When Is a Second Representation Worth Activating? Information Wave–Particle Duality for Sequential Learning

Anonymous authors
Paper under double-blind review

Abstract

Modern sequence models increasingly combine spectral, temporal, local, and global views, implicitly treating more representation as better. We show why that assumption can fail under finite model capacity and ask a more basic question: *when is a second representation worth activating?* We introduce *Information Wave–Particle Duality* (IWPD), where a distributed spectral Wave view and a localized Packet view are alternative lossy representations of the same observed history. Under Bayes log loss, the context-specific value of adding one view after the other is exactly a directional conditional mutual information. For restricted predictors, we derive a finite-capacity decomposition showing that positive population information can still be offset by approximation burden; adding representation cost yields an explicit activation law. We instantiate the Wave branch with a constrained maximum-entropy spectral prior and deploy the theory through history-only prediction of the task risk of Base, Wave, Packet, and Dual actions. Experiments across five datasets reject universal Dual superiority. Most importantly, under an identical-input, identical-split four-action trajectory protocol, the history-only selector attains lower five-seed mean ADE than the condition-wise best fixed action in all 12 highD/NGSIM horizon conditions. Failures outside this setting are retained explicitly. The resulting principle is not "fuse everything," but *activate a representation only when its conditional value survives model and cost constraints*.

1 Introduction

The dominant design pattern in multi-view sequence learning is additive: construct another representation, fuse it with the existing one, and let a sufficiently expressive model decide whether to use it. That strategy is appealing, but it leaves a fundamental question unanswered. A second view can carry additional population information and still fail to improve a finite predictor because it increases approximation burden, optimization difficulty, or deployment cost. The relevant question is therefore not only *how to fuse multiple representations*, but **when an additional representation is worth entering the computation at all**.

We study this question through **Information Wave–Particle Duality (IWPD)**. Here, "Wave" denotes information distributed across the observation window and naturally expressed through global spectral structure, whereas "Particle" is operationalized as a localized information packet capturing transient, shock-like, or event-centered structure. The terminology describes two information organizations of the same sequence, not physical wave–particle dynamics. Formally, from an observed history $\mathbf{X}$, these two lossy views are written as $\mathbf{W} = \phi_w(\mathbf{X})$ and $\mathbf{P} = \phi_p(\mathbf{X})$. The central idea is to separate three levels that are often conflated: *population value* (does one view contain target information beyond the other?), *finite-model value* (does that information survive a restricted predictor and its cost?), and *deployment* (can the useful action be chosen before the future is observed?). Figure 1 summarizes this chain.

The theoretical answer is compact. At a realized auxiliary pre-target context c, the Bayes log-risk reduction from adding Packet after Wave is exactly $\mathrm{I}(\mathbf{Y}; \mathbf{P} \mid \mathbf{W}, \mathbf{C}_h = c)$, and symmetrically in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

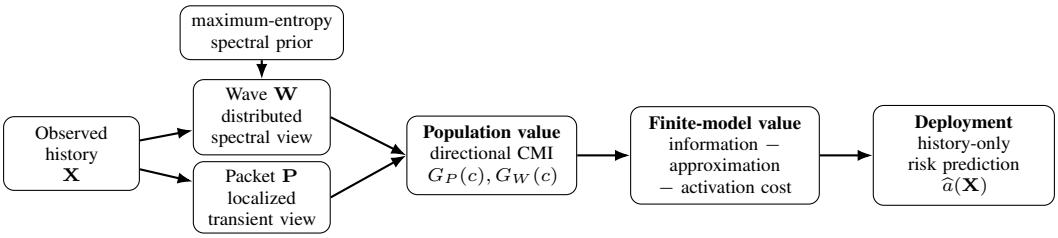

Figure 1: **IWPD in one view.** Directional conditional information defines population complementarity; a finite-capacity decomposition and representation cost determine whether that value is worth using; a history-only selector predicts the task risk of Base/Wave/Packet/Dual before the future is observed. Maximum entropy is an inductive prior for the Wave construction, not the definition of a representation regime.

the other direction. For a restricted predictor family, this population gain is shifted by the change in approximation gap. Once representation cost is included, an additional view is preferred only when its information gain exceeds both the finite-model burden and the incremental cost. For ADE, MSE, or other task-native losses, we use the corresponding conditional action-risk objective rather than relabeling those losses as mutual information.

IWPD is encoder-agnostic at the level of this principle. In one concrete instantiation, the Wave branch receives a spectral prior derived from a constrained maximum-entropy problem, yielding a Planck-shaped allocation with an information-temperature parameter. This is a construction choice, not a universal optimality theorem: uniform and exponential priors are retained as controls. At deployment, observable quantities such as spectral entropy, local difference-energy entropy, concentration, effective event support, and shock magnitude are used only as *history-side routing features*; they are not the theoretical regime axes.

The experiments are designed to falsify the stronger story rather than protect it. Full Dual activation is not uniformly superior across highD, NGSIM, finance, and Weather. Frozen predictive probes show that the two directions of incremental utility can be asymmetric. The strongest controlled result appears only after the problem is posed as conditional selection: with identical inputs, splits, four trained actions, and five seeds, the history-only selector has lower mean ADE than the condition-wise best fixed Base/Wave/Packet/Dual action in all 12 highD/NGSIM horizon conditions. Outside trajectories, that best-fixed claim does not generally hold, which delimits rather than weakens the conclusion.

Contributions.

- **A representation-value principle.** We identify the context-specific value of a missing view with directional conditional mutual information under Bayes log loss, then separate that population quantity from finite-model approximation error.
- **An activation law and deployable formulation.** We derive when an additional representation lowers a risk–cost objective and extend the decision to arbitrary task losses through conditional action risk, enabling history-only Base/Wave/Packet/Dual selection without access to future labels.
- **A falsifiable IWPD instantiation and evidence boundary.** A maximum-entropy Wave prior, branch/prior controls, frozen directional probes, and leakage-separated selection experiments show both where Dual helps and where it does not. The headline result is deliberately restricted to the controlled trajectory action bank rather than presented as universal SOTA.

2 Related Work

Information and representation value. Shannon information and maximum entropy provide the classical foundations for uncertainty and constrained allocation (Shannon, 1948; Jaynes, 1957; Cover & Thomas, 2006). The information bottleneck studies relevance-preserving compression (Tishby et al., 1999; Alemi et al., 2017); modern neural MI estimators are flexible but can be loose in high dimension (Belghazi et al., 2018; Poole et al., 2019; Tschannen et al., 2020). We therefore use mutual

information only in exact population identities, not as a name for held-out probe improvements. Partial-information decompositions distinguish redundancy, uniqueness and synergy (Williams & Beer, 2010); IWPD does not claim to identify PID atoms.

Spectral, local, and multi-expert sequence models. N-BEATS, Informer, Autoformer, FEDformer, TimesNet, RevIN, PatchTST, iTransformer, FreqMoE and M^{2}FMoE expose complementary temporal, spectral, patch, and expert biases (Oreshkin et al., 2020; Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022a; Wu et al., 2023; Kim et al., 2022; Nie et al., 2023; Liu et al., 2024; Liu, 2025; Huang et al., 2026). Their success shows that multiple views can be useful; our question is different: under what condition does one view add value *beyond* another, and should it be activated for a given input regime?

Task-specific forecasting and routing. Motion forecasting includes social, graph, map-aware and spectral models (Alahi et al., 2016; Gupta et al., 2018; Gao et al., 2020; Salzmann et al., 2020; Liang et al., 2020; Zhou et al., 2022b; 2023; Wong et al., 2022; Neumeier et al., 2022; 2023); APE is closely related in learning reliability-aware expert selection under shift (Li et al., 2024a). Finance-specific systems such as StockMixer and MASTER exploit cross-stock or market-side information (Fan & Shen, 2024; Li et al., 2024b). Mixture-of-experts and PAC-Bayesian/exponential-weight views motivate adaptive routing (Jacobs et al., 1991; Shazeer et al., 2017; Cesa-Bianchi & Lugosi, 2006; McAllester, 1999; Catoni, 2007). We do not compare published numbers from incompatible input protocols; richer task-specific reproductions remain a limitation.

