LAWFUL: Law-Aligned Witness for Faithful Use of Latents

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

Abstract

When a neural network predicts a physical system accurately, has it learned the governing law as formal, structured knowledge, and if so, does the network's internal computation actually use that representation throughout the law's domain of validity? We identify four interpretability gaps that limit answering these questions for *physics laws over continuous variables*: the absence of a coverage-aware causal-consistency measure over continuous counterfactuals; of a domain-of-validity test for the identified circuit; of a verification of the law's invariants and forbidden behaviors; and of a quantification of how a derived physical quantity flows through the circuit. We develop a foundational framework, LAWFUL, that closes the first two and lays groundwork for the remaining two, and illustrate it on the Mocap2Radar transformer, validating whether it learns and internally uses the Doppler frequency law $f(t) = \frac{2v(t)}{\lambda}$ from motion-capture and radar data in which neither $f(t)$ nor $v(t)$ appears. Our source code is available at: `https://github.com/aciculachen/LAWFUL`

1. Introduction

When a neural network predicts a physical system accurately, has it actually learned the governing physical law as formal, structured knowledge over the continuous variables that law involves? If so, is that representation causally used by the network's internal computation, or is the representation merely something an external probe can decode (Belinkov, 2022; Hewitt & Liang, 2019)? And does the representation hold throughout the domain of validity of the law, or only within the narrow distribution from which the training or validation data were drawn? These are not three rephrasings of "is the model accurate"; they are three distinct claims about what the network has internalized, and answering them requires mechanistic interpretability rather than behavioral evaluation, and a formal framework that can do so would furnish the principled basis on which the model can be trusted to behave lawfully wherever the law does.

Despite recent progress in mechanistically interpreting how neural models learn physical phenomena, the state of the art for laws expressed over continuous physical variables leaves several interpretability gaps that no existing framework closes. First, although causal abstraction with distributed alignment search and interchange-intervention accuracy (Geiger et al., 2025; 2024; Wu et al., 2023) supplies a formal notion of faithfulness, current instantiations quantify causal consistency only on narrowly sampled, discrete counterfactual pairs (Sutter et al., 2025; Makelov et al., 2024; Méloux et al., 2025); they provide no measure designed to certify consistency over counterfactual families that cover the *continuous* space of the law's variables. Second, no published mechanistic procedure establishes whether an identified physics circuit operates consistently across the full range over which the law actually holds (Friedman et al., 2024; Nanda et al., 2023; Vafa et al., 2025; Liu et al., 2026)—that is, whether it generalizes over the law's *domain of validity* rather than over the empirical training or validation distribution. Circuits with provable guarantees over continuous input domains have been considered for robustness over input perturbations (Hadad et al., 2026), but not for consistency with a known physics law over a derived-feature space.

We develop a foundational treatment that closes both gaps within a single coherent formal framework, LAWFUL, supplying a coverage-aware causal-consistency measure defined over continuous counterfactual families together with a domain-of-validity test for the circuit; see Figure 1. We illustrate the framework on a problem in radar physics: validating whether the sequence-to-sequence transformer MoCap-to-Radar (Chen et al., 2026) explicitly learns, and internally uses, the Doppler frequency law, $f(t) = \frac{2v(t)}{\lambda}$, where $f(t)$ and $v(t)$ are the Doppler frequency and radial velocity at time t, and λ the radar wavelength. The model maps three-dimensional motion-capture (MoCap) trajectories to radar spectrograms via a spatial transformer composed sequentially with a temporal transformer, trained on MoCap

[1]Computer Science & Engineering, The Ohio State University, Columbus, Ohio, USA [2]Independent Researcher, Waterford, Virginia, USA. Correspondence to: Kevin Chen <chen.11020@osu.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

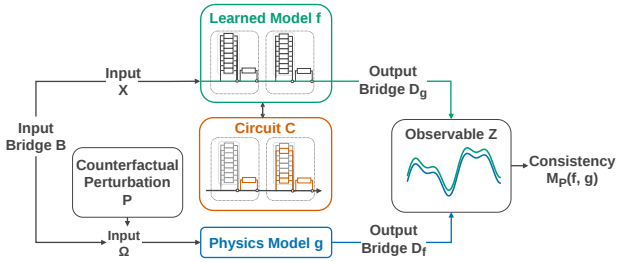


Figure 1: In LAWFUL, an input bridge maps the learned model input to the physics model input, which allows considering counterfactual perturbations over physically meaningful continuous domains. The output bridges project both models' outputs into a shared observable space, from which the physical consistency score is computed. This score serves as the metric for circuit identification and for checking invariances in the physics model.

recordings of a human subject wearing 53 reflective markers paired with measurements from a Bumblebee (The Samraksh Company, 2024) homodyne radar. Crucially, neither $f(t)$ nor $v(t)$ appears in the training data, and the trained model generalizes to out-of-distribution targets — for example, random walks drawn from a class disjoint from the fixed-direction walks used during training.

The same framework lays the groundwork for redressing two further interpretability gaps, which we illustrate only partially through the MoCap-to-Radar model. The first concerns the *negative* content of a physical law, i.e., its associated invariants, conservation principles, and forbidden behaviors (Marks et al., 2025; Belrose et al., 2023; Barbiero et al., 2026; Crabbé & van der Schaar, 2023), for which no existing procedure verifies mechanistic enforcement inside the circuit, since concept-erasure and causal-scrubbing tools target spurious correlations rather than physically mandated invariances. The second concerns the *flow of the physical variables* themselves through the circuit: while path-patching, attribution-patching, and attribution-graph methods (Wang et al., 2023; Conmy et al., 2023; Syed et al., 2024; Heimersheim & Nanda, 2024) identify which edges matter, they do not quantify how the continuous signal of a derived physical quantity is written, transformed, transmitted, and read by the heads and MLPs that constitute the law-implementing pathway. Combined with the causal-consistency measure and the domain-of-validity test, the invariant-enforcement certificate and signal-flow quantification enabled by our formal framework yield a principled measure of how faithfully a neural model embodies the physical law it appears to predict — closing a measurement gap that no current framework addresses for laws over continuous variables.

Our contributions are threefold:

1. We propose a physics-grounded interpretability framework, LAWFUL, that aligns a learned model with an explicit physics law and witnesses their agreement over possibly continuous counterfactual physical states in the region of validity of the law.
2. LAWFUL further defines a physical consistency score over a family of counterfactual perturbations and identifies physically consistent circuits as subgraphs that recover the score under activation patching.
3. Instantiating LAWFUL on a MoCap-to-Radar Doppler task, we characterize its consistency over the region of validity with respect to the Doppler Frequency Law and its invariance to tangential velocity, and we identify a circuit that recovers 91% of the model's Doppler consistency using half of its components, with the response to velocity scaling carried by temporal attention patterns rather than value transformation, suggesting an adaptive form of temporal differentiation.

2. Law-Aligned Interpretability Framework

Our framework formally defines a measure of how consistently a model behaves with respect to a known physical law over continuous variables based on families of counterfactual states, as well as an attribution of that consistency to subsets of internal components in the model. The formalization is in terms of two concepts: *physical consistency*, a score that measures how closely the model's behavior agrees with the physical law under a family of controlled perturbations (§2.2); and *physically consistent circuit*, which is a subgraph of f that recovers this score under activation patching with the family of counterfactual perturbations (§2.3).

