TEMPORAL SHEAF NEURAL NETWORKS: COORDINATE-CONSISTENT MEMORY FOR LINK PREDICTION ACROSS TEMPORAL GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal link prediction requires persistent memory that incorporates new interactions while retaining useful history. We introduce Temporal Sheaf Neural Networks (TSNN), one memory architecture for homogeneous, knowledge-graph, heterogeneous, and sequential temporal graphs under a single typed event schema. TSNN pairs node memories with evolving orthogonal frames: each event updates the endpoint frames, exact carry-over preserves the represented vectors across that coordinate change, and recurrent updates and sheaf diffusion over the retained history graph then update the content. All queries at a timestamp are scored from one strict-history snapshot before any update is committed. Our analysis identifies how frames can influence predictions of the base scorer: only through coordinate-dependent content updates or anisotropic diffusion. With exact carry, equivariant content updates, and a scalar diffusion gain, its ambient trajectories and scores coincide with those of a matched frame-free model; with a fixed gain and a small step, its diffusion maps are non-expansive in a weighted norm as the graph evolves. Across 21 TGB, TGB 2.0, and TGB-Seq datasets, TSNN exceeds every listed comparator on 18, including all eight TGB-Seq datasets, and improves thgl-forum MRR by 0.1075; the same architecture also has the best mean rank in transductive AP on 13 DGB datasets (1.23 among 13 methods, against 1.96 for TPNet). In matched ablations on five datasets, evolving frames have higher mean MRR than frame-free, fixed-frame, shared-frame, side-state, and FiLM controls, with significant gains of 0.0091–0.0245 over the frame-free model (Welch t-test, $p < 0.005$).

1 INTRODUCTION

Temporal graphs describe systems in which interactions—such as transactions, messages, and changing knowledge-base relations—unfold over time (Huang et al., 2023; Gastinger et al., 2024a). A central task is *temporal link prediction*: ranking legal candidates for a query using only events strictly before the query time, under benchmark-specific candidate-set and evaluation protocols (Poursafaei et al., 2022; Gastinger et al., 2024a). Models must consequently retain or reconstruct historical state. Recurrent architectures maintain per-node memories (Kumar et al., 2019; Rossi et al., 2020), history encoders summarise sampled interactions (Xu et al., 2020; Yu et al., 2023), and memorisation baselines exploit recurring interactions (Poursafaei et al., 2022; Gastinger et al., 2024b).

These methods are also fragmented by graph family. Memory and walk models such as TGN and TPNet target homogeneous interaction streams (Rossi et al., 2020; Lu et al., 2024); relational models such as CEN and TLogic target temporal knowledge graphs (Li et al., 2022b; Liu et al., 2022); type-aware models such as STHN target heterogeneous graphs (Li et al., 2023); and session recommenders target sequential user–item streams (Yi et al., 2025b). Each is typically evaluated on one or two families, and several exceed their time or memory budget on the largest TGB 2.0 datasets (Gastinger et al., 2024a). We instead treat every family as one typed event stream.

Most learned decoders compare node representations in one shared coordinate system, which training can adapt to the task. We study a different design, in which each node keeps its memory in its own local coordinate frame, and that frame changes as the node interacts. A frame is a point of view:

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.

changing it should change how a memory is written down, but not what the memory says. This raises two questions: can frames move with events without altering memory or seeing the future, and, if exact carry hides every frame change from the scorer, how can frames help at all? They act only through coordinate-dependent content updates and anisotropic diffusion (Section 5), and matched controls test that channel (Section 6).

Prior work provides pieces of such a design. Geometric models provide local-frame invariance and transport (Cohen et al., 2019; Singer & Wu, 2012), and sheaf Laplacians provide learned diffusion through per-edge maps (Hansen & Ghrist, 2019; Bodnar et al., 2022; Barbero et al., 2022a), including signal-conditioned maps for forecasting (Mostafa et al., 2026). LoCS combines time-indexed frames with a GRU to forecast object trajectories, recomputing each frame from observed positions and velocities (Kofinas et al., 2021). Bundle Neural Networks provide flat orthogonal connections and Householder maps on static graphs (Bamberger et al., 2025), and Householder parameterisations and relation-specific orthogonal maps are established (Mhammedi et al., 2017; Sun et al., 2019; Li et al., 2022a); Appendix A.1 gives a fuller catalogue, and Table 4 separates what TSNN inherits from what it adds. What is missing is a memory architecture for temporal link prediction whose frames persist and change with events, with an exact rule that keeps a memory unchanged when its frame changes, and a scoring order that uses only earlier events, on every family of temporal graphs.

To address this gap, we introduce Temporal Sheaf Neural Networks (TSNN)[1], a causal architecture that represents each node's persistent memory in evolving orthogonal frames parameterised by Householder reflections and exchanges information through node-factored transport. Because every family is read through one typed event schema, the same architecture serves homogeneous, knowledge-graph, heterogeneous, and sequential streams. The resulting contributions are fourfold:

- We introduce TSNN, a temporal memory architecture in which each node stores its memory in its own evolving frame. When a frame turns, exact carry-over rewrites the memory in the new frame without changing the vector it represents. The frame update writes event information into the frame state, the recurrent update and sheaf diffusion then change the content, and each wave is scored before it is committed (Section 3).
- We prove that, for the invariant base scorer, frames can change TSNN's predictions in only two ways: through a recurrent update that depends on the frame, or through diffusion with a direction-dependent gain. With equivariant content updates and a scalar gain, TSNN produces exactly the same scores as the same model with frames removed. We also prove that test scores never use same-time or future events, and that with a fixed gain and a small step diffusion never amplifies a perturbation in a weighted norm, even as the graph changes (Sections 4–5).
- With one configuration per dataset, selected on validation data, TSNN exceeds every listed comparator on 18 of 21 TGB, TGB 2.0, and TGB-Seq datasets, including all eight TGB-Seq datasets, and has the best DGB mean rank. In matched controls, evolving frames have higher mean MRR than frame-free, fixed-frame, shared-frame, side-state, and FiLM variants on all five datasets, significantly so over the frame-free model ($p < 0.005$).
- We show that one architecture covers four families of temporal graphs: homogeneous, knowledge-graph, heterogeneous, and sequential. All families share the core update equations, with task-specific embeddings, features, and candidate domains, so TSNN runs on 34 benchmark settings with up to 5.9M nodes and 67M events, unlike most prior methods (Appendix F.2).

2 PRELIMINARIES

The formulation separates two objects: the directed sequence of timed events, which says what happened, and an undirected graph, which says whose memories exchange information during diffusion. Table 5 (Appendix A.2) collects the notation.

2.1 TYPED TEMPORAL EVENTS AND PREDICTION QUERIES

Let $\mathcal{V}$ be a finite node set, $\mathcal{T}$ a finite node-type vocabulary with type map $\nu : \mathcal{V} \to \mathcal{T}$, and $\mathcal{R}$ a finite edge relation vocabulary. Node identifiers index persistent state; identifier-derived features and metadata enter prediction only when the task provides them. A node need not have appeared earlier to be a legal candidate. An observed event is $e_i = (u_i, v_i, t_i, r_i, \mathbf{x}_i)$, where $u_i, v_i \in \mathcal{V}$ are the source

[1] Code: `https://anonymous.4open.science/r/ICLR-TSNN-CODE-D479/README.md`

and destination nodes, t_i is a scalar timestamp, $r_i \in \mathcal{R}$ is an edge relation, and $\mathbf{x}_i$ is an event-feature vector of fixed width $d_x \geq 0$. The index i is the event identifier, and $u_i = v_i$ denotes a self-interaction. A stream of $N \geq 0$ events is the *sequence* $\mathcal{E} = (e_i)_{i=1}^{N}$ with $t_i \leq t_{i+1}$, not a set: repeated identical observations retain their multiplicity and identifiers. Events with equal timestamps form one wave that updates the state in a single transition that does not depend on their order (Block C), and timestamps use one fixed recorded unit before differences or encodings. If event features are omitted ($d_x = 0$), every computation consuming $\mathbf{x}_i$ drops that term rather than substituting a placeholder. When node types constrain interactions, a task-level schema $\mathcal{S}_{\text{type}} \subseteq \mathcal{T} \times \mathcal{R} \times \mathcal{T}$ specifies the allowed source-type, relation, and destination-type triples. Every benchmark family is an instance of this one schema: homogeneous streams (TGB `tgbl`, DGB) set $|\mathcal{T}| = |\mathcal{R}| = 1$, knowledge graphs (`tkgl`) have $|\mathcal{R}| > 1$, heterogeneous graphs (`thgl`) have $|\mathcal{T}| > 1$ and $|\mathcal{R}| > 1$ with a typed domain from $\mathcal{S}_{\text{type}}$, and sequential streams (TGB-Seq) have long per-source histories.

The strict history $\mathcal{E}_{<t}$ is the subsequence of events with timestamp less than t and defines the information available to the scorer. A tail query $q = (u, ?, r, t)$ asks which node c completes (u, c, t, r), whereas a head query $q^{\text{head}} = (?, v, r, t)$ asks for the missing source. Henceforth, q denotes either direction and c_q^+ its true missing endpoint. The task fixes an admissible domain $\mathcal{D}_q \subseteq \mathcal{V}$, including any type or bipartite constraints, and an ordered candidate list $\mathcal{C}_q = (c_{q1}, \ldots, c_{qM_q}) \in \mathcal{D}_q^{M_q}$ with $M_q \geq 1$. The labelled positive satisfies $c_q^+ = c_{qj_q^+}$ for $j_q^+ \in \{1, \ldots, M_q\}$; the remaining positions contain supplied or sampled negatives. Given q, $\mathcal{E}_{<t}$, and $\mathcal{C}_q$, the prediction task is to learn a scorer $s_q : \mathcal{D}_q \to \mathbb{R}$ that, at query time, uses only the strict history and permitted query information to rank c_q^+ above the remaining candidate-list entries. The admissible domain follows from permitted schema information and known query fields, never from the hidden positive, and current event features $\mathbf{x}_i$ enter the state update only after observation.

2.2 Graph and sheaf notation

The spatial operator is defined on a finite, nonempty, undirected multigraph $G = (V, E)$ with $V \subseteq \mathcal{V}$. Each $e \in E$ is a distinct edge occurrence with unordered endpoints $\partial e = \{a_e, b_e\}$, so parallel edges count separately. Loops are permitted: a loop $\{a, a\}$ has two identical incidences, so its coboundary row vanishes and it adds 2 to the degree of a (Block E specialises G to distinct retained pairs). We use the standard static-graph sheaf formalism (Hansen & Ghrist, 2019; Hansen & Gebhart, 2020; Bodnar et al., 2022; Barbero et al., 2022a), stated for one fixed G and specialised to the equal stalk width $d \geq 1$ used by TSNN. Appendix A.2 retains the general varying-width form. A Euclidean cellular sheaf $\mathcal{F}$ assigns node and edge stalks $\mathcal{F}(a) = \mathcal{F}(e) = \mathbb{R}^d$ (coordinate spaces, not stored values), together with a linear restriction map $\rho_{a,e} : \mathcal{F}(a) \to \mathcal{F}(e)$ at every incidence $a \in \partial e$.

Let $\mathbf{z} = \operatorname{col}_{a \in V} \mathbf{h}_a \in \mathbb{R}^{|V|d}$ stack the node vectors in a fixed vertex order, and assign each edge occurrence an arbitrary orientation $a_e \to b_e$. The coboundary $\delta_{\mathcal{F}} : \bigoplus_{a \in V} \mathbb{R}^d \to \bigoplus_{e \in E} \mathbb{R}^d$ acts as $(\delta_{\mathcal{F}} \mathbf{z})_e = \rho_{b_e,e} \mathbf{h}_{b_e} - \rho_{a_e,e} \mathbf{h}_{a_e}$, where $\bigoplus$ denotes a direct sum. The degree-zero sheaf Laplacian is $\mathbf{L}_{\mathcal{F}} = \delta_{\mathcal{F}}^\top \delta_{\mathcal{F}}$, with Dirichlet energy $\mathscr{E}_{\mathcal{F}}(\mathbf{z}) = \frac{1}{2} \mathbf{z}^\top \mathbf{L}_{\mathcal{F}} \mathbf{z} = \frac{1}{2} \sum_{e \in E} \|(\delta_{\mathcal{F}} \mathbf{z})_e\|_2^2$. Both are independent of the chosen orientations (Hansen & Ghrist, 2019), with each edge occurrence summed once. Sheaf neural networks diffuse node signals with $\mathbf{L}_{\mathcal{F}}$ in place of the graph Laplacian and learn the restriction maps end to end (Hansen & Gebhart, 2020; Bodnar et al., 2022). For general nonorthogonal restrictions, $\rho_{a_e,e}^\top \rho_{b_e,e}$ need not be an invertible coordinate change; the orthogonal choice in Section 3.5 is what makes it one. Norms and other conventions follow Appendix A.2.

3 Temporal Sheaf Neural Networks

TSNN processes time in waves, one timestamp at a time. It first freezes what is known (Block A) and scores every query from that frozen state (Block B). Only then does it read the new events: it turns each endpoint's frame (Block C), carries the stored memory into the new frame and updates it with a recurrent step (Block D), and diffuses the update over the recent history graph (Block E).

3.1 Block A: Freeze history

What may the model know when it scores? The equal-time wave $\mathcal{W}_t = \{i \in \{1, \ldots, N\} : t_i = t\}$ collects all events at timestamp t, and $b(t) = |\{i \in \{1, \ldots, N\} : t_i < t\}|$ counts events strictly before t. Define Ξ^i as the snapshot after the first i committed events, with Ξ^0 denoting the initialisation.

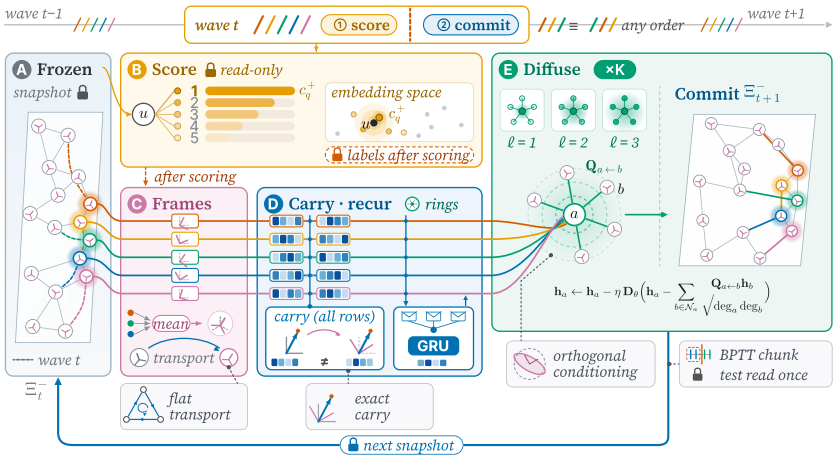


Figure 1: One wave of TSNN. (A) Freeze the snapshot Ξ_t^-; (B) score every query from it; then commit the wave as one transition: (C) turn the endpoint frames, (D) carry memories exactly and apply the GRU, and (E) run K synchronous diffusion rounds on the retained history graph.

For any node field a_w, write $a_w^-(t) = a_w^{b(t)}$ for its value in $\Xi_t^- := \Xi^{b(t)}$. Every query at t reads this snapshot; all its scores are fixed before $\mathcal{W}_t$ updates memory or parameters.

To specify the model state, fix a latent width $d \geq 1$. After event i, node w stores a local vector $\mathbf{h}_w^i \in \mathbb{R}^d$, a frame-parameter matrix $\mathbf{F}_w^i \in \mathbb{R}^{k\times d}$, its last endpoint-interaction time $s_w^i \in \mathbb{R} \cup \{\perp\}$, and a buffer $\mathcal{B}_w^i$ containing the unordered endpoint pairs of its last at most $B_{\max}$ incident events. Here, $k \geq 0$ is the reflector count ($k = 0$: empty product $\mathbf{U} = \mathbf{I}_d$), $B_{\max} \geq 0$ the history cap (counting event *occurrences*, not distinct neighbours), and $\perp$ denotes no endpoint interaction yet. The frozen state Ξ^i contains these fields plus the pair-recency table of Block B. Learned parameters sit outside Ξ^i and remain fixed during wave scoring; a read-only view of the snapshot suffices. Each node's initial state is drawn once, $\mathbf{h}_w^0 \sim \mathcal{N}(0, 10^{-4}\mathbf{I}_d)$ and rows of $\mathbf{F}_w^0 \sim \mathcal{N}(0, 10^{-2}\mathbf{I}_d)$, with $s_w^0 = \perp$ and $\mathcal{B}_w^0 = \varnothing$; random initial states let the model tell nodes apart. The controls of Section 6 instead use a common initialisation or store $\mathbf{h}_w^0 = (\mathbf{U}_w^0)^\top \mathbf{g}_w^0$ for one shared draw $\mathbf{g}_w^0$, so matching seeds give matching ambient memories.

Memory lives in an evolving orthogonal frame. For a raw frame matrix $\mathbf{F}$ write $\mathbf{f}^{(j)} \in \mathbb{R}^d$ for the transpose of its jth row; for $\mathbf{f} \in \mathbb{R}^d$, threshold $\varepsilon_H > 0$, and $\hat{\mathbf{u}} = \mathbf{f}/\|\mathbf{f}\|_2$ when $\|\mathbf{f}\|_2 > \varepsilon_H$, define

$$H_{\varepsilon_H}(\mathbf{f}) = \begin{cases} \mathbf{I}_d - 2\hat{\mathbf{u}}\hat{\mathbf{u}}^\top, & \|\mathbf{f}\|_2 > \varepsilon_H, \\ \mathbf{I}_d, & \|\mathbf{f}\|_2 \leq \varepsilon_H, \end{cases} \qquad \mathcal{U}(\mathbf{F}) = H_{\varepsilon_H}(\mathbf{f}^{(k)}) \cdots H_{\varepsilon_H}(\mathbf{f}^{(1)}), \tag{1}$$

and set $\mathbf{U}_w^i = \mathcal{U}(\mathbf{F}_w^i) \in O(d)$, where $O(d)$ is the orthogonal group (we use $\varepsilon_H = 10^{-8}$). The frame maps w's local coordinates into a common ambient space, where $\mathbf{h}_w$ represents the *ambient vector* $\mathbf{g}_w := \mathbf{U}_w \mathbf{h}_w$. Transport into a's frame is $\mathbf{Q}_{a\leftarrow b}^i = (\mathbf{U}_a^i)^\top \mathbf{U}_b^i$, with $\mathbf{Q}_{b\leftarrow a}^i = (\mathbf{Q}_{a\leftarrow b}^i)^\top$. Depending only on the node frames, transport is defined for any pair, including a candidate with no historical edge to the query source. We omit the state index for a single fixed frame field.

Lemma 1 (Orthogonal transport and low-rank deviation)**.** *Every factor in Eq.* (1) *is orthogonal. If* k_w *rows of* $\mathbf{F}_w$ *have norm above* ε_H*, then* $\operatorname{rank}(\mathbf{U}_w - \mathbf{I}_d) \leq k_w \leq k$ *and* $\operatorname{rank}(\mathbf{Q}_{a\leftarrow b} - \mathbf{I}_d) \leq k_a + k_b \leq 2k$*. Applying either a frame or a transport costs* $O(kd)$ *arithmetic operations, with* $O(kd)$ *stored frame parameters per node.*

Rows are applied one at a time, so no dense $d \times d$ matrix is formed. Each frame has full rank; only its *deviation from the identity* is low-rank, so a small k limits which frames are reachable. The threshold makes $\mathcal{U}$ discontinuous at the branch boundary (Appendix B.1). Because the state is frozen, every candidate is compared before the wave is observed.

3.2 BLOCK B: SCORE AND RANK

How is a candidate scored from the frozen state? The elapsed time since an endpoint timestamp s is $\Delta_w(t; s) = t - s$, and $\Delta_w(t; \perp) = \Delta_\perp$, a large constant; write $\Delta_w^-(t) = \Delta_w(t; s_w^-(t))$. The *pair*

elapsed time $\Delta^{\text{pair}}_{u,c,r}(t)$ is t minus the time of the last event (u, c, t', r) with $t' < t$, and $\Delta_\perp$ if there is none. Relation embeddings $\mathbf{e}_r \in \mathbb{R}^{d_r}$ and type embeddings $\mathbf{e}^{\text{type}}_{\nu(w)} \in \mathbb{R}^{d_{\text{type}}}$ are learned tables, each removable at zero width, and both tables are sized from the dataset's full label vocabularies, so no entry is created during the stream. Beyond the event features, these two tables and the admissible domain $\mathcal{D}_q$ are the only places where the graph family enters; Blocks A–E are otherwise shared.

The score has two parts. A geometric part brings the candidate's memory into the query node's frame and measures how far apart the two memories are; a learned residual part reads that distance together with timing and type information. For $q = (u, ?, r, t)$ and candidate c,

$$D_q(c) = d^{-1}\|\mathbf{h}^-_u - \mathbf{Q}^-_{u\leftarrow c}\mathbf{h}^-_c\|_2^2, \qquad s_{\text{geo},q}(c) = -D_q(c)/\kappa, \tag{2a}$$

$$s_{\text{res},q}(c) = f_{\text{score}}(\boldsymbol{\psi}_q(c)), \qquad s_q(c) = s_{\text{geo},q}(c) + s_{\text{res},q}(c), \tag{2b}$$

with distance scale $\kappa = \max\{0.05, \exp(\text{clip}(\rho_\kappa, -3, 3))\}$ for learned ρ_κ. The shared scalar MLP $f_{\text{score}} : \mathbb{R}^{d_\psi} \to [-10, 10]$ (output clipped) receives, with $\ell(x) = \log(1 + x)$,

$$\begin{aligned}\boldsymbol{\psi}_q(c) = \text{concat}\big(&\ell(D_q(c)), \langle \mathbf{h}^-_u, \mathbf{Q}^-_{u\leftarrow c}\mathbf{h}^-_c\rangle, \ell(\|\mathbf{h}^-_u\|_2^2), \ell(\|\mathbf{h}^-_c\|_2^2), \ell(\Delta^-_u(t)), \ell(\Delta^-_c(t)),\\ &\ell(\Delta^{\text{pair}}_{u,c,r}(t)), \mathbf{e}_r, \mathbf{e}^{\text{type}}_{\nu(u)}, \mathbf{e}^{\text{type}}_{\nu(c)}\big), \qquad d_\psi = 7 + d_r + 2d_{\text{type}}.\end{aligned} \tag{3}$$

The distance is symmetric, but the residual reads source and destination features in order, so the score can be asymmetric; a head query with destination v ranks $c \mapsto s(c, v, r, t)$ over its legal sources. The pair-recency (EB) input $\ell(\Delta^{\text{pair}})$ is on by default and removed in TSNN-core ($d_\psi = 6 + d_r + 2d_{\text{type}}$); optional score components are described in Appendix E.1.

Labels are used only after every score in the wave is fixed. With single positive position j^+_q, set $p_q(j) = \exp(s_q(c_{qj}))/\sum_l \exp(s_q(c_{ql}))$ and $\ell_{\text{CE},q} = -\log p_q(j^+_q)$; training may draw $M_q = 1 + K^{\text{train}}_{\text{neg}}$ candidates from the direction-specific domain, while evaluation uses its own task-specified lists, filtering, and tie policy, and filtering by simultaneous truths never exposes those truths to the scorer or the state updates. The reference training uses $\ell_{\text{CE},q}$ alone, with $K^{\text{train}}_{\text{neg}} = 32$; some benchmark configurations add the auxiliary losses and optional inputs of Appendix E.1 (Table 11), and DGB, scored by AP, uses the binary objective. Under a matched candidate distribution and strict tie-breaking, this loss is consistent with reciprocal rank (Appendix D, Propositions 20 and 21).

3.3 Block C: Update frames

How does the observed event change the endpoint coordinate systems? Each endpoint of a new event turns its frame, based on its own memory, its partner's memory, and the event itself. After every query at timestamp t has been scored, commit the whole wave $\mathcal{W}_t$ in one transition; superscripts $i - 1$ and i denote $\Xi^{b(t)}$ and $\Xi^{b(t)+|\mathcal{W}_t|}$, the states before and after it. Event $j \in \mathcal{W}_t$ gives each endpoint $w \in \{u_j, v_j\}$ a partner $p_j(w)$ (w for a self-interaction) and a role indicator $\boldsymbol{\sigma}_{w,j}$, $(1, 0)$ for the source and $(0, 1)$ for the destination; W_i collects the wave's distinct endpoints, J_w the incidences of w (a self-interaction contributes both roles), and $\Delta_{w,i} = \Delta_w(t; s^{i-1}_w)$. A fixed sinusoidal encoding $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}^{d_{\text{time}}/2}$ (Eq. (14)) enters the contexts directly. All frame increments use the pre-wave state:

$$\begin{aligned}\dot{\mathbf{F}}_{w,j} = f_{\text{frame}}\big(\text{concat}(&\mathbf{h}^{i-1}_w, \mathbf{Q}^{i-1}_{w\leftarrow p_j(w)}\mathbf{h}^{i-1}_{p_j(w)},\\ &\phi(\Delta_{w,i}), \phi(\Delta_{p_j(w),i}), \mathbf{x}_j, \mathbf{e}_{r_j}, \boldsymbol{\sigma}_{w,j})\big), \qquad j \in J_w,\end{aligned} \tag{4}$$

where $f_{\text{frame}} : \mathbb{R}^{d_{\text{ctx}}} \to \mathbb{R}^{k\times d}$ is a shared MLP reshaped rowwise, $d_{\text{ctx}} = 2d + d_{\text{time}} + d_x + d_r + 2$ (plus two type embeddings on heterogeneous graphs). Set $\mathbf{F}^i_w = \text{rn}\left(\mathbf{F}^{i-1}_w + |J_w|^{-1}\sum_{j\in J_w}\dot{\mathbf{F}}_{w,j}\right)$ for $w \in W_i$, where rn maps each row $\mathbf{f} \mapsto \mathbf{f}/\max\{\|\mathbf{f}\|_2, 10^{-12}\}$, leaving other frame parameters unchanged. The partner's memory enters already transported into w's frame. The changed frame requires re-expressing the stored memory before any new information is added.

3.4 Block D: Carry and learn

How is a coordinate change separated from a memory update?

Proposition 2 (Coordinate-consistent carry-over)**.** *For $w \in W_i$, define*

$$\overline{\mathbf{h}}_{w,i} = (\mathbf{U}^i_w)^\top \mathbf{U}^{i-1}_w \mathbf{h}^{i-1}_w. \tag{5}$$

It is the unique vector satisfying $\mathbf{U}^i_w \overline{\mathbf{h}}_{w,i} = \mathbf{U}^{i-1}_w \mathbf{h}^{i-1}_w$.

Turning the frame changes how the memory is written, not what it says (Appendix B.1); only the GRU step and diffusion change the content, and both may depend on the frames (Section 5). All messages are formed before any state is written:

$$\begin{aligned}\mathbf{m}_{w,j} &= f_{\text{msg}}\Big(\text{concat}(\overline{\mathbf{h}}_{w,i}, \mathbf{Q}^i_{w\leftarrow p_j(w)}\overline{\mathbf{h}}_{p_j(w),i}, \phi(\Delta_{w,i}), \phi(\Delta_{p_j(w),i}), \mathbf{x}_j, \mathbf{e}_{r_j}, \boldsymbol{\sigma}_{w,j})\Big)\,,\\ \mathbf{h}^i_{w,0} &= \text{GRU}\left(\overline{\mathbf{h}}_{w,i},\, |J_w|^{-1}\textstyle\sum_{j\in J_w}\mathbf{m}_{w,j}\right),\end{aligned} \tag{6}$$

with shared $f_{\text{msg}} : \mathbb{R}^{d_{\text{ctx}}} \to \mathbb{R}^d$ and the convention $\text{GRU}(\text{previous state}, \text{input})$ (Appendix E). Each distinct endpoint is updated once, including a self-interaction, and a node outside W_i starts diffusion at $\mathbf{h}^i_{a,0} = \mathbf{h}^{i-1}_a$. The role indicator makes these updates orientation-aware; the role-free ablation (Table 19) sets $\boldsymbol{\sigma}_{w,j} = 0$, making the frame and recurrent updates orientation-blind given canonical buffer order (Proposition 17).