3 From Information Value to an Activation Decision

This section turns the paper's central question into mathematics. The logic has three steps: (i) define how much new target information a second view contributes, (ii) account for the fact that a finite model may not be able to exploit all of that information, and (iii) include the extra cost of activating the second view. Each result is followed by its practical meaning.

Let $\mathbf{X} = X_{1:L} \in \mathbb{R}^{L\times d}$ be the observed sequence and $\mathbf{Y}$ the future target at horizon h. Let $\mathbf{W} = \phi_w(\mathbf{X})$ and $\mathbf{P} = \phi_p(\mathbf{X})$ be deterministic, generally lossy views of the same history. Let $\mathbf{C}_h = c_{\text{aux}}(\mathbf{X}, h)$ be an auxiliary pre-target context such as the horizon, metadata, or a coarse history summary. It is deliberately not the full raw history: if $\mathbf{C}_h$ were rich enough to reconstruct $\mathbf{W}$ or $\mathbf{P}$, the corresponding "new information" would be zero by construction. We assume the required regular conditional distributions exist and that the conditional log losses below are finite. A compact symbol table and step-by-step proofs are in Appendices A–B.

3.1 Step 1: How much new information does the second view contain?

For any available representation U, define its conditional Bayes log risk at a realized context c as

$$\mathcal{R}^*(U \mid c) = \inf_{\widetilde{p}} \mathbb{E}[-\log \widetilde{p}(\mathbf{Y} \mid U, \mathbf{C}_h = c) \mid \mathbf{C}_h = c]. \tag{1}$$

Here $\widetilde{p}$ ranges over all valid conditional predictive distributions. In plain language, $\mathcal{R}^*(U \mid c)$ is the smallest average log loss that an ideal predictor could achieve if it were given U in context c.

Theorem 1 (Context-specific conditional information value)**.** *Define*

$$G_P(c) = \mathrm{I}(\mathbf{Y}; \mathbf{P} \mid \mathbf{W}, \mathbf{C}_h = c), \qquad G_W(c) = \mathrm{I}(\mathbf{Y}; \mathbf{W} \mid \mathbf{P}, \mathbf{C}_h = c). \tag{2}$$

Then

$$\mathcal{R}^*(\mathbf{W} \mid c) - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) = G_P(c), \tag{3}$$
$$\mathcal{R}^*(\mathbf{P} \mid c) - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) = G_W(c). \tag{4}$$

Therefore the improvement of the joint view over the better of the two single views is

$$\Delta_{\text{dual}}(c) = \min\{G_P(c), G_W(c)\} \geq 0. \tag{5}$$

Averaging over contexts recovers ordinary conditional mutual information, e.g. $\mathbb{E}_{\mathbf{C}_h}[G_P(\mathbf{C}_h)] = \mathrm{I}(\mathbf{Y}; \mathbf{P} \mid \mathbf{W}, \mathbf{C}_h)$.

What this means. $G_P(c)$ answers a very specific question: after Wave is already known, how much target information is still uniquely supplied by Packet in context c? $G_W(c)$ asks the reverse question. The two numbers need not be equal. Equations (3)–(4) therefore turn "complementarity" from an architectural intuition into a directional, falsifiable quantity.

For a small nonnegative tolerance δ_{reg}, we use the two directions only as a population taxonomy: both above δ_{reg} gives a Dual context; only G_W above it is Wave-dominant; only G_P above it is Packet-dominant; and both at or below it are mutually redundant *conditional on the other view*. "Mutually redundant" does not mean that either view is useless relative to a raw-history Base predictor.

Corollary 1 (Strict complementarity criterion)**.** $\Delta_{\text{dual}}(c) > 0$ *if and only if both* $G_P(c) > 0$ *and* $G_W(c) > 0$.

Why this is useful. The joint representation can beat the better single representation in ideal log risk only when each view contributes something that the other does not already contain. One positive direction alone is not enough for strict two-way complementarity.

3.2 Step 2: Why can a finite network fail even when new information exists?

Theorem 1 describes an ideal predictor. Real neural networks belong to restricted function classes. For each input representation U, let $\mathcal{F}_U$ denote the predictor class actually available when that representation is used, and define

$$\epsilon_{\mathcal{F}}(U \mid c) = \mathcal{R}^*_{\mathcal{F}}(U \mid c) - \mathcal{R}^*(U \mid c) \geq 0, \tag{6}$$

where $\mathcal{R}^*_{\mathcal{F}}(U \mid c)$ is the best log risk achievable inside $\mathcal{F}_U$. Thus $\epsilon_{\mathcal{F}}$ is an *approximation gap*: zero means the available model class is rich enough to match the Bayes optimum for that input; a larger value means more of the theoretically available information is left unused.

Proposition 1 (Finite-capacity decomposition)**.** *For adding Packet to Wave,*

$$\mathcal{R}^*_{\mathcal{F}}(\mathbf{W} \mid c) - \mathcal{R}^*_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) = G_P(c) + \epsilon_{\mathcal{F}}(\mathbf{W} \mid c) - \epsilon_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c). \tag{7}$$

The symmetric identity holds when adding Wave to Packet.

What this means. The first term, $G_P(c)$, is the information that Packet makes available. The remaining two terms measure whether the finite model can exploit that information. If the joint input substantially increases the approximation gap, the observed model can become worse even though $G_P(c) > 0$. This is the precise reason "more information" does not imply "lower error for a particular finite network."

3.3 Step 3: When is the second view worth activating?

Let κ_W and κ_{WP} denote nonnegative deployment costs for Wave-only and joint Wave–Packet, respectively. The scalar $\lambda \geq 0$ converts one unit of representation cost into the same decision scale as log risk. A larger λ means that the deployment places more emphasis on avoiding extra representation cost.

Proposition 2 (Information-to-cost activation law)**.** *Under Bayes log loss, adding Packet to Wave gives a lower risk–cost objective exactly when*

$$G_P(c) > \lambda\big(\kappa_{WP} - \kappa_W\big). \tag{8}$$

The symmetric condition for adding Wave to Packet replaces G_P *by* G_W *and compares* κ_{WP} *with* κ_P.

What this means. At the ideal-predictor level, the extra view is worth activating only if the new information it provides is larger than the price assigned to the extra representation.

Corollary 2 (Finite-model activation criterion)**.** *For the restricted model classes in Proposition 1, adding Packet to Wave is beneficial exactly when*

$$\boxed{G_P(c) > \underbrace{\lambda(\kappa_{WP} - \kappa_W)}_{\textit{extra activation cost}} + \underbrace{\epsilon_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) - \epsilon_{\mathcal{F}}(\mathbf{W} \mid c)}_{\textit{extra finite-model burden}}}. \tag{9}$$

Central interpretation. Equation (9) is the paper's main decision rule: **new information must exceed extra model burden plus extra activation cost**. If the right-hand side is negative, the joint representation is easier for the chosen model class and even a small information gain can justify activation. If it is positive and large, substantial new information is required. The statement is exact for log loss; it is not asserted for ADE or MSE by changing the name of the loss.

For an arbitrary deployment loss ℓ (for example ADE or MSE), we therefore define the decision directly in task-risk space. Let $\mathcal{A} = \{0, W, P, WP\}$ denote Base, Wave, Packet, and Dual, with exposed inputs $Z_0 = \mathbf{X}$, $Z_W = \mathbf{W}$, $Z_P = \mathbf{P}$, and $Z_{WP} = (\mathbf{W}, \mathbf{P})$. Then

$$\mathcal{R}^*_{\ell,\mathcal{F}}(a \mid c) = \inf_{f \in \mathcal{F}_a} \mathbb{E}[\ell(\mathbf{Y}, f(Z_a)) \mid \mathbf{C}_h = c], \qquad a^*_\ell(c) = \arg\min_{a \in \mathcal{A}} \{\mathcal{R}^*_{\ell,\mathcal{F}}(a \mid c) + \lambda \kappa_a\}. \quad (10)$$

Here κ_a is the structural cost assigned to action a. Equation (10) simply says: among the four available actions, choose the one with the smallest expected task loss after adding whatever deployment cost we care about. Our ADE/MSE experiments estimate these relative action risks; they do not estimate conditional mutual information from ADE/MSE.

4 One IWPD Instantiation

The principle above does not require a particular encoder. The experimental IWPD instantiation uses a distributed spectral Wave view and a localized Packet view.