2.1. Aligning a Neural Network and a Physical Law

Our framework pairs a *learned model* f with a *physics model* g that captures a known physical law expected to hold in the domain of f. The learned model $f : X \to Y_f$ maps inputs from a space X to outputs in a space Y_f; the physics model $g : \Omega \to Y_g$ is a function from domain-interpretable quantities in a continuous space Ω to outputs in a space Y_g. To measure how f aligns with g we evaluate the former's behavior against that of the latter. Throughout, we fix an evaluation set $\{x^{(i)}\}_{i=1}^{N} \subseteq X$ of N inputs for which consistency is measured.

2.2. Physical consistency

To compare f and g, we introduce domain-grounded bridges that express each model in terms of a common physical observable, together with a perturbation family on Ω.

Input bridge. Raw inputs of f can carry a mixture of information, including the physical quantities on which g is

defined. We extract these quantities via a map $B : X \to \Omega$ that carries raw inputs into the physics space, taking $\Omega := B(X)$ as the physics space reachable by inputs to f. The choice of B determines the physical variables of interest.

Output bridge. The two models' outputs can live in different spaces and carry different information densities. We choose a common physical observable Z and project each model onto it via $D_f : Y_f \to Z$ and $D_g : Y_g \to Z$. The choice of Z determines which aspect of the physics is being tested.

Perturbation family. We specify a family $\mathcal{P} = \{p_\alpha\}_{\alpha \in A}$ of physically meaningful perturbation operations $p_\alpha : \Omega \to \Omega$ that generate counterfactual physical states at which f and g can be compared. For each (i, α), we assume access to $x'^{(i,\alpha)} \in X$ with $B(x'^{(i,\alpha)}) = p_\alpha(B(x^{(i)}))$, so that f and g can be evaluated at the same perturbed physical state.

Remark. The choice of $\mathcal{P}$ determines which physical relationships in g the analysis probes, and is dictated by the law under study. The choice of A determines the coverage and domain-of-validation in the continuous space Ω. Computing $\mathcal{P}$ for (countably or uncountably) infinite A involves computing p_α for a finite, potentially sparse set of α samples, and then using approximation, interpolation, or signal reconstruction techniques to span all $\alpha \in A$.

Consistency score. We measure the agreement between the bridged outputs of f and g by a scalar functional Φ. The bridged outputs, for every i and $\alpha \in A$, are

$$\begin{aligned} \bar{z}_f^{(i)} &= D_f(f(x^{(i)})), & z_f^{(i,\alpha)} &= D_f(f(x'^{(i,\alpha)})), \\ \bar{z}_g^{(i)} &= D_g(g(B(x^{(i)}))), & z_g^{(i,\alpha)} &= D_g(g(B(x'^{(i,\alpha)}))). \end{aligned} \tag{1}$$

The baselines $\bar{z}_f^{(i)}, \bar{z}_g^{(i)}$ anchor each model, while the perturbed outputs $z_f^{(i,\alpha)}, z_g^{(i,\alpha)}$ reveal how each model responds to $\mathcal{P}$. Collecting these into $\mathbf{z}_f = \{\bar{z}_f^{(i)}, z_f^{(i,\alpha)}\}_{i,\alpha}$ and $\mathbf{z}_g = \{\bar{z}_g^{(i)}, z_g^{(i,\alpha)}\}_{i,\alpha}$, Φ takes the form

$$\Phi : (\mathbf{z}_f, \mathbf{z}_g) \mapsto \mathbb{R}.$$

Φ is a design choice—pointwise error metrics give absolute deviation, similarity-based measures give response alignment—and we adopt the convention that larger Φ indicates closer agreement, with Φ at its supremum when f matches g's ideal response. Since $\mathbf{z}_f, \mathbf{z}_g$ are determined by (f, g) once the evaluation set, bridges, and perturbation family are fixed, we define the *consistency score*

$$M_{\mathcal{P}}(f, g) \coloneqq \Phi(\mathbf{z}_f, \mathbf{z}_g). \tag{2}$$

Across choices of Φ, $M_{\mathcal{P}}(f, g)$ remains physics law-grounded: a change in $M_{\mathcal{P}}$ reflects a change in f's *physically consistent behavior*, i.e., its agreement with g on counterfactual physical states generated by $\mathcal{P}$.

2.3. Physically consistent circuit

We now identify which components of f account for the agreement captured by $M_{\mathcal{P}}(f, g)$. Treating f as a computational graph $G = (V, E)$, where vertices V are model components (e.g., attention heads, sublayers) and directed edges E represent the residual-stream information flow between them (Elhage et al., 2021), we probe subgraphs C of G and demonstrate the framework using activation patching.

Configurations for counterfactual perturbations. Let $f^{C \leftarrow s}(b)$ denote a forward pass on base b in which the activations on C are replaced by those captured during a forward pass on source s, following standard activation patching (Wang et al., 2023; Heimersheim & Nanda, 2024). For each $(x^{(i)}, x'^{(i,\alpha)})$, we apply two configurations, both rooted at the unperturbed base $x^{(i)}$ with the perturbed counterfactual input $x'^{(i,\alpha)}$ as the activation source. The *sufficiency* configuration $f^{C \leftarrow x'^{(i,\alpha)}}(x^{(i)})$ routes the perturbation through C alone, holding the complement $V \setminus C$ at its base activations. The *necessity* configuration $f^{V \setminus C \leftarrow x'^{(i,\alpha)}}(x^{(i)})$ routes the perturbation through the complement, holding C at its base activations. Sufficiency asks whether routing the perturbation through C alone reproduces f's response to the full perturbation; necessity asks whether that response survives when the perturbation reaches every component except C.

Evaluating M under patching. Both configurations modify only the f-side perturbed-state slot $\{z_f^{(i,\alpha)}\}$; the baselines $\{\bar{z}_f^{(i)}\}$ and the physics-side outputs $\{\bar{z}_g^{(i)}, z_g^{(i,\alpha)}\}$ remain unchanged. For the sufficiency configuration, $z_f^{(i,\alpha)}$ is replaced by $D_f(f^{C \leftarrow x'^{(i,\alpha)}}(x^{(i)}))$; for the necessity configuration, by $D_f(f^{V \setminus C \leftarrow x'^{(i,\alpha)}}(x^{(i)}))$. Substituting into Eq. (2) yields the patched consistency scores $M_{\mathcal{P}}^{\mathrm{suf}}(C, g)$ and $M_{\mathcal{P}}^{\mathrm{nec}}(C, g)$, evaluated against the unpatched baseline $M_0 \coloneqq M_{\mathcal{P}}(f, g)$.

Interpreting the patched scores. We measure each patched score relative to the unpatched baseline score, M_0, by

$$\begin{aligned} \Delta_{\mathrm{suf}}(C) &\coloneqq M_0 - M_{\mathcal{P}}^{\mathrm{suf}}(C, g), \\ \Delta_{\mathrm{nec}}(C) &\coloneqq M_0 - M_{\mathcal{P}}^{\mathrm{nec}}(C, g). \end{aligned} \tag{3}$$

A sufficiency deviation $\Delta_{\mathrm{suf}}(C)$ close to zero indicates that routing the perturbation through C alone recovers f's physical consistency. A large necessity deviation $\Delta_{\mathrm{nec}}(C)$ indicates that routing the perturbation through every component except C fails to recover it.

