79**	39.50	39.45	39.28
CoM height m	0.112	**0.119**	0.108	0.123	0.094
Wheel radius r	0.201	**0.205**	0.196	0.212	0.188
Convergence epoch	–	5	10	14	54
Training time/epoch	–	63.57s	22.36	25.07	62.82

Table S2. Comparison of different architectures using the Rover (forced dynamics).

S1.2.2. Multi-modal Ablation Without Audio

Setup. The use of multi-modal input is one of our key contributions, where we extract knowledge from different modalities making EMMA useful for various scenarios. We performed an ablation study on the rover example with video and audio input and compared it against video-only input. The setup was identical except the audio knowledge

was removed.

Parameters	GT	Video+Audio	Video-only
X-arm length (m)	0.178	0.163	0.189
Y-arm length (m)	0.144	0.133	0.203
Mass (kg)	26.88	27.79	39.64
CoM height (m)	0.112	0.129	0.108
Wheel radius (r)	0.201	0.245	0.171
Convergence epoch	–	5	30
Training time (s)	–	54.1	122.3

Table S3. Effect of audio on parameter recovery under forced dynamics.

Table S3 clearly demonstrates the importance of audio knowledge in parameter estimation. When more modalities are available, EMMA can observe and incorporate knowledge that better guides the model to estimate accurate parameters in less time.

S2. Physics Equations

We collect the governing equations for all systems used in the ablations, with parameters and units. These mirror the forms used in the simulator head of EMMA.

S2.1. Damped Pendulum

For a pendulum with angle θ, angular velocity $\omega = \dot{\theta}$, the dynamics follow:

$$\frac{d\theta}{dt} = \omega. \tag{S1}$$

$$\frac{d\omega}{dt} = -\frac{g}{L}\sin\theta - \frac{\tau}{L}\omega. \tag{S2}$$

Parameters: $L \in (0.1, 2.0]$ m (length), $\tau \in (0, 0.5]$ s^{-1} (damping coefficient), $g = 9.81$ m/s^2 (gravity).

S2.2. Torricelli Drainage

For fluid draining through an orifice, height $h(t)$ evolves as:

$$\frac{dh}{dt} = -K\sqrt{h}, \quad K = C_d A_{\text{orifice}}\sqrt{2g}/A_{\text{tank}}. \tag{S3}$$

Parameters: $K \in (0.001, 0.1]$ $\sqrt{\text{m}}$/s (drainage coefficient).

S2.3. LED Exponential Decay

Light intensity $I(t)$ follows first-order decay:

$$\frac{dI}{dt} = -\gamma I(t), \quad I(t) = I_0 e^{-\gamma t}. \tag{S4}$$

Parameters: $\gamma \in (0.01, 5.0]$ s^{-1} (decay rate).

S2.4. Sliding Block with Friction

Block on inclined plane with velocity v:

$$\frac{dv}{dt} = g\sin(\alpha) - \mu g\cos(\alpha). \tag{S5}$$

Parameters: $\alpha \in [10°, 45°]$ (incline), $\mu \in [0.1, 0.5]$ (friction).

S2.5. Free Fall

Vertical velocity v under quadratic drag:

$$\frac{dv}{dt} = g - kv^2\,\text{sign}(v). \tag{S6}$$

Parameters: $k = \frac{C_d \rho A}{2m}$ (drag coefficient).

S2.6. Differential-Drive Rover (9 Parameters)

The rover combines kinematic constraints with dynamic forces:

$$v = \frac{r(\omega_r + \omega_l)}{2}, \quad \dot{\psi} = \frac{r(\omega_r - \omega_l)}{W}. \tag{S7}$$

$$m\dot{v}_x = F_{\text{motor}} - F_{\text{friction}} - F_{\text{drag}}. \tag{S8}$$

Measured Parameters: $a = 0.178$ m (X-arm), $b = 0.144$ m (Y-arm), $r = 0.201$ m (wheel radius), $m = 26.88$ kg, $W = 0.32$ m (wheelbase).
Implicit Parameters: $k_f = 0.15$ (friction), $C_d = 0.42$ (drag), $C_M = 0.112$ m (CoM height).