Time–frequency localization motivation. The Wave/Packet distinction is motivated by the classical Fourier uncertainty principle rather than by quantum mechanics. For a normalized square-integrable signal with finite second moments, temporal and frequency spreads obey a convention-dependent inequality of the form

$$\sigma_t \, \sigma_\omega \geq c_F > 0, \quad (11)$$

where σ_t and σ_ω are the standard deviations of normalized signal energy in time and angular frequency, respectively, and c_F is the positive constant induced by the chosen Fourier convention (Folland & Sitaram, 1997). In plain language: a representation cannot be made arbitrarily concentrated in both time and frequency at once. We use this only as an intuitive reason to construct two lossy views—a more distributed Wave view and a more localized Packet view. It neither implies that the observed data obey quantum dynamics nor guarantees that both views contain unique predictive information.

4.1 A maximum-entropy prior inside the Wave branch

Rather than hand-assign spectral weight, let $N_j \in \mathbb{N}_0$ denote abstract occupancy of positive-frequency mode j with cost $c_j > 0$. Maximizing Shannon entropy of an occupancy distribution q subject to normalization and an expected budget $\mathbb{E}_q[\sum_j c_j N_j] = B$ yields

$$q^*(\mathbf{n}) = \mathcal{Z}(\beta)^{-1} \exp\left(-\beta \sum_j c_j n_j\right), \qquad \mathbb{E}[N_j] = (e^{\beta c_j} - 1)^{-1}, \quad (12)$$

where $\mathcal{Z}(\beta)$ is the partition function and $\beta > 0$ is chosen so that the expected budget equals B. If $c(\omega) = a\omega$ and the mode measure is proportional to ω^ν, the induced normalized spectral prior has the Planck-shaped form

$$\varpi(\omega) \propto \frac{\omega^\nu}{e^{\omega/T_I} - 1}, \qquad T_I = (a\beta)^{-1}. \quad (13)$$

The quantity T_I is an information-temperature parameter: increasing it shifts relative weight toward higher frequencies. Appendix C gives the full constrained-entropy derivation and proves this ordering. Equation (13) is an *inductive prior* for constructing the Wave branch, not a definition of IWPD and not an empirical dominance claim; uniform and exponential alternatives are tested explicitly.

4.2 History-only action-risk prediction: from theory to a usable selector

Equation (10) describes the ideal action, but its true conditional risk depends on the unknown future target. At test time we know the history $\mathbf{X}$ but not $\mathbf{Y}$, so we cannot compute the future error of Base,

Wave, Packet, or Dual directly. The deployment problem is therefore: *use only information available before the future occurs to predict which action is likely to have the smallest task loss.*

We implement this in four leakage-separated steps.

1. **Describe the observed history.** Compute a feature vector $s(\mathbf{X}) \in \mathbb{R}^m$ from the history only. It contains spectral concentration, normalized spectral entropy, detrended concentration, local difference-energy entropy and peak concentration, effective event support, and relative shock magnitude. For normalized positive-frequency power p_j and normalized local difference energy e_t,
$$H_\omega = -\frac{\sum_j p_j \log p_j}{\log J}, \qquad H_t = -\frac{\sum_t e_t \log e_t}{\log(L-1)}. \tag{14}$$
These entropies summarize whether the observed history is spectrally concentrated or temporally localized. They are **routing features only**; they neither define the population regimes nor estimate $G_P(c)$ or $G_W(c)$.
2. **Create supervised risk targets on routefit only.** For each routefit example i and action $a \in \mathcal{A}$, replay the frozen action checkpoint to obtain a prediction $\widehat{\mathbf{Y}}_{i,a}$ and compute its realized task loss
$$L_{i,a} = \ell(\mathbf{Y}_i, \widehat{\mathbf{Y}}_{i,a}). \tag{15}$$
Future labels are used here because routefit is a training split for the selector, not the final test set.
3. **Learn one risk predictor per action.** Fit a ridge regressor g_a that maps history features to the realized loss of action a,
$$\widehat{g}_a = \arg\min_{g_a} \sum_{i \in \text{routefit}} \big(L_{i,a} - g_a(s(\mathbf{X}_i))\big)^2 + \alpha \|w_a\|_2^2, \tag{16}$$
where $g_a(s) = w_a^\top s + b_a$. The ridge strength α and any selector hyperparameters are chosen on calibration, never on test.
4. **Choose using history only at test time.** For a new history x, compute four predicted risks $\widehat{R}_a(x) = \widehat{g}_a(s(x))$ and select
$$\widehat{a}_{\text{acc}}(x) = \arg\min_{a \in \mathcal{A}} \widehat{R}_a(x). \tag{17}$$
No future target is used in this choice. A cost-aware variant instead minimizes $\widehat{R}_a(x) + \lambda \kappa_a$; calibration chooses λ so that predicted structural cost is reduced while the calibration loss stays within a pre-specified 0.5% regret budget relative to AlwaysDual. If no λ is feasible, a pre-specified fallback is used.

What this section is – and is not – claiming. The selector does *not* observe the theoretical CMI regime and does not try to reconstruct it from two entropy numbers. It learns a simpler operational object: which already-trained action is likely to incur the lowest deployment loss for this history. This is the empirical counterpart of Eq. (10). The structural costs $\{0, 1, 1, 2\}$ for Base/Wave/Packet/Dual count active specialized representations only; they are an activation proxy, not measured latency, FLOPs, training cost, or checkpoint memory.

5 Experiments

5.1 Protocol and comparators

We evaluate five datasets spanning three application backgrounds: highD and NGSIM trajectories (Krajewski et al., 2018; U.S. Department of Transportation Federal Highway Administration, 2016), an A-share panel and JPX finance, and Weather. One maximum-horizon model is trained and evaluated at fixed prefixes: $h \in \{1, 5, 10, 15, 20, 30\}$ for trajectories, $h \in \{1, 3, 5, 10, 20\}$ for finance, and $h \in \{1, 96, 192\}$ for Weather. Critical IWPD/GRU-wide/FreqMoE and branch-ablation comparisons use five completed seeds $\{42, 2024, 2025, 3407, 10086\}$ under the same dataset version, split, training budget and learning rate; several source-locked general baselines retain their three completed seeds and are reported as such. The data roles are train/selection/routefit/calibration/test, with normalization and imputation fitted on training only.

Table 1: Trajectory ADE under the unified protocol (lower is better). IWPD, GRU-wide and FreqMoE use five seeds; DLinear, PatchTST and iTransformer retain three completed seeds. Bold marks the best listed standard/full method.

	highD			NGSIM		
Method	h=10	20	30	h=10	20	30
DLinear	.2462	.6629	1.2403	.7572	1.8734	3.3518
PatchTST	.4320	.7688	1.2630	.7332	1.7241	3.0786
iTransformer	.3242	.6448	1.1253	.7674	1.7444	3.0749
FreqMoE	.3948	.7824	1.3344	.8170	1.8512	3.2400
GRU-wide	.2200	.5418	1.0073	**.6366**	**1.5933**	**2.8904**
IWPD	**.2162**	**.5365**	**.9949**	.6511	1.6101	2.9036

Table 2: Frozen predictive-probe directional gains (nats per output dimension; higher is better). $P \mid W$ means the gain from adding Packet after Wave; $W \mid P$ is the reverse direction. These are finite-probe log-score differences, not measured conditional mutual information.

Dataset	Direction	h=1	5	10	15	20	30
highD	$P \mid W$	0.158	0.155	0.095	0.057	0.037	0.030
highD	$W \mid P$	0.054	0.037	0.005	-0.009	-0.015	-0.013
NGSIM	$P \mid W$	0.231	0.084	0.051	0.038	0.031	0.024
NGSIM	$W \mid P$	0.026	0.003	-0.001	-0.006	-0.008	-0.007

Baselines are GRU, a wider capacity-matched GRU, DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), and FreqMoE (Liu, 2025). The latter three are source-locked upstream implementations trained from scratch. IWPD controls remove Wave or Packet and replace the maximum-entropy prior by uniform or exponential priors. No incompatible published number is inserted into the tables. On highD/NGSIM, full IWPD has 114,631 parameters versus 124,860 for GRU-wide, so the positive highD pattern is not explained by greater parameter count.