Definition of physically consistent circuit. For $\tau \in (0,1)$ and $M_0 > 0$, a subgraph C is *τ-consistent* for $(f, g, \mathcal{P})$ if

$$\Delta_{\mathrm{suf}}(C) \le (1-\tau)\, M_0,$$
$$\Delta_{\mathrm{nec}}(C) \ge \tau\, M_0.$$

A τ-consistent C is *minimal* if no $C \setminus \{c\}$ for $c \in C$ is itself τ-consistent. A *τ-physically consistent circuit* is a minimal τ-consistent subgraph.

Remark. The patching targets recovery of the unpatched baseline M_0, so the resulting analysis is specific to physically consistent behavior: a subgraph C that recovers M_0 accounts for f's agreement with g under $\mathcal{P}$. This specializes the standard notion of circuit by taking the recovery target to be physical consistency rather than task output, and the identified circuit therefore depends on $\mathcal{P}$: different perturbation families may expose different subgraphs of f.

3. Instantiation: Doppler Frequency Consistency in MoCap-to-Radar

We instantiate the framework for the physical-consistency evaluation of Chen et al. (2026): we specify the learned transformer model f and physics model g (§3.1); the bridges, perturbation families, and consistency scores (§3.2); and the circuit discovery procedure on the resulting computational graph (§3.3).

3.1. Transformer and Doppler frequency model

We pair a MoCap-to-radar transformer f with a Doppler frequency model g, both instantiated on a single continuous MoCap recording of 3D marker positions, with $x_{m,t} \in \mathbb{R}^3$ denoting the 3D position of marker m at sample index t, with $m = 1, \ldots, M$. The recording is partitioned into N overlapping STFT windows of length L; each window $x^{(i)} \in \mathbb{R}^{M \times L \times 3}$ is fed to f, forming the evaluation set $\{x^{(i)}\}_{i=1}^{N}$ for which consistency is measured. We suppress the window index (i) henceforth, except when aggregating across windows.

The learned model f is a spatio-temporal transformer that composes a spatial transformer over markers with a temporal transformer over frames, each a single self-attention layer followed by an MLP, to predict a dB-scale micro-Doppler spectrum $S^{\mathrm{pred}} = f(x) \in \mathbb{R}^K$, where K is the number of Doppler-frequency bins.

The physics model g applies the Doppler frequency law independently to each MoCap marker:

$$g(v^{\mathrm{rad}}) \;=\; \frac{2\, v^{\mathrm{rad}}}{\lambda}, \tag{4}$$

where λ is the radar wavelength. Its linearity implies that scaling every marker's radial velocity by a common α scales every marker's Doppler frequency by α.

3.2. Doppler consistency

Input bridge. Let $x_{\mathrm{radar}} \in \mathbb{R}^3$ denote the radar location in the MoCap coordinate frame. For each marker m, the range to the radar is $r_{m,t} = \|x_{m,t} - x_{\mathrm{radar}}\|_2$, and radial velocity is approximated by central finite difference on the recording:

$$v^{\mathrm{rad}}_{m,t} \;\approx\; \frac{r_{m,t+1} - r_{m,t-1}}{2\,\Delta t}. \tag{5}$$

The bridge $B : \mathbb{R}^{M \times L \times 3} \to \mathbb{R}^{M \times L}$ maps a window x to the corresponding sequence of $(v^{\mathrm{rad}}_{m,t})$.

Output bridge. Both bridges target the Doppler centroid, a scalar summary of how spectral power is distributed across frequency. The learned-side bridge $D_f : \mathbb{R}^K \to \mathbb{R}$ computes the power-weighted mean Doppler frequency of the predicted spectrum:

$$D_f\big(S^{\mathrm{pred}}\big) = \frac{\sum_{k=0}^{K-1} P_k\, \nu_k}{\sum_{k=0}^{K-1} P_k}, \tag{6}$$

where $P_k = 10^{S^{\mathrm{pred}}_k/10}$, and ν_k is the Doppler frequency at bin k.

Since radar returns aggregate scattering across the body, the physics-side bridge $D_g : \mathbb{R}^{M \times L} \to \mathbb{R}$ aggregates per-marker Doppler frequencies into a body-level centroid with weights w_m approximating each marker's radar cross-section (RCS) via body-surface-area (BSA) proportions:

$$D_g(g(v^{\mathrm{rad}})) \;=\; \frac{1}{L} \sum_{t=1}^{L} \sum_{m=1}^{M} \frac{w_m}{\sum_j w_j}\, g(v^{\mathrm{rad}}_{m,t}). \tag{7}$$

Perturbation family. We perturb radial velocity by uniform scaling: for $\alpha \in \mathbb{R}$, p_α scales all radial velocities by α. Since $v^{\mathrm{rad}}_{m,t}$ is obtained by finite differencing the range, and the first-order change in range equals the line-of-sight component of the frame-to-frame displacement, scaling the radial velocity by α corresponds to scaling the radial component of the displacement, while leaving the tangential component unchanged. We decompose $\Delta x_{m,t} = \Delta x^{\mathrm{rad}}_{m,t} + \Delta x^{\mathrm{tan}}_{m,t}$, where $\Delta x^{\mathrm{rad}}_{m,t} = (\Delta x_{m,t} \cdot \hat{r}_{m,t-1})\, \hat{r}_{m,t-1}$ is the projection of $\Delta x_{m,t}$ onto the line-of-sight direction $\hat{r}_{m,t-1} = (x_{m,t-1} - x_{\mathrm{radar}}) / \|x_{m,t-1} - x_{\mathrm{radar}}\|$ at $x_{m,t-1}$. The counterfactual $x'^{(\alpha)}$ is then generated by

$$\begin{aligned} x'^{(\alpha)}_{m,1} &= x_{m,1}, \\ x'^{(\alpha)}_{m,t} &= x'^{(\alpha)}_{m,t-1} + \alpha\, \Delta x^{\mathrm{rad}}_{m,t} + \Delta x^{\mathrm{tan}}_{m,t}, \end{aligned} \tag{8}$$

attenuating or amplifying $\Delta x^{\mathrm{rad}}_{m,t}$ while preserving tangential motion, and realizing $B(x'^{(\alpha)}) \approx \alpha\, B(x)$.

