The Ohio State University

EDGE Institute

LAWFUL

Law-Aligned Witness for Faithful Use of Latents

Kevin Chen[1], Kenneth W. Parker[2], Anish Arora[1]

[1]The Ohio State University [2]Independent Researcher

How would you know if your model *learned* a physics law — or only appears to?

We provide a framework that leverages *physics counterfactuals* to test whether a law is *causally used*.

On a radar-synthesis transformer, we localize the Doppler law to a compact circuit.

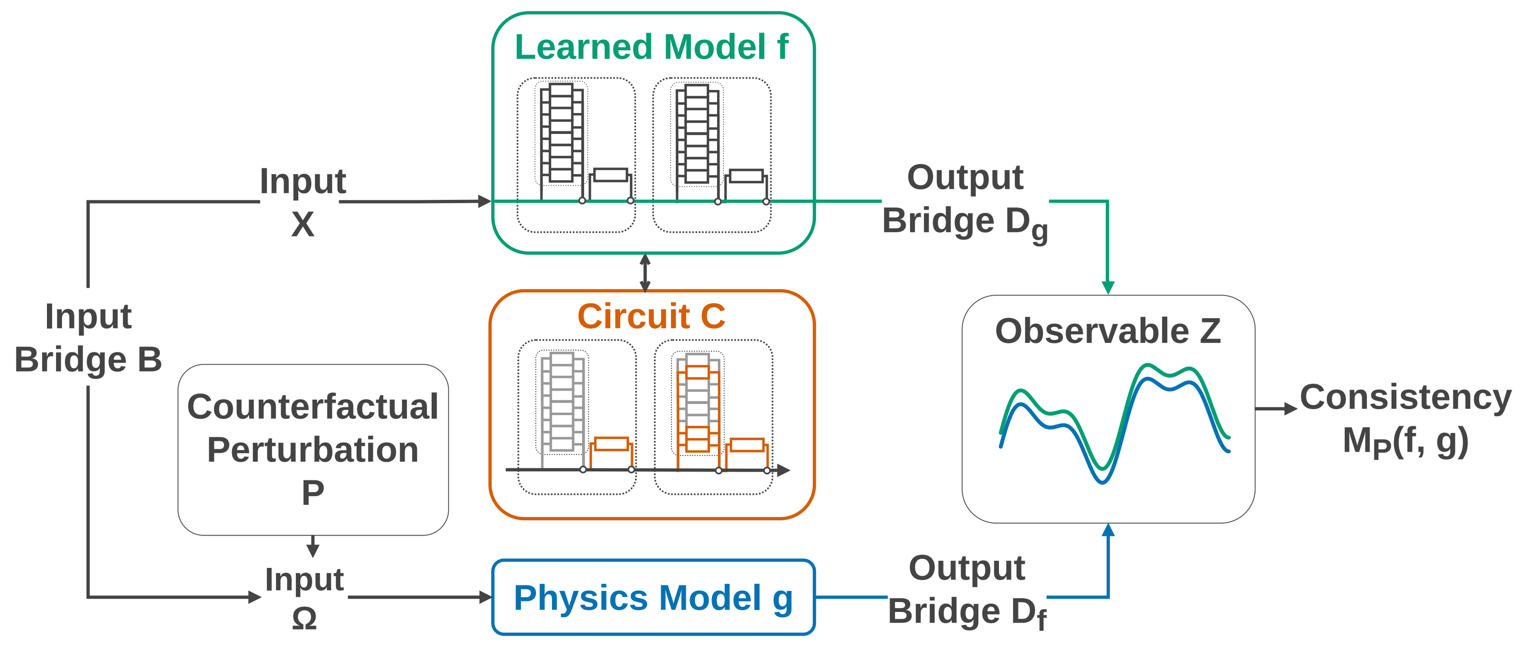
Has the Model Learned the Law?

Accurate prediction $\neq$ learning the law.

An ML model can match a physical system's outputs without internalizing the law that governs it. Establishing that the law is actually *learned* requires evidence about the network's internal computation, not just its predictions.

★ Framework for Aligning Model and Law

Align a learned model f with a physics model g through shared observables and continuous counterfactuals.



The input bridge B maps model inputs into the physics space Ω. Output bridges D_f, D_g project f and g into a shared observable Z, where agreement under perturbations $\mathcal{P}$ is scored.