For mechanism analysis, frozen best checkpoints are evaluated on fixed held-out subsets. Equal-rank linear CKA compares geometry (Kornblith et al., 2019); a ridge predictive probe is fit on routefit, calibrated on calibration, and evaluated on test. We report *predictive log-score gains*, not MI estimates or MI lower bounds.

5.2 Q1: Is mandatory Dual activation better?

Table 1 answers no for representative medium/long horizons, and the full horizon-by-horizon plot is moved to Appendix E. On highD, full IWPD improves five-seed mean ADE over GRU-wide by 2.27%, 1.77%, 0.97%, 0.98%, and 1.23% at $h = 5, 10, 15, 20, 30$, respectively, but is worse at $h = 1$. On NGSIM, GRU-wide is better at every horizon. At $h = 10, 20, 30$, IWPD wins only 3/5 matched highD seeds, so the full-model pattern is not a seed-uniform or significance claim.

The same boundary appears outside trajectories: Weather, A-share and JPX often favor a standard baseline or a single-view ablation in point error (Appendix E.1). This negative evidence is important: it demonstrates why a conditional activation problem is needed in the first place.

5.3 Q2: Do the two views add value in the same direction?

Table 2 shows that they do not. Packet-given-Wave predictive log-score gain is positive at all six horizons on both highD and NGSIM (12/12 trajectory condition means), while Wave-given-Packet decreases and becomes negative at medium/long horizons. Thresholded regime labels are margin-sensitive near zero, so we treat the continuous directional gains as the primary evidence. This is the empirical signature predicted by the theory: complementarity can be asymmetric and context dependent rather than a permanent property of the architecture.

5.4 Q3: Can history identify the useful action before the future?

This is the strongest controlled result. The accuracy selector observes history descriptors only and uses no structural-cost penalty. Relative to mandatory AlwaysDual, it improves five-seed mean

Table 3: Accuracy-selector gain (%) over the condition-wise best fixed action in the same four-action bank. Parentheses give paired seed wins out of five.

Dataset	h=1	5	10	15	20	30
highD	+7.38 (4/5)	+11.47 (4/5)	+6.23 (3/5)	+4.10 (3/5)	+2.67 (3/5)	+1.36 (3/5)
NGSIM	+13.63 (5/5)	+1.96 (5/5)	+0.51 (5/5)	+0.22 (5/5)	+0.19 (5/5)	+0.29 (4/5)

Table 4: Compact selector audit. Counts are positive dataset–horizon means under the stated comparison.

Comparison	Result	Comment
Accuracy selector vs AlwaysDual	$25/25$ positive	cross-domain count
Accuracy selector vs best fixed action	$12/12$ trajectory, $13/25$ overall	strongest controlled claim
Accuracy selector vs HistoryPhaseRule	$12/12$ trajectory means positive	non-learned switching control
Accuracy selector vs RandomFrequencyControl	$12/12$ trajectory means positive	non-learned switching control
Budgeted selector vs AlwaysDual	$23/25$ positive	negative cells: JPX $h = 10, 20$
Budgeted selector mean activation proxy	0.443 vs 2.0	proxy only, not runtime/FLOPs
Budgeted selector fallback	$17/125$ selections	0.5% calibration screen infeasible

loss in **25/25 dataset–horizon cells**. The harder comparator is the *condition-wise best fixed action* among Base, Wave, Packet and Dual. Under the unified trajectory protocol, the selector has lower five-seed mean ADE in **12/12 highD/NGSIM cells**; it also beats the fixed-quantile phase rule and randomized-frequency switching control in all 12 condition means. These horizons are correlated fixed-prefix evaluations, not 12 independent hypothesis tests. The detailed scatter/frontier visuals are moved to Appendix E.

Outside trajectories the selector beats the best fixed action in only **1/13** cells; therefore the 25/25 result is specifically a result against mandatory Dual, not universal cross-domain optimality. The trajectory claim is exactly scoped: same input, same split, same four trained actions, same five seeds, and history-only selection.

The budgeted selector answers a different question. It improves over AlwaysDual in **23/25** cells and has mean activation proxy 0.443 versus 2 for AlwaysDual, but this is not a latency/FLOP or memory reduction because the policy reuses separately trained checkpoints. The 0.5% calibration constraint is infeasible in 17/125 seed–dataset–horizon selections, invoking the pre-specified fallback; those cases remain in the averages.

5.5 What does the maximum-entropy prior contribute?

The prior is principled but not uniformly optimal. Uniform allocation is slightly better on several highD medium/long horizons, whereas the maximum-entropy prior is marginally better on several NGSIM horizons; finance differences are generally small. The routing audit shows the same pattern: objective-dependent gates remain distinct under common priors, yet simpler contextual or global routing can outperform the Gibbs implementation in some domains. We therefore interpret maximum entropy and Gibbs routing as auditable inductive references, not dominance theorems. This placement is intentional: the core contribution is the activation principle, while the entropy construction is one theoretically motivated way to instantiate the Wave side.

6 Discussion and Limitations

The empirical results sharpen the theory rather than simply confirming it. More representation is not monotonically better: the benefit of adding a view depends on direction, horizon, predictor class, and deployment objective. The conditional-information identity is a population statement under log loss; the history selector is a task-risk predictor under ADE/MSE-type losses. Keeping these levels

separate prevents a common but invalid shortcut in which an empirical error reduction is labeled as mutual information.

The study also has clear limits. The unified protocol prioritizes controlled comparability over domain-specific leaderboard reproduction and does not reproduce all richer map/graph inputs of GFTNNv2, learned-vs-rule experts of APE, or market-side inputs of StockMixer/MASTER. Several general baselines retain three completed seeds while targeted IWPD/capacity/branch comparisons use five. The selector is a post-training policy over four separately trained checkpoints, so its activation proxy is not measured runtime or memory. Finally, the strongest best-fixed result is trajectory-specific; the cross-domain result supports rejection of mandatory Dual, not universal selector dominance or zero-shot generalization.

7 Conclusion

We asked one question: **when is a second representation worth activating?** IWPD answers it in three steps. First, directional conditional mutual information identifies the context-specific population value of adding Wave after Packet or Packet after Wave. Second, a finite-capacity decomposition shows why that information may fail to improve a restricted network; adding representation cost yields the activation condition in Eq. (9), where new information must exceed approximation burden plus incremental cost. Third, because the target-dependent population quantities are unavailable at inference time, deployment becomes a history-only conditional-risk problem over Base, Wave, Packet, and Dual.

The maximum-entropy construction now has a precise role in this story: it is a principled, budget-constrained spectral prior for the Wave branch, not the definition of duality and not an empirical guarantee. The experiments likewise define the boundary of the claim. Full Dual and the maximum-entropy prior are not uniformly best across five datasets. Yet, within the controlled trajectory action bank, history-only selection lowers five-seed mean ADE relative to the condition-wise best fixed action in all 12 highD/NGSIM horizon conditions and beats two non-learned switching controls in all 12 condition means. Taken together, the theoretical and empirical results support a shift in model design: rather than asking how to fuse every available representation, sequential learning systems should ask *which representation is conditionally useful enough to justify participating in prediction*.

Reproducibility Statement

The evidence archive accompanying this manuscript contains dataset/protocol identifiers, a sanitized completed-task inventory, raw and aggregate five-seed result tables, selector hyperparameters and occupancy summaries, mechanism-analysis CSVs, and the figure-generation script. Main performance uses full test outputs; mechanism diagnostics use fixed bounded subsets and are explicitly labeled. Failed or missing analyses are not imputed, and incompatible protocol generations are not merged into headline tables. The selector replays completed Base/Wave/Packet/Dual checkpoints read-only, fits action-risk surrogates on routefit, tunes choices on calibration, and evaluates test once; the test-label oracle is only a non-deployable upper bound. For a reproducible submission, the exact anonymized training implementation/configuration used to instantiate ϕ_w, ϕ_p, fusion, and the four action checkpoints should accompany these result artifacts as supplementary code.

AI Use Statement

Generative AI tools were used to assist with conceptual and theoretical framing, proposing and refining mathematical claims and proof sketches, experimental design, software implementation and editing, analysis and interpretation of experiment artifacts, literature discovery and bibliographic organization, and drafting and editing the manuscript. The authors independently checked the mathematical statements and proofs, executed and reviewed the code, verified the cited literature against source records, and derived every reported measurement from saved experiment artifacts. Generative AI was not used to fabricate measurements or to create synthetic data presented as real empirical evidence. The authors take responsibility for the final submission.