S2.7. 6-DOF Quadrotor (12 Parameters)

Full rigid-body dynamics with rotor dynamics:

$$\tau^2\ddot{w}_i + 2\zeta\tau\dot{w}_i + w_i = k_p u_i. \tag{S9}$$

$$T_i = k_{Th} w_i^2, \quad \tau_i = k_{To} w_i^2. \tag{S10}$$

$$m\ddot{\mathbf{p}} = R(\mathbf{q})\mathbf{T} - mg\mathbf{e}_z - \mathbf{F}_{\text{drag}}. \tag{S11}$$

Measured Parameters: $k_{Th} = 1.1 \times 10^{-5}$ N·s^2/rad^2, $k_{To} = 1.3 \times 10^{-7}$ N·m·s^2/rad^2, $d_{xm} = 0.18$ m, $d_{ym} = 0.20$ m, $d_{zm} = 0.07$ m.
Audio-Inferred: $k_p = 0.91$, $\tau = 0.012$ s, $\zeta = 0.7$.

S3. Differentiable Trajectory Rollout

To ensure numerical stability and consistency with the architecture layout in Fig. 2 of the main paper, we employ a differentiable 4th-order Runge-Kutta (RK4) integrator. This provides higher-order error control compared to standard Euler integration, which is critical for stiff dynamical systems like the quadrotor.

Given estimated parameters $\hat{\boldsymbol{\theta}}$ from the LTC network and the continuous physics function $\mathbf{f}$, the state update from time t to $t+1$ is computed as:

$$k_1 = f(x_t, u_t; \hat{\theta}). \tag{S12}$$

$$k_2 = f\left(x_t + \frac{\Delta t}{2} k_1, u_{t+\frac{1}{2}}; \hat{\theta}\right). \tag{S13}$$

$$k_3 = f\left(x_t + \frac{\Delta t}{2} k_2, u_{t+\frac{1}{2}}; \hat{\theta}\right). \tag{S14}$$

$$k_4 = f(x_t + \Delta t k_3, u_{t+1}; \hat{\theta}). \tag{S15}$$

$$x_{t+1} = x_t + \frac{\Delta t}{6}(k_1 + 2k_2 + 2k_3 + k_4). \tag{S16}$$

where the simulation time step is clamped at $\Delta t = \min(0.03, \text{fps}^{-1})$. Intermediate control inputs $\mathbf{u}_{t+\frac{1}{2}}$ are obtained via linear interpolation of the forcing signal. The simulation runs for $T_{\text{sim}} = \min(500, T)$ steps. Parameter physical constraints are enforced via soft clamping: $\theta_i \leftarrow \max(\epsilon, \theta_i)$ with $\epsilon = 10^{-4}$.

S4. Additional Robustness and Ablation Experiments

This section reports five additional experiments validating EMMA's robustness across feature extraction backbones, architecture choices, initialization sensitivity, statistical reproducibility, and audio noise levels.

S4.1. Optical Flow vs. YOLO (Detector Agnosticism)

To verify that EMMA's physics extraction is independent of the object detection front-end, we replaced YOLOv11 with unsupervised Farneback optical flow tracking on both the rover and pendulum systems. Table S4 shows that optical flow achieves comparable accuracy, confirming that EMMA's core contribution lies in the LTC physics layer rather than the feature extractor.

(a) Rover

Parameter	Ground Truth	YOLOv11	Optical Flow
a (m)	0.178	0.196	**0.184**
b (m)	0.144	**0.134**	0.110
r (m)	0.201	0.223	**0.202**

(b) Pendulum (L, m)

Configuration	Ground Truth	YOLOv11	Optical Flow
45 cm	0.45	**0.50±0.04**	0.66±0.00
90 cm	0.90	0.86±0.07	**0.89±0.00**
150 cm	1.50	**1.50±0.00**	1.49±0.00

Table S4. YOLOv11 vs. unsupervised optical flow on rover and pendulum. Optical flow achieves comparable accuracy without any pretrained detector. Best per row in bold.

S4.2. Statistical Validation (Multi-Seed Reproducibility)

Table S5 reports rover parameter estimates over 5 random seeds (42–46), providing standard deviations that quantify reproducibility. Average error across four measurable parameters is 9.5% ± 8.9%.