Physical consistency score. For baseline and perturbed outputs z_f, z_g, LAWFUL scores agreement with a functional Φ:

$$M_{\mathcal{P}}(f, g) := \Phi(z_f, z_g), \qquad M_{\mathcal{P}} \leq 1,$$

where larger values indicate closer agreement with the physics law over the counterfactual family $\mathcal{P}$.

Defining the Circuit

A circuit is a subgraph that recovers physical consistency under patching, while its complement fails to do so.

Treat f as a computational graph and patch a candidate subgraph C with activations from the perturbed input. **Sufficiency** routes the perturbation through C alone; **necessity** routes it through $V \setminus C$:

$$\Delta_{\text{suf}}(C) = M_0 - M_{\mathcal{P}}^{\text{suf}}(C, g), \quad \Delta_{\text{nec}}(C) = M_0 - M_{\mathcal{P}}^{\text{nec}}(C, g).$$

Small Δ_{suf} means C recovers the unpatched consistency M_0; large Δ_{nec} means the complement cannot. The minimal C satisfying both at threshold τ is the **τ-physically consistent circuit**.

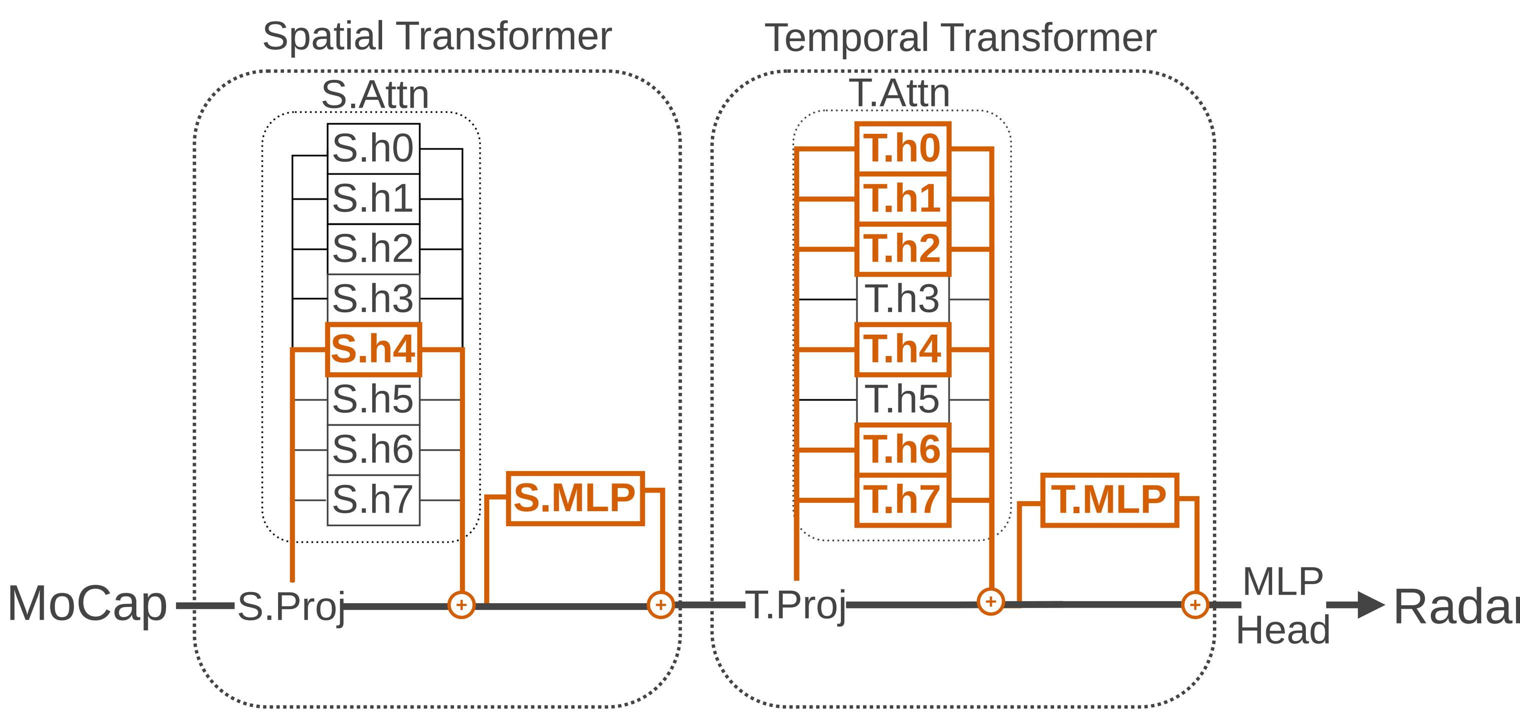
Doppler Law in MoCap2Radar

Test whether the model's Doppler response follows the Doppler law, although neither Doppler frequency nor radial velocity is given as supervision:

$$f_d = 2\, v^{\text{rad}} / \lambda.$$

The learned model f (MoCap2Radar) maps 3D marker trajectories to a micro-Doppler spectrogram. The physics model g applies this law to each marker's radial velocity.