REFERENCES

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016. doi: 10.1109/CVPR.2016.110.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, 2018.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, OH, 2007. ISBN 9780940600720. doi: 10.1214/074921707000000391.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. doi: 10.1002/047174882X.

Jinyong Fan and Yanyan Shen. Stockmixer: A simple yet strong mlp-based architecture for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8389–8397, 2024. doi: 10.1609/aaai.v38i8.28681.

Gerald B. Folland and Alladi Sitaram. The uncertainty principle: A mathematical survey. *Journal of Fourier Analysis and Applications*, 3(3):207–238, 1997. doi: 10.1007/BF02649110.

Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. doi: 10.1109/CVPR.2018.00240.

Yaohui Huang, Runmin Zou, Yun Wang, Laeeq Aslam, and Ruipeng Dong. M2FMoE: Multi-resolution multi-view frequency mixture-of-experts for extreme-adaptive time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22075–22083, 2026. doi: 10.1609/aaai.v40i26.39362.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018. doi: 10.1109/ITSC.2018.8569552.

Jinning Li, Jiachen Li, Sangjae Bae, and David Isele. Adaptive prediction ensemble: Improving out-of-distribution generalization of motion forecasting. *arXiv preprint arXiv:2407.09475*, 2024a.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. Master: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 162–170, 2024b. doi: 10.1609/aaai.v38i1.27767.

Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556, 2020. doi: 10.1007/978-3-030-58536-5_32.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Ziqi Liu. Freqmoe: Enhancing time series forecasting through frequency decomposition mixture of experts. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3430–3438. PMLR, 2025.

David A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. doi: 10.1023/A:1007618624809.

Marion Neumeier, Andreas Tollkuhn, Michael Botsch, and Wolfgang Utschick. A multidimensional graph fourier transformation neural network for vehicle trajectory prediction. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 687–694, 2022. doi: 10.1109/ITSC55140.2022.9922419.

Marion Neumeier, Sebastian Dorn, Michael Botsch, and Wolfgang Utschick. Prediction and interpretation of vehicle trajectories in the graph spectral domain. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1172–1179, 2023. doi: 10.1109/ITSC57777.2023.10422530.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180, 2019.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700, 2020. doi: 10.1007/978-3-030-58523-5_40.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

U.S. Department of Transportation Federal Highway Administration. Next generation simulation (ngsim) vehicle trajectories and supporting data. ITS DataHub dataset, 2016.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010.

Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, et al. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *European Conference on Computer Vision*, pages 682–700. Springer, 2022.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27268–27286, 2022a.

Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022b. doi: 10.1109/CVPR52688.2022.00862.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023. doi: 10.1109/CVPR52729.2023.01713.

A NOTATION

Table 5 lists the symbols used in the theoretical statements. We intentionally avoid reusing the same symbol for unrelated objects: U denotes a generic representation in Bayes-risk formulas, while $\mathcal{Z}(\beta)$ denotes the maximum-entropy partition function; $\widetilde{p}$ denotes a candidate predictive distribution, while q denotes the spectral occupancy distribution.

Table 5: Core notation. Spectral prior ϖ and routing prior ρ are distinct.

Symbol	Meaning	Symbol	Meaning
$\mathbf{X}, \mathbf{Y}$	observed history / future target	$h, \mathbf{C}_h, c$	horizon / auxiliary pre-target context / realized context
$\mathbf{W}, \mathbf{P}$	Wave / Packet representations	U	generic available representation in a risk formula
$G_P(c)$	information added by Packet after Wave	$G_W(c)$	information added by Wave after Packet
$\mathcal{R}^*(U \mid c)$	ideal Bayes log risk	$\mathcal{R}^*_{\mathcal{F}}(U \mid c)$	best log risk in the available finite model class
$\epsilon_{\mathcal{F}}(U \mid c)$	finite-model approximation gap	$\Delta_{\mathrm{dual}}(c)$	joint-view gain over the better single view
δ_{reg}	optional regime-label tolerance	κ_a, λ	action cost / cost tradeoff weight
ℓ	deployment loss such as ADE or MSE	$\mathcal{A}, a^*_\ell$	action set / ideal task-risk action
$s(\mathbf{X})$	history-only routing feature vector	$L_{i,a}, \widehat{R}_a$	realized routefit loss / predicted action risk
p_j, e_t, J	normalized spectral power / difference energy / number of modes	H_ω, H_t	normalized spectral / local-difference entropy
N_j, c_j, B	spectral occupancy / occupancy cost / expected budget	$q(\mathbf{n})$	occupancy probability mass function
$\beta, \mathcal{Z}(\beta), T_I$	Lagrange multiplier / partition function / information temperature	$\varpi(\omega)$	maximum-entropy spectral prior
M, π, ρ, τ	number of experts / routing distribution / routing prior / routing temperature	$r_k, \widehat{r}_k, \varepsilon_r$	true expert risk / surrogate risk / uniform surrogate-error bound

B STEP-BY-STEP PROOFS

The proofs below show every algebraic step needed for the main claims. Each proof starts with the idea, then carries out the calculation, and ends with the practical meaning.

B.1 PROOF OF THEOREM 1: INFORMATION VALUE EQUALS IDEAL LOG-RISK REDUCTION

Idea. Under log loss, the best possible probabilistic predictor is the true conditional distribution. Its expected log loss is conditional entropy. Therefore, subtracting two ideal log risks is the same as subtracting two conditional entropies, which is exactly conditional mutual information.

Step 1: Bayes log risk equals conditional entropy. Fix a realized context c. By definition,

$$\mathcal{R}^*(U \mid c) = \inf_{\widetilde{p}} \mathbb{E}[-\log \widetilde{p}(\mathbf{Y} \mid U, \mathbf{C}_h = c) \mid \mathbf{C}_h = c].$$

The minimizer is the true conditional distribution $p(\mathbf{Y} \mid U, \mathbf{C}_h = c)$, so

$$\mathcal{R}^*(U \mid c) = \mathrm{H}(\mathbf{Y} \mid U, \mathbf{C}_h = c). \tag{18}$$

Step 2: add Packet after Wave. Using Eq. (18),

$$\begin{aligned}
&\mathcal{R}^*(\mathbf{W} \mid c) - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) \\
&= \mathrm{H}(\mathbf{Y} \mid \mathbf{W}, \mathbf{C}_h = c) - \mathrm{H}(\mathbf{Y} \mid \mathbf{W}, \mathbf{P}, \mathbf{C}_h = c) \\
&= \mathrm{I}(\mathbf{Y}; \mathbf{P} \mid \mathbf{W}, \mathbf{C}_h = c) = G_P(c),
\end{aligned}$$

where the last equality is the standard identity $\mathrm{I}(A; B \mid D) = \mathrm{H}(A \mid D) - \mathrm{H}(A \mid B, D)$.

Step 3: reverse the direction. Swapping Wave and Packet gives

$$\mathcal{R}^*(\mathbf{P} \mid c) - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) = G_W(c).$$

Step 4: compare the joint view with the better single view. The better single view has risk $\min\{\mathcal{R}^*(\mathbf{W} \mid c), \mathcal{R}^*(\mathbf{P} \mid c)\}$. Subtracting the joint-view risk gives

$$\begin{aligned}&\min\{\mathcal{R}^*(\mathbf{W} \mid c), \mathcal{R}^*(\mathbf{P} \mid c)\} - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c)\\ &= \min\{G_P(c), G_W(c)\} = \Delta_{\text{dual}}(c).\end{aligned}$$

Both conditional mutual informations are nonnegative, so $\Delta_{\text{dual}}(c) \geq 0$.

Step 5: connect local and aggregate quantities. The law of total conditional mutual information gives

$$\mathbb{E}_{\mathbf{C}_h}[G_P(\mathbf{C}_h)] = \mathrm{I}(\mathbf{Y}; \mathbf{P} \mid \mathbf{W}, \mathbf{C}_h),$$

and symmetrically for G_W.

Meaning. The theorem does not say that a finite neural network must improve when a view is added. It says something more basic: for an ideal probabilistic predictor under log loss, the amount of improvement available from the second view is exactly its conditional information contribution. □

B.2 Proof of Corollary 1: when is strict two-way complementarity possible?