Consistency score. For counterfactuals $x'^{(\alpha)}$ constructed by scaling radial velocity by α, physical consistency demands that f's response track g's across the perturbation

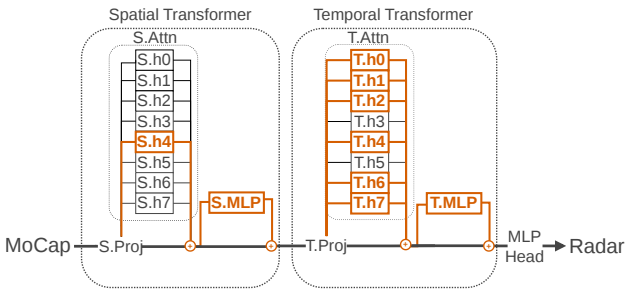


Figure 2: The architecture of the ST(Flat) spatio-temporal transformer. Components highlighted in red form an example of an identified $\tau = 0.9$ physical consistency circuit, as reported in Table 1.

family $\mathcal{P} = \{p_\alpha\}_{\alpha \in A}$, indexed by a set of scaling factors $A \subset \mathbb{R}$. We summarize each model's response at α by the least-squares slope of the perturbed bridged output against the unperturbed baseline over the evaluation set,

$$\hat{m}_f(\alpha) \;=\; \frac{\sum_{i=1}^{N} \bar{z}_f^{(i)}\, z_f^{(i,\alpha)}}{\sum_{i=1}^{N} \big(\bar{z}_f^{(i)}\big)^2}. \tag{9}$$

Replacing f with g in Eq. (9) yields the physics-side slope $\hat{m}_g(\alpha)$. The consistency score quantifies how closely $\hat{m}_f(\alpha)$ matches $\hat{m}_g(\alpha)$ across $\mathcal{P}$.

We aggregate the squared deviations between $\hat{m}_f$ and $\hat{m}_g$ across A:

$$M_{\mathcal{P}}(f,g) \;=\; 1 \;-\; \frac{\sum_{\alpha \in A} \big(\hat{m}_f(\alpha) - \hat{m}_g(\alpha)\big)^2}{\sum_{\alpha \in A} \big(\hat{m}_g(\alpha) - \bar{m}_g\big)^2}, \tag{10}$$

where $\bar{m}_g = \frac{1}{|A|} \sum_{\alpha \in A} \hat{m}_g(\alpha)$. $M_{\mathcal{P}} \leq 1$, with 1 attained when $\hat{m}_f$ matches $\hat{m}_g$ across $\mathcal{P}$.

3.3. Doppler consistency circuit

We identify circuits in f following the τ-physical consistency circuit definition of §2.3, beginning with the computational graph $G = (V, E)$ on the spatio-temporal transformer. We analyze the model at component granularity: vertices V comprise the attention heads and MLPs of the spatial transformer and those of the temporal transformer; edges E trace the residual-stream flow within each transformer together with the linear projection bridging the spatial and temporal streams. As a first instantiation aimed at demonstrating the framework, this granularity supports circuit identification at the level of specific heads and MLPs and is well-matched to the modest component count of f in this study.

4. Physics-Based Interpretation of the MoCap-to-Radar Transformer

We now identify circuits in f and analyze their physical consistency with the Doppler model g. We follow the experimental setup of Chen et al. (2026), using their ST(Flat) MoCap-to-Radar architecture which exposes $|V| = 18$ patchable components (Figure 2). We use the RandomWalk dataset as our primary evaluation set.

Domain of validity for perturbation family. Ideally, consistency scoring and circuit identification should use continuous perturbations within the radial velocity bounds that the radar can resolve. The Bumblebee radar used in this study resolves radial velocities only up to 2.6 m/s (The Samraksh Company, 2024). Using the same RCS weights w_m as in Eq. (7), the maximum training-set body-level radial-velocity magnitude is 2.24 m/s. We accordingly estimate the *radar-resolvable* scaling regime to be $\alpha \in A = [-\alpha_{\max}, \alpha_{\max}]$ with $\alpha_{\max} = 2.6/2.24 \approx 1.16$, within which every frame's body-level radial velocity stays below the radar velocity limit; α outside this range exceeds the radar's resolvable velocity.

4.1. Physically consistent circuit identification

Unpatched consistency baseline. The consistency score of the unpatched model is $M_0 = 0.988$, reflecting near-perfect agreement with g before any patching. This baseline establishes that physical consistency is present in f, motivates the search for specific components that carry it, and defines the reference against which patched scores are measured: subgraph contributions are quantified by the degradation they induce relative to M_0.

Incremental search procedure. We construct a family $\{C(\tau)\}_\tau$ of τ-consistent circuits by greedy bottom-up search, processing τ in increasing order. For each τ, we initialize C from the previous circuit (or $\emptyset$ for the smallest τ), iteratively add the component c that most reduces $\Delta_{\text{suf}}(C \cup \{c\})$ until C is τ-consistent, then prune the components added at the current τ to satisfy minimality. By construction, $C(\tau_i) \subseteq C(\tau_j)$ whenever $\tau_i < \tau_j$, and each $C(\tau)$ is minimal among τ-consistent supersets of its predecessor. We sweep $\tau \in \{0.3, 0.4, \ldots, 0.9\}$.

Controls. We compare $C(\tau)$ against a *source-randomized control*, applied in both patching configurations. In the sufficiency configuration, the control patches the components in $C(\tau)$ but draws source activations from a random evaluation window $x'^{(j,\alpha)}$, with $j \neq i$, independently for each (i, α) pair, yielding $\Delta_{\text{suf}}^{\text{rand}}$. In the necessity configuration, the control patches the components in $V \setminus C(\tau)$ and draws source activations from random evaluation windows in the same manner, yielding $\Delta_{\text{nec}}^{\text{rand}}$. These controls test whether the effect of patching depends on the matched counterfactual source, rather than only on patching the same locations. Reported $\pm$ values denote standard errors over ten random seeds, with each seed independently sampling $j \neq i$ for every (i, α) pair.

Table 1: Reported τ-physical consistency circuits. Component notation: T / S denote temporal / spatial transformer branches; hk the k-th attention head; MLP the feed-forward block. Source-randomized controls $\Delta_{\text{suf}}^{\text{rand}}$ and $\Delta_{\text{nec}}^{\text{rand}}$ patch the same components as C but draw source activations from a random evaluation window.

$\lvert C\rvert$	τ	circuit C	Δ_{suf}	Δ_{nec}	$\Delta_{\text{suf}}^{\text{rand}}$	$\Delta_{\text{nec}}^{\text{rand}}$
3	0.3–0.5	T:{MLP, h6, h7}	0.49	1.89	1.82 ± 0.04	1.11 ± 0.02
4	0.6	T:{MLP, h1, h6, h7}	0.35	2.43	2.11 ± 0.04	1.29 ± 0.02
5	0.7	T:{MLP, h1, h6, h7}, S:{MLP}	0.24	3.08	2.39 ± 0.04	2.17 ± 0.02
6	0.8	T:{MLP, h1, h4, h6, h7},S:{MLP}	0.18	3.19	2.52 ± 0.04	2.03 ± 0.02
9	0.9	T:{MLP, h0, h1, h2, h4, h6, h7}, S:{MLP, h4}	0.09	3.50	3.05 ± 0.05	3.23 ± 0.02

Search results. Doppler consistency is concentrated in a temporal-block core. As shown in Table 1, the smallest τ-consistent circuit ($|C| = 3$) consists of three temporal-block components, T:{MLP, h6, h7}, and circuits at higher τ expand almost entirely within the temporal block; the spatial block contributes only marginally, with S:{MLP} entering at $\tau = 0.7$ and an additional spatial head only at $\tau = 0.9$. The source-randomized controls satisfy $\Delta_{\text{suf}}^{\text{rand}} > \Delta_{\text{suf}}$ and $\Delta_{\text{nec}}^{\text{rand}} < \Delta_{\text{nec}}$ at every $|C|$, so this concentration is specific to the matched counterfactual source, not an artifact of the patched locations. For the analysis and discussion that follows, we anchor on the *primary circuit*

$$C^\dagger = T : \{\text{MLP}, h_0, h_1, h_2, h_4, h_6, h_7\},\ S : \{\text{MLP}, h_4\},$$

which, at $\tau = 0.9$, recovers 91% of M_0 using only half ($|C^\dagger|/|V| = 9/18$) of the components, most of which lie in the temporal block (7 out of 9).