3.5 Block E: Diffuse and commit

Which historical interactions couple the updated memories, and which states are persisted? Each node's buffer keeps its most recent events; together the buffers form a small graph of recent interactions, over which updated memories are shared with their neighbours. The wave is first appended to its endpoints' buffers, in lexicographic (source, destination, relation) order, each buffer keeping its last $B_{\max}$ records. The retained history graph is the union of all buffers,

$$E_i = \textstyle\bigcup_{a\in\mathcal{V}} \mathcal{B}^i_a, \qquad V_i = \bigcup_{\{a,b\}\in E_i}\{a,b\}, \qquad G_i = (V_i, E_i),$$

a set of distinct unordered pairs: repeated events between the same nodes occupy buffer slots but give one edge, and a self-interaction occupies two slots of its node's buffer and gives a loop $\{a, a\}$ (Section 2.2). Direction, relation, and features are omitted.

On the edges of G_i, instantiate the sheaf of Section 2.2 by setting $\rho^i_{a,e} = \mathbf{U}^i_a$ at every incidence; identify all edge stalks with the ambient space and hold these frames fixed for all rounds; then $(\rho^i_{a,e})^\top \rho^i_{b,e} = \mathbf{Q}^i_{a\leftarrow b}$, a special case of a connection Laplacian (Barbero et al., 2022a), with sheaf Laplacian $\mathbf{L}_i$ from Section 2.2. The neighbor multiset $\mathcal{N}_{a,i}$ holds one adjacent vertex per incident edge occurrence (a loop lists a twice), giving true degree $\deg_{a,i} = |\mathcal{N}_{a,i}|$ and safe degree $\overline{\deg}_{a,i} = \max\{\deg_{a,i}, 1\}$, with $\mathbf{D}_i = \text{diag}_{a\in V_i}(\overline{\deg}_{a,i})$, $\boldsymbol{\Gamma}_i = \mathbf{D}_i \otimes \mathbf{I}_d$ for the Kronecker product $\otimes$, and the symmetric normalization $\mathbf{L}^{\text{sym}}_i = \boldsymbol{\Gamma}_i^{-1/2}\mathbf{L}_i\boldsymbol{\Gamma}_i^{-1/2}$ (Chung, 1997). Choose rounds $K \geq 0$, step $\eta > 0$ (reference: $K = 1$, $\eta = 0.05$), and positive diagonal feature gain $\mathbf{D}_\theta = \text{diag}(\text{softplus}(\theta_1), \ldots, \text{softplus}(\theta_d))$ shared across nodes; with $\mathbf{z}_{i,\ell} = \text{col}_{a\in V_i}\mathbf{h}^i_{a,\ell}$, $\mathbf{P}_i = \mathbf{I}_{|V_i|} \otimes \mathbf{D}_\theta$, and $\ell = 0, \ldots, K-1$, the synchronous update is

$$\mathbf{h}^i_{a,\ell+1} = \mathbf{h}^i_{a,\ell} - \eta\mathbf{D}_\theta\Big(\mathbf{h}^i_{a,\ell} - \sum_{b\in\mathcal{N}_{a,i}} \frac{\mathbf{Q}^i_{a\leftarrow b}\mathbf{h}^i_{b,\ell}}{\sqrt{\overline{\deg}_{a,i}\overline{\deg}_{b,i}}}\Big), \tag{7}$$

i.e. the stacked map $\mathbf{z}_{i,\ell+1} = \mathbf{z}_{i,\ell} - \eta\mathbf{P}_i\mathbf{L}^{\text{sym}}_i\mathbf{z}_{i,\ell}$ written per node. Every round reads round ℓ for *all* active nodes, and diffusion is skipped when $E_i = \varnothing$. Persist $\mathbf{h}^i_a = \mathbf{h}^i_{a,K}$ for *all* $a \in V_i \cup W_i$ (with $\mathbf{h}^i_{a,K} = \mathbf{h}^i_{a,0}$ off the graph), retaining the old memory elsewhere; set $s^i_w = t$ for $w \in W_i$ and insert the wave into the pair-recency table. Other last-interaction times are unchanged. The reference training minimises the candidate cross-entropy,

$$\mathcal{L} = \frac{1}{N_{\text{ev}}}\sum_q \ell_{\text{CE},q} \tag{8}$$

per chunk of two whole waves (other chunk lengths: Appendix F.3), with N_{ev} the chunk's event count. Gradients flow through the frame and recurrent updates within a chunk and are cut between chunks, and parameters change only after a chunk's last commit, so no score sees a parameter update from its own wave. The full per-wave cost and the $|\mathcal{V}|(k+1)d$ persistent state are derived in Appendix E.2.

4 What the Construction Guarantees

We prove that no test score can see the future and that diffusion is stable.

Lemma 3 (Causality of a frozen scoring wave)**.** *Assume Ξ^0 uses only permitted initial metadata, every committed transition uses its previous state and its observed wave, and for every t all scores use Ξ_t^- and parameters fixed before that wave. Each candidate score is then a function only of $\mathcal{E}_{<t}$, the query, the candidate, permitted metadata, and fixed model randomness, provided the parameters are fixed or learned from permitted earlier data.*

It covers frozen checkpoints fit on earlier splits, not multi-epoch training replay (Appendix B.2).

For diffusion, fix a graph $G = (V, E)$ with the node-factored orthogonal restrictions of Section 3.5 and a fixed gain, and suppress the event index. Reusing Block E's degree convention on the multiplicity-counting adjacency $\mathbf{A}_G$, set $\mathbf{D}_G^0 = \mathrm{diag}_a(\deg_a)$, $\mathbf{D}_G = \mathrm{diag}_a(\overline{\deg}_a)$, $\mathbf{\Gamma}_G = \mathbf{D}_G \otimes \mathbf{I}_d$, $\mathbf{L}_G = \mathbf{D}_G^0 - \mathbf{A}_G$, $\mathbf{L}_{\mathcal{F}}^{\mathrm{rw}} = \mathbf{\Gamma}_G^{-1}\mathbf{L}_{\mathcal{F}}$, $\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}} = \mathbf{\Gamma}_G^{-1/2}\mathbf{L}_{\mathcal{F}}\mathbf{\Gamma}_G^{-1/2}$, and $\mathbf{S} = \mathrm{blkdiag}_{a\in V}(\mathbf{U}_a)$.

Proposition 4 (Flat transport and lifted graph operator)**.** *For any walk $a_0, \dots, a_l$ at this fixed frame field, $\mathbf{Q}_{a_0 \leftarrow a_1} \cdots \mathbf{Q}_{a_{l-1} \leftarrow a_l} = \mathbf{U}_{a_0}^\top \mathbf{U}_{a_l}$, so every closed walk has identity transport. Moreover $\mathbf{L}_{\mathcal{F}} = \mathbf{S}^\top(\mathbf{L}_G \otimes \mathbf{I}_d)\mathbf{S}$, $\quad \mathrm{spec}(\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}}) \subseteq [0, 2]$, the same conjugation carries $\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}}$ to the normalized graph Laplacian, and $\mathbf{L}_{\mathcal{F}}^{\mathrm{rw}}$ has that same spectrum.*

The sheaf operator is a rotated graph Laplacian, with spectrum in $[0, 2]$ for any degrees or types (Appendix B.3). Set $\mathscr{E}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^\top \mathbf{L}_{\mathcal{F}}\mathbf{z}$, $\mathbf{\Omega} = \mathbf{D}_G \otimes \mathbf{D}_\theta^{-1}$, $\|\mathbf{z}\|_{\mathbf{\Omega}}^2 = \mathbf{z}^\top \mathbf{\Omega}\mathbf{z}$, and $\mathbf{L}_\Omega = \mathbf{\Omega}^{-1/2}\mathbf{L}_{\mathcal{F}}\mathbf{\Omega}^{-1/2}$.

Theorem 5 (Conditional descent and stability)**.** *Suppose $\mathbf{D}_\theta$ is positive definite and held fixed. The simultaneous full-active random-walk map $T\mathbf{z} = \mathbf{z} - \eta\mathbf{\Omega}^{-1}\mathbf{L}_{\mathcal{F}}\mathbf{z}$ is a gradient step for $\mathscr{E}$ in the $\mathbf{\Omega}$ inner product and $\lambda_{\max}(\mathbf{L}_\Omega) \le 2\lambda_{\max}(\mathbf{D}_\theta)$. For $0 < \eta \le 1/\lambda_{\max}(\mathbf{D}_\theta)$ the energy does not increase (quantitatively, Eq.* (10)*) and $\|T\mathbf{x} - T\mathbf{y}\|_{\mathbf{\Omega}} \le \|\mathbf{x} - \mathbf{y}\|_{\mathbf{\Omega}}$; for $0 < \eta \le 1/(2\lambda_{\max}(\mathbf{D}_\theta))$ the map is firmly non-expansive and its repeated application converges to the $\mathbf{\Omega}$-orthogonal projection onto $\ker \mathbf{L}_{\mathcal{F}}$.*

A small step never raises the energy or separates two states, for all K rounds of Eq. (7) (Remark 1); at $\eta = 0.05$ this means $\max_j \mathrm{softplus}(\theta_j) \le 20$, checked every epoch (Appendix B.4). Since the symmetric metric is graph-free, this extends to changing graphs.

Corollary 6 (Diffusion across changing graphs)**.** *Fix $\mathbf{D}_\theta$ positive definite and $0 < \eta \le 1/\lambda_{\max}(\mathbf{D}_\theta)$, and let $\mathcal{A}_i$ be wave i's K rounds of Eq.* (7)*, extended by the identity to nodes off G_i. For every sequence of retained graphs and frame fields, $\|\mathcal{A}_T \cdots \mathcal{A}_1\|_{\mathbf{\Omega}_{\mathrm{sym}} \to \mathbf{\Omega}_{\mathrm{sym}}} \le 1$ with $\mathbf{\Omega}_{\mathrm{sym}} = \mathbf{I}_{|\mathcal{V}|} \otimes \mathbf{D}_\theta^{-1}$, and replacing G_i by another graph $\widehat{G}_i$ changes $\mathcal{A}_i$ by at most $K\eta\lambda_{\max}(\mathbf{D}_\theta)\|\mathbf{L}_i^{\mathrm{sym}} - \widehat{\mathbf{L}}_i^{\mathrm{sym}}\|_2$ in the same norm.*

Diffusion never amplifies differences in this norm as the graph evolves (Appendix B.4); the bound is stated for fixed-gain diffusion, since carry and the recurrent update can rescale the weighted norm; with the halved step, on a fixed graph it converges to a degree-scaled consensus (Proposition 15).

Frames enter diffusion only through the gain: writing $\mathbf{g}_a = \mathbf{U}_a\mathbf{h}_a$, one round reads $\mathbf{g}_a^+ = \mathbf{g}_a - \eta(\mathbf{U}_a\mathbf{D}_\theta\mathbf{U}_a^\top)(\mathbf{g}_a - \sum_{b\in\mathcal{N}_a} \mathbf{g}_b/\sqrt{\overline{\deg}_a\overline{\deg}_b})$. An isotropic gain $c\mathbf{I}_d$ gives scalar graph diffusion; an anisotropic gain can retain frame dependence (Proposition 13).

Frames are Lipschitz in raw rows of norm at least $\tau_H > \varepsilon_H$ (Lemma 7), and a change $\delta_U = \max_a \|\mathbf{U}_a - \widehat{\mathbf{U}}_a\|_2$ moves the map T by at most $2\eta\lambda_{\max}(\mathbf{D}_\theta)\delta_U$ in $\mathbf{\Omega}$-operator norm (Theorem 8).

5 What a Frame Channel Can and Cannot Distinguish

Section 3 lets frames move; this section asks when that movement can change a prediction. Independent orthogonal maps $\mathbf{R}_v \in O(d)$ act on $(\mathbf{h}_v, \mathbf{U}_v)$ as $(\mathbf{R}_v^\top \mathbf{h}_v, \mathbf{U}_v\mathbf{R}_v)$. Two configurations share an orbit exactly when their ambient vectors agree (Proposition 9); hence every invariant readout factors through those vectors. In particular, $\|\mathbf{h}_u - \mathbf{Q}_{u\leftarrow v}\mathbf{h}_v\|_2 = \|\mathbf{g}_u - \mathbf{g}_v\|_2$. The score thus sees the ambient vectors, never the frames (optional inputs: Appendix E.1): a pure frame change with exact carry leaves every score unchanged with zero gradient (Proposition 16), and under a common initialisation cold candidates without history or type tie (Proposition 18).

Frame state can still act through coordinate-dependent dynamics. In ambient coordinates, with content map Φ_ξ (message MLP, then GRU), the wave keeps the raw frame state $\mathbf{F}$, updates $\mathbf{g}_w^+ =$

$\mathbf{U}_w^+ \Phi_\xi\big((\mathbf{U}_w^+)^\top \mathbf{g}_w^-, ((\mathbf{U}_w^+)^\top \mathbf{g}_p^-)_p\big)$, and diffuses with gain $\mathbf{U}_a \mathbf{D}_\theta \mathbf{U}_a^\top$ (Proposition 11). Equivariant content maps with scalar gain make the ambient trajectory frame-independent (Proposition 12); Φ_ξ is not equivariant, a finite-parameter example gives nonzero frame dependence (Proposition 14), and Proposition 13 bounds the anisotropic diffusion term. Block C of Table 3 and Panels A–B of Table 19 test this dependence, and Appendix F.5.2 adds a synthetic drift study. Two small constructions complete the picture. A homogeneous zero-message control never leaves its common initial ambient vector (Proposition 10), and frames separate streams that a restricted comparator cannot. Let node c first contact a in one stream and b in the other, and let a and then b contact v in both. A receiver-only shared-coordinate memory whose gate is closed at rest, with no ungated side state, stays at rest on both streams, so it assigns v the same state and cannot tell the streams apart. With evolving frames, the first contact turns the frame of a or of b, the transported messages to v differ, and for suitable parameters v ends in different ambient states (Theorem 19; Appendices B.6, C, and F.6).

6 Experiments and Evidence Scope

We evaluate one TSNN architecture on 34 settings: thirteen TGB/TGB 2.0 and eight TGB-Seq datasets scored by mean reciprocal rank (MRR), and thirteen DGB datasets scored by average precision (AP). Comparators are reported results, not matched reruns: TGB leaderboard entries, EdgeReMIND (Pollard, 2026), CRAFT/CRAFT-R (Yi et al., 2025a), and on TGB-Seq also SDG, SGNN-HN, and JODIE. Per dataset, one configuration is selected by mean validation score, frozen, and evaluated once per seed (Appendix F.3). Counting every listed result, TSNN is best on 10 of 13 TGB datasets (Table 1): all four knowledge graphs, three of four heterogeneous graphs, and three of five `tgbl` datasets, with the largest margins on thgl-forum ($+0.1075$) and thgl-github ($+0.0494$). The other three are led by STHN (thgl-software), TPNet and CRAFT-R (tgbl-coin), and CRAFT-R (tgbl-flight).

Table 1: Temporal link prediction, test MRR. Best, Second, Third: selected TGB leaderboard and author-reported EdgeReMIND results (Pollard, 2026); Additional: CRAFT (review/comment) or CRAFT-R (coin/flight) (Yi et al., 2025a), three runs; – where none is reported (sources: Appendix F.3). Val.: mean validation MRR; Test: mean±s.d. over seeds 42–46. Bold: best excluding the Additional column; †: its result is higher.

	TSNN		Leaderboard + EdgeReMIND, test MRR (method)			Additional
Dataset	Val.	Test	Best	Second	Third	(CRAFT/-R)
tkgl-smallpedia	0.6605	**0.6242**$_{\pm.0025}$	0.6136 (EdgeReMIND)	0.612 (CEN)	0.605 (RecB)	–
tkgl-polecat	0.2498	**0.2303**$_{\pm.0007}$	0.228 (TLogic)	0.198 (RecB)	0.184 (CEN)	–
tkgl-icews	0.2921	**0.2871**$_{\pm.0003}$	0.2694 (EdgeReMIND)	0.211 (RecB)	0.187 (CEN)	–
tkgl-wikidata	0.7236	**0.6502**$_{\pm.0022}$	0.6397 (EdgeReMIND)	0.535 (EdgeBank$_{\text{tw}}$)	0.535 (EdgeBank$_{\infty}$)	–
thgl-software	0.5659	0.6449$_{\pm.0018}$	**0.731** (STHN)	0.5013 (EdgeReMIND)	0.449 (EdgeBank$_{\infty}$)	–
thgl-forum	0.8394	**0.8381**$_{\pm.0013}$	0.7306 (EdgeReMIND)	0.729 (TGN, edge type)	0.649 (TGN)	–
thgl-github	0.7507	**0.8336**$_{\pm.0001}$	0.7842 (EdgeReMIND)	0.413 (EdgeBank$_{\infty}$)	0.374 (EdgeBank$_{\text{tw}}$)	–
thgl-myket	0.9227	**0.9219**$_{\pm.0009}$	0.9056 (EdgeReMIND)	0.456 (EdgeBank$_{\infty}$)	0.245 (EdgeBank$_{\text{tw}}$)	–
tgbl-wiki	0.8471	**0.8290**$_{\pm.0001}$	0.827 (TPNet)	0.821 (LocalGlobal)	0.810 (HyperEvent)	–
tgbl-review	0.4859	**0.5400**$_{\pm.0006}$	0.521 (GraphMixer)	0.405 (CTAN)	0.377 (TNCN)	0.4177
tgbl-coin	0.7992	0.8294$_{\pm.0003}$	**0.832** (TPNet)	0.774 (LocalRec. LocalPop.)	0.773 (HyperEvent)	0.8847
tgbl-comment	0.9334	**0.9276**$_{\pm.0001}$	0.825 (TPNet)	0.760 (HyperEvent)	0.723 (GlobalRecency)	0.9172
tgbl-flight	0.9099	**0.9040**$_{\pm.0013}$†	0.884 (TPNet)	0.878 (LocalRec. LocalPop.)	0.877 (HyperEvent)	0.9139

Table 2: TGB-Seq test MRR, fixed splits, 100 negatives. TSNN: mean±s.d. over seeds 42–46; best published: highest result among the cited reports (Yi et al., 2025b;a; Duc & Ta, 2026). The SGNN-HN result is from Yi et al. (2025a); its architecture is due to Pan et al. (2020).

Method	ML-20M	Taobao	Yelp	GoogleLocal	Flickr	YouTube	Patent	WikiLink
TSNN	**0.4438**$_{\pm.0003}$	**0.7071**$_{\pm.0002}$	**0.7385**$_{\pm.0001}$	**0.6638**$_{\pm.0015}$	**0.6943**$_{\pm.0003}$	**0.6541**$_{\pm.0006}$	**0.7138**$_{\pm.0003}$	**0.7664**$_{\pm.0017}$
Best published method	0.3663 SDG	0.7068 CRAFT	0.7269 CRAFT	0.6288 SGNN-HN	0.6234 CRAFT	0.6054 SDG	0.2460 JODIE	0.7548 CRAFT

On TGB-Seq, TSNN exceeds the best published result on all eight datasets (Table 2), 18 of 21 overall, by margins from $+0.0003$ on Taobao to $+0.4678$ on Patent, where strict-past co-citation history alone reaches 0.62 MRR (Appendix F.3). On DGB, trained with the benchmark's binary objective and optional EdgeBank and common-neighbour inputs, it has the highest transductive AP on 11 of 13 datasets and the best mean rank, 1.23 against 1.96 for TPNet (Table 13). Across all 34 settings, one architecture covers every graph family and is best on 29 (Table 10).

Table 3: Ablations of TSNN, block by block: test MRR (higher is better), three-seed mean±s.d. over seeds 42–44. MRR and its standard deviation are reported as fractions, not percentages. Paired intervals: Appendix F.5.2; additional controls: Table 19.

Intervention	tkgl-icews	tgbl-wiki	tkgl-smallpedia	YouTube	tkgl-polecat
Full TSNN, evolving frames (reference)	0.2748±0.00255	0.8223±0.00136	0.6221±0.00053	0.5207±0.00228	0.1849±0.00131
Block A: initialisation					
Common initial memory and frame	0.2678±0.0004	0.8208±0.0001	0.6137±0.0008	0.5143±0.0017	0.1832±0.0004
Block B: score					
Distance-only score (no residual MLP)	0.0105±0.0035	0.0686±0.0014	0.0178±0.002	0.1483±0.0032	0.1136±0.0017
Ambient-vector MLP decoder	0.2610±0.0018	0.8145±0.0011	0.6002±0.0028	0.5010±0.0021	0.0517±0.0023
Block C: frame update and matched frame controls					
Frame generator frozen at initialisation	0.2750±0.0021	0.8233±0.003	0.6130±0.0033	0.5230±0.0023	0.1841±0.002
Frame-free, same scorer ($\mathbf{U}=\mathbf{I}$)	0.2503±0.0012	0.8094±0.0008	0.6089±0.0008	0.5096±0.002	0.1758±0.0019
Frames held fixed for the whole run	0.2698±0.001	0.8199±0.0013	0.6103±0.0003	0.5108±0.0012	0.1835±0.0001
Shared, one global frame	0.2698±0.0009	0.8201±0.001	0.6095±0.0013	0.5099±0.0002	0.1831±0.0008
Side state, elementwise GRU	0.0135±0.0018	0.2427±0.0015*	0.4610±0.0022	0.3806±0.0020	0.1439±0.0017
FiLM, state modulates messages	0.2228±0.0003	0.8201±0.0002	0.6097±0.0035	0.5045±0.0023	0.1829±0.0014
Block D: carry-over and recurrent update					
Without carry-over	0.2650±0.0015	0.8193±0.0013	0.6139±0.0022	0.5249±0.0017	0.1646±0.0027
Block E: diffusion and diagonal gain					
Without diffusion ($K=0$)	0.2625±0.0007	0.8209±0.0007	0.6176±0.0025	0.3169±0.0017	0.1744±0.0006
Isotropic gain, $\mathbf{D}_\theta=\mathbf{I}_d$	0.2697±0.001	0.8214±0.0003	0.6174±0.0024	0.5179±0.0017	0.1772±0.0034

The benchmarks evaluate the complete predictor; Table 3 removes one component at a time to measure its contribution. Its three-seed configuration is separate from the benchmark runs (YouTube reference 0.5207 vs. 0.6541), and its reference was fixed after an earlier cohort was evaluated on test (Appendix F.5), so all arms are compared within this configuration. Component removals lower MRR on at least four of the five datasets, with the largest loss from removing the residual scorer: distance-only scoring, which removes the residual MLP together with its recency and type inputs, collapses MRR on every dataset (Wiki $0.8223 \rightarrow 0.0686$), and an ambient-vector decoder costs up to 0.1332 (Polecat). Diffusion removal has the next largest effect: removing it costs 0.2038 on YouTube and 0.0105–0.0123 on Polecat and icews. Carry-over matters most on the knowledge graphs, where removing it costs 0.0203 on Polecat and 0.0098 on icews, with YouTube the exception (0.5249). A diagonal rather than isotropic gain and seeded per-node initial states improve all five datasets, by up to 0.0077 and 0.0084.

Evolving frames improve mean MRR on every dataset: the full model exceeds every arm that removes the frames, stops their evolution, or replaces them with coordinate-free state. Removing frames costs 0.0245 on icews and 0.0091–0.0132 elsewhere, significant on every dataset (Welch t-test, $p < 0.005$; Table 20). Fixed per-node frames cost up to 0.0118 and one shared frame up to 0.0126, yet both still beat the frame-free model on all five datasets. The elementwise-GRU side state underperforms both the full and frame-free models on all five datasets, most sharply on icews (0.0135) and Wiki (0.2427); FiLM costs up to 0.0520 (icews). Evolution matters more than learning its generator: a generator frozen at initialisation, whose frames still evolve, matches or exceeds the full model on three datasets and stays above the frame-free model on all five. The additional controls of Table 19 all stay below the full model except a learned scalar gain on Polecat (0.1854 vs. 0.1849).

7 CONCLUSIONS

In TSNN, events write into per-node frame state, exact carry-over preserves what a memory represents when its frame turns, and recurrence and sheaf diffusion update the content. For the invariant base scorer, frames act only through coordinate-dependent updates or anisotropic diffusion. One architecture covers all four graph families and leads every listed comparator on 18 of 21 TGB and TGB-Seq datasets and on DGB mean rank. The residual scorer, diffusion, and carry-over have the largest effects among component removals; evolving frames beat the frame-free model significantly on all five ablation datasets and exceed fixed, shared, side-state, and FiLM controls in mean MRR. Comparisons use published results, and extending the stability guarantee from diffusion to the recurrent update is a natural next step.

AI USE STATEMENT

We used large language models (LLMs) in four ways. *Writing:* to aid and polish the writing, including drafting revisions of parts of the text such as the experimental discussion and the conclusion. *Research execution:* to write and debug the implementation and evaluation code and to orchestrate experiment runs. *Synthetic data:* to generate the synthetic data of the drift study and the constructed mechanism checks (Appendices F.5.2 and F.6). *Checking:* to verify arithmetic, cross-references, citations, and proof steps, and to revise the wording of some proofs. The authors reviewed and verified all AI-assisted text, code, data, and proofs, and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Code: `https://anonymous.4open.science/r/ICLR-TSNN-CODE-D479/README.md`. The repository provides the TSNN implementation, training and evaluation code, dependency versions, dataset-loading instructions, and runnable configurations. Its README documents setup and execution, and unit tests cover core components, temporal ordering, and transport invariants. Appendix B gives proofs and supporting derivations, Appendix E specifies the algorithms, and Appendix F describes datasets and evaluation protocols.

REFERENCES

Jacob Bamberger, Federico Barbero, Xiaowen Dong, and Michael M. Bronstein. Bundle neural network for message diffusion on graphs. In *International Conference on Learning Representations*, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ac879d1865475a7abc8dfc7a9c15c27-Abstract-Conference.html`.

Federico Barbero, Cristian Bodnar, Haitz Sáez de Ocáriz Borde, Michael Bronstein, Petar Veličković, and Pietro Liò. Sheaf neural networks with connection Laplacians. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pp. 28–36. PMLR, 2022a.

Federico Barbero, Cristian Bodnar, Haitz Sáez de Ocáriz Borde, and Pietro Liò. Sheaf attention networks. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022b.

Claudio Battiloro, Zhiyang Wang, Hans Riess, Paolo Di Lorenzo, and Alejandro Ribeiro. Tangent bundle convolutional learning: From manifolds to cellular sheaves and back. *IEEE Transactions on Signal Processing*, 2024.

Silvia Beddar-Wiesing and Alice Moallemy-Oureh. Weisfeiler–Lehman meets events: An expressivity analysis for continuous-time dynamic graph neural networks. *arXiv preprint arXiv:2508.18052*, 2025.

Torben Berndt, Benjamin Walker, Tiexin Qin, Jan Stühmer, and Andrey Kormilitzin. Permutation equivariant neural controlled differential equations for dynamic graph representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.20324.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, and Michael M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1346.

Alessio Borgi, Fabrizio Silvestri, and Pietro Liò. Polynomial neural sheaf diffusion: A spectral filtering approach on cellular sheaves. *arXiv preprint arXiv:2512.00242*, 2025. doi: 10.48550/arXiv.2512.00242.

Luke Braithwaite, Alessio Borgi, Gabriele Onorato, Kristjan Tarantelli, Francesco Restuccia, Fabrizio Silvestri, and Pietro Liò. Heterogeneous sheaf neural networks. *arXiv preprint arXiv:2409.08036*, 2024.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Dmitri Burago, Sergei Ivanov, Yaroslav Kurylev, and Jinpeng Lu. Approximations of the connection Laplacian spectra. *Mathematische Zeitschrift*, 301(3):3185–3206, 2022.