Theorem 1 gives $\Delta_{\text{dual}}(c) = \min\{G_P(c), G_W(c)\}$ with $G_P(c), G_W(c) \geq 0$. A minimum of two nonnegative numbers is strictly positive exactly when both numbers are strictly positive. Hence $\Delta_{\text{dual}}(c) > 0$ if and only if $G_P(c) > 0$ and $G_W(c) > 0$.

Meaning. If only one direction is positive, one view adds information after the other, but the joint representation cannot strictly improve on the better single view in both directions. □

B.3 Proof of Proposition 1: separating information from finite-model burden

Idea. Restricted-model risk equals ideal risk plus an approximation gap. Substitute this decomposition for Wave-only and for Wave–Packet, then collect terms.

By Eq. (6),

$$\mathcal{R}^*_{\mathcal{F}}(\mathbf{W} \mid c) = \mathcal{R}^*(\mathbf{W} \mid c) + \epsilon_{\mathcal{F}}(\mathbf{W} \mid c),$$

and

$$\mathcal{R}^*_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) = \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) + \epsilon_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c).$$

Subtracting the second line from the first gives

$$\begin{aligned}&\mathcal{R}^*_{\mathcal{F}}(\mathbf{W} \mid c) - \mathcal{R}^*_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c)\\ &= \big[\mathcal{R}^*(\mathbf{W} \mid c) - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c)\big] + \epsilon_{\mathcal{F}}(\mathbf{W} \mid c) - \epsilon_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c).\end{aligned}$$

The bracketed term equals $G_P(c)$ by Theorem 1, yielding Eq. (7).

Meaning. The finite model receives two effects at once: Packet makes new information available through $G_P(c)$, but the joint input may be harder or easier for the chosen model class. The net observed benefit is their sum. □

B.4 Proof of Proposition 2: ideal information versus activation cost

Idea. Write the risk–cost objective for Wave-only and for Wave–Packet, require the joint objective to be smaller, and rearrange.

At fixed c, define

$$J_W = \mathcal{R}^*(\mathbf{W} \mid c) + \lambda\kappa_W, \qquad J_{WP} = \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) + \lambda\kappa_{WP}.$$

Adding Packet is beneficial exactly when $J_{WP} < J_W$. Therefore

$$\begin{aligned}&\mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) + \lambda\kappa_{WP} < \mathcal{R}^*(\mathbf{W} \mid c) + \lambda\kappa_W\\ \iff\quad &\mathcal{R}^*(\mathbf{W} \mid c) - \mathcal{R}^*(\mathbf{W}, \mathbf{P} \mid c) > \lambda(\kappa_{WP} - \kappa_W).\end{aligned}$$

The left side is $G_P(c)$ by Theorem 1, giving Eq. (8).

Meaning. Even an ideal predictor should not activate a costly second representation when the new information is too small relative to the assigned cost. □

B.5 Proof of Corollary 2: the main finite-model activation rule

Start from the restricted-model risk–cost comparison:

$$\begin{aligned} & \mathcal{R}^*_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) + \lambda \kappa_{WP} < \mathcal{R}^*_{\mathcal{F}}(\mathbf{W} \mid c) + \lambda \kappa_W \\ \iff \quad & \mathcal{R}^*_{\mathcal{F}}(\mathbf{W} \mid c) - \mathcal{R}^*_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) > \lambda(\kappa_{WP} - \kappa_W). \end{aligned}$$

Insert Proposition 1:

$$G_P(c) + \epsilon_{\mathcal{F}}(\mathbf{W} \mid c) - \epsilon_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) > \lambda(\kappa_{WP} - \kappa_W).$$

Move the two approximation-gap terms to the right:

$$G_P(c) > \lambda(\kappa_{WP} - \kappa_W) + \epsilon_{\mathcal{F}}(\mathbf{W}, \mathbf{P} \mid c) - \epsilon_{\mathcal{F}}(\mathbf{W} \mid c),$$

which is Eq. (9).

Meaning. This proof shows exactly where every term in the boxed rule comes from. The left side is available population information. The first right-side term is extra deployment cost. The second right-side term is the change in finite-model difficulty. Activation is justified only when the left side is larger than their sum. □

C Optional Inductive Mechanisms: Formal Statements

These mechanisms instantiate the Wave branch or provide a routing reference. They are not part of the definition of a representation regime and are not required by the strongest history-selector result.

C.1 Maximum-entropy spectral allocation

Let $\mathbf{N} = (N_1, \ldots, N_J) \in \mathbb{N}_0^J$ be an abstract occupancy vector over positive spectral modes. Each occupancy of mode j costs $c_j > 0$, and the expected total budget is fixed to $B > 0$. Let $q(\mathbf{n})$ be a probability mass function over occupancy states $\mathbf{n}$.

Theorem 2 (Maximum-entropy occupancy)**.** *Among all q satisfying $\sum_{\mathbf{n}} q(\mathbf{n}) = 1$ and $\mathbb{E}_q[\sum_j c_j N_j] = B$, the entropy-maximizing distribution is*

$$q^*(\mathbf{n}) = \mathcal{Z}(\beta)^{-1} \exp\left[-\beta \sum_j c_j n_j\right], \qquad \mathcal{Z}(\beta) = \sum_{\mathbf{n}} \exp\left[-\beta \sum_j c_j n_j\right], \tag{19}$$

with $\beta > 0$ chosen to satisfy the budget. Its mean occupancy is

$$\mathbb{E}[N_j] = (e^{\beta c_j} - 1)^{-1}. \tag{20}$$

If $c(\omega) = a\omega$ and the mode measure is proportional to ω^ν with $a, \nu > 0$, the induced normalized allocation has shape $\varpi(\omega) \propto \omega^\nu / (e^{\omega/T_I} - 1)$ with $T_I = (a\beta)^{-1}$.

Proof, step 1: write the constrained objective. Maximize Shannon entropy $-\sum_{\mathbf{n}} q(\mathbf{n}) \log q(\mathbf{n})$ subject to normalization and the expected-budget constraint. Introduce Lagrange multipliers α and β:

$$\mathcal{L}(q) = -\sum_{\mathbf{n}} q(\mathbf{n}) \log q(\mathbf{n}) - \alpha\left(\sum_{\mathbf{n}} q(\mathbf{n}) - 1\right) - \beta\left(\sum_{\mathbf{n}} q(\mathbf{n}) \sum_j c_j n_j - B\right).$$

Step 2: differentiate with respect to each probability. Setting $\partial \mathcal{L} / \partial q(\mathbf{n}) = 0$ gives

$$-\log q(\mathbf{n}) - 1 - \alpha - \beta \sum_j c_j n_j = 0,$$

so

$$q(\mathbf{n}) = C \exp\left[-\beta \sum_j c_j n_j\right].$$

Normalization makes $C = 1/\mathcal{Z}(\beta)$, proving Eq. (19).

Step 3: compute one mode's mean occupancy. Because the exponent is a sum over modes, the distribution factorizes. For mode j,

$$q_j(n) = (1 - e^{-\beta c_j})e^{-\beta c_j n}, \qquad n = 0, 1, 2, \ldots,$$

a geometric distribution. Using $\sum_{n\geq 0} nr^n = r/(1-r)^2$ with $r = e^{-\beta c_j}$,

$$\mathbb{E}[N_j] = \frac{e^{-\beta c_j}}{1 - e^{-\beta c_j}} = \frac{1}{e^{\beta c_j} - 1},$$

which proves Eq. (20).

Step 4: obtain the spectral shape. Set $c(\omega) = a\omega$ and $T_I = (a\beta)^{-1}$. Then the mean occupancy contributes the factor $(e^{\omega/T_I} - 1)^{-1}$. Multiplying by a mode-density factor proportional to ω^ν gives Eq. (13). Entropy is strictly concave in q, so the feasible maximizer is unique. □

Meaning. The theorem does not claim that nature follows this spectrum or that it must outperform every alternative. It supplies a least-committal spectral allocation under one explicit expected-budget constraint, which is why we treat it as an inductive prior and test uniform/exponential alternatives.