Visualizing physical consistency of $C^\dagger$. The primary circuit $C^\dagger$ reproduces the physics-consistent Doppler centroid; its complement does not. As shown in Figure 3, the Counterfactual Model panel shows f's Doppler centroid compressed in amplitude relative to Unperturbed. This *compression is consistent with the $\alpha = 0.3$ scaling* applied to the radial velocities of x'; the Counterfactual Circuit panel reproduces this compressed trajectory, while the Counterfactual Complement panel *retains the original amplitude* seen in the Unperturbed Model panel.

4.2. Counterfactual behavior over domain of validity

We evaluate both the learned model f and the circuit $C^\dagger$ by comparing their responses to counterfactual physical states against the responses predicted by the physics model g. The evaluation focuses on two physically meaningful aspects of their behavior: (i) how they behave across the domain of validity, with $\alpha \in [-1.16, 1.16]$; this includes characterizing the response to sign-reversed perturbations that invert the radial motion direction, and then beyond to $\alpha \in [-2, 2]$; and (ii) how they behave across perturbations that affect the Doppler frequency (radial) versus perturbations that should not (tangential).

Radial velocity scaling over radar-resolvable domain. Both f and $C^\dagger$ exhibit the behavior prescribed by the Doppler equation: an approximately linear response to radial scaling within the radar-resolvable range. As shown in Fig. 4a, the *learned model f tracks g closely throughout the radar-resolvable domain*, accounting for the high M_0 reported earlier. The circuit $C^\dagger$ also tracks g across this range and is particularly tight on $\alpha \in [0, 1]$, while the complement $V \setminus C^\dagger$ remains nearly flat and shows little response to scaling. Beyond the radar-resolvable range, both f and $C^\dagger$ no longer closely follow g. This discrepancy should not be interpreted as evidence of a single failure mode: it may arise from physical extrapolation beyond the radar's resolvable velocity range, distributional extrapolation beyond the support of the training data, or an interaction between the two. On the negative-α side, $C^\dagger$ loses tracking earlier, which is suggestive: negative scaling combines magnitude scaling with a sign flip, and components outside the identified circuit may be responsible for the latter. Identifying such a sign-flip substructure would require either a perturbation family that isolates the sign operation or a consistency metric sensitive to it, which we leave to future work.

Tangential velocity scaling as a Doppler-invariance test. Tangential motion is orthogonal to the radar line-of-sight and contributes no first-order Doppler shift. A Doppler-consistent model is therefore expected to be invariant under tangential scaling: scaling the tangential component of motion should not alter the predicted radar response. Because Doppler responds to radial but not tangential motion, the radial counterfactual in Eq. (8) can preserve the tangential component without affecting $\hat{m}(\alpha)$, but the mirror design for the tangential test would leave a Doppler-active radial background and contaminate the slope. We therefore zero the radial displacement, so that $\hat{m}(\alpha)$ reflects the model's response to tangential motion alone:

$$x_{m,t}'^{(\alpha)} = x_{m,t-1}'^{(\alpha)} + \alpha\, \Delta x_{m,t}^{\text{tan}}.$$

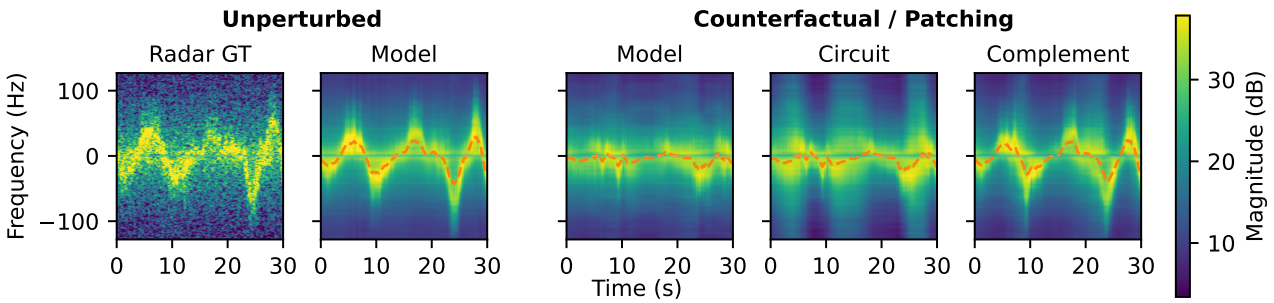


Figure 3: GT, predicted, and patched Doppler spectrograms for a representative 30s evaluation clip with perturbation $\alpha = 0.3$. Orange dashed curves denote Doppler centroids computed by the output bridge D_f in Eq. (6). The left *Unperturbed* group contains the ground-truth radar spectrogram (*Radar GT*) and the model prediction $f(x)$ (*Model*). The right *Counterfactual* group contains the model prediction $f(x')$ (*Model*) and two patching-based outputs: the sufficiency configuration $f^{C^\dagger \leftarrow x'}(x)$ (*Circuit*) and the necessity configuration $f^{V \setminus C^\dagger \leftarrow x'}(x)$ (*Complement*).

As shown in Fig. 4b, under tangential scaling, f, $C^\dagger$, and $V \setminus C^\dagger$ *all exhibit small responses, consistent with the expected near-invariance under tangential scaling*. The point $\alpha = 0$ corresponds to a stationary body and provides the strongest physics constraint on this panel: the radar response should vanish exactly. Both f and g meet this constraint, while $C^\dagger$ shows a slight deviation and the complement $V \setminus C^\dagger$ deviates substantially.

4.3. Component-level mechanism

The circuit identification analysis establishes that $C^\dagger$ recovers physical consistency at the output, but does not characterize what each component contributes internally. Because the input in this case study does not directly encode radial velocity as an explicit feature, we ask: (i) where in $C^\dagger$ does a representation of radial velocity emerge, and how is it propagated to downstream components; and (ii) within these components, what mechanism produces the Doppler consistency response to scaling perturbations?

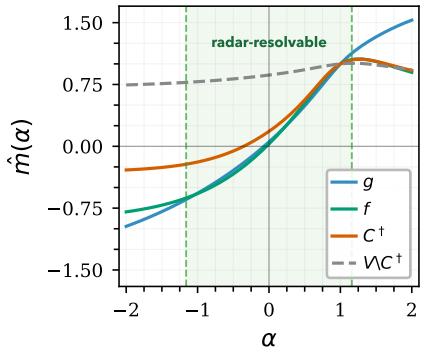
Emergence and propagation of radial-velocity representations. We use linear probing methods (Alain & Bengio, 2016) to characterize the components of $C^\dagger$ and the projection layer (T.Proj). Table 2 reports linear-probe R^2 scores for decoding per-marker radial velocity, from the activations of components in $C^\dagger$. The pre-attention temporal activations, T.Proj, contain little linearly decodable radial-velocity information ($R^2 = 0.019$). Likewise, no head individually decodes v_{rad} above $R^2 = 0.213$. In contrast, their joint representation reaches $R^2 = 0.761$, indicating that v_{rad} becomes linearly available through the combined temporal-attention output. T.MLP also decodes v_{rad} ($R^2 = 0.556$), but replacing the attention output with its evaluation-set mean collapses this to $R^2 = 0.0004$. Together with the per-head and $C^\dagger$-level results, these observations establish radial-velocity emergence in the combined output of $C^\dagger$ temporal heads and its propagation through T.MLP.