Bridges & perturbation. B extracts radial velocities; D_f, D_g map model and physics outputs to a shared power-weighted Doppler centroid. The perturbation family $\mathcal{P} = \{p_\alpha\}$ scales radial velocity by α, and $M_{\mathcal{P}}$ scores agreement between $\hat{m}_f(\alpha)$ and $\hat{m}_g(\alpha)$ across α.



Architecture of MoCap2Radar with $|V| = 18$ patchable components. Highlighted components mark the τ=0.9 circuit $C^\dagger$.

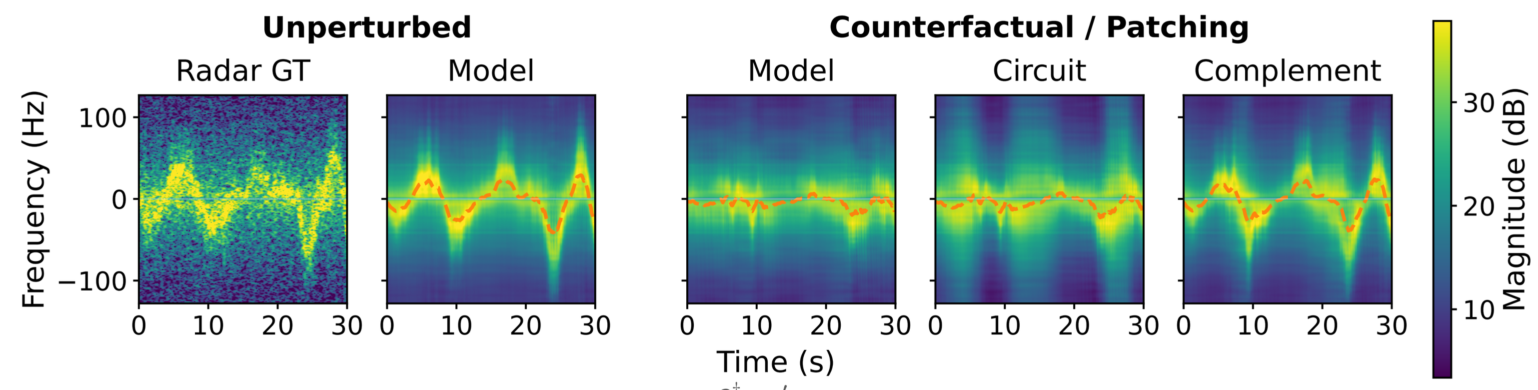
Searching for the Circuit

Doppler consistency is concentrated in a temporal-block core.

The unpatched model is highly Doppler-consistent ($M_0 = 0.988$). Greedy search over $\tau \in \{0.3, \ldots, 0.9\}$ first identifies a three-component temporal core, T:{MLP, h6, h7}. At $\tau = 0.9$, the primary circuit $C^\dagger$ recovers **91%** of M_0 using 9/18 components, with 7/9 in the temporal block.

$\lvert C \rvert$	τ	circuit C	Δ_{suf}	Δ_{nec}
3	0.3–0.5	T:{MLP,h6,h7}	0.49	1.89
4	0.6	+ T:{h1}	0.35	2.43
5	0.7	+ S:{MLP}	0.24	3.08
6	0.8	+ T:{h4}	0.18	3.19
9	0.9	$C^\dagger$	**0.09**	3.50