Proposition 3 (Information-temperature ordering)**.** *For $w_T(\omega) = \omega^\nu/(e^{\omega/T} - 1)$ and $0 < \omega_1 < \omega_2$, the ratio $w_T(\omega_2)/w_T(\omega_1)$ increases strictly with T. It tends to 0 as $T \downarrow 0$ and to $(\omega_2/\omega_1)^{\nu-1}$ as $T \to \infty$.*

Proof. Let $R(T) = w_T(\omega_2)/w_T(\omega_1)$. Then

$$\log R(T) = \nu \log \frac{\omega_2}{\omega_1} + \log(e^{\omega_1/T} - 1) - \log(e^{\omega_2/T} - 1).$$

Differentiating,

$$\frac{d}{dT} \log R(T) = \frac{1}{T^2} \left[\frac{\omega_2}{1 - e^{-\omega_2/T}} - \frac{\omega_1}{1 - e^{-\omega_1/T}} \right].$$

For fixed $T > 0$, the function $x/(1 - e^{-x/T})$ is increasing for $x > 0$, so the bracket is positive and $R(T)$ increases with T. As $T \downarrow 0$, $e^{\omega/T} - 1 \sim e^{\omega/T}$, giving $R(T) \to 0$. As $T \to \infty$, $e^{\omega/T} - 1 \sim \omega/T$, giving $R(T) \to (\omega_2/\omega_1)^{\nu-1}$. □

Meaning. Raising T_I shifts relative prior mass toward higher frequencies; lowering it emphasizes lower frequencies. This gives the "information temperature" a precise monotonic interpretation.

C.2 KL-REGULARIZED ROUTING REFERENCE

Let $f_1, \ldots, f_M$ be experts with conditional risks $r_k(c)$. Let $\pi \in \Delta^{M-1}$ be a routing distribution, $\rho_k > 0$ a normalized prior over experts, and $\tau > 0$ a routing temperature.

Proposition 4 (Gibbs router and regret)**.** *The minimizer of $\sum_k \pi_k r_k + \tau \mathrm{KL}(\pi\|\rho)$ is*

$$\pi_k^* = \frac{\rho_k e^{-r_k/\tau}}{\sum_j \rho_j e^{-r_j/\tau}}, \tag{21}$$

and if $k^* \in \arg\min_k r_k(c)$, then

$$\sum_k \pi_k^* r_k - r_{k^*} \leq \tau \log(1/\rho_{k^*}). \tag{22}$$

Proof. Add a multiplier η for $\sum_k \pi_k = 1$. Differentiating $\sum_k \pi_k r_k + \tau \sum_k \pi_k \log(\pi_k/\rho_k) + \eta(\sum_k \pi_k - 1)$ with respect to π_k gives $r_k + \tau[\log(\pi_k/\rho_k) + 1] + \eta = 0$. Solving for π_k and normalizing yields Eq. (21). For the regret bound, optimality implies the regularized objective at π^* is no larger than at the point mass on the best expert k^*. Dropping the nonnegative term $\tau \mathrm{KL}(\pi^*\|\rho)$ from the left gives Eq. (22). □

Meaning. Low-risk experts receive exponentially larger routing weight, but the prior ρ and temperature τ prevent the router from becoming an unrestricted winner-take-all rule.

Proposition 5 (Learned-score stability)**.** *If surrogate risks satisfy* $\max_k |\widehat{r}_k - r_k| \le \varepsilon_r$*, pairwise log-routing odds deviate from ideal Gibbs odds by at most* $2\varepsilon_r/\tau$*, and true-risk regret is at most* $\tau \log(1/\rho_{k^*}) + 2\varepsilon_r$.

Proof. For experts i and j, the Gibbs log-odds based on true risks are $\log(\pi_i^*/\pi_j^*) = \log(\rho_i/\rho_j) - (r_i - r_j)/\tau$. Replacing r by $\widehat{r}$ changes this by $[(\widehat{r}_i - r_i) - (\widehat{r}_j - r_j)]/\tau$, whose absolute value is at most $2\varepsilon_r/\tau$. The regret bound follows by comparing the learned and ideal regularized objectives: each replacement of a true risk by its surrogate changes an expectation by at most ε_r, giving at most $2\varepsilon_r$ total on top of Proposition 4. □

Meaning. The router is stable when the risk predictor is accurate relative to the routing temperature. Small τ makes routing sharper but also amplifies surrogate-risk errors.

Proposition 6 (Objective-conditioned routing)**.** *Under a common prior* ρ *and temperature* τ*, two objectives with different expert-risk vectors change pairwise Gibbs odds only through their pairwise risk differences.*

Proof. For objective o, Eq. (21) gives $\log(\pi_i^{(o)}/\pi_j^{(o)}) = \log(\rho_i/\rho_j) - [r_i^{(o)} - r_j^{(o)}]/\tau$. Subtracting this expression for two objectives cancels the common prior term. □

Meaning. If two tasks prefer different experts, their routing distributions can differ even when the prior is identical; conversely, observed routing differences should not automatically be attributed to different priors.

D Unified Protocol Details and Full Result Audit

The accumulated unified experiment report contains 221 completed fresh neural trials after the targeted seed extension. Critical IWPD/GRU-wide/FreqMoE and noWave/noPacket comparisons use five matched seeds, while several source-locked general baselines retain three completed seeds; failed/incomplete tasks are not imputed. PatchTST, iTransformer and FreqMoE are source-locked upstream implementations; no pretrained weights are used. Tables in the main text are generated from `fresh_v7_summary.csv`. Mechanism analysis uses frozen best checkpoints and recorded prediction banks. One Weather seed did not pass the earlier strict same-backend replay tolerance for latent analysis and is excluded only from that representation/probe aggregation; its main performance trial remains valid. The corrected history-only regime experiment completes 25 dataset-seed jobs (five datasets by five common seeds) using read-only replay of Base/Wave/Packet/Dual checkpoints.

D.1 Financial ranking boundary

Table 6: DailyRankIC in the unified protocol. "Best standard" is the strongest of GRU, GRU-wide, DLinear, PatchTST, iTransformer and FreqMoE; higher is better.

Dataset	h	IWPD	Best standard
A-share	1	+0.2460	GRU-wide (+0.2658)
A-share	3	+0.1387	GRU-wide (+0.1744)
A-share	5	+0.0763	GRU-wide (+0.1073)
A-share	10	+0.0620	GRU-wide (+0.0823)
A-share	20	+0.0691	GRU-wide (+0.0668)
JPX	1	-0.0496	iTransformer (-0.0117)
JPX	3	+0.0046	GRU-wide (+0.0082)
JPX	5	-0.0013	GRU (+0.0271)
JPX	10	+0.0323	GRU-wide (+0.0151)
JPX	20	+0.0078	DLinear (-0.0098)

D.2 Corrected history-selector audit

The corrected selector report contains 25 complete dataset–seed jobs and 1125 per-seed/per-horizon action rows. The accuracy-only and budgeted policies are intentionally distinct. The accuracy selector improves over AlwaysDual in 25/25 dataset–horizon cells and over the best fixed four-action choice in 12/12 trajectory cells (13/25 overall); the budgeted selector improves over AlwaysDual in 23/25 cells. Representative five-seed means are shown below; cost is the structural branch-activation proxy with AlwaysDual cost 2.

Table 7: Representative corrected history-selector results. Gain is the reduction in five-seed mean test loss relative to AlwaysDual (higher is better).

Dataset	h	Acc. gain	Acc. cost	Budget gain	Budget cost
highD	1	23.04%	1.080	23.54%	0.631
highD	10	6.23%	1.177	5.88%	0.951
highD	30	1.36%	1.378	0.35%	1.046
NGSIM	1	29.85%	0.559	29.69%	0.450
NGSIM	10	2.73%	0.525	2.33%	0.016
NGSIM	30	0.74%	0.786	0.47%	0.004
Weather	1	6.52%	0.914	9.04%	0.240
A-share	10	1.05%	1.271	1.45%	0.580
JPX	10	0.14%	0.805	-0.07%	0.0004

The budgeted selector selects Base on approximately 98.5% of NGSIM $h = 10$ windows and 99.6% at $h = 30$, whereas highD $h = 10$ remains mixed across Base/Dual/Packet/Wave. This distinction prevents a cheap dataset-level fallback from being misinterpreted as fine-grained regime switching.

E Visual Intuition and Moved Main-Text Figures

Time–frequency intuition from an earlier version. Figure 2 preserves the earlier schematic population-regime picture. Its role is descriptive only: the axes are local directional conditional-information gains at a realized auxiliary context c, not raw spectral energy or hand-crafted descriptors.