Table 2: Linear probe R^2 for per-marker radial velocity v^{rad}; QK and value patching in sufficiency-configuration evaluated on the attention heads of $C^\dagger$.

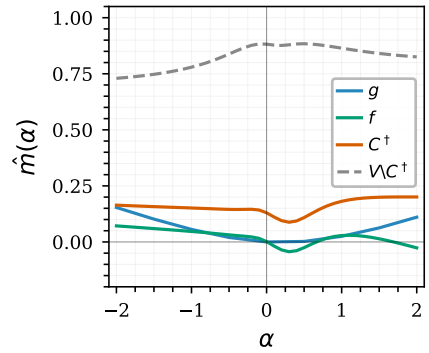
	R^2 (per-marker v^{rad})	$\Delta^{\text{QK}}_{\text{suf}}$	$\Delta^{\text{VAL}}_{\text{suf}}$
Non-attention			
T.Proj	0.019	—	—
T.MLP	0.556	—	—
T.MLP (Attn ablated)	0.0004	—	—
Attention heads — in $C^\dagger$			
T.h0	0.078	3.06	3.48
T.h1	0.042	2.68	3.56
T.h2	0.213	3.14	3.54
T.h4	0.057	3.45	3.28
T.h6	0.152	1.88	3.47
T.h7	0.066	2.34	3.33
S.h4	0.032	3.55	2.92
All attention heads combination — in $C^\dagger$			
$C^\dagger_{\text{attn}}$	0.761	0.35	1.31

Mechanism within temporal attention. We have established that temporal heads dominate Doppler frequency consistency and that radial velocity emerges from their combined output. We now ask what mechanism gives rise to this emergence: is the scaling response carried by attention pattern formation (QK) or value transformation (VAL)? Within our framework, we extend sufficiency patching to these two pathways: for each attention head in $C^\dagger$, we replace either its attention pattern (QK) or its value activations (VAL) with those captured during a counterfactual forward pass on $x'^{(\alpha)}$, holding the complementary pathway at base.

For a sub-circuit $c \subseteq C^\dagger_{\text{attn}}$, where $C^\dagger_{\text{attn}}$ denotes all the



(a)



(b)

Figure 4: Doppler frequency consistency under counterfactual velocity perturbations, comparing f (learned model), g (physics model), $C^\dagger$ (Circuit), and $V \setminus C^\dagger$ (Complement). Each panel plots the slope $\hat{m}(\alpha)$ versus scaling factor α. **(a)** Radial velocity perturbations, shaded band indicating the domain of validity $|\alpha| \leq \alpha_{\max} \approx 1.16$. **(b)** Tangential velocity perturbations, illustrating invariance.

attention heads in $C^\dagger$, we define

$$\Delta_{\text{suf}}^{\text{QK}}(c) := M_0 - M_{\mathcal{P}}^{\text{QK-suf}}(c, g),$$
$$\Delta_{\text{suf}}^{\text{VAL}}(c) := M_0 - M_{\mathcal{P}}^{\text{VAL-suf}}(c, g),$$

in direct analogy to Eq. (3). We use the same perturbation family $\mathcal{P}$ as in circuit identification.

As shown in Table 2, $\Delta_{\text{suf}}^{\text{QK}}(C_{\text{attn}}^\dagger)$ is substantially smaller than $\Delta_{\text{suf}}^{\text{VAL}}(C_{\text{attn}}^\dagger)$ (0.35 vs. 1.31), indicating that replacing the attention pattern alone is sufficient to recover physical consistency, whereas replacing value activations alone is not. At the single-head level, six of the seven temporal heads in $C^\dagger$ are more QK-sufficient than VAL-sufficient (the only exception is T.h4). The spatial head S.h4 is the only head where VAL is more sufficient than QK, but S.h4 is also the least QK-sufficient ($\Delta_{\text{suf}}^{\text{QK}} = 3.55$ is the largest deviation in the table, indicating the weakest QK contribution), so its VAL contribution is marginal. Within the temporal block, T.h1, T.h6, and T.h7 are the three most QK-sufficient heads, and they coincide with the heads that saturate sufficiency earliest in circuit identification (Table 1). Taken together, the QK contribution is concentrated and structurally aligned with the circuit's sufficiency ordering, while the VAL contribution is diffuse and small.

We interpret this as evidence that f adaptively computes temporal differences across frames. Radial velocity is recovered by comparing range measurements between frames, and the most useful time gap depends on how fast the target is moving. Short gaps capture fast motion, whereas longer gaps are more suitable for slow motion. The QK-dominant and temporally localized attribution suggests that the attention pattern controls this choice by selecting the frame pairs used for the comparison. This is similar to multi-scale temporal differentiation in classical signal processing.

5. Limitations

We instantiate LAWFUL on a single case study: one linear law (the Doppler relation) in a spatio-temporal transformer. Nonlinear or multi-variable laws, and larger models, may require different bridges, perturbation families, and search procedures than the ones we use here, and characterizing the bridged outputs at scale remains open. LAWFUL also takes the grounding law as given, and does not address how it should be chosen—among candidate laws or levels of description—or whether it could be discovered rather than supplied.

A subtler limit concerns coverage of the domain of validity. LAWFUL's perturbations generate out-of-distribution physical states, and we observe that the model continues to satisfy the Doppler law across them—evidence of physical generalization beyond the training distribution. These perturbations rescale the speed of recorded motions uniformly across the body. Perturbations that change the motion itself, or that rescale only part of the body, remain untested. Finally, our output bridge reduces each spectrogram to its Doppler centroid, so the analysis characterizes the bulk shift of the distribution rather than its finer structure—spread, harmonics, micro-Doppler—and circuits carrying those features need not coincide with the one we identify.

6. Conclusion

This paper develops physics-aware mechanistic interpretability: in its LAWFUL framework a known physical law, rather than task accuracy or language-model-shaped benchmarks, supplies the ground truth against which a model's internal computation is evaluated. We expect the combination of physics-grounded perturbations and mechanism-level attribution to extend to other physics-based interpretability settings, especially for the emerging world models of physical AI.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Barbiero, P., Espinosa Zarlenga, M., Giannini, F., Termine, A., Bonchi, F., Jamnik, M., and Marra, G. Actionable interpretability must be defined in terms of symmetries. *arXiv preprint arXiv:2601.12913*, 2026.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chen, K., Parker, K. W., and Arora, A. What physics do data-driven MoCap-to-Radar models learn? In *2026 IEEE Radar Conference (RadarConf26)*, Phoenix, AZ, USA, May 2026. IEEE.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Crabbé, J. and van der Schaar, M. Evaluating the robustness of interpretability methods through explanation invariance and equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Friedman, D., Lampinen, A. K., Dixon, L., Chen, D., and Ghandeharioun, A. Interpretability illusions in the generalization of simplified models. In *International Conference on Machine Learning (ICML)*, 2024.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning (CLeaR)*, 2024.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N. D., Potts, C., and Icard, T. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26, 2025.