Yoonhyuk Choi, Jiho Choi, Taewook Ko, JongWook Kim, and Chong-Kwon Kim. Sheaf graph neural networks via PAC-Bayes spectral optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20570–20578, 2026.

Fan R. K. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997. doi: 10.1090/cbms/092.

Hsing-Huan Chung, Shravan Chaudhari, Xing Han, Yoav Wald, Suchi Saria, and Joydeep Ghosh. Between linear and sinusoidal: Rethinking the time encoder in dynamic graph learning. *Transactions on Machine Learning Research*, 2025.

Taco S. Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2019.

Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. Do we really need complicated model architectures for temporal networks? In *International Conference on Learning Representations (ICLR)*, 2023.

Filip Cornell, Oleg Smirnov, Gabriela Zarzar Gandler, and Lele Cao. On the power of heuristics in temporal graphs. In *Proceedings on "I Can't Believe It's Not Better: Challenges in Applied Deep Learning" at ICLR 2025 Workshops*, volume 296 of *Proceedings of Machine Learning Research*, pp. 37–46, 2025.

Pim de Haan, Maurice Weiler, Taco Cohen, and Max Welling. Gauge equivariant mesh CNNs: Anisotropic convolutions on geometric graphs. In *International Conference on Learning Representations (ICLR)*, 2021.

Zifeng Ding, Yifeng Li, Yuan He, Antonio Norelli, Jingcheng Wu, Volker Tresp, Michael Bronstein, and Yunpu Ma. DyGMamba: Efficiently modeling long-term temporal dependency on continuous-time dynamic graphs with state space models. *Transactions on Machine Learning Research*, 2025.

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. SE(3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning (ICML)*, 2022.

Weitao Du, Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla P. Gomes, and Zhi-Ming Ma. A new perspective on building efficient and expressive 3D equivariant graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nguyen Minh Duc and Viet Cuong Ta. Sequence diffusion model for temporal link prediction in continuous-time dynamic graph. *arXiv preprint arXiv:2601.23233*, 2026.

ZhengZhao Feng, Rui Wang, TianXing Wang, Mingli Song, Sai Wu, and Shuibing He. A comprehensive survey of dynamic graph neural networks: Models, frameworks, benchmarks, experiments and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 38(1):26–46, 2026. doi: 10.1109/TKDE.2025.3621291.

Jian Gao, Jianshe Wu, and JingYi Ding. HyperEvent: A strong baseline for dynamic link prediction via relative structural encoding. *arXiv preprint arXiv:2507.11836*, 2025. doi: 10.48550/arXiv.2507.11836. URL https://arxiv.org/abs/2507.11836.

Julia Gastinger, Shenyang Huang, Mikhail Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt, Reihaneh Rabbany, and Guillaume Rabusseau. TGB 2.0: A benchmark for learning on temporal knowledge graphs and heterogeneous graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Julia Gastinger, Christian Meilicke, Federico Errica, Timo Sztyler, Anett Schülke, and Heiner Stuckenschmidt. History repeats itself: A baseline for temporal knowledge graph forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4016–4024, 2024b. doi: 10.24963/ijcai.2024/444. arXiv:2404.16726.

Julia Gastinger, Christian Meilicke, and Heiner Stuckenschmidt. CountTRuCoLa: Rule learning for interpretable temporal knowledge graph forecasting. *arXiv preprint arXiv:2509.09474*, 2025.

Alessio Gravina, Giulio Lovisotto, Claudio Gallicchio, Davide Bacciu, and Claas Grohnfeldt. Long range propagation on continuous-time dynamic graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16206–16225. PMLR, 2024. URL https://proceedings.mlr.press/v235/gravina24a.html.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.12537.

Jakob Hansen and Thomas Gebhart. Sheaf neural networks. In *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*, 2020.

Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

Lingshen He, Yiming Dong, Yisen Wang, Dacheng Tao, and Zhouchen Lin. Gauge equivariant transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Franziska Heeg, Jonas Sauer, Petra Mutzel, and Ingo Scholtes. Weisfeiler and Leman follow the arrow of time: Expressive power of message passing in temporal event graphs. *arXiv preprint arXiv:2505.24438*, 2025.

Chuan-Shen Hu. A sheaf-theoretic and topological perspective on complex network modeling and attention mechanisms in graph neural models. *arXiv preprint arXiv:2601.21207*, 2026.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pengfei Jiao, Hongjiang Chen, Xuan Guo, Zhidong Zhao, Dongxiao He, and Di Jin. A survey on temporal interaction graph representation learning: Progress, challenges, and opportunities. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, pp. 10499–10507, 2025. doi: 10.24963/ijcai.2025/1166.

Jinwoo Kim, Olga Zaghen, Ayhan Suleymanzade, Youngmin Ryou, and Seunghoon Hong. Revisiting random walks for learning on graphs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.01214.

Miltiadis Kofinas, Naveen Shankar Nagaraja, and Efstratios Gavves. Roto-translated local coordinate frames for interacting dynamical systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019.

Veronica Lachi, Francesco Ferrini, Antonio Longa, Bruno Lepri, Andrea Passerini, and Manfred Jaeger. Bridging theory and practice in link representation with graph neural networks. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-4090.

Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. Tensor decompositions for temporal knowledge base completion. In *International Conference on Learning Representations (ICLR)*, 2020.

Moritz Lampert, Christopher Blöcker, and Ingo Scholtes. From link prediction to forecasting: Addressing challenges in batch-based temporal graph learning. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=iZPAykLE3l.

Ce Li, Rongpei Hong, Xovee Xu, Goce Trajcevski, and Fan Zhou. Simplifying temporal heterogeneous network for continuous-time link prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1288–1297, 2023.

Dongyuan Li, Shiyin Tan, Ying Zhang, Ming Jin, Shirui Pan, Manabu Okumura, and Renhe Jiang. DyG-Mamba: Continuous state space modeling on dynamic graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2408.06966.

Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, Xing Xie, and Qi Zhang. HousE: Knowledge graph embedding with Householder parameterization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13209–13224, 2022a. URL https://proceedings.mlr.press/v162/li22ab.html.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 408–417, 2021. doi: 10.1145/3404835.3462963.

Zixuan Li, Saiping Guan, Xiaolong Jin, Weihua Peng, Yajuan Lyu, Yong Zhu, Long Bai, Wei Li, Jiafeng Guo, and Xueqi Cheng. Complex evolutional pattern learning for temporal knowledge graph reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 290–296, 2022b. doi: 10.18653/v1/2022.acl-short.32.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. TLogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4120–4127, 2022. doi: 10.1609/aaai.v36i4.20330.

Xiaodong Lu, Leilei Sun, Tongyu Zhu, and Weifeng Lv. Improving temporal link prediction via temporal walk matrix projection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuhong Luo and Pan Li. Neighborhood-aware scalable temporal network representation learning. In *Learning on Graphs Conference (LoG)*, 2022.

Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using Householder reflections. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2401–2409. PMLR, 2017. URL https://proceedings.mlr.press/v70/mhammedi17a.html.

Abeer Mostafa, Raneen Younis, and Zahra Ahmadi. Dynamic sheaf diffusion networks with adaptive local structure for heterogeneous spatio-temporal graph learning. *arXiv preprint arXiv:2604.11275*, 2026.

Sajad Movahedi, Felix Sarnthein, Nicola Muca Cirone, and Antonio Orvieto. Fixed-point RNNs: Interpolating from diagonal to dense. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1495.

Bao Nguyen, Lorenzo Sani, Xinchi Qiu, Pietro Liò, and Nicholas D. Lane. Sheaf HyperNetworks for personalized federated learning. *arXiv preprint arXiv:2405.20882*, 2024.

Zhiqiang Pan, Fei Cai, Wanyu Chen, Honghui Chen, and Maarten de Rijke. Star graph neural networks for session-based recommendation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1195–1204, 2020. doi: 10.1145/3340531.3412014.

Jung Yeon Park, Lawson L. S. Wong, and Robin Walters. Modeling dynamics over meshes with gauge equivariant nonlinear message passing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jie Peng, Zhewei Wei, and Yuhang Ye. TIDFormer: Exploiting temporal and interactive dynamics makes a great dynamic graph transformer. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. arXiv:2506.00431.

Katarina Petrović, Shenyang Huang, Farimah Poursafaei, and Petar Veličković. Temporal graph rewiring with expander graphs. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling (GRaM)*, 2024.

Bryant Pollard. EdgeReMIND: A scalable, top-ranked memorization baseline for temporal multi-relational link prediction. *arXiv preprint arXiv:2609.17916*, 2026.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

André Ribeiro, Ana Luiza Tenório, Juan Belieni, Amauri H. Souza, and Diego Mesquita. Cooperative sheaf neural networks. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.00647.

Raphaël Romero, Maarten Buyl, Tijl De Bie, and Jefrey Lijffijt. Exploring the performance of continuous-time dynamic link prediction algorithms. *Applied Sciences*, 14(8):3516, 2024.

Emanuele Rossi, Benjamin Paul Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020. URL https://grlplus.github.io/papers/58.pdf.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023. doi: 10.48550/arXiv.2303.10993. URL https://arxiv.org/abs/2303.10993.

Amit Singer and Hau-Tieng Wu. Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012.

Amauri H. Souza, Diego Mesquita, Samuel Kaski, and Vikas Garg. Provably expressive temporal graph networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Li Sun, Zhongbao Zhang, Jiawei Zhang, Feiyang Wang, Hao Peng, Sen Su, and Philip S. Yu. Hyperbolic variational graph neural network for modeling dynamic graphs. In *AAAI Conference on Artificial Intelligence*, 2021.

Li Sun, Zhenhao Huang, Silei Chen, Lanxu Yang, Junda Ye, Sen Su, and Philip Yu. Multi-domain riemannian graph gluing for building graph foundation models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.00618.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations (ICLR)*, 2019.

Aleksandar Terzic, Nicolas Menet, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. Structured sparse transition matrices to enable state tracking in state-space models. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-2784.

Yuxing Tian, Yiyan Qi, and Fan Guo. FreeDyG: Frequency enhanced continuous-time dynamic graph model for link prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

Katherine Tieu, Dongqi Fu, Zihao Li, Ross Maciejewski, and Jingrui He. Learnable spatial-temporal positional encoding for link prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59570–59597, 2025. URL https://proceedings.mlr.press/v267/tieu25a.html. arXiv:2506.08309.

Benedict Aaron Tjandra, Federico Barbero, and Michael Bronstein. Enhancing the expressivity of temporal graph networks through source-target identification. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations (NeurReps)*, 2024. arXiv:2411.03596.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. DyRep: Learning representations over dynamic graphs. In *International Conference on Learning Representations (ICLR)*, 2019.

Siddharth Viswanath, Panayiotis Ketonis, Chen Liu, Michael Perlmutter, Dhananjay Bhaskar, and Smita Krishnaswamy. BrainDyn: A sheaf neural ODE for generative brain dynamics. *arXiv preprint arXiv:2605.19324*, 2026.

Benjamin Walker, Lingyi Yang, Nicola Muca Cirone, Cristopher Salvi, and Terry Lyons. Structured linear CDEs: Maximally expressive and parallel-in-time sequence models. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-5669.

Lu Wang, Xiaofu Chang, Shuang Li, Yunfei Chu, Hui Li, Wei Zhang, Xiaofeng He, Le Song, Jingren Zhou, and Hongxia Yang. TCL: Transformer-based dynamic graph modelling via contrastive learning. *arXiv preprint arXiv:2105.07944*, 2021a.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations (ICLR)*, 2021b.

Zihui Wang, Peizhen Yang, Xiaoliang Fan, Xu Yan, Zonghan Wu, Shirui Pan, Longbiao Chen, Yu Zang, Cheng Wang, and Rongshan Yu. ConTIG: Continuous representation learning on temporal interaction graphs. *Neural Networks*, 172:106151, 2024.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Lu Yi, Runlin Lei, Fengran Mo, Yanping Zheng, Zhewei Wei, and Yuhang Ye. Future link prediction without memory or aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. doi: 10.52202/085713-0084. arXiv:2505.19408.

Lu Yi, Jie Peng, Yanping Zheng, Fengran Mo, Zhewei Wei, Yuhang Ye, Zixuan Yue, and Zengfeng Huang. TGB-Seq benchmark: Challenging temporal GNNs with complex sequential dynamics. In *International Conference on Learning Representations (ICLR)*, 2025b. arXiv:2502.02975.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhongyi Yu, Jianqiu Wu, Zhenghao Wu, Shuhan Zhong, Weifeng Su, Chul-Ho Lee, and Weipeng Zhuo. TAMI: Taming heterogeneity in temporal interactions for temporal graph link prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.23577.

Olga Zaghen. Nonlinear sheaf diffusion in graph neural networks. *arXiv preprint arXiv:2403.00337*, 2024.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Xiaohui Zhang, Yanbo Wang, Xiyuan Wang, and Muhan Zhang. Efficient neural common neighbor for temporal graph link prediction. In *Proceedings of the Fourth Learning on Graphs Conference*, 2025a. URL `https://openreview.net/forum?id=ffsDVGiKRm`.

Xin Zhang, Fei Cai, Jianming Zheng, Zhiqiang Pan, Wanyu Chen, Honghui Chen, and Chonghao Chen. Triangle matters! TopDyG: Topology-aware transformer for link prediction on dynamic graphs. In *Proceedings of the ACM Web Conference (WWW)*, pp. 3607–3617, 2025b. doi: 10.1145/3696410.3714564.

Tao Zou, Yuhao Mao, Junchen Ye, and Bowen Du. Repeat-aware neighbor sampling for dynamic graph learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

A Extended Related Work and Supplementary Notation

A.1 Extended related work

Geometric learning studies symmetries (Bronstein et al., 2021) and local-frame invariance (Cohen et al., 2019), with architectures including LoCS, mesh models, ClofNet, LEFTNet, and Hermes (Kofinas et al., 2021; de Haan et al., 2021; He et al., 2021; Du et al., 2022; 2023; Park et al., 2023); transport also appears in vector diffusion and connection Laplacians (Singer & Wu, 2012; Burago et al., 2022). Adaptive orthogonal frames support graph-domain integration (Sun et al., 2026).

Sheaf Laplacians use incidence restrictions (Hansen & Ghrist, 2019; Hansen & Gebhart, 2020), generalising the graph-Laplacian setting (Chung, 1997). Learned sheaf diffusion and attention address heterophily and over-smoothing (Bodnar et al., 2022; Barbero et al., 2022b); Rusch et al. (2023) surveys over-smoothing more broadly. Further work develops connection and bundle diffusion, nonlinear and spectral operators, heterogeneous and cooperative models, federated hypernetworks, and sheaf-theoretic perspectives (Barbero et al., 2022a; Bamberger et al., 2025; Battiloro et al., 2024; Ribeiro et al., 2026; Borgi et al., 2025; Choi et al., 2026; Zaghen, 2024; Braithwaite et al., 2024; Nguyen et al., 2024; Hu, 2026). Dynamic sheaf diffusion uses signal-conditioned restrictions for forecasting (Mostafa et al., 2026); BrainDyn combines LSTM encoding with sheaf neural-ODE evolution (Viswanath et al., 2026). Other methods study continuous-time propagation, controlled differential equations, and hyperbolic representations (Gravina et al., 2024; Berndt et al., 2025; Sun et al., 2021); TGR connects temporally distant nodes through expanders (Petrović et al., 2024).

Temporal models maintain recurrent memory (Kumar et al., 2019; Trivedi et al., 2019; Rossi et al., 2020), distinguish source and target roles (Tjandra et al., 2024), model continuous node-state trajectories (Wang et al., 2024), or encode histories through walks, patches, MLP mixing, and neighborhood summaries (Xu et al., 2020; Wang et al., 2021b;a; Luo & Li, 2022; Souza et al., 2022; Yu et al., 2023; Cong et al., 2023; Zhang et al., 2025a; Lu et al., 2024; Tian et al., 2024; Li et al., 2025; Ding et al., 2025; Zou et al., 2024; Peng et al., 2025; Chung et al., 2025; Tieu et al., 2025; Zhang et al., 2025b; Yi et al., 2025a; Yu et al., 2025). Memorization exploits recurring interactions (Poursafaei et al., 2022; Gastinger et al., 2024b; 2025). Weisfeiler–Leman and walk analyses address structural distinguishability (Xu et al., 2019; Souza et al., 2022; Heeg et al., 2025; Beddar-Wiesing & Moallemy-Oureh, 2025); representation power is distinct from predictive usefulness (Kim et al., 2025; Lachi et al., 2025). State tracking depends on transition, injection, precision, and readout assumptions (Grazzi et al., 2025; Terzic et al., 2025; Walker et al., 2025; Movahedi et al., 2025). Knowledge-graph decoders also use quaternion factors (Zhang et al., 2019), and TComplEx uses timestamp-modulated complex factors without a unitary constraint (Lacroix et al., 2020). Benchmark comparisons require specified candidates, samplers, splits, and temporal batching (Huang et al., 2023; Gastinger et al., 2024a; Yi et al., 2025b; Poursafaei et al., 2022; Romero et al., 2024; Lampert et al., 2026). Surveys organise the wider model and evaluation landscape (Feng et al., 2026; Jiao et al., 2025).

At the model level, relative to TGN, TSNN adds a second per-node state, the frame, and three steps with no counterpart there: transport before scoring, exact carry-over when a frame changes, and the frame update itself; TGN also scores each batch from one shared memory state without an equal-timestamp guarantee (Rossi et al., 2020). Relative to TPNet and TNCN, whose per-node or pairwise encodings summarise walk or common-neighbour structure (Lu et al., 2024; Zhang et al., 2025a), the frame is a coordinate system: it governs how other state is read, not what is counted. Relative to LoCS, whose frames are recomputed at each step from observed positions and velocities with no persistent frame state (Kofinas et al., 2021), TSNN's frames are latent, persist across events, and are learned from the interactions. Relative to neural sheaf diffusion and bundle neural networks, which act on static graphs (Bodnar et al., 2022; Bamberger et al., 2025), and to dynamic sheaf diffusion, whose maps are recomputed from the current signals for forecasting (Mostafa et al., 2026), TSNN keeps its frames as memory, carries stored vectors across frame changes, and scores links under a strict-history transaction.

Table 4: What TSNN inherits and what it adds. Each component is listed with where it appears in this paper and its origin.

Component	Origin	Status
Per-node memory with a GRU update (Eq. 6)	Rossi et al. (2020)	inherited
Sinusoidal time encoding (Eq. 14)	Xu et al. (2020)	inherited
Capped recent-event buffer (Blocks A and E)	Rossi et al. (2020); Luo & Li (2022)	inherited
Pair recency and EdgeBank mixing (Eqs. 3, 19)	Poursafaei et al. (2022)	inherited, learned gate
Common-neighbour features (Appendix E.1)	Zhang et al. (2025a); Lu et al. (2024)	inherited, optional
Neighbour attention (Appendix E.1)	Xu et al. (2020); Yu et al. (2023)	inherited, after transport
Householder orthogonal product and its $O(kd)$ cost (Eq. 1, Lemma 1)	Mhammedi et al. (2017)	inherited
Orthogonal restriction maps and sheaf Laplacian (Section 2.2)	Hansen & Ghrist (2019); Bodnar et al. (2022); Barbero et al. (2022a)	inherited
Normalised sheaf diffusion on flat bundles (Eq. 7)	Bodnar et al. (2022); Bamberger et al. (2025)	inherited
Degree-independent $[0, 2]$ spectrum (Proposition 4)	Chung (1997); Bamberger et al. (2025)	inherited
Persistent per-node frame state $\mathbf{F}_w$, updated by events (Eq. 4)	this work	new
Scoring in transported coordinates from a frozen pre-wave snapshot (Eqs. 2–3)	this work	new
Transported messages into the GRU (Eq. 6)	this work	new
Exact carry-over across a frame change (Proposition 2)	coordinate re-expression	applied here
Wave transaction: all equal-time queries scored before one aggregated commit (Blocks A–E, Lemma 3, Algorithm 1)	this work	new
When frames can change predictions: ambient equivalence, cancellation, anisotropy (Propositions 11–13)	this work	new
Energy descent and non-expansiveness of frozen diffusion (Theorem 5)	gradient-descent analysis	adapted
Frame-field perturbation bound (Theorem 8)	first-order perturbation	adapted
Consensus limit under frame-conjugated gains (Proposition 15)	graph diffusion	adapted
Gauge redundancy: invariant readouts see only ambient vectors (Proposition 9)	gauge equivariance	adapted
Frame drift away from the safeguard (Lemma 7)	Lipschitz analysis	adapted
Orientation blindness of the role-free update and tied cold candidates (Propositions 17, 18)	this work	new, supporting
Carry-only invariance: a pure frame change leaves every score unchanged, with zero gradient (Proposition 16)	this work	new
Raw frame state carries information beyond the frame (Proposition 14)	this work	new
Zero-message control and closed-gate separation witness (Proposition 10, Theorem 19)	this work	new, supporting
Loss–reciprocal-rank consistency (Propositions 20, 21)	Pinsker's inequality	adapted
Diffusion stability across changing graphs (Corollary 6)	this work	new
One typed event schema and one architecture for four graph families (Section 2)	this work	new

The closest single method is BuNN (Bamberger et al., 2025), which diffuses node features over flat orthogonal bundles on static graphs. TSNN inherits its flat-bundle structure: at a fixed frame field the sheaf Laplacian is the graph Laplacian conjugated by block-diagonal orthogonal frames, which gives the degree-independent $[0, 2]$ spectrum (Proposition 4). What TSNN adds is temporal. Frames persist and are updated by events; each stored memory is carried exactly across every frame change (Proposition 2); all queries at a timestamp are scored from one frozen snapshot before any commit (Lemma 3); diffusion stays non-expansive as the graph changes (Corollary 6); and the analysis identifies when frames can change predictions, through a non-equivariant content update or an anisotropic gain whose frame sensitivity scales with $\lambda_{\max}(\mathbf{D}_\theta) - \lambda_{\min}(\mathbf{D}_\theta)$ (Propositions 11–13).

A.2 SUPPLEMENTARY NOTATION

Section 2.2 states the sheaf apparatus directly at the equal stalk width d that TSNN uses. In the general Euclidean cellular sheaf, the node and edge stalks are $\mathcal{F}(a) = \mathbb{R}^{d_a^{\text{stalk}}}$ and $\mathcal{F}(e) = \mathbb{R}^{d_e^{\text{stalk}}}$ for positive integer dimensions $d_a^{\text{stalk}}, d_e^{\text{stalk}}$, and the restriction maps $\rho_{a,e} : \mathcal{F}(a) \to \mathcal{F}(e)$ are rectangular. The coboundary, Laplacian, and energy of Section 2.2 are unchanged in form; the coordinate matrices and their transposes use the stated Euclidean inner products. TSNN sets $d_a^{\text{stalk}} = d_e^{\text{stalk}} = d$ for every node and edge, which is what makes the node-factored identification $\rho_{a,e}^i = \mathbf{U}_a^i$ with a common ambient space available.

The norm $\|\cdot\|_2$ is Euclidean for vectors and spectral for matrices, diag forms a diagonal matrix, and $\mathbf{I}_m$ is the $m \times m$ identity. The inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}$ is Euclidean unless a metric is specified; $\mathbf{1}_m$ is the all-ones vector in $\mathbb{R}^m$, the subscript on $\mathbf{I}_m$ is omitted only when the ambient dimension is fixed, $\mathrm{spec}(\mathbf{A})$ is the spectrum of $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ its largest eigenvalue for symmetric $\mathbf{A}$, and $\ker \mathbf{A} = \{\mathbf{z} : \mathbf{A}\mathbf{z} = 0\}$ its kernel. For symmetric positive-definite $\mathbf{\Omega}$, the induced operator norm is $\|\mathbf{A}\|_{\mathbf{\Omega}\to\mathbf{\Omega}} = \|\mathbf{\Omega}^{1/2}\mathbf{A}\mathbf{\Omega}^{-1/2}\|_2$. Here MLP means a multilayer perceptron whose hidden layers have the form $\mathbf{x} \mapsto \mathrm{softplus}(\mathrm{LN}(W\mathbf{x} + \mathbf{b}))$, followed by a linear output layer, with layer normalization LN and its own learned parameters. The frame network has one hidden layer and the message and score networks have two; the frame and message maps use hidden width p and the score map p_s, with $p = p_s = 192$ in every reported run, and Eq. (20) counts these layers. Lemma 1 concerns applying frames, not evaluating the MLP that produces their rows. Throughout, $\mathrm{sigm}(x) = (1 + e^{-x})^{-1}$ and $\mathrm{softplus}(x) = \log(1 + e^x)$. Table 5 lists the symbols.

Table 5: Notation.

Symbol	Meaning
Events, queries, and time	
$\mathcal{V}$, $\mathcal{T}$, $\mathcal{R}$	node set, node-type vocabulary, relation vocabulary
$e_i = (u_i, v_i, t_i, r_i, \mathbf{x}_i)$, N	event i; stream length
$\mathcal{E}_{<t}$, $\mathcal{W}_t$, $b(t)$	strict history; equal-time wave; number of events before t
q, $\mathcal{C}_q$, M_q, m_q	query; candidate list; list length; negatives ($M_q = 1 + m_q$)
$M_{\mathrm{ev}} = \sum_q M_q$	candidate positions of one event's queries (cost bound)
$\Delta_w^-(t)$, $\Delta_{w,i}$, $\Delta_{u,c,r}^{\mathrm{pair}}(t)$	elapsed time since w's last endpoint interaction (at scoring; at commit); since the last (u, c, r) event
$\Delta_\perp$	large constant gap used when there is no earlier interaction
ϕ, d_{time}, n_ω	sinusoidal time encoding; its width parameter; number of frequencies
State, frames, and scoring	
d, k, k_w	latent width; reflector count; reflectors of w above ε_H
$\mathbf{h}_w$, $\mathbf{g}_w = \mathbf{U}_w \mathbf{h}_w$	local memory; its ambient vector
$\mathbf{F}_w$, $\mathbf{U}_w = \mathcal{U}(\mathbf{F}_w)$	raw frame parameters; orthogonal frame
$\dot{\mathbf{F}}_{w,j}$, $\overline{\mathbf{h}}_{w,i}$	frame increment of incidence j; carried memory
$\boldsymbol{\sigma}_{w,j}$	role indicator: $(1, 0)$ for the source, $(0, 1)$ for the destination
$\hat{\mathbf{u}} = \mathbf{f}/\|\mathbf{f}\|_2$	unit reflector direction of a frame row $\mathbf{f}$
$\mathbf{Q}_{a \leftarrow b} = \mathbf{U}_a^\top \mathbf{U}_b$	transport into a's frame
ε_H, τ_H	Householder safeguard threshold; reflector-norm bound assumed in Lemma 7
$B_{\max}$, $\mathcal{B}_w$, B_{attn}	event-buffer cap; w's buffer; attention-history cap ($B_{\mathrm{attn}} \le B_{\max}$)
Ξ^i, Ξ_t^-	state after i committed events; scoring snapshot at t
$D_q(c)$, κ, $\boldsymbol{\psi}_q(c)$	aligned squared distance; distance scale; residual-score inputs
d_{cd}, P_{cd}	width and map of the optional coordinate-difference projection
Diffusion and analysis	
$G_i = (V_i, E_i)$, $\mathcal{N}_{a,i}$	retained history graph; neighbour multiset
$\deg_a$, $\overline{\deg}_a$	graph degree; safe degree $\max\{\deg_a, 1\}$ (distinct from the width d)
$\mathscr{E}_{\mathcal{F}}$, $\mathscr{E}$	Dirichlet (disagreement) energy
$\mathbf{A}_G$, $\mathbf{D}_G$, $\mathbf{L}_G$	adjacency; safe-degree matrix; graph Laplacian
$\boldsymbol{\Gamma}_G = \mathbf{D}_G \otimes \mathbf{I}_d$	lifted degree matrix ($\boldsymbol{\Gamma}_i$ on G_i)
$\mathbf{L}_{\mathcal{F}}$, $\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}}$, $\mathbf{L}_{\mathcal{F}}^{\mathrm{rw}}$	sheaf Laplacian; symmetric and random-walk normalisations
K, η, $\mathbf{D}_\theta$	diffusion rounds; step; positive diagonal gain
$\boldsymbol{\Omega} = \mathbf{D}_G \otimes \mathbf{D}_\theta^{-1}$, $\mathbf{L}_\Omega$	metric of Theorem 5; whitened Laplacian $\boldsymbol{\Omega}^{-1/2} \mathbf{L}_{\mathcal{F}} \boldsymbol{\Omega}^{-1/2}$
δ_U	largest frame difference between two frame fields
Constructions and tables	
D_{sep}	ambient separation of the two witness streams
$\Delta\mathrm{AP}$	AP difference, variant minus reference (DGB tables)

B Proofs and Supporting Derivations

Figure 2 maps the results of this appendix onto the wave transaction. Panel (a) is Proposition 2 (Eq. (5)): a frame change re-expresses $\mathbf{h}$ without changing $\mathbf{g}$. Panel (b) is the flatness of Proposition 4, which gives the $[0, 2]$ spectral bound used by Theorem 5. Panel (c) is the ambient gain of Eq. (12): only its anisotropic part can make diffusion frame-dependent (Proposition 13), which supports the main-text claim that, for the invariant base scorer, frame state acts only through coordinate-dependent operations. The dependency order is: frame orthogonality and carry-over; frozen transport factorization; the spectral bound; feature-scaled descent; and same-graph frame perturbation. Causality is an independent statement about information flow. Throughout a frozen-graph proof, $G, \mathbf{U}_a, \mathbf{D}_\theta$ are fixed, $n = |V|$, and the sheaf definitions are those of Section 2.2, specialized by the construction in Block E of Section 3 and the fixed-graph notation in Section 4. In particular E is an undirected edge multiset, with each occurrence counted once. Table 6 states which operations and configurations each family of results covers.