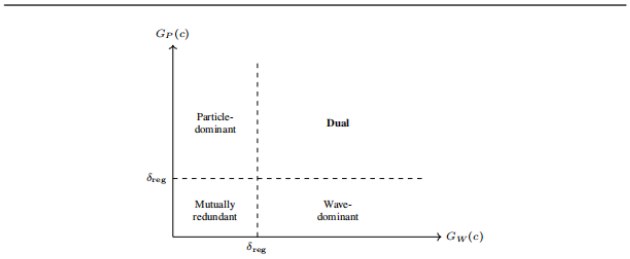

Figure 1: **Context-specific population representation regimes.** The axes are local directional conditional-information gains at a realized auxiliary context c, not raw Fourier energy or hand-designed features. History-only descriptors may predict an operational action but do not define these population quantities.

Figure 2: Earlier schematic retained for intuition. Both axes are population quantities: $G_P(c) = I(Y; P \mid W, C_h = c)$ and $G_W(c) = I(Y; W \mid P, C_h = c)$. History-only descriptors may predict an operational action but do not define these quantities.

Figures moved out of the main text. The next two figures are kept only as supplementary visuals because the main text now uses tables for the same evidence.

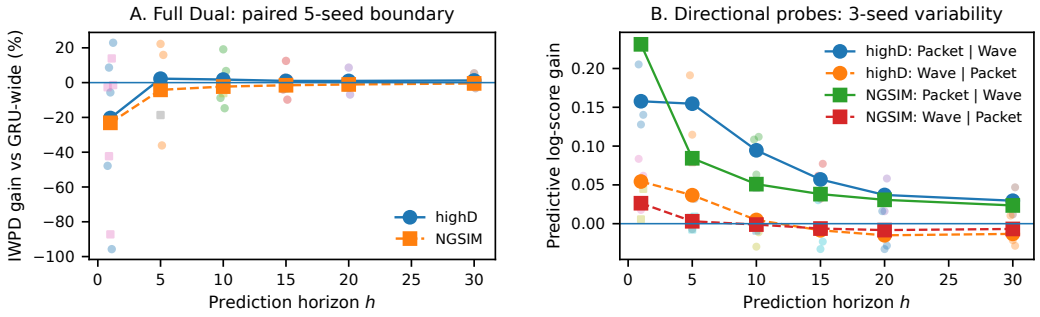

Figure 3: Main-text Figure 2 moved to the appendix. Mandatory Dual is not uniformly better, and directional probe gains are asymmetric.

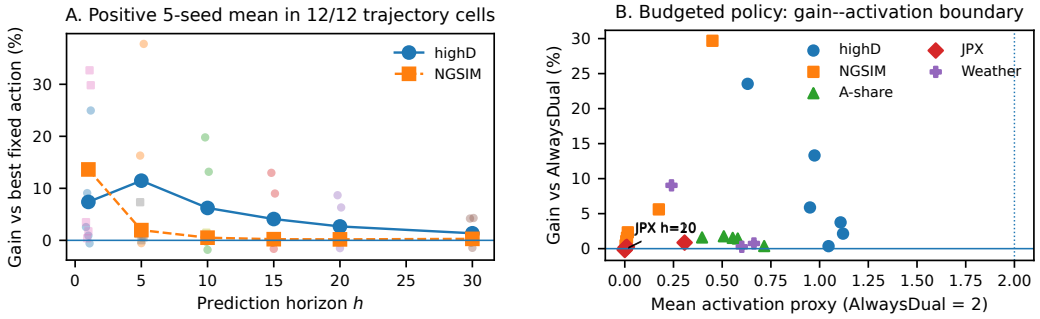

Figure 4: Main-text Figure 3 moved to the appendix. The accuracy selector beats the condition-wise best fixed action in all 12 trajectory condition means, while the budgeted selector trades gain against the activation proxy.

E.1 Cross-domain performance and selector boundaries

Table 8: Representative non-trajectory RMSE boundaries (lower is better). IWPD/GRU-wide/noWave/noPacket use five seeds.

Dataset	h	IWPD	Best standard	Best ablation
Weather	1	0.20020	PatchTST (0.18403)	noWave (0.19198)
Weather	96	0.40485	GRU-wide (0.40313)	noWave (0.40055)
Weather	192	0.45609	GRU-wide (0.45590)	noWave (0.45381)
A-share	1	0.13935	GRU-wide (0.13892)	noWave (0.13839)
A-share	10	0.13980	GRU-wide (0.13904)	noWave (0.13898)
A-share	20	0.14122	GRU-wide (0.14124)	noWave (0.14063)
JPX	1	0.02483	DLinear (0.02465)	noWave (0.02479)
JPX	10	0.02404	FreqMoE (0.02397)	noPacket (0.02402)
JPX	20	0.02391	FreqMoE (0.02387)	noPacket (0.02390)

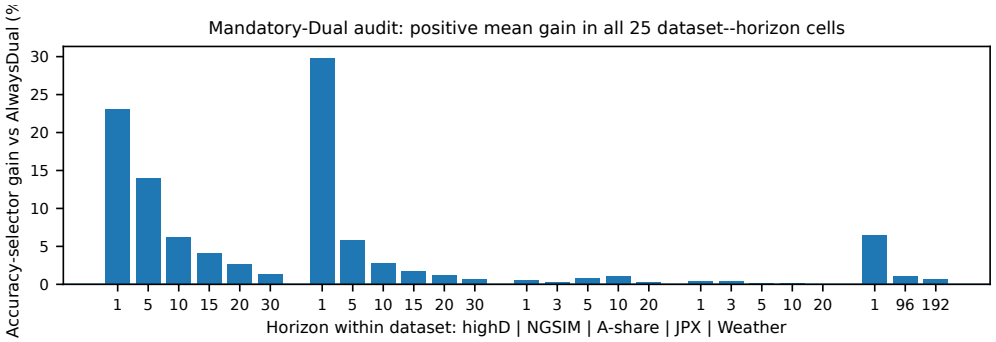

Figure 5: Cross-domain audit of the accuracy-only history selector. Every bar is positive relative to AlwaysDual, giving 25/25 positive dataset–horizon means. This does not imply superiority to the best fixed action outside the 12 trajectory cells.

E.2 Selector controls, seed direction, and margin sensitivity

The learned accuracy selector is not merely exploiting the existence of any switching rule. Averaged over horizons, its five-seed mean loss is lower than the fixed-quantile HistoryPhaseRule by 9.13% on highD and 4.40% on NGSIM, and lower than RandomFrequencyControl by 7.37% and 4.22%, respectively; the direction is favorable in all six horizons of both datasets. These are descriptive reductions of condition means, not additional independent significance tests.

Seed direction is heterogeneous despite all 12 positive condition means. The paired seed-win counts against the condition-wise best fixed action are highD $\{4, 4, 3, 3, 3, 3\}/5$ and NGSIM $\{5, 5, 5, 5, 5, 4\}/5$ for horizons $\{1, 5, 10, 15, 20, 30\}$. This is why the main figure displays seed-level points and why the paper does not claim seed-uniform superiority.

The thresholded regime labels are also descriptive rather than invariant. Using mean frozen-probe gains, highD $h = 10$ and NGSIM $h = 5$ are labeled Dual at $\delta_{\text{reg}} = 0$ but Particle-dominant once $\delta_{\text{reg}} \geq 0.005$; the continuous directional gains and the downstream action-risk experiment are therefore the primary evidence.

E.3 Experimental evidence files

The supplementary evidence directory includes raw and aggregate five-seed performance outputs, frozen-router comparisons, representation CKA, predictive log-score gains, prior/attention audits, perturbation sensitivity, prediction-health checks, selector hyperparameters/occupancy, a sanitized task-completion inventory, and the explicit remaining-gap inventory. These artifacts are intended to make positive and negative outcomes equally auditable.

F Protocol Safeguards

All normalization and imputation statistics are fit on training data only. Selection chooses neural checkpoints, routefit fits frozen routing/probe models, calibration estimates held-out residual quantities or empirical guard statistics, and test is used once for reporting. Dataset IDs and scientific-source hashes prevent silent reuse across changed data or code. Time-series dependence means the calibration screens are empirical held-out checks rather than distribution-free safety certificates.