Hadad, I., Katz, G., and Bassan, S. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations (ICLR)*, 2026.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. doi: 10.48550/arXiv.2404.15255.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Liu, Z., Sanborn, S., Ganguli, S., and Tolias, A. From Kepler to Newton: Inductive biases guide learned world models in transformers. *arXiv preprint arXiv:2602.06923*, 2026.

Makelov, A., Lange, G., and Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Méloux, M., Maniu, S., Portet, F., and Peyrard, M. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *arXiv preprint arXiv:2502.20914*, 2025.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

Sutter, D., Minder, J., Jaeger, S., and Schölkopf, B. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery. In *Workshop on Mechanistic Interpretability at the International Conference on Machine Learning*, 2024.

The Samraksh Company. BumbleBee radar: A low-power, mote-scale pulsed Doppler radar sensor board. `https://samraksh.com/index.php/products/sensors/32-product-pages/products-sensors/71-bumblebee-radar`, 2024.

Vafa, K., Chang, P. G., Rambachan, A., and Mullainathan, S. What has a foundation model found? using inductive bias to probe for world models. In *International Conference on Machine Learning (ICML)*, 2025.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 Small. In *International Conference on Learning Representations (ICLR)*, 2023.

Wu, Z., Geiger, A., Icard, T., Potts, C., and Goodman, N. D. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A. Notation Summary

Table 3 summarizes the notation used throughout the paper.

Table 3: Summary of notation.

Symbol	Description
LAWFUL framework	
$f : X \to Y_f$	learned model, from input space X to output space Y_f
$g : \Omega \to Y_g$	physics model, from continuous physics space Ω to output space Y_g
$B : X \to \Omega$	input bridge; $\Omega := B(X)$ is the physics space reachable by inputs to f
Z	common physical observable space
$D_f : Y_f \to Z,\ D_g : Y_g \to Z$	output bridges projecting the learned and physics models onto Z
$\{x^{(i)}\}_{i=1}^{N}$	evaluation set of N inputs on which consistency is measured
$\mathcal{P} = \{p_\alpha\}$	perturbation family of physically meaningful operations $p_\alpha : \Omega \to \Omega$
α	perturbation index
$x'^{(i,\alpha)}$	counterfactual input satisfying $B(x'^{(i,\alpha)}) = p_\alpha(B(x^{(i)}))$
$\bar{z}_f^{(i)},\ z_f^{(i,\alpha)}$	bridged learned-model outputs: unperturbed baseline and perturbed response
$\bar{z}_g^{(i)},\ z_g^{(i,\alpha)}$	bridged physics-model outputs: unperturbed baseline and perturbed response
Φ	scalar agreement functional on the collected bridged outputs; larger indicates closer agreement
$M_{\mathcal{P}}(f, g)$	physical consistency score, aggregating the agreement measured by Φ
M_0	unpatched baseline score, the value of $M_{\mathcal{P}}(f, g)$ before patching
Physically consistent circuit	
$G = (V, E)$	computational graph of f: components V, residual-stream edges E
$C \subseteq V$	candidate circuit, a subgraph of G
c	a single component of f
$f^{C \leftarrow s}(b)$	forward pass on base b with activations on C patched from source s
$M_{\mathcal{P}}^{\mathrm{suf}}(C, g),\ M_{\mathcal{P}}^{\mathrm{nec}}(C, g)$	consistency score under sufficiency / necessity patching
$\Delta_{\mathrm{suf}}(C),\ \Delta_{\mathrm{nec}}(C)$	deviations of the patched scores from M_0
$\Delta_{\mathrm{suf}}^{\mathrm{rand}},\ \Delta_{\mathrm{nec}}^{\mathrm{rand}}$	source-randomized control counterparts
τ	consistency threshold defining τ-consistent circuits
$C(\tau)$	circuit identified at threshold τ by the incremental greedy search
$C^\dagger$	primary circuit for experimental analysis
$C_{\mathrm{attn}}^\dagger$	the attention heads of $C^\dagger$
$\Delta_{\mathrm{suf}}^{\mathrm{QK}}(c),\ \Delta_{\mathrm{suf}}^{\mathrm{VAL}}(c)$	sufficiency deviation under attention-pattern-only (QK) / value-only (VAL) patching
Doppler instantiation	
$x_{m,t}$	3D position of marker m at sample index t
M	number of MoCap markers; distinct from the score $M_{\mathcal{P}}$
T.h_k, S.h_k; T.MLP, S.MLP	the k-th attention head and the feed-forward block of the temporal (T) / spatial (S) transformer
T.Proj, S.Proj; T.Out, S.Out	input-projection and output layers of the temporal (T) / spatial (S) block
L	STFT window length
λ	radar wavelength in the Doppler law $f(t) = 2v(t)/\lambda$
$v_{m,t}^{\mathrm{rad}}$	radial velocity of marker m, by central finite difference of the range to the radar
w_m	RCS weight of marker m (body-surface-area proportion)
$\hat{m}_f(\alpha),\ \hat{m}_g(\alpha)$	least-squares response slopes of f and g at scaling α
$\alpha_{\max}$	radar-resolvable bound, $\alpha \in [-\alpha_{\max}, \alpha_{\max}]$

B. Training Setup

The MoCap-to-Radar model maps each window of 256 MoCap frames to the Doppler spectrum of the STFT frame spanning the same interval, with windows advancing at the STFT's 32-frame stride. Marker coordinates and target spectra are standardized (zero mean, unit variance) using training-set statistics only. The train/validation split is at the recording level (six/two walking recordings); the random-walk recordings used for all counterfactual analyses are held out entirely. Table 4 lists the optimization hyperparameters. All experimental results use the lowest-validation-loss checkpoint.

Table 4: Training hyperparameters for the MoCap-to-Radar model.

Parameter count	3.2 M
Loss	L1 on standardized log-magnitude spectra
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	0.01
Batch size	32
LR schedule	halve on validation-loss plateau (patience 10)
Early stopping	patience 40 epochs
Maximum epochs	200

C. Circuit Search Procedure

Algorithm 1 formalizes the greedy bottom-up search of Section 4. We specify here the pruning setup, the role of necessity, and the exact minimality it guarantees.