B.1 Frame parameterization, carry-over, and drift

Proof of Lemma 1. For a nondegenerate row, write $\mathbf{q} = \mathbf{f}/\|\mathbf{f}\|_2$. Then $(I - 2\mathbf{q}\mathbf{q}^\top)^\top (I - 2\mathbf{q}\mathbf{q}^\top) = I$, because $\mathbf{q}^\top \mathbf{q} = 1$. A degenerate row gives I and is also orthogonal. Products and transposes of

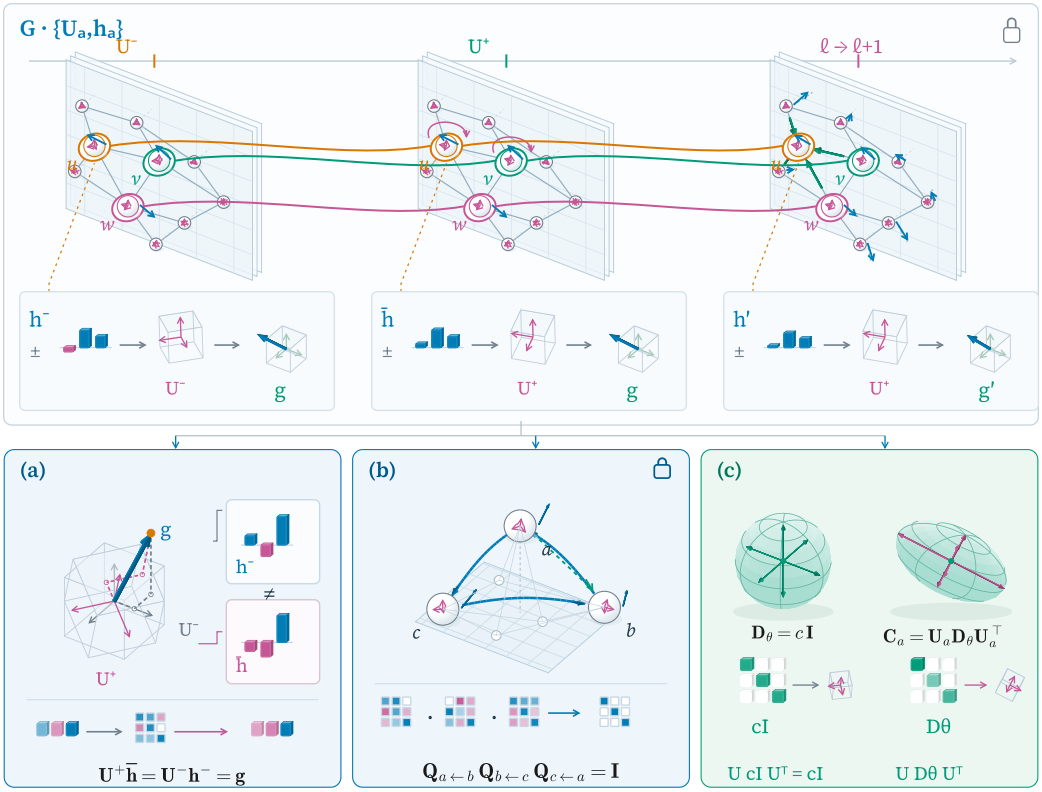


Figure 2: Overview of the analysis. Top: node states $\{\mathbf{U}_a, \mathbf{h}_a\}$ before a wave $(\mathbf{U}^-, \mathbf{h}^-)$, after the frame update and exact carry $(\mathbf{U}^+, \overline{\mathbf{h}}$, same ambient $\mathbf{g})$, and after a diffusion round $\ell \to \ell + 1$ $(\mathbf{h}', \mathbf{g}')$. (a) Carry-over, $\mathbf{U}^+\overline{\mathbf{h}} = \mathbf{U}^-\mathbf{h}^- = \mathbf{g}$. (b) Flat transport, $\mathbf{Q}_{a\leftarrow b}\mathbf{Q}_{b\leftarrow c}\mathbf{Q}_{c\leftarrow a} = \mathbf{I}$. (c) An isotropic gain $c\mathbf{I}$ is unchanged by a frame, whereas $C_a = \mathbf{U}_a\mathbf{D}_\theta\mathbf{U}_a^\top$ can depend on it.

Table 6: Coverage of the theoretical results.

Results	Applicable operation or configuration	Conditions and exclusions
Orthogonality, carry (Lemma 1, Prop. 2)	Orthogonal frames with exact coordinate conversion	Exact arithmetic
Causality (Lemma 3)	Frozen strict-history scoring with permitted parameter history	Does not certify multi-epoch training chronology or split boundaries
Spectrum, stability, consensus (Prop. 4, Thm. 5, Cor. 6, Remark 1, Prop. 15)	Frozen retained graph; synchronous symmetric diffusion; positive gain; stated step bound	Diffusion across changing graphs covered (Cor. 6); no GRU or training-stability guarantee
Ambient equivalence, cancellation, anisotropy (Props. 11–13, Thm. 8)	Unconditioned core with matched metadata, randomness, and inputs	Cancellation also needs an equivariant content map and scalar gain
Carry-only and readout invariance (Props. 9, 16)	Base scalar decoder under joint reexpression	Excludes coordinate features, attention, and frame-only changes
Orientation, cold candidates (Props. 17, 18)	The role-free and common-initialisation ablations	The roles-on, seeded-initialisation reference does not meet their premises
Ranking (Props. 20, 21)	Stated candidate distribution, tie policy, and margin	No automatic transfer to benchmark average-tie MRR
Separation (Prop. 10, Thm. 19)	Supplementary receiver-only constructions	Not the trained GRU

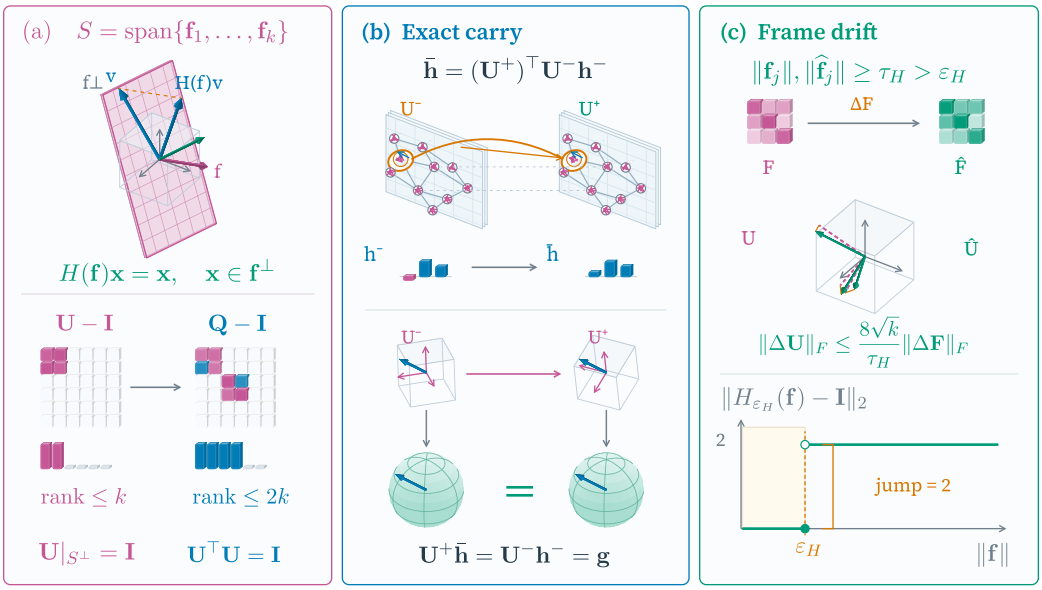


Figure 3: Lemma 1, Proposition 2, and Lemma 7: low-rank reflectors, the unique carry-over, and drift away from the safeguard.

orthogonal matrices are orthogonal. The subspace S_w is the span of the k_w nondegenerate row vectors. Every reflector fixes every vector in $S_w^{\perp}$, so $S_w^{\perp} \subseteq \ker(\mathbf{U}_w - I)$. Rank-nullity yields $\mathrm{rank}(\mathbf{U}_w - I) \leq \dim S_w \leq k_w$. Also $\mathbf{Q}_{a \leftarrow b} - I = \mathbf{U}_a^{\top}(\mathbf{U}_b - \mathbf{U}_a)$, so subadditivity of rank gives the bound $k_a + k_b$. One reflector acts by $\mathbf{y} \mapsto \mathbf{y} - 2\mathbf{f}(\mathbf{f}^{\top}\mathbf{y})/(\mathbf{f}^{\top}\mathbf{f})$ when nondegenerate, requiring $O(d)$ work. A frame needs k such operations and a pairwise transport at most $2k$. The parameter matrix contains kd scalars. These operation counts exclude the network that produces the rows. □

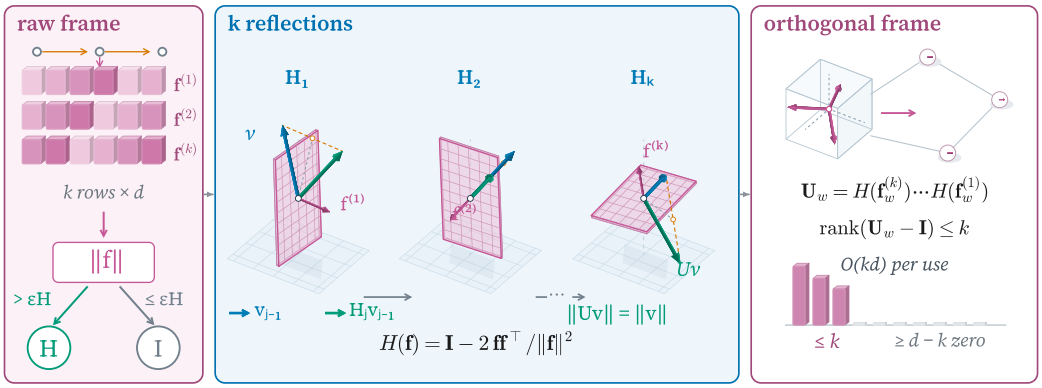


Figure 4: Householder frames of Eq. (1). Left: the raw frame $\mathbf{F} \in \mathbb{R}^{k \times d}$ with rows $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(k)}$; a row with $\|\mathbf{f}\|_2 > \varepsilon_H$ gives a reflection H, otherwise $\mathbf{I}$. Middle: $H(\mathbf{f}) = \mathbf{I} - 2\mathbf{f}\mathbf{f}^{\top}/\|\mathbf{f}\|_2^2$ reflects $\mathbf{v}_{j-1}$ across the hyperplane $\mathbf{f}^{\perp}$, and the product preserves $\|\mathbf{U}\mathbf{v}\|_2 = \|\mathbf{v}\|_2$. Right: $\mathbf{U}_w = H(\mathbf{f}_w^{(k)}) \cdots H(\mathbf{f}_w^{(1)})$ has $\mathrm{rank}(\mathbf{U}_w - \mathbf{I}) \leq k$ and costs $O(kd)$ per application.

Figure 4 illustrates Lemma 1: each factor fixes $\mathbf{f}^{\perp}$, so the product fixes the orthogonal complement of the row span, and $\mathbf{U}_w - \mathbf{I}$ has rank at most k although $\mathbf{U}_w$ has full rank. This is what makes applying $\mathbf{Q}_{a \leftarrow b} = \mathbf{U}_a^{\top}\mathbf{U}_b$ and the carry of Eq. (5) cost $O(kd)$ without forming a $d \times d$ matrix. Row normalisation in Block C uses a clamped denominator: ordinary rows become unit vectors, while a sufficiently small row can still reach the safeguard branch.

Proof of Proposition 2. Multiplying Eq. (5) by $\mathbf{U}_w^i$ gives $\mathbf{U}_w^i(\mathbf{U}_w^i)^\top \mathbf{U}_w^{i-1}\mathbf{h}_w^{i-1} = \mathbf{U}_w^{i-1}\mathbf{h}_w^{i-1}$. Conversely any vector with this ambient representation becomes exactly Eq. (5) after multiplication by $(\mathbf{U}_w^i)^\top$. This proves existence and uniqueness. □

The next lemma is the frame-drift statement summarized in Section 4. Its hypothesis excludes the safeguard branch. It is a condition on the rows involved, not a margin the update enforces: an updated row has norm $r/\max\{r, 10^{-12}\}$ for pre-normalised norm r, a row not yet updated keeps its Gaussian initialisation, and a proposal can cancel a row exactly.

Lemma 7 (Frame drift away from the safeguard)**.** *Suppose every row of* $\mathbf{F}, \mathbf{F}' \in \mathbb{R}^{k\times d}$ *has norm at least* $\tau_H > \varepsilon_H$*. Then*

$$\|\mathcal{U}(\mathbf{F}') - \mathcal{U}(\mathbf{F})\|_F \le \frac{8\sqrt{k}}{\tau_H}\|\mathbf{F}' - \mathbf{F}\|_F. \tag{9}$$

Under the lemma's norm assumptions the pre-normalised rows stay above the safeguard, and the renormalisation of Block C preserves their directions and hence $\mathcal{U}$, so for TSNN's consecutive frames the squared bound is $(64k/\tau_H^2)\|\dot{\overline{\mathbf{F}}}_w\|_F^2$ with $\dot{\overline{\mathbf{F}}}_w = |J_w|^{-1}\sum_{j\in J_w}\dot{\mathbf{F}}_{w,j}$, applied to $\mathbf{F}_w^{i-1}$ and $\mathbf{F}_w^{i-1} + \dot{\overline{\mathbf{F}}}_w$; the DGB setup penalises this increment through the smoothness term of Eq. (17). Both matrices must meet the norm assumption; τ_H is a hypothesis, not an additional safeguard or clipping threshold.

Proof of Lemma 7. For $\|\mathbf{f}\|_2, \|\mathbf{f}'\|_2 \ge \tau_H$ the reverse triangle inequality gives $\|\hat{\mathbf{u}}' - \hat{\mathbf{u}}\|_2 \le 2\|\mathbf{f}' - \mathbf{f}\|_2/\tau_H$. Writing the difference of two outer products as $(\mathbf{q}' - \mathbf{q})\mathbf{q}'^\top + \mathbf{q}(\mathbf{q}' - \mathbf{q})^\top$ gives

$$\|H_{\varepsilon_H}(\mathbf{f}') - H_{\varepsilon_H}(\mathbf{f})\|_F \le 4\|\hat{\mathbf{u}}' - \hat{\mathbf{u}}\|_2 \le \frac{8}{\tau_H}\|\mathbf{f}' - \mathbf{f}\|_2.$$

Telescope the difference of the two ordered products, replacing one factor at a time. All factors to its left and right are orthogonal, and therefore preserve the Frobenius norm. Summing the single-factor bounds and applying Cauchy–Schwarz to the k row differences gives

$$\|\mathcal{U}(\mathbf{F}') - \mathcal{U}(\mathbf{F})\|_F \le \frac{8}{\tau_H}\sum_{j=1}^{k}\|\mathbf{f}'^{(j)} - \mathbf{f}^{(j)}\|_2 \le \frac{8\sqrt{k}}{\tau_H}\|\mathbf{F}' - \mathbf{F}\|_F.$$

Multiplication by either frame's transpose shows that this is also the norm of $I - \mathcal{U}(\mathbf{F}')^\top\mathcal{U}(\mathbf{F})$. □

The assumption excludes crossing the threshold branch. For example, along a fixed unit direction, radii just below and just above ε_H produce I and a reflection separated by operator norm 2, while the row difference tends to zero. A global Lipschitz assertion for the safeguarded map is false. When all k rows are nondegenerate, $\det\mathcal{U}(\mathbf{F}) = (-1)^k$; fixed small k does not parameterize every element of $O(d)$.

B.2 FROZEN-WAVE CAUSALITY

This subsection proves that no score can use an event from its own or a later timestamp (Lemma 3).

Proof of Lemma 3. Condition on permitted metadata, model parameters, and any independent random seed. By induction over waves, the post-wave state is a function of the initialized state and the events committed so far, because each wave transition reads only the pre-wave state and the events of that wave. For time t, $\Xi^{b(t)}$ therefore depends only on $\mathcal{E}_{<t}$. Each scoring input in Block B of Section 3 is computed from this state, the query, and its candidate's permitted metadata. No scorer reads $\mathbf{x}_i$ or the positive indicator of an event at t. Freezing the snapshot, count tables, and parameters until the last query in the wave is scored ensures that the same argument holds for every query in that wave. If parameters are adapted online, induction applies also to their updates when those updates use only completed earlier waves. If parameters were fit on future labels, the last assumption fails and the lemma does not apply; this includes training scores in the second and later epochs of multi-epoch replay (Appendix E). □

The candidate lists and filtering rules belong to the evaluation task. A score for one candidate must not exploit its position as a positive in an evaluator's array. Offline filtering by simultaneous truths is not permission to read those truths into node states. Freezing only node vectors while updating counts, samplers, neighbor buffers, or learned routing weights can violate the premise.

B.3 Frozen transport and the graph operator

Proposition 4 states the unnormalized factorization and the spectral bound. The normalized forms obey the same conjugation,

$$\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}} = \mathbf{S}^{\top}((\mathbf{D}_G^{-1/2}\mathbf{L}_G\mathbf{D}_G^{-1/2}) \otimes \mathbf{I}_d)\mathbf{S},$$

and $\mathbf{L}_{\mathcal{F}}^{\mathrm{rw}} = \mathbf{\Gamma}_G^{-1}\mathbf{L}_{\mathcal{F}}$ is similar to $\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}}$, hence has the same spectrum.

Proof of Proposition 4. In the walk product, every adjacent pair $\mathbf{U}_{a_j}\mathbf{U}_{a_j}^{\top}$ cancels to I, giving the endpoint map. For an edge $e = \{a, b\}$,

$$\|\mathbf{h}_a - \mathbf{Q}_{a\leftarrow b}\mathbf{h}_b\|_2^2 = \|\mathbf{U}_a\mathbf{h}_a - \mathbf{U}_b\mathbf{h}_b\|_2^2.$$

Summing once over every edge occurrence yields $\mathbf{z}^{\top}\mathbf{L}_{\mathcal{F}}\mathbf{z} = \mathbf{z}^{\top}\mathbf{S}^{\top}(\mathbf{L}_G \otimes I_d)\mathbf{S}\mathbf{z}$. Both matrices are symmetric, so equality of their quadratic forms for every $\mathbf{z}$ proves equality of matrices. Equivalently, with Block E's loop convention, the blocks are

$$(\mathbf{L}_{\mathcal{F}})_{aa} = \big(\deg_a - (\mathbf{A}_G)_{aa}\big)I_d, \qquad (\mathbf{L}_{\mathcal{F}})_{ab} = -(\mathbf{A}_G)_{ab}\mathbf{U}_a^{\top}\mathbf{U}_b \quad (a \neq b).$$

The scalar diagonal blocks of $\mathbf{\Gamma}_G$ commute with $\mathbf{S}$. Conjugating the matrix identity by $\mathbf{\Gamma}_G^{-1/2}$ proves the symmetric normalized identity; left multiplication by $\mathbf{\Gamma}_G^{-1}$ gives the random-walk version. Moreover $\mathbf{\Gamma}_G^{1/2}\mathbf{L}_{\mathcal{F}}^{\mathrm{rw}}\mathbf{\Gamma}_G^{-1/2} = \mathbf{L}_{\mathcal{F}}^{\mathrm{sym}}$, so these two operators have the same spectrum.

For any $\mathbf{x} \in \mathbb{R}^n$, set $y_a = x_a/\sqrt{\deg_a}$. Then

$$0 \le \mathbf{x}^{\top}\mathbf{D}_G^{-1/2}\mathbf{L}_G\mathbf{D}_G^{-1/2}\mathbf{x} = \sum_{e=\{a,b\}} (y_a - y_b)^2 \le 2\sum_a \deg_a y_a^2 \le 2\|\mathbf{x}\|_2^2.$$

The middle inequality uses $(y_a - y_b)^2 \le 2y_a^2 + 2y_b^2$. Isolated vertices contribute zero. The Rayleigh quotient and orthogonal similarity establish the spectral interval $[0, 2]$. □

B.4 Descent, non-expansiveness, and convergence

Figure 5 shows the operator that Theorem 5 analyses: every active node reads the same round ℓ, the frames and gain are fixed across the K rounds, and the degree-normalised transported disagreements are the gradient, in the gain-weighted metric, of the degree-normalised sheaf energy (Remark 1). This is why the theorem needs the full-active synchronous update of Eq. (7) and not an endpoint-only variant. The quantitative energy decrease referenced in Theorem 5 is

$$\mathscr{E}(T\mathbf{z}) \le \mathscr{E}(\mathbf{z}) - \eta\left(1 - \frac{\eta}{2}\lambda_{\max}(\mathbf{L}_\Omega)\right)\|\mathbf{L}_{\mathcal{F}}\mathbf{z}\|_{\mathbf{\Omega}^{-1}}^2, \tag{10}$$

and the firm non-expansiveness inequality in the smaller step range is $\|T\mathbf{w}\|_{\mathbf{\Omega}}^2 + \|(I - T)\mathbf{w}\|_{\mathbf{\Omega}}^2 \le \|\mathbf{w}\|_{\mathbf{\Omega}}^2$ for every $\mathbf{w}$.

Proof of Theorem 5. Since $\nabla\mathscr{E}(\mathbf{z}) = \mathbf{L}_{\mathcal{F}}\mathbf{z}$ and $\mathbf{\Omega}^{-1} = (I_n \otimes \mathbf{D}_\theta)\mathbf{\Gamma}_G^{-1}$, the update is $\mathbf{z} - \eta\,\mathrm{grad}_{\mathbf{\Omega}}\mathscr{E}(\mathbf{z})$, where $\mathrm{grad}_{\mathbf{\Omega}}\mathscr{E} = \mathbf{\Omega}^{-1}\nabla\mathscr{E}$. In whitened coordinates $\mathbf{y} = \mathbf{\Omega}^{1/2}\mathbf{z}$,

$$\mathbf{\Omega}^{1/2}T\mathbf{\Omega}^{-1/2} = I - \eta\mathbf{L}_\Omega, \qquad \mathscr{E}(\mathbf{z}) = \tfrac{1}{2}\mathbf{y}^{\top}\mathbf{L}_\Omega\mathbf{y}.$$

The symmetric positive-semidefinite matrix $\mathbf{L}_\Omega$ satisfies

$$\mathbf{L}_\Omega = (I_n \otimes \mathbf{D}_\theta^{1/2})\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}}(I_n \otimes \mathbf{D}_\theta^{1/2}) \preceq 2(I_n \otimes \mathbf{D}_\theta) \preceq 2\lambda_{\max}(\mathbf{D}_\theta)I.$$

Here $\mathbf{X} \preceq \mathbf{Y}$ means $\mathbf{Y} - \mathbf{X}$ is positive semidefinite. Choose an orthonormal eigenbasis $\mathbf{v}_j$ of $\mathbf{L}_\Omega$, with eigenvalues $\lambda_j \ge 0$, and write $\mathbf{y} = \sum_j a_j\mathbf{v}_j$. Direct substitution gives

$$\mathscr{E}(T\mathbf{z}) - \mathscr{E}(\mathbf{z}) = -\eta\sum_j \lambda_j^2(1 - \eta\lambda_j/2)a_j^2 \le -\eta(1 - \eta\lambda_{\max}(\mathbf{L}_\Omega)/2)\|\mathbf{L}_{\mathcal{F}}\mathbf{z}\|_{\mathbf{\Omega}^{-1}}^2,$$

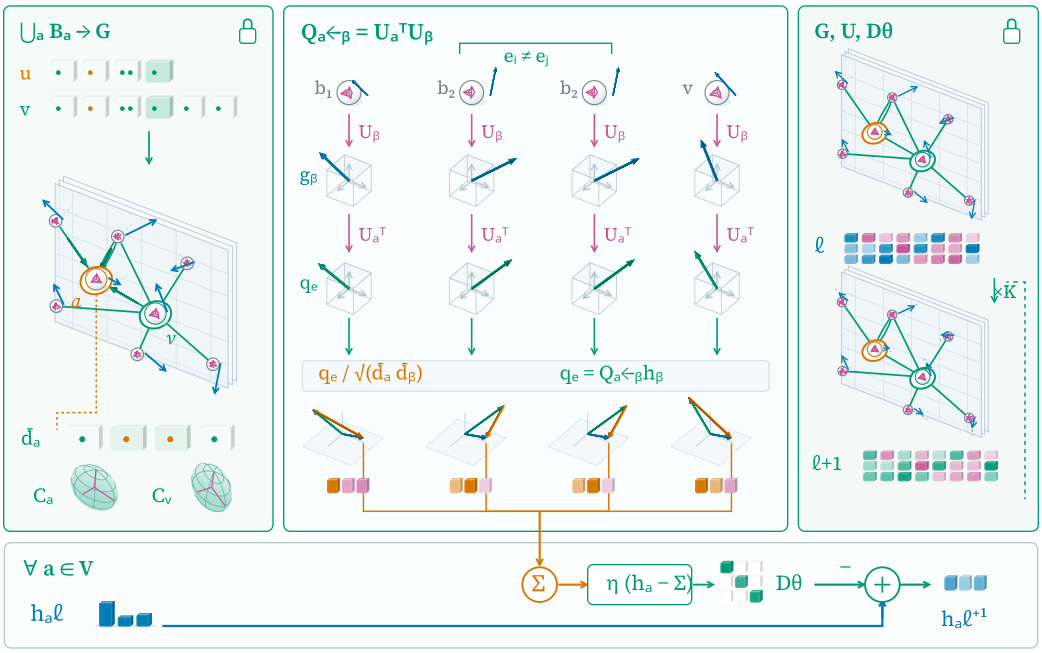


Figure 5: One diffusion round. Left: the retained history graph, the union of all buffers, with safe degree $\overline{\deg}_a$ and ambient gains $C_a = \mathbf{U}_a \mathbf{D}_\theta \mathbf{U}_a^\top$. Middle: each neighbour b is transported by $\mathbf{Q}_{a\leftarrow b} = \mathbf{U}_a^\top \mathbf{U}_b$ to $\mathbf{q}_e = \mathbf{Q}_{a\leftarrow b}\mathbf{h}_b$, weighted by $(\overline{\deg}_a \overline{\deg}_b)^{-1/2}$, and summed. Right and bottom: $\mathbf{h}_a - \sum$ is scaled by $\eta \mathbf{D}_\theta$ and applied synchronously to every $a \in V$, K times (Eq. (7)). The middle panel draws the general multigraph operator of Section 2.2; Block E's retained graph keeps one edge per distinct pair.

because the last squared norm is $\sum_j \lambda_j^2 a_j^2$. If $0 < \eta \le 1/\lambda_{\max}(\mathbf{D}_\theta)$, every multiplier $1 - \eta\lambda_j$ lies in $[-1, 1]$. Thus the energy is non-increasing and the update is non-expansive in the $\mathbf{\Omega}$ norm. For the smaller bound $0 < \eta \le 1/(2\lambda_{\max}(\mathbf{D}_\theta))$, put $x_j = \eta\lambda_j \in [0, 1]$. The scalar inequality $(1 - x_j)^2 + x_j^2 \le 1$ proves firm non-expansiveness after summation. For $\lambda_j > 0$, $0 \le 1 - \eta\lambda_j < 1$, so repeated iterations send that component to zero; components with $\lambda_j = 0$ are fixed. Transforming back proves convergence to the metric-orthogonal projection onto $\ker \mathbf{L}_\mathcal{F}$. If $\mathbf{L}_\Omega = 0$, the update is the identity and every state already belongs to this kernel. □

At the larger admissible boundary a multiplier may equal -1, so non-expansiveness alone does not imply convergence. The proof uses one fixed metric and operator, not a sequence of changing event graphs. It also uses simultaneous full-active updates; reading stale neighbors or updating only endpoints defines a different matrix.