★ Circuit Reproduces the Physics Law

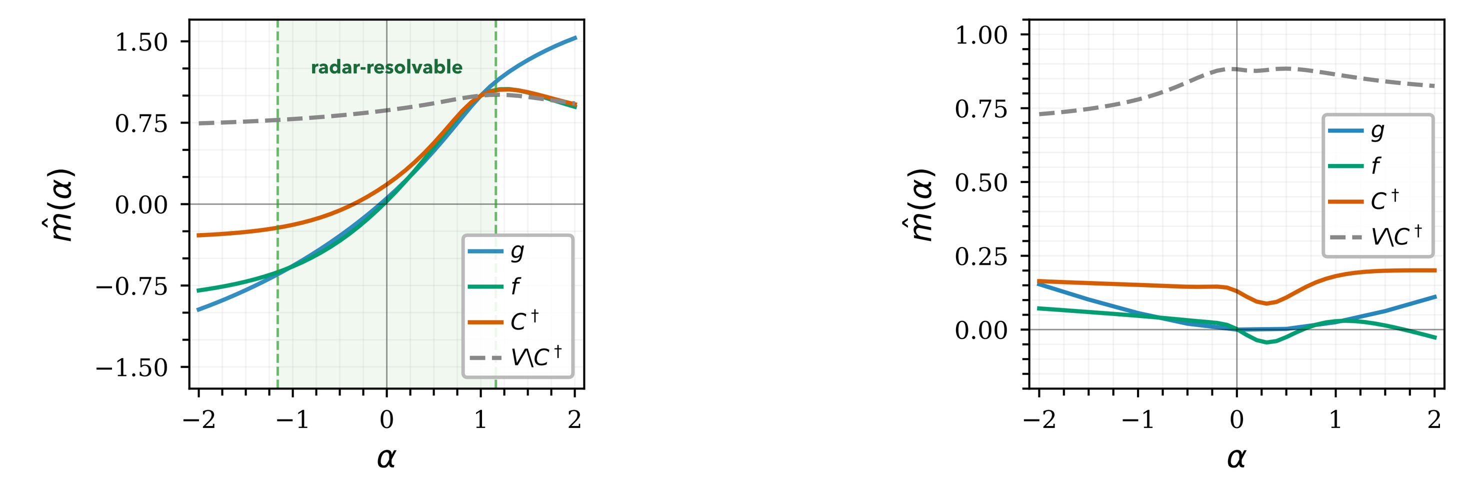


Doppler spectrograms, α=0.3. The **Circuit** ($f^{C^\dagger \leftarrow x'}$) reproduces the amplitude-compressed centroid of the counterfactual Model; the **Complement** ($f^{V \setminus C^\dagger \leftarrow x'}$) retains the unperturbed amplitude. Orange dashed = Doppler centroid from D_f.

★ Testing the Domain of Validity

f and $C^\dagger$ follow the Doppler law inside the radar-resolvable regime and remain near-invariant to tangential motion.

The radar resolves $|v^{\text{rad}}| \leq 2.6$ m/s. With maximum body-level $|v^{\text{rad}}| = 2.24$ m/s, this gives $\alpha \in [-1.16, 1.16]$ as the resolvable scaling regime.



(a) Radial scaling. f and $C^\dagger$ track g within the resolvable band; $V \setminus C^\dagger$ remains nearly flat.

(b) Tangential scaling. All responses remain small, consistent with the absence of first-order Doppler shift.

Emergence of Radial Velocity

Radial velocity emerges jointly across temporal heads and is passed to T.MLP through the attention output.

No single head in $C^\dagger$ linearly decodes v^{rad} well ($R^2 \leq$ **0.21**), but their joint output reaches $R^2 =$ **0.76**. T.MLP also decodes v^{rad} ($R^2 =$ **0.56**), which collapses to $R^2 =$ **0.0004** when the upstream attention output is mean-ablated. Thus, T.MLP inherits radial-velocity information from temporal attention.

Inside the Attention Heads

The Doppler-scaling response is carried primarily by attention patterns (QK) rather than values (VAL).

Path patching within $C^\dagger_{\text{attn}}$ separates the attention-pattern pathway (QK) from the value pathway (VAL). Patching QK leaves a small consistency gap ($\Delta^{\text{QK}}_{\text{suf}} =$ **0.35**), while patching VAL leaves a much larger gap ($\Delta^{\text{VAL}}_{\text{suf}} =$ **1.31**). One mechanism consistent with this is adaptive temporal differentiation, where heads may compare frame pairs at shorter intervals for faster motion.

Limitations

- **Single case study.** Broader instantiations and comparisons across identified circuits are needed.
- **Centroid-level observable.** Spread, harmonics, and limb-level micro-Doppler structure may involve different circuits.
- **Partial mechanism.** How the circuit enforces invariants and how downstream layers consume v^{rad} remain open.