Algorithm 1 Incremental greedy search with multi-ordering pruning

Require: component set V; threshold set $\mathcal{T} = \{0.3, 0.4, \ldots, 0.9\}$; baseline score M_0; pruning trials R
Ensure: nested family $\{C(\tau)\}_{\tau \in \mathcal{T}}$
$C \leftarrow \emptyset$; $C_{\text{prev}} \leftarrow \emptyset$
for $\tau \in \mathcal{T}$ in ascending order **do**
  while $\Delta_{\text{suf}}(C) > (1-\tau)\, M_0$ **and** $C \neq V$ **do**
    $c^\star \leftarrow \arg\min_{c \in V \setminus C} \Delta_{\text{suf}}(C \cup \{c\})$
    $C \leftarrow C \cup \{c^\star\}$ {ties broken by a fixed component order}
  end while
  for $r = 1, \ldots, R$ **do**
    $C_r \leftarrow C$
    draw a random ordering of $C \setminus C_{\text{prev}}$
    repeat
      remove from C_r the first $c \in C_r \setminus C_{\text{prev}}$ in the drawn ordering satisfying $\Delta_{\text{suf}}(C_r \setminus \{c\}) \leq (1-\tau)\, M_0$, if any
    until no such component exists
  end for
  $r^\star \leftarrow \arg\min_r \big(|C_r|, \Delta_{\text{suf}}(C_r)\big)$ {lexicographic order}
  $C \leftarrow C_{r^\star}$
  $C(\tau) \leftarrow C$; $C_{\text{prev}} \leftarrow C$
end for

Pruning. Pruning uses $R = 8$ random deletion orderings under a fixed seed, and returns the smallest resulting circuit (ties broken by lower Δ_{suf}).

Role of necessity. Necessity is not part of the search: components are selected and pruned using Δ_{suf} alone. We evaluate $\Delta_{\text{nec}}(C(\tau))$ post hoc on each returned circuit and report it in Table 1.

Minimality. Pruning yields circuits that are irredundant relative to the circuit found at the previous threshold—no single further component is removable—rather than globally inclusion-minimal.

D. Linear Probing on Circuit Components

We probe each component with a ridge regression (α^{rid}=1) against the per-marker signed radial velocity (the projection of the marker's 3D velocity onto the radar line-of-sight unit vector). Probe feature dimensions are 128 for spatial-side components, 32 for each individual temporal head, and 256 for T.Proj, T.Out, and T.MLP; combinations concatenate the component features. The corresponding Δ_{suf} and Δ_{nec} are additionally reported in Table 5.

#	Component / Probe set	R^2 (per-marker v^{rad})	Δ_{suf}	Δ_{nec}
Spatial side				
1	S.Proj	0.0104	—	—
2	S.Attn	0.1200	3.679	0.106
3	S.MLP	0.0005	3.648	0.123
4	S.Out	0.0185	—	—
Temporal entry				
5	T.Proj	0.0188	—	—
Temporal heads — in $C^\dagger$				
6	h0	0.0784	4.543	0.114
7	h1	0.0422	3.982	0.155
8	h2	0.2134	4.547	0.126
9	h4	0.0567	4.452	0.120
10	h6	0.1524	2.996	0.251
11	h7	0.0655	3.306	0.264
Temporal heads — not in $C^\dagger$				
12	h3	0.0172	4.706	0.085
13	h5	0.0145	4.181	0.089
Circuit combinations				
14	$C^\dagger$ heads only (no MLPs)	0.7614	—	—
15	full $C^\dagger$ (9 components)	0.7699	0.092	3.503
16	complement $V \setminus C^\dagger$	0.3469	3.503	0.092
T.MLP				
17	T.MLP (baseline)	0.5556	1.295	0.202
Temporal exit				
18	T.Out	0.7072	—	—
Causal intervention				
19	T.MLP (Attn ablated)	0.0004	—	—

Table 5: Linear probing R^2 on RandomWalk eval (mean across 5 time-contiguous CV folds, gap=8 windows) and sufficiency/necessity deviations.

E. Temporal Attention Heatmaps

We include the temporal attention heatmaps on the held-out random walk recording in Figure 5. The heatmaps reveal a complementary key-selection structure between T.h6 and T.h7: T.h6 concentrates attention on the earliest frames of the window, most strongly for queries in the first half, while T.h7 concentrates on the latest frames, most strongly for queries in the second half. This complementarity could explain why h6 and h7 are the only attention heads in the smallest τ-consistent circuit, T:{MLP, h6, h7}, which persists across $\tau \in \{0.3, 0.4, 0.5\}$ (Table 1): a temporal contrast requires both an early and a late reference point, and the two heads supply these endpoints in a query-dependent division of labor, so neither head alone carries the difference signal. However, we note that T.h2 exhibits a similar early-key preference, but it enters the circuit only at $\tau = 0.9$ (Table 1), suggesting possible redundancy among heads anchored to early frames.

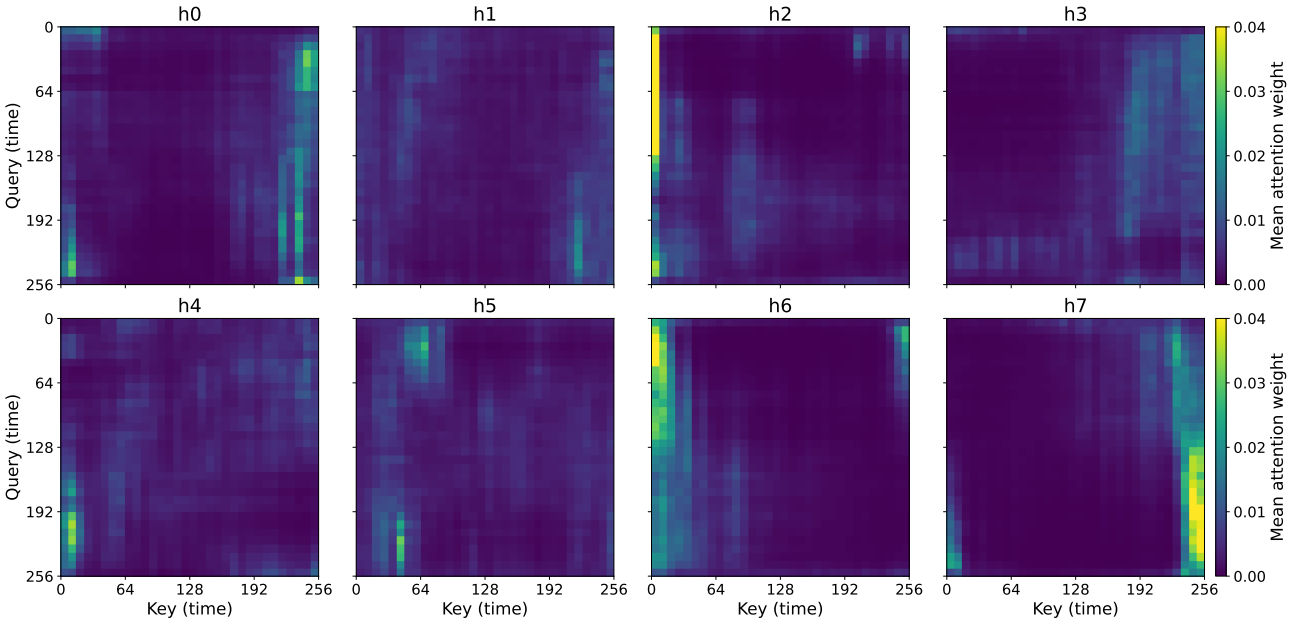


Figure 5: Per-head attention patterns of the temporal transformer. Each panel shows the mean attention map of one head (h_0–h_7), averaged over 200 evaluation windows. Maps are down-sampled from $L \times L = 256 \times 256$ to 32×32 via 8×8 block-mean.