Remark 1 (The implemented symmetric normalisation)**.** *Eq.* (7) *applies* $T_{\mathrm{sym}}\mathbf{y} = \mathbf{y} - \eta\mathbf{P}\mathbf{L}_\mathcal{F}^{\mathrm{sym}}\mathbf{y}$ *with* $\mathbf{P} = \mathbf{I}_n \otimes \mathbf{D}_\theta$*. Theorems 5 and 8 hold for* T_{sym} *with the same step conditions and constants, with* $\mathbf{\Omega}_{\mathrm{sym}} = \mathbf{I}_n \otimes \mathbf{D}_\theta^{-1}$ *in place of* $\mathbf{\Omega}$ *and* $\mathscr{E}_{\mathrm{sym}}(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top \mathbf{L}_\mathcal{F}^{\mathrm{sym}}\mathbf{y}$ *in place of* $\mathscr{E}$*; the limit is the* $\mathbf{\Omega}_{\mathrm{sym}}$*-orthogonal projection onto* $\ker \mathbf{L}_\mathcal{F}^{\mathrm{sym}} = \mathbf{\Gamma}_G^{1/2} \ker \mathbf{L}_\mathcal{F}$*. A loop* $\{a, a\}$ *adds* 2 *to* $\deg_a$ *and* 2 *to* $(\mathbf{A}_G)_{aa}$*, with* $\mathbf{Q}_{a\leftarrow a} = \mathbf{I}_d$*, so the diagonal block of* $\mathbf{L}_\mathcal{F}$ *becomes* $(\deg_a - (\mathbf{A}_G)_{aa})\mathbf{I}_d$*; it leaves the unnormalised energy unchanged, changes the normalised one only through* $\overline{\deg}_a$*, and the* $[0, 2]$ *bound still holds.*

Remark 1 carries Theorems 5 and 8, proved for the random-walk form, over to the symmetric normalisation used by Eq. (7); the step conditions quoted in Section 4 are therefore the ones that apply to the reported runs. The only change is the metric and the energy, which are rescaled by the fixed degrees. The step condition concerns the $\mathbf{D}_\theta$ in force at each wave; the per-epoch check certifies the frozen evaluation checkpoint.

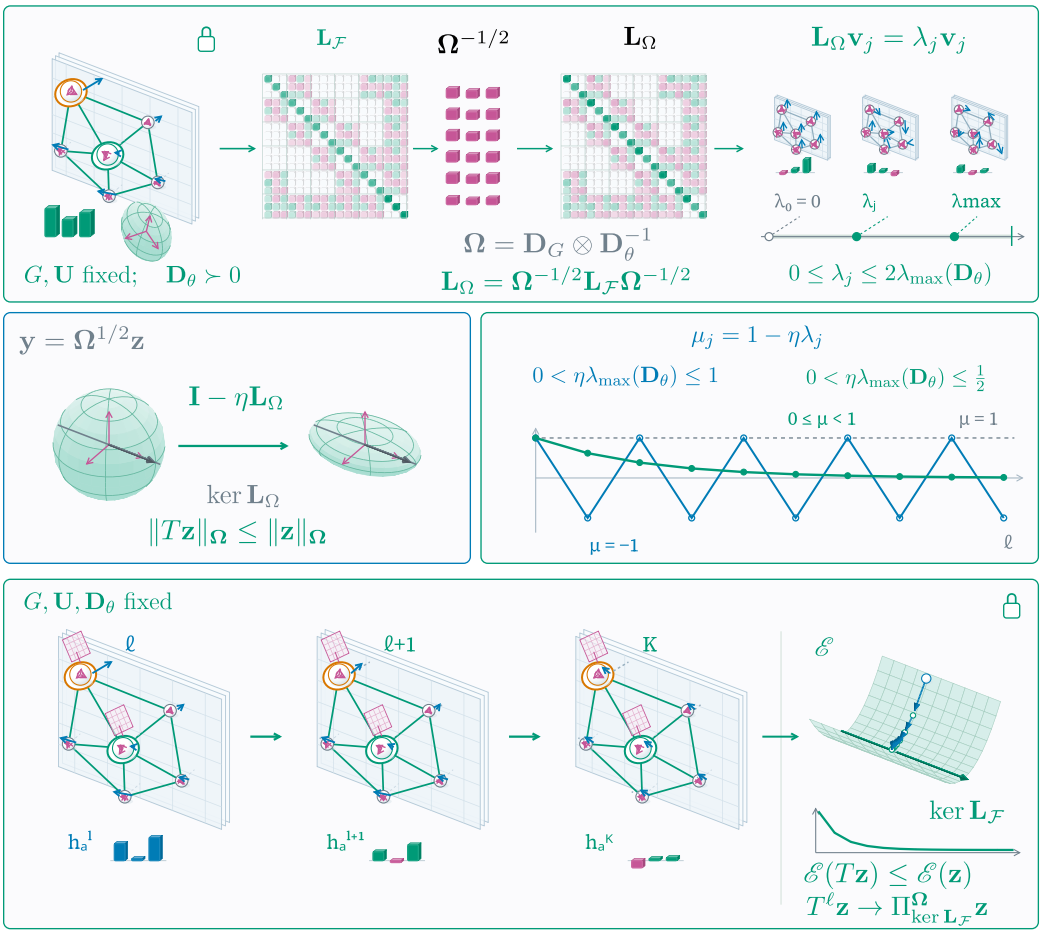


Figure 6: Theorem 5. Top: whitening $\mathbf{L}_{\mathcal{F}}$ by $\mathbf{\Omega}^{-1/2}$ gives $\mathbf{L}_\Omega$ with eigenvalues $0 \le \lambda_j \le 2\lambda_{\max}(\mathbf{D}_\theta)$. Middle: each round multiplies whitened eigencomponent j by $\mu_j = 1 - \eta\lambda_j$; $\eta\lambda_{\max}(\mathbf{D}_\theta) \le 1$ gives $|\mu_j| \le 1$ (non-expansive, possibly oscillating), and $\eta\lambda_{\max}(\mathbf{D}_\theta) \le \frac{1}{2}$ gives $\mu_j \in [0, 1]$ (convergent). Bottom: K rounds on a fixed graph lower the energy $\mathscr{E}$ toward $\ker \mathbf{L}_{\mathcal{F}}$.

Proof. Both $\mathbf{\Gamma}_G$ and $\mathbf{P}$ are block diagonal, and every block of $\mathbf{\Gamma}_G$ is a scalar multiple of $\mathbf{I}_d$, so they commute. Hence $\mathbf{\Gamma}_G^{1/2} T \mathbf{\Gamma}_G^{-1/2} = \mathbf{I} - \eta\mathbf{P}\mathbf{\Gamma}_G^{1/2}\mathbf{\Gamma}_G^{-1}\mathbf{L}_{\mathcal{F}}\mathbf{\Gamma}_G^{-1/2} = T_{\text{sym}}$, where $T\mathbf{z} = \mathbf{z} - \eta\mathbf{\Omega}^{-1}\mathbf{L}_{\mathcal{F}}\mathbf{z}$ is the map of Theorem 5. With $\mathbf{y} = \mathbf{\Gamma}_G^{1/2}\mathbf{z}$, $\mathscr{E}(\mathbf{z}) = \mathscr{E}_{\text{sym}}(\mathbf{y})$ and $\|\mathbf{z}\|_{\mathbf{\Omega}}^2 = \mathbf{y}^\top \mathbf{\Gamma}_G^{-1/2}(\mathbf{D}_G \otimes \mathbf{D}_\theta^{-1})\mathbf{\Gamma}_G^{-1/2}\mathbf{y} = \|\mathbf{y}\|_{\mathbf{\Omega}_{\text{sym}}}^2$. The change of variables is linear and invertible and maps T to T_{sym}, so every energy, non-expansiveness, firm non-expansiveness, and convergence statement transfers verbatim, and the $\mathbf{\Omega}$-orthogonal projection onto $\ker \mathbf{L}_{\mathcal{F}}$ becomes the $\mathbf{\Omega}_{\text{sym}}$-orthogonal projection onto its image. For Theorem 8, the proof already bounds $\|\mathbf{L}_{\mathcal{F}}^{\text{sym}} - \widehat{\mathbf{L}}_{\mathcal{F}}^{\text{sym}}\|_2 \le 2\delta_U$, and $T_{\text{sym}} - \widehat{T}_{\text{sym}} = -\eta\mathbf{P}(\mathbf{L}_{\mathcal{F}}^{\text{sym}} - \widehat{\mathbf{L}}_{\mathcal{F}}^{\text{sym}})$ gives $\|\mathbf{\Omega}_{\text{sym}}^{1/2}(T_{\text{sym}} - \widehat{T}_{\text{sym}})\mathbf{\Omega}_{\text{sym}}^{-1/2}\|_2 \le 2\eta\lambda_{\max}(\mathbf{D}_\theta)\delta_U$. For a loop, the edge term $\|\mathbf{h}_a - \mathbf{Q}_{a \leftarrow a}\mathbf{h}_a\|_2^2$ vanishes and, in the proof of Proposition 4, $x^\top \mathbf{D}_G^{-1/2}\mathbf{L}_G\mathbf{D}_G^{-1/2}x$ sums only over non-loop edges, which is at most $2\sum_a \deg_a y_a^2$. □

Proof of Corollary 6. Extend every operator to the node universe $\mathcal{V}$, with zero blocks of $\mathbf{L}_i^{\text{sym}}$ off V_i, and write $\mathbf{P} = \mathbf{I}_{|\mathcal{V}|} \otimes \mathbf{D}_\theta$ and $\mathbf{\Omega}_{\text{sym}} = \mathbf{P}^{-1}$. One round is $\mathbf{M}_i = \mathbf{I} - \eta\mathbf{P}\mathbf{L}_i^{\text{sym}}$, and $\mathbf{\Omega}_{\text{sym}}^{1/2}\mathbf{M}_i\mathbf{\Omega}_{\text{sym}}^{-1/2} = \mathbf{I} - \eta\mathbf{P}^{1/2}\mathbf{L}_i^{\text{sym}}\mathbf{P}^{1/2}$ is symmetric. By Proposition 4, $\mathbf{0} \preceq \mathbf{L}_i^{\text{sym}} \preceq 2\mathbf{I}$ for every graph, so $\mathbf{0} \preceq \mathbf{P}^{1/2}\mathbf{L}_i^{\text{sym}}\mathbf{P}^{1/2} \preceq 2\lambda_{\max}(\mathbf{D}_\theta)\mathbf{I}$, and the eigenvalues of the whitened round lie in $[1 - 2\eta\lambda_{\max}(\mathbf{D}_\theta), 1] \subseteq [-1, 1]$. Hence $\|\mathbf{M}_i\|_{\mathbf{\Omega}_{\text{sym}} \to \mathbf{\Omega}_{\text{sym}}} \le 1$; nodes off V_i are left unchanged, so $\mathcal{A}_i = \mathbf{M}_i^K$ and every product of such maps also has norm at most 1. For the second claim, write

$\widehat{\mathbf{M}}_i$ for the round built from $\widehat{G}_i$; telescoping gives $\mathbf{M}_i^K - \widehat{\mathbf{M}}_i^K = \sum_{j=0}^{K-1} \mathbf{M}_i^j(\mathbf{M}_i - \widehat{\mathbf{M}}_i)\widehat{\mathbf{M}}_i^{K-1-j}$, each power has norm at most 1, and $\|\mathbf{M}_i - \widehat{\mathbf{M}}_i\|_{\mathbf{\Omega}_{\mathrm{sym}} \to \mathbf{\Omega}_{\mathrm{sym}}} = \eta \|\mathbf{P}^{1/2}(\mathbf{L}_i^{\mathrm{sym}} - \widehat{\mathbf{L}}_i^{\mathrm{sym}})\mathbf{P}^{1/2}\|_2 \le \eta\lambda_{\max}(\mathbf{D}_\theta)\|\mathbf{L}_i^{\mathrm{sym}} - \widehat{\mathbf{L}}_i^{\mathrm{sym}}\|_2$. □

The product bound concerns the diffusion maps. Between waves, exact carry acts on an endpoint as $\mathbf{h} \mapsto \mathbf{R}\mathbf{h}$ with $\mathbf{R} = (\mathbf{U}_w^i)^\top \mathbf{U}_w^{i-1}$; it preserves $\|\cdot\|_{\mathbf{D}_\theta^{-1}}$ when $\mathbf{R}$ commutes with $\mathbf{D}_\theta$ (for example, under a scalar gain) and in general changes it by a factor of at most $\sqrt{\lambda_{\max}(\mathbf{D}_\theta)/\lambda_{\min}(\mathbf{D}_\theta)}$, so trajectories that interleave carry under an anisotropic gain obey the bound up to this factor per carry.

B.5 Frame-field perturbations

This subsection bounds how much one diffusion step can change when one frame field replaces another on the same graph.

Theorem 8 (Same-graph frame perturbation)**.** *Consider two orthogonal frame fields $\{\mathbf{U}_a\}$ and $\{\widehat{\mathbf{U}}_a\}$ on the same graph, with identical degrees and feature gain. Write $\delta_U = \max_a \|\mathbf{U}_a - \widehat{\mathbf{U}}_a\|_2$ and use a hat for the second field's operators. Then*

$$\|\mathbf{L}_{\mathcal{F}}^{\mathrm{rw}} - \widehat{\mathbf{L}}_{\mathcal{F}}^{\mathrm{rw}}\|_{\mathbf{\Gamma}_G \to \mathbf{\Gamma}_G} \le 2\delta_U, \qquad \|(T - \widehat{T})\mathbf{z}\|_{\mathbf{\Omega}} \le 2\eta\lambda_{\max}(\mathbf{D}_\theta)\delta_U\|\mathbf{z}\|_{\mathbf{\Omega}}.$$

Each update is separately non-expansive under Theorem 5's step-size condition.

This bounds diffusion sensitivity to replacing one frame field by another on the same graph. Both updates use the same coordinate input, so comparing equal ambient memories requires carry-over between frames; stored frames remain current until updated, so age alone creates no discrepancy.

Proof of Theorem 8. Define $\widehat{\mathbf{S}} = \mathrm{blkdiag}_a \widehat{\mathbf{U}}_a$, $\mathbf{N}_G = \mathbf{D}_G^{-1/2}\mathbf{A}_G\mathbf{D}_G^{-1/2}$, and $\mathbf{\Pi}_G = \mathrm{diag}_a(\mathbf{1}[\deg_a > 0])$. Then the normalized graph Laplacian is $\mathbf{\Pi}_G - \mathbf{N}_G$. The $\mathbf{\Pi}_G \otimes I_d$ term commutes with both frame fields and cancels in their difference. The edgewise inequality $|2y_a y_b| \le y_a^2 + y_b^2$ shows $\|\mathbf{N}_G\|_2 \le 1$. Block diagonality gives $\|\mathbf{S} - \widehat{\mathbf{S}}\|_2 = \delta_U$. Adding and subtracting one mixed product therefore yields

$$\|\mathbf{L}_{\mathcal{F}}^{\mathrm{sym}} - \widehat{\mathbf{L}}_{\mathcal{F}}^{\mathrm{sym}}\|_2 = \|\widehat{\mathbf{S}}^\top(\mathbf{N}_G \otimes I_d)\widehat{\mathbf{S}} - \mathbf{S}^\top(\mathbf{N}_G \otimes I_d)\mathbf{S}\|_2 \le 2\delta_U.$$

Whitening by $\mathbf{\Gamma}_G$ proves the first stated bound. Define $\mathbf{A} = \mathbf{L}_{\mathcal{F}}^{\mathrm{rw}} - \widehat{\mathbf{L}}_{\mathcal{F}}^{\mathrm{rw}}$ and $\mathbf{A}_{\mathrm{sym}} = \mathbf{\Gamma}_G^{1/2}\mathbf{A}\mathbf{\Gamma}_G^{-1/2}$. With $\mathbf{P} = I_n \otimes \mathbf{D}_\theta$,

$$\mathbf{\Omega}^{1/2}\mathbf{P}\mathbf{A}\mathbf{\Omega}^{-1/2} = (I_n \otimes \mathbf{D}_\theta^{1/2})\mathbf{A}_{\mathrm{sym}}(I_n \otimes \mathbf{D}_\theta^{1/2}).$$

Its norm is at most $2\lambda_{\max}(\mathbf{D}_\theta)\delta_U$. Since $T - \widehat{T} = -\eta\mathbf{P}\mathbf{A}$, the second bound follows. Applying Theorem 5 separately to each field proves the last assertion. □

B.6 Gauge invariants and the zero-message control

This subsection shows what the core score can and cannot see: it depends on memories only through their ambient vectors.

Proposition 9 (Gauge redundancy of the frame-state configuration)**.** *On $\prod_{v \in \mathcal{V}}(\mathbb{R}^d \times O(d))$, independent matrices $\mathbf{R}_v \in O(d)$ act as $(\mathbf{h}_v, \mathbf{U}_v) \mapsto (\mathbf{R}_v^\top \mathbf{h}_v, \mathbf{U}_v\mathbf{R}_v)$. Two configurations are related by this action if and only if their ambient vectors $\{\mathbf{U}_v\mathbf{h}_v\}_v$ agree. Consequently every invariant readout of this configuration factors through those ambient vectors.*

This identifies the information boundary of the unconditioned geometric score, $\|\mathbf{h}_u - \mathbf{Q}_{u \leftarrow v}\mathbf{h}_v\|_2 = \|\mathbf{g}_u - \mathbf{g}_v\|_2$. It does not make ambient dynamics constant or identify raw reflector parameters with frames, and the action need not preserve a fixed small-k family.

Proof of Proposition 9. The ambient map is invariant since $(\mathbf{U}_v\mathbf{R}_v)(\mathbf{R}_v^\top \mathbf{h}_v) = \mathbf{U}_v\mathbf{h}_v$. Conversely, suppose $(\mathbf{h}_v, \mathbf{U}_v)$ and $(\mathbf{h}'_v, \mathbf{U}'_v)$ have the same ambient vectors. Choosing $\mathbf{R}_v = \mathbf{U}_v^\top \mathbf{U}'_v$ gives $\mathbf{U}_v\mathbf{R}_v = \mathbf{U}'_v$ and $\mathbf{R}_v^\top \mathbf{h}_v = (\mathbf{U}'_v)^\top \mathbf{U}_v\mathbf{h}_v = (\mathbf{U}'_v)^\top \mathbf{U}'_v\mathbf{h}'_v = \mathbf{h}'_v$. Thus the fibers of the ambient map are exactly the gauge orbits. An invariant readout is constant on each orbit and therefore defines a function on the ambient vectors. No assumption about learning or semantic identifiability is used. □

Consider a receiver-only control with exact carry-over, message $\mathrm{MSG}(\mathbf{a}, \mathbf{b}) = \psi(\mathbf{b}) - \psi(\mathbf{a})$ for $\psi : \mathbb{R}^d \to \mathbb{R}^d$, and update $\mathrm{UPD}(\mathbf{h}, \mathbf{m}) = \mathbf{h} + \mathbf{m}$. Both message arguments are in the receiver's new frame, and there are no other memory updates or readout inputs.

Proposition 10 (Homogeneous zero-message control)**.** *If all initial ambient vectors equal $\mathbf{a}_0$, every ambient vector remains $\mathbf{a}_0$ on every event stream under this control, for any frame updates. It cannot distinguish streams by their ambient configurations.*

This is a negative control for the mechanism analysis: frame motion alone cannot change homogeneous ambient memory under this difference-message update.

Proof of Proposition 10. Induct on events, assuming every node's ambient vector equals $\mathbf{a}_0$. Exact carry-over places the receiver's stored vector at $(\mathbf{U}_v^{\mathrm{new}})^\top \mathbf{a}_0$. Transporting the sender into that frame gives the same vector. Thus both inputs to ψ agree, the message is zero, and the additive memory update changes nothing in ambient coordinates. The frame update may be arbitrary; every unaffected node retains its ambient vector. This proves the induction from the homogeneous initialization. □

B.7 Ambient implementation, frame cancellation, and frame sensitivity

These results use the implemented wave transaction of Section 3: mean frame proposals with row renormalisation, one carry and one GRU step per endpoint with mean messages, and synchronous diffusion with the symmetric normalisation on the retained history graph. Write $U_a^-, \mathbf{g}_a^-$ for pre-wave frames and ambient memories, U_a^+ for installed frames, and $\xi_{w,j}$ for the non-coordinate inputs of incidence j of w (gaps, features, relation, types, role).

Proposition 11 (Ambient implementation with raw frame state retained)**.** *Fix the unconditioned* TSNN *core, its initialisation, parameters, and event stream. Replacing each stored $\mathbf{h}_a$ by $\mathbf{g}_a = \mathcal{U}(\mathbf{F}_a)\mathbf{h}_a$, while retaining $\mathbf{F}_a$ and all other state, and recovering every local network input by the conversions below, gives an exactly equivalent implementation: in exact arithmetic both have identical raw frames, metadata, represented memories, and base scores after every wave.*

Proof. Assume agreement before a wave. For incidence j of endpoint w with partner p, the memory inputs of Eq. (4) are $\mathbf{h}_w^- = (U_w^-)^\top \mathbf{g}_w^-$ and $\mathbf{Q}_{w\leftarrow p}^- \mathbf{h}_p^- = (U_w^-)^\top \mathbf{g}_p^-$, which equal the original inputs; the same frame network, mean, and row renormalisation therefore give the same $\mathbf{F}_w^+$ and U_w^+. The carry is $\overline{\mathbf{h}}_w = (U_w^+)^\top \mathbf{g}_w^-$, and once all endpoint frames are installed, $\mathbf{Q}_{w\leftarrow p}^+ \overline{\mathbf{h}}_p = (U_w^+)^\top U_p^+ (U_p^+)^\top \mathbf{g}_p^- = (U_w^+)^\top \mathbf{g}_p^-$. With $\Phi_\xi(\mathbf{x}, (\mathbf{y}_j)_j) = \mathrm{GRU}\left(\mathbf{x}, |J_w|^{-1} \sum_j f_{\mathrm{msg}}(\mathrm{concat}(\mathbf{x}, \mathbf{y}_j, \xi_{w,j}))\right)$, the ambient recurrent state is

$$\mathbf{g}_{w,0}^+ = U_w^+ \, \Phi_\xi\big((U_w^+)^\top \mathbf{g}_w^-, ((U_w^+)^\top \mathbf{g}_{p_j(w)}^-)_{j \in J_w}\big), \tag{11}$$

and every endpoint uses pre-wave partner memories. The buffers, and hence the retained graph, depend only on the observed events. Multiplying Eq. (7) by the fixed U_a^+ and using $U_a^+ \mathbf{Q}_{a\leftarrow b}^+ \mathbf{h}_b = \mathbf{g}_b$ gives

$$\mathbf{g}_{a,\ell+1} = \mathbf{g}_{a,\ell} - \eta \, C_a \Big(\mathbf{g}_{a,\ell} - \sum_{b \in \mathcal{N}_a} \frac{\mathbf{g}_{b,\ell}}{\sqrt{\overline{\deg}_a \overline{\deg}_b}} \Big), \qquad C_a = U_a^+ \mathbf{D}_\theta (U_a^+)^\top, \tag{12}$$

so each synchronous round agrees. At a scoring wave, $d^{-1}\|\mathbf{h}_u - \mathbf{Q}_{u\leftarrow c}\mathbf{h}_c\|_2^2 = d^{-1}\|\mathbf{g}_u - \mathbf{g}_c\|_2^2$, $\mathbf{h}_u^\top \mathbf{Q}_{u\leftarrow c}\mathbf{h}_c = \mathbf{g}_u^\top \mathbf{g}_c$, and $\|\mathbf{h}_a\|_2 = \|\mathbf{g}_a\|_2$; the remaining inputs of Eq. (3) are metadata. Induction over waves completes the proof. □

Proposition 11 shows that the local-coordinate specification of Blocks C–E can be run entirely on ambient memories: frames then affect ambient content only through the conjugated content map of Eq. (11) and the conjugated gain C_a of Eq. (12); they also enter the frame network's inputs, which shape later frames. This makes precise the main-text claim of Section 5 that frame state can act only through coordinate-dependent operations. The frame channel is not removed: $\mathbf{F}$ remains persistent state.

Proposition 12 (Frame-independent content dynamics under equivariant updates)**.** *Compare waves with the same initial ambient memories, events, and non-coordinate inputs, but arbitrary orthogonal frame trajectories with exact carry. Suppose the content map has no raw-frame input and satisfies $\Phi_\xi(R^\top \mathbf{x}, (R^\top \mathbf{y}_j)_j) = R^\top \Phi_\xi(\mathbf{x}, (\mathbf{y}_j)_j)$ for every $R \in O(d)$, and that diffusion uses the same scalar*

gain $\mathbf{D}_\theta = c\mathbf{I}_d$. Then the ambient content trajectories and base scores are identical for every frame trajectory, and equal those of a frame-free implementation with the same Φ_ξ.

Proof. Assume equal ambient memories before a wave. Applying the equivariance with $R = U_w^+$ to Eq. (11) gives $\mathbf{g}_{w,0}^+ = \Phi_\xi(\mathbf{g}_w^-, (\mathbf{g}_{p_j(w)}^-)_j)$, which does not depend on the frame. In Eq. (12), $C_a = c\mathbf{I}_d$, so every round is frame-free, and the retained graph depends only on the events. Induction over waves and the score identities of Proposition 11 complete the proof. □

Proposition 12 identifies the two places where evolving frames can matter: a non-equivariant content map and an anisotropic gain. It is the theoretical basis of the frame-free and isotropic-gain controls in Tables 3 and 19 (Blocks C and E, Panel A): under equivariant content updates *and* the same scalar gain, the framed and frame-free implementations coincide. Isotropic gain alone does not cancel the frame dependence of a non-equivariant GRU. The condition is sufficient, not necessary; the coordinatewise GRU is not equivariant, and Proposition 14 gives valid finite parameters for which frames change the next ambient memory.