Parameter	Ground Truth	Mean ± Std	Error (%)
a (m)	0.178	0.184 ± 0.020	3.4
b (m)	0.144	0.176 ± 0.031	22.2
r (m)	0.201	0.222 ± 0.027	10.4
CM (m)	0.112	0.114 ± 0.008	1.8

Table S5. Rover statistical validation over 5 random seeds (42–46). ± denotes standard deviation across seeds.

S4.3. Drone Statistical Validation (Multi-Seed Reproducibility)

Table S6 reports drone parameter estimates over 5 random seeds (42–46), providing standard deviations that quantify reproducibility. Average error across all measurable parameters is 17.5%.

Parameter	**Ground Truth**	**Mean ± Std**	**Error (%)**
k_{Th}	1.1	1.076 ± 0.000	2.2
k_{To}	1.3	1.632 ± 0.000	25.5
k_p	0.91	1.000 ± 0.000	9.9
τ_2	0.012	0.015 ± 0.000	25.0
d_{xm}	0.18	0.160 ± 0.000	11.1
d_{ym}	0.20	0.160 ± 0.000	20.0
d_{zm}	0.07	0.050 ± 0.000	28.6

Table S6. Drone statistical validation over 5 random seeds (42–46). ± denotes standard deviation across seeds.

S4.4. LTC vs. Neural ODE vs. CT-GRU

We compare LTC against two continuous-time alternatives: Neural ODE and CT-GRU. Table S7 shows that all architectures perform comparably on the unforced pendulum. Under forcing inputs (rover), LTC outperforms Neural ODE by approximately 25% and CT-GRU by approximately 5% in average parameter error, validating that input-dependent time constants are critical for modeling forced dynamics.

S4.5. Initialization Sensitivity

To evaluate robustness to poor initialization, we expand parameter bounds by 200% and initialize far from ground truth. Table S8 shows EMMA achieves <10% error in 5 out of 6 configurations, confirming that accurate estimation does not require initialization close to ground truth.

S4.6. Audio Noise Robustness

We inject additive Gaussian noise at SNR levels of 20, 10, and 5 dB into the rover audio stream. Table S9 shows that

(a) Rover (Forcing Input)

Parameter	Ground Truth	LTC	Neural ODE	CT-GRU
a (m)	0.178	0.196	0.238	**0.179**
b (m)	0.144	**0.134**	0.212	0.186
r (m)	0.201	**0.223**	0.244	0.226
Avg. Error		**9.3%**	34.1%	14.1%

(b) Pendulum (No Forcing)

Configuration	Ground Truth	LTC	Neural ODE	CT-GRU
45 cm	0.45	**0.50**	0.59	0.54
90 cm	0.90	0.86	0.99	**0.94**
150 cm	1.50	**1.50**	1.56	1.48
Avg. Error		**5.2%**	15.0%	8.6%

Table S7. Continuous-time architecture comparison. All architectures perform comparably without forcing; under forcing inputs, LTC outperforms Neural ODE by approximately 25% and CT-GRU by approximately 5%. Best per row in bold.

(a) Sliding Block (α)

Config	GT	Init	Est	Error (%)
Low	20°	25°	20.69°	**3.45**
Mid	25°	30°	25.48°	**1.92**
High	30°	45°	28.25°	**5.84**

(b) Pendulum (L, m)

Config	GT	Init	Est	Error (%)
45 cm	0.45	0.85	0.447	**0.73**
90 cm	0.90	1.30	0.683	24.08
150 cm	1.50	1.10	1.620	**8.02**

Table S8. Initialization sensitivity with 200% expanded bounds and distant initialization. Bold indicates $<$10% error. EMMA converges accurately in 5 of 6 configurations.

parameter estimates vary by less than 1.1% across all noise levels, demonstrating that EMMA's audio pipeline degrades gracefully under realistic acoustic interference.

Parameter	Ground Truth	SNR 20 dB	SNR 10 dB	SNR 5 dB	Var (%)
a (m)	0.178	0.205	0.205	0.205	0.15
b (m)	0.144	0.165	0.165	0.165	0.55
r (m)	0.201	0.184	0.186	0.186	1.08
CM (m)	0.112	0.115	0.115	0.115	0.04

Table S9. Rover audio noise robustness. Additive Gaussian noise at three SNR levels causes $<$1.1% variation in all estimated parameters.