Proposition 13 (The scalar gain component cancels)**.** *Fix one retained graph with safe degrees $\overline{\deg}_a$, a symmetric positive-definite gain $\mathbf{D}$, and $\eta > 0$. For orthogonal frame fields $U, \widehat{U}$ let $\mathcal{T}_U = \mathbf{I} - \eta\,\mathrm{blkdiag}_a(U_a\mathbf{D}U_a^\top)(\mathbf{D}_G^{-1/2}\mathbf{L}_G\mathbf{D}_G^{-1/2} \otimes \mathbf{I}_d)$ be one ambient round of Eq.* (12) *and $\delta_U = \max_a \|U_a - \widehat{U}_a\|_2$. At the same ambient input $\mathbf{g}$,*

$$\|(\mathcal{T}_U - \mathcal{T}_{\widehat{U}})\mathbf{g}\|_2 \le 2\eta\big(\lambda_{\max}(\mathbf{D}) - \lambda_{\min}(\mathbf{D})\big)\delta_U\|\mathbf{g}\|_2,$$

and the difference is exactly zero for $\mathbf{D} = c\mathbf{I}_d$. The same bound holds in the $\mathbf{\Gamma}_G$ norm for the random-walk form.

Proof. Put $c = (\lambda_{\max} + \lambda_{\min})/2$ and $\mathbf{A} = \mathbf{D} - c\mathbf{I}_d$, so $\|\mathbf{A}\|_2 = (\lambda_{\max} - \lambda_{\min})/2$. Orthogonal conjugation fixes $c\mathbf{I}_d$, and adding and subtracting $\widehat{U}\mathbf{A}U^\top$ gives $\|U\mathbf{D}U^\top - \widehat{U}\mathbf{D}\widehat{U}^\top\|_2 = \|U\mathbf{A}U^\top - \widehat{U}\mathbf{A}\widehat{U}^\top\|_2 \le 2\|\mathbf{A}\|_2\|U - \widehat{U}\|_2$. The block-diagonal difference therefore has norm at most $(\lambda_{\max} - \lambda_{\min})\delta_U$, and the normalised graph factor has norm at most 2 by the proof of Proposition 4 (loops included, Remark 1). Multiplying by η proves the bound. For the random-walk form, the block difference commutes with $\mathbf{\Gamma}_G^{1/2}$, and whitening by $\mathbf{\Gamma}_G^{1/2}$ reduces it to the symmetric case. For $\mathbf{D} = c\mathbf{I}_d$ every block equals $c\mathbf{I}_d$. □

Proposition 13 quantifies the gain channel of Proposition 12: the frame sensitivity of one round scales with the spread $\lambda_{\max}(\mathbf{D}) - \lambda_{\min}(\mathbf{D})$ and vanishes for scalar gain. This is the algebraic distinction tested by the isotropic-versus-diagonal comparisons of Table 3 (Block E) and Table 19 (Panel A). It bounds one frozen round at equal ambient input and complements Theorem 8, which compares equal local inputs.

Proposition 14 (Raw reflector state is not determined by the current frame)**.** *Under the implemented row-renormalised frame update, two raw states with the same frame U and the same ambient memory can produce different next frames and, for valid finite GRU parameters with every bias zero and both gates equal to $1/2$, different next ambient memories.*

Proof. Take $d = 2$, $k = 1$, $\varepsilon_H < 1$, the unit raw rows $\mathbf{f}_1 = \mathbf{e}_1$ and $\mathbf{f}_2 = -\mathbf{e}_1$, both giving $U = H(\pm\mathbf{e}_1) = \mathrm{diag}(-1, 1)$, and the same ambient memory $\mathbf{g} = \mathbf{e}_1$. Let a self-interaction arrive, with frame-network output the constant $(0, 2)$ (zero weights, final bias $(0, 2)$), so both incidences propose $(0, 2)$. The new rows are $\mathrm{rn}(\pm\mathbf{e}_1 + (0, 2)) \propto (\pm 1, 2)$, and

$$U_1^+ = H((1,2)^\top) = \tfrac{1}{5}\begin{pmatrix} 3 & -4 \\ -4 & -3 \end{pmatrix}, \qquad U_2^+ = H((-1,2)^\top) = \tfrac{1}{5}\begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix} \neq U_1^+.$$

For the GRU of Appendix E set every gate weight and bias to zero, $A_h = 0$, $W_h = 2\mathbf{I}_2$, and $\mathbf{b}_h = \mathbf{b}'_h = 0$; then $\boldsymbol{\zeta} = \boldsymbol{\varrho} = \frac{1}{2}\mathbf{1}$, $\mathbf{h}_{\mathrm{cand}} = \tanh(\mathbf{h})$, and $\Phi(\mathbf{h}) = \frac{1}{2}\mathbf{h} + \frac{1}{2}\tanh(\mathbf{h})$ for every message. Take $K = 0$. The carries are $U_j^+\mathbf{e}_1 = (3/5, \mp 4/5)^\top$, so with $\alpha = \frac{1}{2}(\frac{3}{5} + \tanh\frac{3}{5})$ and $\beta = \frac{1}{2}(\frac{4}{5} + \tanh\frac{4}{5})$, $\mathbf{g}_1^+ = U_1^+(\alpha, -\beta)^\top$ and $\mathbf{g}_2^+ = U_2^+(\alpha, \beta)^\top$ have second coordinates $\mp(4\alpha - 3\beta)/5$. Since $\tanh$ is strictly concave on $(0, \infty)$ with $\tanh 0 = 0$, $\tanh\frac{3}{5} > \frac{3}{4}\tanh\frac{4}{5}$, so $4\alpha - 3\beta = 2\tanh\frac{3}{5} - \frac{3}{2}\tanh\frac{4}{5} > 0$ and $\|\mathbf{g}_1^+ - \mathbf{g}_2^+\|_2 = \frac{2}{5}(4\alpha - 3\beta) > 0$. □

Proposition 14 shows two things for the implemented update. First, the raw rows $\mathbf{F}$ carry information beyond the frame $\mathbf{U}$, because a reflector is unchanged by the sign of its row while the additive update of Block C is not; the gauge quotient of Proposition 9 therefore does not summarise the state. Second, the GRU of Eq. (6) makes the next ambient memory depend on the frame, so the equivariance premise of Proposition 12 fails for real parameters.

Proposition 15 (Consensus limit of frozen diffusion)**.** *Under the conditions of Theorem 5 for its convergence range, with the implemented normalisation (Remark 1), let $C_a = U_a \mathbf{D}_\theta U_a^\top$ and let $\mathcal{C}$ be a connected component of the retained graph. Repeated rounds send the ambient memories on $\mathcal{C}$ to $\mathbf{g}_a^\infty = \sqrt{\overline{\deg}_a}\, \mathbf{c}_\mathcal{C}$ with*

$$\mathbf{c}_\mathcal{C} = \Big(\sum_{a \in \mathcal{C}} \overline{\deg}_a C_a^{-1} \Big)^{-1} \sum_{a \in \mathcal{C}} \sqrt{\overline{\deg}_a}\, C_a^{-1} \mathbf{g}_a^{(0)}.$$

For the random-walk form the limit is the common vector $(\sum_a \overline{\deg}_a C_a^{-1})^{-1} \sum_a \overline{\deg}_a C_a^{-1} \mathbf{g}_a^{(0)}$.

Proof. By Remark 1, the limit is the $\mathbf{\Omega}_{\mathrm{sym}}$-orthogonal projection of the initial state onto $\ker \mathbf{L}_\mathcal{F}^{\mathrm{sym}} = \mathbf{\Gamma}_G^{1/2} \ker \mathbf{L}_\mathcal{F}$. By Proposition 4, $\ker \mathbf{L}_\mathcal{F}$ holds the states whose ambient vectors are constant on each component, so the kernel states are $\mathbf{h}_a = \sqrt{\overline{\deg}_a}\, U_a^\top \mathbf{c}$. Their squared $\mathbf{\Omega}_{\mathrm{sym}}$-distance to the initial state is $\sum_{a \in \mathcal{C}} (\sqrt{\overline{\deg}_a}\mathbf{c} - \mathbf{g}_a^{(0)})^\top C_a^{-1} (\sqrt{\overline{\deg}_a}\mathbf{c} - \mathbf{g}_a^{(0)})$, a strictly convex quadratic in $\mathbf{c}$ because every C_a is positive definite. Setting its gradient to zero gives the formula. The random-walk case is identical with metric $\mathbf{\Omega}$ and kernel states $\mathbf{h}_a = U_a^\top \mathbf{c}$. □

Proposition 15 describes what repeated rounds of Eq. (7) converge to on a frozen retained graph. With the implemented symmetric weights, the fixed point is a degree-scaled consensus $\sqrt{\overline{\deg}_a}\, \mathbf{c}$ rather than a common vector, so high-degree nodes keep larger ambient memories; with isotropic gain $\mathbf{c} = \sum_a \sqrt{\overline{\deg}_a}\, \mathbf{g}_a^{(0)} / \sum_a \overline{\deg}_a$. With the reference $K = 1$ and $\eta = 0.05$ each wave takes one small step towards this limit, and the limit concerns a frozen operator, not changing graphs with recurrent injections.

Proposition 16 (Carry-only invariance and gradient cancellation)**.** *Fix ambient memories $\mathbf{g}_a$ and metadata, let frames $U_a(\alpha)$ vary, and set $\mathbf{h}_a(\alpha) = U_a(\alpha)^\top \mathbf{g}_a$. Every unconditioned base score of Eqs.* (2)–(3) *is constant in α, and its derivative in α is zero wherever the frame path is differentiable.*

Proof. With $\mathbf{h}_a = U_a^\top \mathbf{g}_a$ and U_a orthogonal, the coordinate-dependent inputs become $d^{-1}\|\mathbf{g}_u - \mathbf{g}_c\|_2^2$, $\mathbf{g}_u^\top \mathbf{g}_c$, and $\|\mathbf{g}_a\|_2^2$, as in Proposition 11; the others are metadata. The feature vector, hence the score, is constant, and a constant function has zero derivative. □

Proposition 16 is the scoring half of the carry argument of Block D: a frame update followed by the exact carry of Eq. (5) leaves every score of Eq. (3) unchanged, so prediction losses cannot reward a pure change of coordinates. Gradients reach the frame network only through the coordinate-dependent content and diffusion updates of Proposition 11. The statement excludes the optional non-invariant inputs of Appendix E.1.

Proposition 17 (Orientation blindness of the role-free ablation)**.** *In the role-free ablation ($\boldsymbol{\sigma}_{w,j} = 0$), fix initialisation, parameters, timestamps, relations, and features, and reverse the orientation of admissible events without changing, within any wave, the relative insertion order of each node's buffer records (for example, when every node has at most one incidence per wave). Then the state $(\mathbf{h}, \mathbf{F}, s, \mathcal{B})$ after every wave is unchanged.*

Proof. Induct over waves. Reversal preserves each endpoint's incidences, partners, gaps, features, relation, and types, and with $\boldsymbol{\sigma} = 0$ the contexts of Eqs. (4) and (6) list the endpoint's own inputs first whatever its role, so every proposal, mean, carry, message, and GRU step agrees. Buffers store unordered pairs in the same relative order, so the retained graph and diffusion agree, and endpoint times are set identically. □

Proposition 17 explains the "roles removed from frame and message contexts" row of Table 19 (Panel A), subject to its buffer-order hypothesis. Without $\boldsymbol{\sigma}$, the frame and recurrent updates of Eqs. (4) and (6) use endpoint-relative inputs with no source/destination role. Reversing events can still change the lexicographic buffer insertion, and hence the retained graph, when buffers are truncated: with $B_{\max} = 1$, the wave $(1,2),(1,3),(2,3)$ retains $\{\{1,3\},\{2,3\}\}$, but reversing its first event retains $\{\{1,2\},\{2,3\}\}$. Directional information can therefore enter through that graph, the ordered decoder, and the oriented pair-recency input. The reference configuration also has role inputs.

Proposition 18 (Indistinguishable cold candidates)**.** *Under the common initialisation of the Block A control, fix a query source, relation, and time, and let all $M_q \geq 1$ candidates be distinct nodes with no earlier event, equal type, and no pair history with the source. All base scores are equal, and the average-tie reciprocal rank of the positive is $2/(M_q+1)$.*

Proof. Every candidate has the same memory and frame, gap $\Delta_\perp$, pair elapsed time $\Delta_\perp$, and type, so its feature vector, hence score, is the same. With $M_q - 1$ tied negatives, $r = 1 + (M_q-1)/2 = (M_q+1)/2$. □

Proposition 18 is an information-boundary check for Lemma 3: without history, identity, or type, the scorer of Eq. (3) cannot rank cold candidates better than an average tie. The reference configuration's seeded per-node initialisation breaks this premise, which is why the common-initialisation row of Table 3, Block A, is the relevant control for identity information.

B.8 Temporal composition of transport

For a fixed node v and two committed-state indices $i \leq j$, define the waiting map $\mathbf{W}_v^{j \leftarrow i} = (\mathbf{U}_v^j)^\top \mathbf{U}_v^i$. Composing spatial maps and waiting maps along a spacetime path cancels each intermediate frame with its transpose. For common initial and terminal node-state-index pairs the result depends only on those endpoints. Products of spatial event maps at different times, without waiting maps, need not telescope. For example, if an intermediate frame changes from I to a reflection H (reachable with $k = 1$), the uncancelled pair contributes $H \neq I$. The frozen-graph flatness statement does not assert flatness of an arbitrary product that omits these coordinate changes.

C The Conditional Separation Construction

This appendix defines every operation in the witness. It uses a receiver-only frame update based on the sender's raw local state; the full methodology instead uses aligned endpoint context and updates both endpoints. Their common ingredients are persistent frame state, carry-over, and transported messages.

The restricted shared-coordinate comparator has one state per node, common initialization $\mathbf{h}_0$, receiver-only messages, and

$$\mathrm{UPD}(\mathbf{h}, \mathbf{m}) = \mathbf{h} + g(\mathbf{h})\mathbf{m}, \qquad g : \mathbb{R}^d \to \mathbb{R} \text{ continuous}, \quad g(\mathbf{h}_0) = 0.$$

Messages depend continuously on current states and observed metadata; readouts use states alone, with no ungated side state or diffusion. The framed comparison adds persistent reflector parameters, ungated frame updates, exact carry-over, and transported messages. For nodes $\{v, a, b, c\}$, times $t_0 < t_1 < t_2$, and identical features and relation labels (omitted), compare the event sequences

$$\mathcal{G}_1 = ((c,a,t_0),(a,v,t_1),(b,v,t_2)), \qquad \mathcal{G}_2 = ((c,b,t_0),(a,v,t_1),(b,v,t_2)). \tag{13}$$

Theorem 19 (Conditional closed-gate separation)**.** *(a) Every shared-coordinate comparator just defined stays at $\mathbf{h}_0$ on every stream, regardless of its continuous message function. (b) There exist a continuous gate, frame update, message function, and initialization for the framed comparison with $d = 2, k = 1$ such that its final ambient state at v differs between $\mathcal{G}_1$ and $\mathcal{G}_2$ by Euclidean norm $137950/104223 > 0$.*

Part (a) is universal within its comparator and part (b) is existential; an ungated side-state model can also separate the streams.

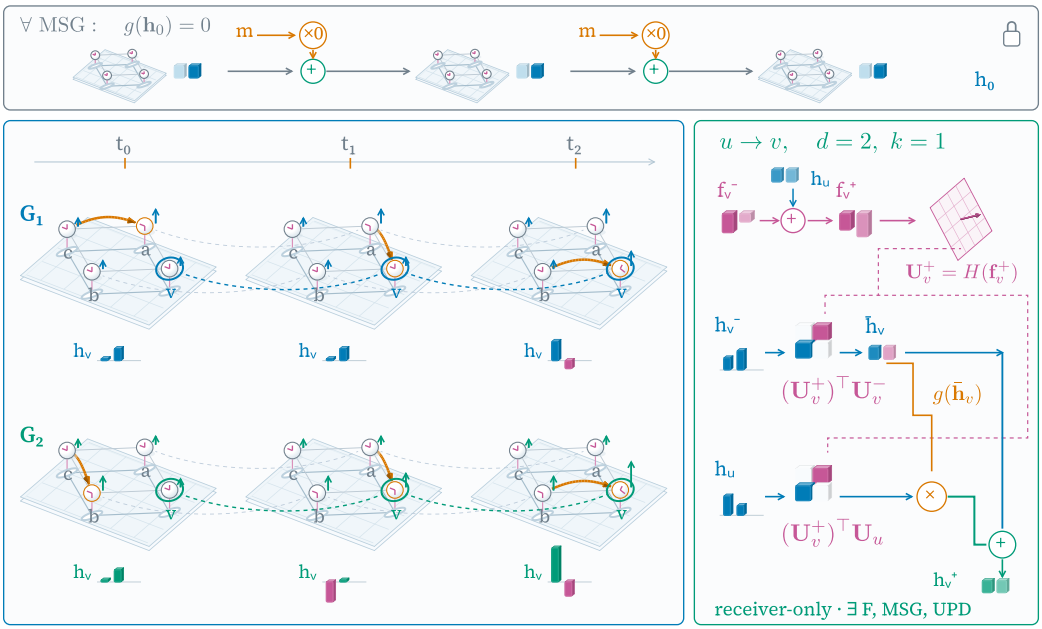


Figure 7: Theorem 19. Top: the closed-gate comparator ($g(\mathbf{h}_0) = 0$) never leaves $\mathbf{h}_0$. Left: the streams $\mathcal{G}_1, \mathcal{G}_2$ of Eq. (13); bars show v's local memory. Right: one framed receiver update ($d = 2$, $k = 1$). The final frames agree while the final ambient states differ.

Proof of Theorem 19(a). If all states equal $\mathbf{h}_0$ before an event, its receiver becomes $\mathbf{h}_0 + g(\mathbf{h}_0)\mathbf{m} = \mathbf{h}_0$ for any finite message. The other states do not change. Induction from the common initialization proves the assertion for all streams. A readout of those states alone cannot distinguish the two streams. □

Exact witness for Theorem 19(b). Use the standard basis $\mathbf{e}_1 = (1, 0)^\top$, $\mathbf{e}_2 = (0, 1)^\top$, and initialize $\mathbf{f}_w = \mathbf{e}_1$, $\mathbf{h}_w = \mathbf{e}_2$ at all four nodes. Set $k = 1$ and choose any $0 < \varepsilon_H < 1/2$. All reflector norms below exceed that threshold. Define

$$H(\mathbf{f}) = I - \frac{2\mathbf{f}\mathbf{f}^\top}{\mathbf{f}^\top\mathbf{f}}, \qquad g(\mathbf{h}) = \frac{\|\mathbf{h} - \mathbf{e}_2\|_2^2}{1 + \|\mathbf{h} - \mathbf{e}_2\|_2^2}.$$

The gate is continuous and $g(\mathbf{e}_2) = 0$. On event $u \to v$, superscripts "old" and "new" denote the two sides of this one receiver update:

$$\begin{aligned}
\mathbf{f}_v^{\text{new}} &= \mathbf{f}_v^{\text{old}} + \mathbf{h}_u^{\text{old}}, \\
\overline{\mathbf{h}}_v &= H(\mathbf{f}_v^{\text{new}})H(\mathbf{f}_v^{\text{old}})\mathbf{h}_v^{\text{old}}, \\
\mathbf{m}_v &= H(\mathbf{f}_v^{\text{new}})H(\mathbf{f}_u^{\text{old}})\mathbf{h}_u^{\text{old}}, \\
\mathbf{h}_v^{\text{new}} &= \overline{\mathbf{h}}_v + g(\overline{\mathbf{h}}_v)\mathbf{m}_v.
\end{aligned}$$

All other node states and frames remain unchanged. This is a shared linear frame increment and a continuous identity message in the transported sender argument. At the priming event, the receiver has $\mathbf{f} = (1, 1)^\top$, $\overline{\mathbf{h}} = (-1, 0)^\top$, $g(\overline{\mathbf{h}}) = 2/3$, and $\mathbf{m} = (-1, 0)^\top$. Its new local vector is $(-5/3, 0)^\top$, with ambient vector $(0, 5/3)^\top$. For completeness, the receiver v evolves as follows; the table gives its frame vector and ambient vector after each of the last two events:

Stream	Event	$\mathbf{f}_v$	$\mathbf{g}_v$
$\mathcal{G}_1$	$a \to v$	$(-2/3, 0)^\top$	$(0, 1)^\top$
$\mathcal{G}_1$	$b \to v$	$(-2/3, 1)^\top$	$(0, 85/49)^\top$
$\mathcal{G}_2$	$a \to v$	$(1, 1)^\top$	$(0, 5/3)^\top$
$\mathcal{G}_2$	$b \to v$	$(-2/3, 1)^\top$	$(0, 6505/2127)^\top$

To verify the last step explicitly, both final frames use $\mathbf{f}_v = (-2/3, 1)^\top$, for which $H(\mathbf{f}_v)\mathbf{e}_2 = (12/13, -5/13)^\top =: \mathbf{q}_*$. In $\mathcal{G}_1$, both the carried vector and the transported message are $\mathbf{q}_*$, and $g(\mathbf{q}_*) = 36/49$. Its final ambient vector is therefore $(1 + 36/49)\mathbf{e}_2 = (85/49)\mathbf{e}_2$. In $\mathcal{G}_2$, both arguments are $(5/3)\mathbf{q}_*$, and $g((5/3)\mathbf{q}_*) = 592/709$. Its final ambient vector is $(5/3)(1 + 592/709)\mathbf{e}_2 = (6505/2127)\mathbf{e}_2$. Thus the final frames agree and the final second coordinates differ by $6505/2127 - 85/49 = 137950/104223 > 0$. This proves the claimed ambient separation. □

An independent rational-arithmetic reconstruction recomputes every witness update. The numerical experiment of Section F.6 uses a different gate and message.

D Candidate Cross-Entropy and Reciprocal Rank

This result connects the candidate cross-entropy of Eq. (8) to reciprocal rank at the population level.

Proposition 20 (Matched-distribution ranking consistency)**.** *Fix a finite list of distinct candidates with one positive. Denote the true conditional probability that candidate c is positive by $\pi(c)$ and a predicted probability distribution on the same list by $p(c) > 0$. For a total order r, define $U_\pi(r) = \sum_c \pi(c)/\operatorname{rank}_r(c)$. Descending π order r_π maximizes this utility. Descending p order r_p, with ties resolved by a fixed rule independent of the true positive, obeys*

$$U_\pi(r_\pi) - U_\pi(r_p) \le \|\pi - p\|_1 \le \sqrt{2\,\mathrm{KL}(\pi\|p)}.$$

Here $\mathrm{KL}(\pi\|p) = \sum_c \pi(c)\log(\pi(c)/p(c))$, with $0\log 0 = 0$. Thus vanishing excess expected candidate cross-entropy implies vanishing expected reciprocal-rank regret under a matched candidate distribution and this total-order tie policy.

Proof. If $\pi(a) \ge \pi(b)$ but a follows b, swapping the pair changes utility by $(\pi(a) - \pi(b))(1/\operatorname{rank}(b) - 1/\operatorname{rank}(a)) \ge 0$. Removing inversions proves optimality. Add and subtract utilities under p:

$$U_\pi(r_\pi) - U_\pi(r_p) = [U_\pi(r_\pi) - U_p(r_\pi)] + [U_p(r_\pi) - U_p(r_p)] + [U_p(r_p) - U_\pi(r_p)].$$

The middle term is nonpositive. Each remaining term is bounded above by total variation $\|\pi - p\|_1/2$, since $U_\pi(r) - U_p(r) = \sum_c(\pi(c) - p(c))/\operatorname{rank}_r(c)$ has weights in $(0, 1]$ and $\sum_c(\pi(c) - p(c)) = 0$. For completeness, the KL bound follows by grouping coordinates into $A = \{c : \pi(c) \ge p(c)\}$ and its complement. Write $a = \pi(A)$ and $b = p(A)$. The log-sum inequality gives $\mathrm{KL}(\pi\|p) \ge a\log(a/b) + (1 - a)\log((1 - a)/(1 - b))$. For $b \in (0, 1)$ this binary divergence, as a function of a, has value and first derivative zero at $a = b$ and second derivative $1/a + 1/(1 - a) \ge 4$. It is therefore at least $2(a - b)^2$ by integration, including the endpoint limits. Since $a - b = \|\pi - p\|_1/2$, this proves $\|\pi - p\|_1 \le \sqrt{2\,\mathrm{KL}(\pi\|p)}$. If $\pi = p$, the bound is immediate. Candidate cross-entropy minus the true conditional entropy equals KL; taking expectations and using Jensen's inequality proves the population statement. □

An average-tie evaluator uses a different utility. At $p = \pi = (1/2, 1/2)$, a total order gives expected reciprocal rank $3/4$, while average-tie ranks give $2/3$, despite zero KL. Therefore the result cannot be applied unchanged to average-tie MRR. Likewise, candidate-distribution shift between training and evaluation requires a separate argument. Neither the result nor its bound implies monotone finite-sample validation MRR.

Proposition 21 (A strict posterior margin removes tie ambiguity)**.** *For a fixed list of at least two distinct candidates, suppose the posterior probabilities are distinct with $\gamma = \min_{i \ne j} |\pi_i - \pi_j| > 0$. If $\mathrm{KL}(\pi\|p) < \gamma^2/2$, descending p is exactly the strict posterior order, which maximises expected reciprocal rank under both strict-order and average-tie evaluation.*

Proof. Pinsker's inequality, proved within Proposition 20, gives $\|p - \pi\|_1 < \gamma$. If $\pi_i > \pi_j$, then $p_i - p_j \ge \pi_i - \pi_j - |p_i - \pi_i| - |p_j - \pi_j| > \gamma - \gamma = 0$, so the order is correct and p has no ties. The posterior order is optimal among strict orders by Proposition 20. A tied block S occupying positions R_S earns $(\sum_{j \in S} \pi_j)/\bar{r}_S$ under average ties, with $\bar{r}_S$ the mean of R_S, while a uniformly random strict order inside the block earns $(\sum_{j \in S} \pi_j)|S|^{-1}\sum_{r \in R_S} 1/r \ge (\sum_{j \in S} \pi_j)/\bar{r}_S$ by convexity of $1/r$. Some strict order is therefore at least as good as any tied prediction, which proves the claim. □

Proposition 21 complements the tie counterexample above: when the posterior has a strict margin and the candidate cross-entropy of Eq. (8) is small enough, the learned order is the Bayes order and the average-tie convention of the reported MRR cannot change the conclusion.

E Algorithmic Details and Optional Components

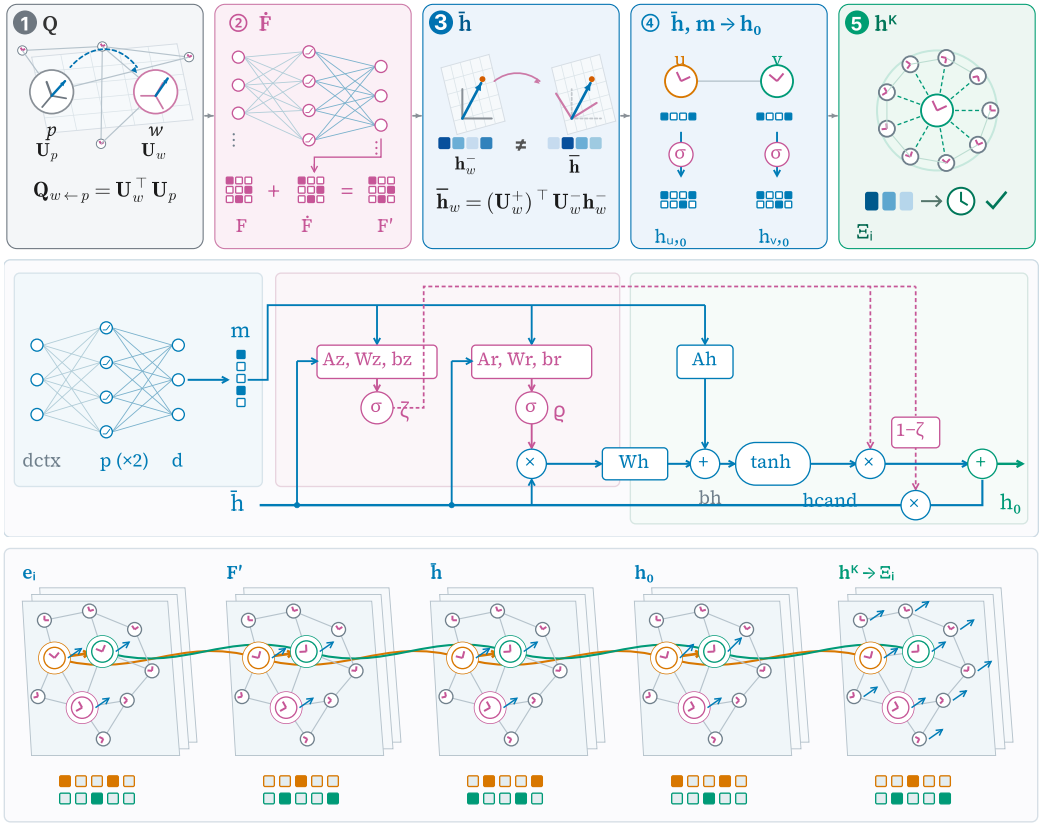


Figure 8: Inside one commit, for one incidence. Top: (1) transport $\mathbf{Q}_{w\leftarrow p} = \mathbf{U}_w^\top \mathbf{U}_p$; (2) frame increment $\dot{\mathbf{F}}$ from f_{frame} added to $\mathbf{F}$; (3) exact carry $\bar{\mathbf{h}}_w = (\mathbf{U}_w^+)^\top \mathbf{U}_w^- \mathbf{h}_w^-$; (4) messages $\mathbf{m}$ and GRU states $\mathbf{h}_{u,0}, \mathbf{h}_{v,0}$; (5) K diffusion rounds and the commit to Ξ^i. Middle: f_{msg} ($d_{\text{ctx}} \to p \to p \to d$, two hidden layers of width p) and the GRU gates $\boldsymbol{\zeta}$, $\boldsymbol{\varrho}$; the model applies the reset gate after W_h (Appendix E). Bottom: the state after each step. A wave averages the proposals and messages of all its incidences (Eqs. (4) and (6)).

Figure 8 expands Blocks C–E of Section 3 for one incidence: the carry of Eq. (5) comes before the content update of Eq. (6), so learned changes act on a vector that already represents the old ambient memory. Figure 9 shows the information boundary of Lemma 3: scores use only Ξ_t^- and fixed parameters, and labels of a wave reach the state only through its commit. Panel (c) is the truncated backpropagation through time (BPTT) over connected chunks of Algorithm 1 and Eq. (8), and panel (d) is the frozen replay behind every reported validation and test MRR. The algorithms specify the states, graph construction, and event updates of Section 3. The supplementary comparisons include variants with different recurrence and diffusion rules, as detailed below.

For previous local state $\mathbf{h} \in \mathbb{R}^d$ and message $\mathbf{m} \in \mathbb{R}^d$, the shared gated recurrent unit (GRU) is

$$\begin{aligned}
\boldsymbol{\zeta} &= \operatorname{sigm}(A_z \mathbf{m} + W_z \mathbf{h} + \mathbf{b}_z), \\
\boldsymbol{\varrho} &= \operatorname{sigm}(A_r \mathbf{m} + W_r \mathbf{h} + \mathbf{b}_r^{\text{GRU}}), \\
\mathbf{h}_{\text{cand}} &= \tanh(A_h \mathbf{m} + \mathbf{b}_h + \boldsymbol{\varrho} \odot (W_h \mathbf{h} + \mathbf{b}_h')), \\
\text{GRU}(\mathbf{h}, \mathbf{m}) &= \boldsymbol{\zeta} \odot \mathbf{h} + (\mathbf{1}_d - \boldsymbol{\zeta}) \odot \mathbf{h}_{\text{cand}}.
\end{aligned}$$

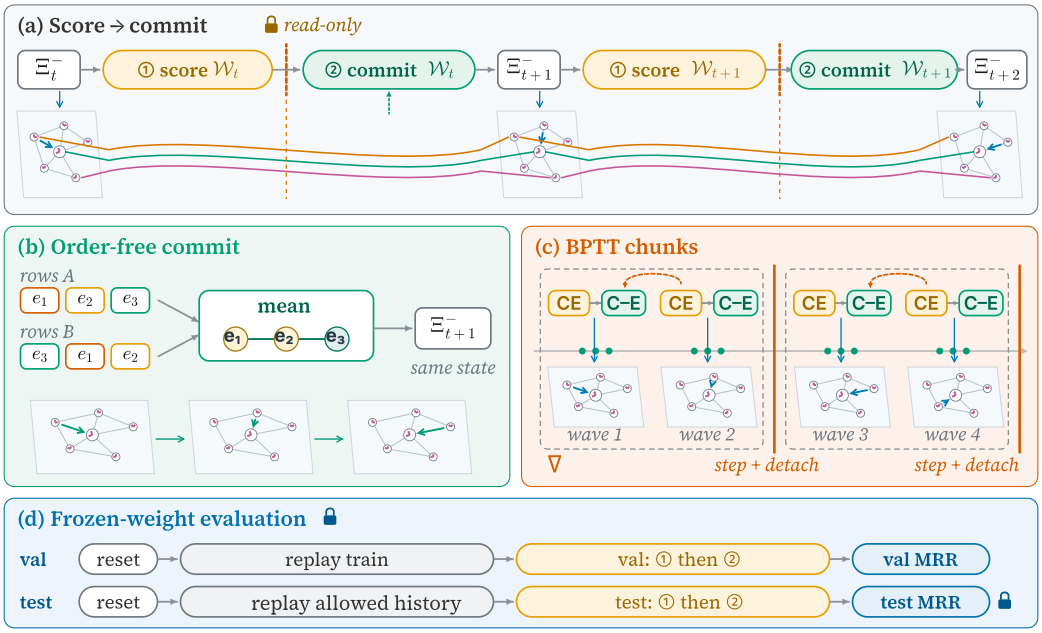


Figure 9: Protocol. (a) Each wave $\mathcal{W}_t$ is scored from the read-only snapshot Ξ_t^-, then committed to give the post-wave state read at the next observed timestamp. (b) The commit averages over incidences and inserts buffer records in canonical order, so any order of the events within a wave gives the same state (Block C). (c) Truncated back-propagation over chunks of two waves: the cross-entropy of each wave accumulates, then the optimiser steps and the state is detached. (d) Frozen-weight evaluation: reset, replay the allowed history, then score and observe every validation or test wave.

Algorithm 1 A complete scoring wave and its state commits.

Require: State $\Xi^{b(t)}$; queries and candidate lists at t; fixed parameters; the events of $\mathcal{W}_t$.
Ensure: Frozen scores for every query and state after all events at t.

1: Create a read-only snapshot $\Xi_t^- = \Xi^{b(t)}$, including all histories and the pair-recency table.
2: **for** each query q at t **do**
3: **for** each candidate c in $\mathcal{C}_q$ **do**
4: Compute $s_q(c)$ from Eqs. (2)–(3) using Ξ_t^-.
5: **end for**
6: **end for**
7: Finish all wave scores before using any wave label for an update.
8: If training, accumulate the candidate cross-entropy of Eq. (8) from these scores and the revealed positives.
9: Compute every incidence's frame increment from Ξ_t^- by Eq. (4); average them per endpoint, add, and renormalise the rows.
10: Install each endpoint's $\mathbf{F}_w^i$ and carry its memory once by Eq. (5).
11: Compute every incidence message, average them per endpoint, and apply one GRU step per endpoint by Eq. (6).
12: Append the wave to the endpoint buffers (truncating to $B_{\max}$ records) and rebuild the retained history graph G_i.
13: **for** $\ell = 0, \ldots, K-1$ **do**
14: Compute every round-$\ell+1$ state of G_i from round-ℓ states by Eq. (7).
15: **end for**
16: Commit $\mathbf{h}_a^i = \mathbf{h}_{a,K}^i$ for every $a \in V_i \cup W_i$, with $\mathbf{h}_{a,K}^i = \mathbf{h}_{a,0}^i$ off the graph (so every endpoint is persisted even when $B_{\max} = 0$ and G_i is empty), and every endpoint frame; update endpoint times and the pair-recency table.
17: At a training chunk boundary only: complete the gradient of Eq. (8) with the actual event count, update parameters, and detach the retained state.

Here $A_z, A_r, A_h, W_z, W_r, W_h \in \mathbb{R}^{d\times d}$ and all biases are in $\mathbb{R}^d$; $\odot$ denotes elementwise multiplication, and sigmoid and tanh act elementwise; the reset gate acts after W_h.

At the beginning of an epoch reset node memories, frames, timestamps, buffers, and the pair-recency table to the declared initialization; learned weights remain at their current values. This multi-epoch policy preserves the within-wave barrier, but does not in general satisfy the parameter-history hypothesis of Lemma 3 for replayed training timestamps. Form chunks of L_c complete waves ($L_c = 2$ in the reference training of Section 3; the configurations of Tables 11 and 14 list their own L_c). Within a chunk, the state of every touched node, including neighbours updated by diffusion, stays differentiable, so a later query can backpropagate through earlier commits in the same chunk. Candidate losses are accumulated after each wave's scores are fixed, and parameters stay fixed throughout the chunk. After its final commit, the loss of Eq. (8) is averaged over the chunk's events, the parameters are updated, and the state is detached; this is truncated backpropagation through time (BPTT) over chunks.

At evaluation, disable optimization and dropout. Reset state and replay the allowed earlier history using the selected checkpoint before scoring the next split; test replays the training and validation events. Preserve the equal-time barrier during replay and scoring, and separate validation-based model selection from test evaluation.

E.1 Optional scoring components

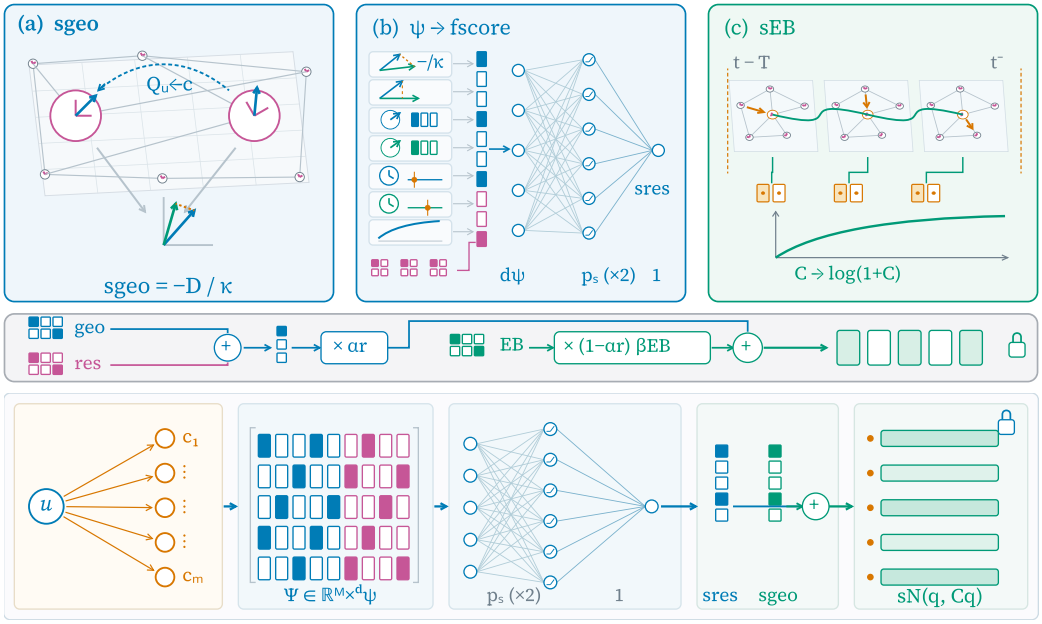


Figure 10: Scoring paths. (a) The geometric score $s_{\text{geo}} = -D/\kappa$ compares $\mathbf{h}_u$ with $\mathbf{Q}_{u\leftarrow c}\mathbf{h}_c$. (b) The residual MLP f_{score} maps $\boldsymbol{\psi}_q(c) \in \mathbb{R}^{d_\psi}$ through hidden width p_s to s_{res} from the inputs of Eq. (3), with two hidden layers. (c) The opt-in EdgeBank count mixture uses $\log(1+C)$ of an oriented pair count, drawn over a window $[t-T, t)$; every reported run uses unlimited counts. Middle: $s_{\text{N}} = s_{\text{geo}} + s_{\text{res}}$ is mixed with s_{EB} by α_r and $(1-\alpha_r)\beta_{\text{EB}}$ (Eq. (19)). Bottom: all M_q candidates of a query are scored in one batch $\Psi \in \mathbb{R}^{M_q\times d_\psi}$ from the frozen snapshot.

Figure 10 separates the three scoring paths of Block B. The geometric and residual paths use only the scalar inputs of Eq. (3), which are invariant under the gauge action of Proposition 9; this is the invariance boundary stated at the end of this subsection. The EdgeBank count mixture of Eq. (19) is enabled in every DGB benchmark run (Table 14) and in the tgbl-wiki and tgbl-review runs of Table 11, always with unlimited history; otherwise the pair history enters only through $\log(1+\Delta^{\text{pair}})$. Throughout, "EB" denotes this pair-recency input, and "EdgeBank" in the configuration tables denotes the count mixing.

Block C uses a fixed sinusoidal encoding of the elapsed endpoint gap. Choose a width d_{time} divisible by four (32 in the main configuration) and $n_\omega = d_{\text{time}}/4$; for $j = 0, \dots, n_\omega - 1$,

$$\phi(\Delta) = \big(\sin(\omega_j \log(1+\Delta))\big)_{j<n_\omega} \,\big\|\, \big(\cos(\omega_j \log(1+\Delta))\big)_{j<n_\omega}, \qquad \omega_j = 10000^{-4j/d_{\text{time}}}, \tag{14}$$

so $\phi(\Delta) \in \mathbb{R}^{d_{\text{time}}/2}$, and the frame and message contexts concatenate the encodings of both endpoint gaps directly. For $d_{\text{time}} = 0$ the encoding is omitted. The fixed timestamp unit makes Δ the numerical gap in that unit, and subtraction must precede any lossy cast. This is a chosen encoding related to Xu et al. (2020), not a guarantee of predictive usefulness. A matrix output of a network, as in Eq. (4), is a rowwise reshape of its final vector.

The reference training of Section 3 uses the candidate cross-entropy alone and none of the auxiliary terms below. The DGB setup of Table 18 and the DGB benchmark tables uses binary classification. With $y_{qj} = \mathbf{1}[j = j_q^+]$ and $\widetilde{D}_q(c)$ the scorer's mean squared distance after any enabled attention, its objective is the average over the events of an optimiser chunk of

$$\begin{aligned} &\ell_{\text{BCE},q} + \lambda_{\text{geo}}\ell_{\text{geo},q} + \lambda_{\text{res}}\ell_{\text{res},q} + \lambda_s\ell_{\text{smooth},q} + \lambda_e\ell_{\text{energy},q}, \\ &\ell_{\text{BCE},q} = \frac{1}{M_q}\sum_j \big[\,\text{softplus}(s_q(c_{qj})) - y_{qj}s_q(c_{qj})\big]. \end{aligned} \tag{15}$$

With $a_{qj} = -\operatorname{clip}(\widetilde{D}_q(c_{qj})/\tau_{\text{geo}}, -50, 50)$ and temperature $\tau_{\text{geo}} > 0$,

$$\ell_{\text{geo},q} = -\log\frac{\exp(a_{qj_q^+})}{\sum_j \exp(a_{qj})}, \qquad \ell_{\text{res},q} = \frac{1}{M_q}\sum_j s_{\text{res},q}(c_{qj})^2, \tag{16}$$

and, for the two endpoint proposals and post-diffusion states,

$$\ell_{\text{smooth},q} = \frac{\|\dot{\mathbf{F}}_{u,q}\|_F^2 + \|\dot{\mathbf{F}}_{v,q}\|_F^2}{kd}, \qquad \ell_{\text{energy},q} = \frac{1}{d}\big\|(\mathbf{U}_v^+)^\top\mathbf{U}_u^+\mathbf{h}_u^+ - \mathbf{h}_v^+\big\|_2^2. \tag{17}$$

with $\ell_{\text{smooth},q} := 0$ when $k = 0$. In the compound no-transport intervention of Table 18, the energy term uses identity transport. The weight λ_{res} is the reported λ_{bias}. These penalties encourage their stated quantities; they neither enforce a hard norm bound nor guarantee nonzero gradients in every configuration.

The core score of Eq. (2) uses the unconditioned transport $\mathbf{Q}^-_{u\leftarrow c}$. An optional relation-conditioned comparison, enabled in no reported run, uses learned $\gamma_r \in \mathbb{R}$ and $\mathbf{b}_r \in \mathbb{R}^{k\times d}$:

$$\mathbf{F}^-_{u|r} = \mathbf{F}^-_u + \text{sigm}(\gamma_r)\mathbf{b}_r, \quad \mathbf{U}^-_{u|r} = \mathcal{U}(\mathbf{F}^-_{u|r}), \quad \mathbf{Q}^{-,r}_{u\leftarrow c} = (\mathbf{U}^-_{u|r})^\top\mathbf{U}^-_c. \tag{18}$$

When this option is enabled, every occurrence of $\mathbf{Q}^-_{u\leftarrow c}$ in Eqs. (2) and (3) is replaced by $\mathbf{Q}^{-,r}_{u\leftarrow c}$; with it disabled, $\mathbf{Q}^{-,r} = \mathbf{Q}^-$ exactly. The time argument is suppressed within a single query. This is a scoring transformation: it does not change the stored source frame, does not require carry-over, is not merely a re-expression of the same source vector, is not the diffusion transport, and its reverse-source version need not be the transpose.

The pair-recency (EB) channel and EdgeBank count mixing are separate; the subscript in s_{EB} and β_{EB} refers to the latter. Pair recency is the single input $\log(1 + \Delta^{\text{pair}}_{u,c,r}(t))$ of Eq. (3), on by default and removed in TSNN-core. EdgeBank count mixing, enabled where the configuration tables say so, uses the unlimited count $C^-_{u,c,r}(t)$ of earlier events with this oriented triple; finite-window counts are not used by any reported run. Both read one table frozen with Ξ_t^- and updated only after the wave is scored. With $s_{\text{EB},q}(c) = \log(1 + C^-_{u,c,r}(t))$, $\alpha_r = \text{sigm}(a_r)$, and learned $a_r, \beta_{\text{EB}} \in \mathbb{R}$ (β_{EB} initialised at 1), the mixed score replaces $s_q(c)$ of Eq. (2b) by

$$s_q(c) = \alpha_r s_{\text{N},q}(c) + (1-\alpha_r)\beta_{\text{EB}}s_{\text{EB},q}(c), \qquad s_{\text{N},q}(c) = s_{\text{geo},q}(c) + s_{\text{res},q}(c). \tag{19}$$

The disabled setting is $\alpha_r = 1$ exactly. For a head query, substitute $u = c$ and destination v in every pair feature, including the source-conditioned relation bias and the oriented count $C^-_{c,v,r}(t)$.

Figure 11 shows the three optional reads defined in this paragraph and in the EB paragraph above. They read the frozen snapshot Ξ_t^-; the attention output and coordinate features are not gauge-invariant, so the invariance boundary below does not cover them. The main training setup excludes the attention and common-neighbour extensions; the pair-recency channel is enabled for the matched controls in Table 3, Block C, and the additional controls in Table 19, Panels A–B. The base residual

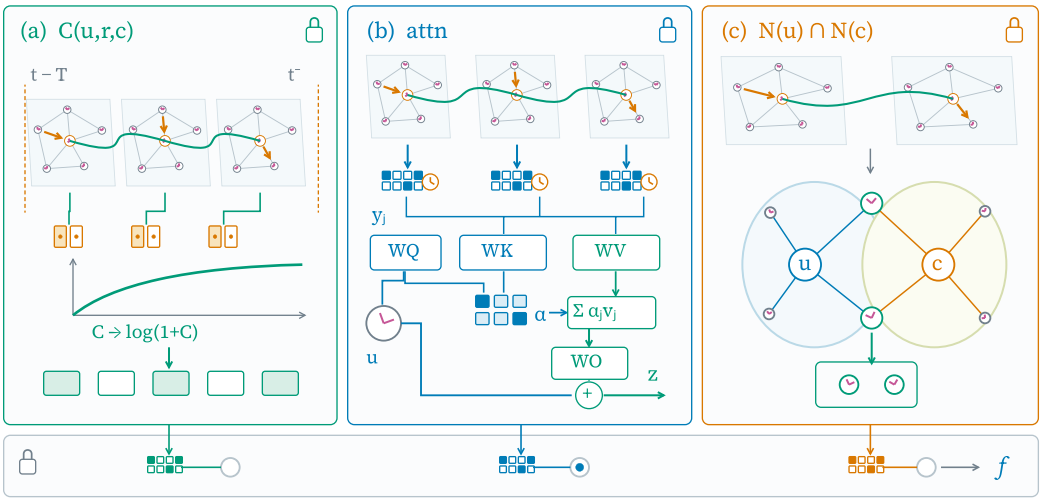


Figure 11: Optional history reads from the DGB experiments' pre-event state. (a) An oriented count $C(u, r, c)$ over a window $[t - T, t)$, mapped to $\log(1 + C)$; reported runs use unlimited counts. (b) The attention head over tokens $\mathbf{y}_j$ of the source's recent occurrences, with W_Q, W_K, W_V, weights α_j, output projection W_O, and residual result $\mathbf{z}$. (c) Common-neighbour features from $\mathcal{N}(u) \cap \mathcal{N}(c)$. Each read adds inputs to the scorer.

inputs are exactly Eq. (3), whose endpoint elapsed times are scalar features. The implemented optional scorer extensions are projected coordinate differences and common-neighbour features. The coordinate channel (DGB setup; tgbl-review in Table 11) appends $P_{\mathrm{cd}}\big((\mathbf{U}_c^-)^\top \mathbf{U}_u^- \widetilde{\mathbf{h}}_u - \mathbf{h}_c^-\big)$ with $P_{\mathrm{cd}} = \mathrm{GELU} \circ \mathrm{LN} \circ \mathrm{Affine}_{d \to d_{\mathrm{cd}}}$, where $\widetilde{\mathbf{h}}_u$ includes any enabled attention output and the projection applies dropout; it adds d_{cd} inputs, which are coordinate-dependent, not invariant scalars.

An optional attention head reads the last at most $B_{\mathrm{attn}} \le B_{\max}$ source history occurrences in the query snapshot. Occurrence j at source u has partner w_j and, in the DGB setup, age argument $a_j = \log(1 + \max\{t - t_j, 0\})$. Define tokens

$$\mathbf{y}_j = \mathbf{Q}^-_{u \leftarrow w_j}(t)\mathbf{h}^-_{w_j}(t) + A_t \phi(a_j) + \mathbf{b}_t + \mathbf{e}^{\mathrm{attn}}_{r_j}, \qquad A_t \in \mathbb{R}^{d \times d_{\mathrm{time}}/2},$$

with a separate relation embedding $\mathbf{e}^{\mathrm{attn}}_r \in \mathbb{R}^d$; since ϕ itself applies $\log(1 + \cdot)$, the implemented age input is a double logarithm. Past occurrences can repeat the same neighbour; their tokens use the neighbor's current pre-query memory and the occurrence's own time and relation. Choose a head count H_{attn} dividing d, with head width $d_h = d/H_{\mathrm{attn}}$. For head h, matrices $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d_h \times d}$ give

$$q^{(h)} = W_Q^{(h)} \mathbf{h}_u^-, \quad k_j^{(h)} = W_K^{(h)} \mathbf{y}_j, \quad v_j^{(h)} = W_V^{(h)} \mathbf{y}_j, \qquad a_j^{(h)} = \frac{\exp((q^{(h)})^\top k_j^{(h)} / \sqrt{d_h})}{\sum_l \exp((q^{(h)})^\top k_l^{(h)} / \sqrt{d_h})}.$$

Concatenate $\sum_j a_j^{(h)} v_j^{(h)}$ across heads, multiply by $W_O \in \mathbb{R}^{d \times d}$, and add $\mathbf{h}_u^-$. The projections are affine, training applies dropout to the attention weights, and the result is always normalised, $\mathbf{h}_u^{\mathrm{attn}} = \mathrm{LN}(\mathbf{z})$, where $\mathbf{z}$ is that sum and $\mathrm{LN}(\mathbf{z}) = \boldsymbol{\gamma} \odot (\mathbf{z} - \mu \mathbf{1}_d)/\sqrt{v + c_{\mathrm{LN}}} + \boldsymbol{\beta}$, $\mu = d^{-1} \sum_l z_l$, $v = d^{-1} \sum_l (z_l - \mu)^2$, $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^d$, and $c_{\mathrm{LN}} > 0$. With no recorded neighbours, the module attends to one learned null memory and null frame, without age or relation inputs. Replacing the source vector in scoring by this result does not write node memory.

The DGB common-neighbour (CN) channel has seven features: $\log(1 + |C_1|)$, the one-hop Jaccard similarity, $\sum_{w \in C_1} 1/\log(d_w + 2)$, a direct-neighbour indicator, $\log(1 + |C_2|)$, an approximate two-hop Jaccard similarity, and $\sum_{w \in C_1} 1/(d_w + 1)$. Here C_1 intersects the endpoints' distinct buffered neighbours, d_w counts occupied buffer slots, and the two-hop sets truncate each neighbour list to ten entries. The features carry no explicit recency weighting and are used only by the DGB setup, the tgbl-wiki configuration of Table 11, and the $d = 192$ edge-action run of Table 12.

Under $\mathbf{h}'_a = \mathbf{R}_a^\top \mathbf{h}_a$ and $\mathbf{U}'_a = \mathbf{U}_a \mathbf{R}_a$, unconditioned transport obeys $\mathbf{Q}'_{a\leftarrow b} = \mathbf{R}_a^\top \mathbf{Q}_{a\leftarrow b} \mathbf{R}_b$. Aligned inner products, distances, and norms in the base residual are therefore invariant, while metadata is unchanged. A scalar MLP of these invariant inputs preserves that invariance. This establishes an invariant score for the unconditioned base decoder and frame-free count mixture. This null test changes frames and memories jointly; changing $\mathbf{U}_a$ while holding $\mathbf{h}_a$ fixed generally changes the represented memory and is not a coordinate-invariance test, and arbitrary rotations may need a matrix-form decoder test because a fixed reflector count need not represent them.

Here $\mathbf{e}_1, \mathbf{e}_2$ denote the first two standard basis vectors of $\mathbb{R}^d$, for $d \geq 2$, and H denotes the nonsafeguarded reflection formula in Appendix C. The raw relation bias in Eq. (18) has no such automatic law: $H(\mathbf{e}_1) = H(2\mathbf{e}_1)$ but $H(\mathbf{e}_1 + \mathbf{e}_2) \neq H(2\mathbf{e}_1 + \mathbf{e}_2)$ for $d \geq 2$. Thus identical base frames can produce different biased frames from different raw parameters. A relation-conditioned scalar score is invariant only if its conditioned transport transforms coherently. Generic attention projections, coordinatewise normalization, raw vector features, frame MLPs, and GRUs need not satisfy this law. The full model consequently has explicit coordinate dependence; the scalar result is not a proof of whole-model equivariance.

E.2 Compute and memory derivation

Updating the endpoint buffers costs $O(B_{\max})$ per incidence, and rebuilding the retained graph G_i from its pair counts costs $O(|E_i| \log |E_i|)$ per wave, with $|E_i| \leq B_{\max}|V_i|$ because every edge holds at least one buffer record; the rebuild is skipped when $K = 0$. Reading and writing active node vectors adds $O(|V_i|d)$ work. The scoring snapshot can be a read-only view, so no full-state copy is charged per query or event. For $M_{\mathrm{ev}} = \sum_q M_q$ candidate positions across an event's queries, applying source and destination frames costs $O(M_{\mathrm{ev}}kd)$ and the candidate distances $O(M_{\mathrm{ev}}d)$, which remains when $k = 0$; carry-over and the two frame-update contexts add $O(kd)$ transports per event. Each of K synchronous rounds traverses both directions of every retained edge and applies every active frame, so diffusion costs $O(K(|V_i|k + |E_i|)d)$ per wave, which remains positive when $k = 0$. The one-hidden-layer frame MLP with hidden width p maps d_{ctx} inputs to kd outputs at cost $O(p(d_{\mathrm{ctx}} + kd))$. The corresponding message MLP costs $O(p(d_{\mathrm{ctx}} + p + d))$ with its two hidden layers, and the GRU costs $O(d^2)$. The two-hidden-layer residual MLP costs $O(M_{\mathrm{ev}}p_s(d_\psi + p_s + 1))$, and candidate-loss evaluation adds $O(M_{\mathrm{ev}})$; these are absorbed by the displayed terms. For one event of a wave, the following expected bound adds the per-wave graph terms once; for a frame network with one hidden layer and a message network with two, both of hidden width p, and a two-hidden-layer scoring MLP of hidden width p_s:

$$\begin{aligned} O\big(&B_{\max} + |E_i| \log |E_i| + |V_i|d + M_{\mathrm{ev}}d + (M_{\mathrm{ev}} + 1)kd + K(|V_i|k + |E_i|)d \\ &+ p[2d_{\mathrm{ctx}} + p + (k+1)d] + d^2 + M_{\mathrm{ev}}p_s(d_\psi + p_s + 1)\big). \end{aligned} \tag{20}$$

The terms account for buffer updates and graph construction, active-memory access, transport, the frame and message MLPs, the GRU, and the residual scorer; the bound includes candidate scoring but not full sorted rankings. Computing one positive's rank needs only a linear scan of its candidate scores, whereas a full comparison sort adds $O(M_q \log M_q)$ per query. Candidate generation and task-specific filtering are separate costs. An identity transport still leaves these dense network costs unless the variant also removes those networks.

Optional attention adds $O(B_{\mathrm{attn}}kd + B_{\mathrm{attn}}d^2 + d^2)$ work, plus its temporal/relation projections; optional common-neighbor features require candidate-dependent history traversal. Edge-count lookup is expected $O(1)$ per candidate, and unlimited counts store one entry per distinct oriented triple. These costs are additional to the core bound.

Persistent node memories and raw frames require $(k + 1)d$ scalars per stored node. Histories require at most $B_{\max}$ pair records per node. A chunk touches every vertex of its retained graphs through diffusion; candidates can require additional read access. Training memory also grows with the chunk length, the number of candidates, the widths, and the diffusion depth.

F Datasets, Baselines, and Supplementary Experiments

Figure 12 lists the alternatives that a frame-specific claim must be compared with. The frame-free arm removes the frame channel while keeping the scorer of Eq. (3); fixed and shared frames keep frames

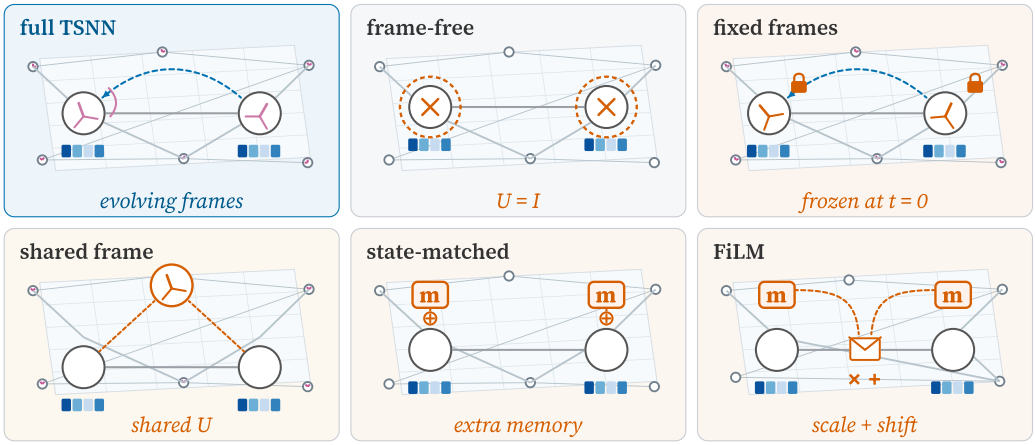


Figure 12: Matched controls of Table 3, Block C, and additional controls in Table 19: full TSNN with evolving frames; frame-free ($\mathbf{U} = \mathbf{I}$); fixed frames, frozen at initialisation for the whole run (distinct from a frozen frame *generator*, whose frames still evolve); one shared global frame; a side memory $\mathbf{m}$ of the same state size, not matched in parameters or compute; and FiLM, in which $\mathbf{m}$ scales and shifts the messages. The Block C arms share the scorer, data, initialisation of ambient memory, and protocol.

but remove their per-node evolution; the side-state and FiLM arms add persistent state of the same size without coordinates. Section 6 reports these arms, and Proposition 12 states when the frame-free arm must coincide exactly with the full-frame arm. This appendix follows the evidence in Section 6 in sequence: dataset and task definitions; comparator and metric conventions; TGB protocols, configurations, and supplementary results; DGB benchmark results; DGB ablations, sensitivity analyses, and diagnostics; and the constructed mechanism checks in Appendix F.6. Coverage comprises thirteen TGB/TGB 2.0 datasets, eight TGB-Seq datasets, and thirteen DGB datasets. The DGB benchmark tables (Tables 13–17) and the DGB control table (Table 18) report distinct experiment sets. Differences of reported ablation means and the separate drift study are in Appendix F.5.2.

F.1 Datasets and prediction tasks

Tables 7 and 9 give whole-dataset counts. Nodes are distinct observed identifiers, events are timestamped interaction records, and timestamps are distinct recorded time values. These are not interchangeable: $\mathcal{T}$ denotes the node-type vocabulary, $\mathcal{R}$ the relation vocabulary, $\mathcal{V}$ the node universe, and N the stream length (Section 2). The active diffusion vertex set V_i and neighbor multiset $\mathcal{N}_{a,i}$ are defined in Block E of Section 3; the latter is not a dataset node count. Absent explicit labels, the model uses a singleton vocabulary; bipartite candidate restrictions can still distinguish endpoint roles. DGB unique-pair counts discard time; the cited TGB summaries do not provide an equivalent column. TKG temporal counts exclude optional static relations.

The five `tgbl` tasks predict destinations: wiki and review concern editor–page and user–product interactions, coin concerns transfers, comment concerns user replies, and flight concerns airport traffic. For `tkgl`, smallpedia and wikidata contain yearly knowledge facts; polecat and icews contain daily political events. They evaluate head and tail prediction with a specified relation. The four `thgl` tasks evaluate typed destinations: software and github contain GitHub activity, forum contains user replies and subreddit posts, and myket contains user–application installations and updates. Their timestamps have second-level resolution (Huang et al., 2023; Gastinger et al., 2024a). Table 8 details the three TGB datasets whose TSNN configurations are listed in Table 11. The wiki-v2 and review-v2 updates increased evaluation-negative counts; the v2 suffix does not denote TGB 2.0 or establish a reshuffled split.[2]

The thirteen DGB datasets span interaction, communication, proximity, transportation, voting, and trade networks. Wikipedia/Reddit, MOOC, and LastFM have bipartite roles; the remaining tasks use

[2]TGB version history: `https://tgb.complexdatalab.com/docs/update/`.

Table 7: Published TGB/TGB 2.0 dataset statistics (Huang et al., 2023; Gastinger et al., 2024a). Counts refer to complete datasets, not active diffusion graphs or evaluation queries. A dash denotes no explicit node-type count in the benchmark summary; it does not mean zero node types.

Family	Dataset	Domain	# nodes	# events	# relation types	# node types	# timestamps
TGB link	tgbl-wiki	interaction	9,227	157,474	1	–	152,757
TGB link	tgbl-review	rating	352,637	4,873,540	1	–	6,865
TGB link	tgbl-coin	transaction	638,486	22,809,486	1	–	1,295,720
TGB link	tgbl-comment	social	994,790	44,314,507	1	–	30,998,030
TGB link	tgbl-flight	traffic	18,143	67,169,570	1	–	1,385
TGB 2.0 TKG	tkgl-smallpedia	knowledge	47,433	550,376	283	–	125
TGB 2.0 TKG	tkgl-polecat	political	150,931	1,779,610	16	–	1,826
TGB 2.0 TKG	tkgl-icews	political	87,856	15,513,446	391	–	10,224
TGB 2.0 TKG	tkgl-wikidata	knowledge	1,226,440	9,856,203	596	–	2,025
TGB 2.0 THG	thgl-software	software	681,927	1,489,806	14	4	689,549
TGB 2.0 THG	thgl-forum	social	152,816	23,757,707	2	2	2,558,457
TGB 2.0 THG	thgl-github	software	5,856,765	17,499,577	14	4	2,510,415
TGB 2.0 THG	thgl-myket	interaction	1,530,835	53,632,788	2	2	14,828,090

Table 8: Modalities of the three TGB datasets of the TSNN neural runs in Table 11. Dataset-version suffixes are distinct from the TGB 2.0 benchmark name.

Dataset	Structure	Event attributes	Dataset detail
tgbl-wiki-v2	editor–page bipartite	172-d LIWC text	v2 expands evaluation negatives
tgbl-review-v2	user–product bipartite	1-d star rating	ratings in $\{1, \ldots, 5\}$
thgl-software	heterogeneous GitHub activity	no edge-attribute vector	4 node types, 14 relation types, second-wise time

Table 9: DGB dataset statistics and temporal-profile indices (Poursafaei et al., 2022). Unique pairs discard timestamps and follow the stored endpoint convention. Feature widths describe dataset attributes, not necessarily model inputs. Nov., Reo., and Sur. are defined in the text.

Dataset	Domain	Bipartite	# nodes	# events	# unique pairs	# timestamps	Duration / features	Nov.	Reo.	Sur.
Wikipedia	editing	yes	9,227	157,474	18,257	152,757	1 month / 172-d LIWC	0.46	0.26	0.42
Reddit	posting	yes	10,984	672,447	78,516	669,065	1 month / 172-d LIWC	0.26	0.52	0.18
MOOC	education	yes	7,144	411,749	178,443	345,600	17 months / 4-d	0.75	0.02	0.79
LastFM	music	yes	1,980	1,293,103	154,993	1,283,614	1 month / –	0.28	0.30	0.35
Enron	email	no	184	125,235	3,125	22,632	3 years / –	0.30	0.22	0.27
Social Evo.	proximity	no	74	2,099,519	4,486	565,932	8 months / 2-d	0.11	0.51	0.02
UCI	messaging	no	1,899	59,835	20,296	58,911	196 days / –	0.73	0.01	0.56
Flights	transport	no	13,169	1,927,145	395,072	122	4 months (day) / 1-d	0.21	0.60	0.19
CanParl	politics	no	734	74,478	51,331	14	14 years (year) / 1-d	0.69	0.01	0.57
USLegis	politics	no	225	60,396	26,423	12	12 congresses / 1-d	0.44	0.08	0.45
UN Trade	economics	no	255	507,497	36,182	32	32 years (year) / 1-d	0.07	0.87	0.04
UN Vote	politics	no	201	1,035,742	31,516	72	72 years (year) / 1-d	0.03	0.93	0.01
Contact	proximity	no	694	2,426,280	79,531	8,065	1 month (5 min) / 1-d RSSI	0.42	0.44	0.12

general node pairs. DGB Flights is distinct from `tgbl-flight`, as are Reddit and `thgl-forum`. Novelty is the timestamp-average fraction of pairs not seen earlier. Reoccurrence is the fraction of distinct training pairs also present at test; surprise is the fraction of distinct test pairs absent from training. The latter two depend on the split used in the published profile, not on a newly chosen evaluation mask. These are pair-level indices, not fractions of unseen nodes (Poursafaei et al., 2022).

F.2 COMPARATORS, VARIANTS, AND METRICS

Table 10 compares the graph families on which TSNN and the strongest comparators are evaluated. TSNN runs the same architecture of Blocks A–E on all five collections (34 benchmark settings); node types and relations enter only through their embeddings and the schema $\mathcal{S}_{\text{type}}$ of Section 2. Each comparator other than TGN, which later benchmarks rerun on four families, has results on at most two families in Table 10, and several relational and heterogeneous methods exceed the time or memory budget on the largest TGB 2.0 datasets (Gastinger et al., 2024a).

Table 10: Evaluation coverage by graph family: datasets with a reported result / datasets on which the method has the best test score among all methods in Tables 1, 2, and 13. OOM/OOT: datasets on which the source reports running out of memory or time; –: no result under the protocol of these tables. Sources: TGB and TGB 2.0 (Huang et al., 2023; Gastinger et al., 2024a) and their leaderboards, TGB-Seq (Yi et al., 2025b), TPNet (Lu et al., 2024), CRAFT (Yi et al., 2025a), SDG (Duc & Ta, 2026), EdgeReMIND (Pollard, 2026), and STHN (Li et al., 2023).

Family (collection, datasets)	TSNN	TPNet	CRAFT	SDG	EdgeReMIND	STHN	TGN	SGNN-HN
Homogeneous (TGB `tgbl`, 5)	5 / 3	4 / 0	4 / 2	–	–	–	5 / 0	–
Knowledge graph (TGB 2.0 `tkgl`, 4)	4 / 4	–	–	–	4 / 0	–	–	–
Heterogeneous (TGB 2.0 `thgl`, 4)	4 / 3	–	–	–	4 / 0	1 / 1 (3 OOM)	2 / 0 (2 OOM)	–
Sequential (TGB-Seq, 8)	8 / 8	–	7 / 0	5 / 0	–	–	8 / 0	7 / 0
Dynamic graphs (DGB, 13)	13 / 11	13 / 2	–	–	–	–	13 / 0	–
Total (34)	34 / 29	17 / 2	11 / 2	5 / 0	8 / 0	1 / 1	28 / 0	7 / 0

JODIE projects recurrent node representations forward in time; DyRep models interaction intensities; TGN combines event memory with temporal neighborhood embeddings (Kumar et al., 2019; Trivedi et al., 2019; Rossi et al., 2020). TGAT uses temporal attention, CAWN uses causal anonymous walks, TCL uses a temporal transformer, and PINT uses expressive temporal message passing (Xu et al., 2020; Wang et al., 2021b;a; Souza et al., 2022). GraphMixer summarizes histories with MLP mixing, NAT maintains neighborhood caches, and DyGFormer uses patched histories and co-occurrence encodings (Cong et al., 2023; Luo & Li, 2022; Yu et al., 2023). TNCN scores temporal common neighbors, TPNet projects temporal-walk matrices, DyG-Mamba uses a state-space encoder, and CTAN models continuous-time dynamics (Zhang et al., 2025a; Lu et al., 2024; Li et al., 2025; Gravina et al., 2024). STHN targets heterogeneous temporal interactions; HyperEvent encodes relative structural patterns in event sequences (Li et al., 2023; Gao et al., 2025). EdgeBank memorizes prior pairs, either all of them (EdgeBank$_\infty$) or those in a recent time window (EdgeBank$_{\mathrm{tw}}$) (Poursafaei et al., 2022), while LocalGlobal combines local and global history statistics (Cornell et al., 2025). For TKGs, RE-GCN evolves relational graph representations (Li et al., 2021), CEN models evolution patterns, TLogic learns temporal rules, and RecB scores recurring facts. These TKG comparators complement the THG models and typed TGN variant (Gastinger et al., 2024a; Li et al., 2022b; Liu et al., 2022; Gastinger et al., 2024b).

CN denotes common-neighbor features and EB the pair-recency channel, which differs from the binary EdgeBank baseline. Every benchmark result in Tables 1, 2, and 13–17 uses the same TSNN architecture of Blocks A–E; on DGB it trains with the benchmark's binary objective (Appendix E.1). No result in Tables 1–3 and 19 uses the edge-dependent transport $\mathbf{U}_v^\top A_e \mathbf{U}_u$ of the earlier edge-action variant (Table 12). In the DGB tables, TSNN denotes the full model with the EB and CN channels enabled, and TSNN-core the same model without pair recency (Section 3.2) and, on DGB, without CN; the label "reference" names the comparison row of each table.

For a positive score s_q^+ and m_q negative scores s_{qj}^-, the reported average-tie convention is

$$r_q = 1 + \tfrac{1}{2}\left(\sum_{j=1}^{m_q} \mathbf{1}[s_{qj}^- > s_q^+] + \sum_{j=1}^{m_q} \mathbf{1}[s_{qj}^- \geq s_q^+]\right), \qquad \mathrm{MRR} = \frac{1}{N_q}\sum_{q=1}^{N_q} \frac{1}{r_q},$$

where N_q is the number of scored queries. Hits@10 is the mean of $\mathbf{1}[r_q \leq 10]$. For a list containing one positive and m_q negatives, $M_q = 1 + m_q$; N_q counts queries, not candidates. Head and tail queries are separate observations when both are evaluated.

AP denotes average precision. For distinct score thresholds in descending order, with precision P_l and recall R_l, its non-interpolated form is $\sum_l (R_l - R_{l-1})P_l$, with $R_0 = 0$. ROC AUC is the fraction of positive-negative pairs for which the positive has the larger score, with half credit for ties. Global AP over a concatenated split can differ from the mean of batch APs. "New-node AP" denotes the new-node/inductive test subset. AP/AUC with sampled negatives and filtered candidate MRR are different workloads. We write $\Delta\mathrm{AP} = \mathrm{AP}_{\mathrm{variant}} - \mathrm{AP}_{\mathrm{reference}}$; a difference reported in percentage points is $100\,\Delta\mathrm{AP}$, not a relative percentage change.

F.3 TGB PROTOCOLS AND SUPPLEMENTARY RESULTS

Table 1 combines selected official TGB leaderboard entries (Gastinger et al., 2024a), checked 26 September 2026 (tgbl: v2 evaluation files; tgbl-comment: v1), with author-reported EdgeReMIND results (Pollard, 2026). EdgeReMIND's numbers come from its paper; they are not listed on the official pages checked for this revision. The Additional column reports CRAFT on review/comment and CRAFT-R on coin/flight under the benchmark negatives (Yi et al., 2025a). Recency and popularity heuristics (LocalGlobal, LocalRec., LocalPop., GlobalRecency) follow Cornell et al. (2025).

The published TGB 2.0 protocol uses time-aware filtered MRR, evaluating both endpoints for TKGs and typed tail candidates for THGs. Smallpedia, polecat, and ICEWS use all-entity evaluation; wikidata and software use 1,000 sampled negatives, forum 100, and github and myket 20 (Gastinger et al., 2024a). Table 1 uses the official TGB negative samples and evaluator, with each timestamp wave scored before observing its positives. On thgl-software and thgl-github, TSNN's test MRR exceeds its validation MRR (0.5659 to 0.6449; 0.7507 to 0.8336), and more mildly on tgbl-review (0.4859 to 0.5400) and tgbl-coin (0.7992 to 0.8294). The published baselines show the same direction on these splits: on thgl-software, EdgeBank$_\infty$ rises from 0.399 to 0.449, TGN from 0.299 to 0.324, and TGN with edge types from 0.376 to 0.424; on thgl-github, EdgeBank$_{\mathrm{tw}}$ rises from 0.355 to 0.374 and EdgeBank$_\infty$ from 0.403 to 0.413 (Gastinger et al., 2024a). The test periods of these two datasets are therefore easier for history-based scoring, and every configuration is selected on validation alone.

For TGB-Seq Patent, an audit on 5,000 test queries shows the source of the large margin. On these queries, strict-past co-citation history alone reaches MRR 0.620 (Hits@10 0.618), and adding reverse citations and candidate in-degree reaches 0.744, against 0.246 for the best published method and 0.020 for an all-tied ranking. Raw candidate identifiers alone reach 0.194, because identifiers follow time; no negative appears among the source's earlier citations. The Patent margin in Table 2 therefore reflects benchmark structure that strict-past history exploits.

Table 11 gives the settings of the TSNN runs on wiki, review, and software, which used the DGB experiments' auxiliary losses (Appendix E.1). These configurations differ from the reference training of Section 3 in their losses, negatives, and chunk lengths, and their optional coordinate features and attention fall outside the base invariance analysis. The review input has two channels for its one-dimensional rating. d_{cd} is the output width of the coordinate-difference projection P_{cd}, and λ_{bias} is the λ_{res} of Eq. (16). The recent-neighbour budget counts distinct neighbours, whereas the core cap $B_{\max}$ counts event occurrences.

The tgbl-wiki edge-action rows of Table 12 are *not* the core model of Section 3: they come from an earlier, non-flat variant that replaces $\mathbf{Q}_{v\leftarrow u}$ by the edge-dependent transport $\mathbf{Q}^e_{v\leftarrow u} = \mathbf{U}_v^\top A_e \mathbf{U}_u$ in the frame proposals, messages, energy penalty, and scoring. Here A_e is a product of $n_{\mathrm{rot}} = 2$ rotations in learned orthonormalised 2-planes, each angle a relation offset plus $\pi \tanh$ of a linear map of the source's pre-event memory, log elapsed time, and relation and type embeddings; it is candidate-independent. Because A_e sits between the two frames, closed walks need not have identity transport, so Proposition 4 and the results built on it do not apply. The $d = 192$ run also enables pair recency and CN features; each width has one run (seed 42). No result in Tables 1–3 and 19 uses this variant.

Table 11: TSNN configurations for wiki, review, and software. "EdgeBank" denotes the count mixing of Eq. (19) with unlimited history, not the pair-recency input.

Symbol / Hyperparameter	tgbl-wiki	tgbl-review	thgl-software
Model hyperparameters (architecture)			
d (latent width)	256	256	256
Input edge-feature width	172	2	–
k (Householder reflectors)	4	4	4
d_{cd} (coordinate-difference projection width)	32	32	32
d_r (relation embedding dim)	–	–	64
d_{type} (node-type embedding width)	–	–	32
MLP hidden width	192	192	192
$\mathbf{1}[\text{CN feats}]$	T	F	F
$\mathbf{1}[\text{transported coordinate-difference feats}]$	F	T	F
$\mathbf{1}[\text{nbr. attn.}]$	F	F	T
$\mathbf{1}[\text{EdgeBank}]$ (mode)	T (unlim.)	T (unlim.)	F
Training hyperparameters			
K (diffusion steps)	3	2	3
η (diffusion step size)	0.01	0.005	0.01
Epochs	15	30	30
Early-stop patience	5	8	8
α_{lr} (AdamW base LR)	3×10^{-4}	1×10^{-4}	3×10^{-4}
$K_{\mathrm{neg}}^{\mathrm{train}}$ (neg. per pos. in $\mathcal{L}_{\mathrm{CE}}$)	200	500	400
Chunk length L_c	600	1024	300
Recent-neighbor budget	30	60	60
λ_{geo} (geo-branch reg. coeff.)	0.3	0.05	0.3
λ_{bias} (clipped residual-score penalty weight)	0.1	0.02	0.1
λ_s (smoothness weight)	10^{-4}	10^{-3}	10^{-4}
λ_e (energy weight)	10^{-4}	5×10^{-3}	10^{-4}
Temperature	0.1	0.1	0.1

Table 12: Earlier runs of the non-flat edge-action variant (not the core model) on tgbl-wiki, seed 42, 23,621 queries per split. Table 1 reports TSNN (0.8471 / 0.8290).

Model	Learned parameters	Val. MRR	Test MRR	Val. Hits@10	Test Hits@10
Edge-action variant (non-flat), $d = 128$, $k = 4$	638,898	0.8398	0.8199	0.9028	0.8862
Edge-action variant (non-flat), $d = 192$, $k = 4$	—	0.8447	0.8243	0.9054	0.8896

F.4 DGB BENCHMARK CONFIGURATIONS AND FULL RESULTS

This subsection reports TSNN on the thirteen DGB datasets against twelve published models, with transductive AP in Table 13. The experiments use a chronological 70/15/15 split, random negatives, and five-seed summaries, with baseline results from Lu et al. (2024). Mean ranks weight datasets equally and average tied ranks. Table 15 presents transductive ROC AUC; Tables 16 and 17 present the corresponding inductive results. The displayed means give TSNN 11 of 13 wins in both transductive metrics, with mean rank 1.23 in each; TPNet leads on UCI and Flights. In the inductive setting, TSNN leads on 12 of 13 datasets for both AP and ROC AUC, with mean ranks 1.15 and 1.08, respectively; TPNet leads on Flights. The transductive AP counts agree with Tables 10 and 1–2: 29 wins across the 34 benchmark settings. These counts compare reported means, not statistical significance.

Table 14 lists three training negatives for MOOC, CanParl, USLegis, and UCI, and one elsewhere. The LastFM and Enron input widths are model-input settings; these datasets have no native edge attributes.

Table 13: Transductive DGB test AP (%) across thirteen datasets and thirteen models. Five-seed means ± standard deviations; baselines from Lu et al. (2024).

Dataset	JODIE	DyRep	TGAT	TGN	CAWN	EdgeBank	TCL	GraphMixer	NAT	PINT	DyGFormer	TPNet	TSNN (EB+CN)
Wikipedia	96.50±0.14	94.86±0.06	96.94±0.06	98.45±0.06	98.76±0.03	90.37±0.00	96.47±0.16	97.25±0.03	98.03±0.07	98.45±0.04	99.03±0.02	99.32±0.03	**99.41±0.02**
Reddit	98.31±0.14	98.22±0.04	98.52±0.02	98.63±0.06	99.11±0.01	94.86±0.00	97.53±0.02	97.31±0.01	99.13±0.10	99.15±0.02	99.22±0.01	99.27±0.00	**99.72±0.02**
MOOC	80.23±2.44	81.97±0.49	85.84±0.15	89.15±1.60	80.15±0.25	57.97±0.00	82.38±0.24	82.78±0.15	85.88±0.55	88.08±0.86	87.52±0.49	96.39±0.09	**99.38±0.01**
LastFM	70.85±2.13	71.92±2.21	73.42±0.21	77.07±3.97	86.99±0.06	79.29±0.00	67.27±2.16	75.61±0.24	88.02±1.94	89.66±1.81	93.00±0.12	94.50±0.08	**94.51±0.03**
Enron	84.77±0.30	82.38±3.36	71.12±0.97	86.53±1.11	89.56±0.09	83.53±0.00	79.70±0.71	82.25±0.16	90.60±0.66	92.20±0.15	92.47±0.12	92.90±0.17	**97.73±0.02**
Social Evo.	89.89±0.55	88.87±0.30	93.16±0.17	93.57±0.17	84.96±0.09	74.95±0.00	93.13±0.16	93.37±0.07	88.92±3.45	94.42±0.03	94.73±0.01	94.73±0.02	**96.45±0.40**
UCI	89.43±1.09	65.14±2.30	79.63±0.70	92.34±1.04	95.18±0.06	76.20±0.00	89.57±1.63	93.25±0.57	93.40±0.26	96.45±0.11	95.79±0.17	**97.35±0.04**	97.06±0.02
Flights	95.60±1.73	95.29±0.72	94.03±0.18	97.95±0.14	98.51±0.01	89.35±0.00	91.23±0.02	90.99±0.05	98.57±0.12	98.80±0.02	98.91±0.01	**98.93±0.02**	98.83±0.01
Can. Parl.	69.26±0.31	66.54±2.76	70.73±0.72	70.88±2.34	69.82±2.34	64.55±0.00	68.67±2.67	77.04±0.46	79.72±1.76	68.36±1.43	97.36±0.45	90.28±0.37	**98.42±0.07**
US Legis.	75.05±1.52	75.34±0.39	68.52±3.16	75.99±0.58	70.58±0.48	58.39±0.00	69.59±0.48	70.74±1.02	78.71±0.87	74.85±0.97	71.11±0.59	80.58±0.23	**87.46±0.02**
UN Trade	64.94±0.31	63.21±0.93	61.47±0.18	65.03±1.37	65.39±0.12	60.41±0.00	62.21±0.03	62.61±0.27	73.95±1.16	70.20±0.58	66.46±1.29	87.24±0.65	**91.55±0.68**
UN Vote	63.91±0.81	62.81±0.80	52.21±0.98	65.72±2.17	52.84±0.10	58.49±0.00	51.90±0.30	52.11±0.16	70.45±0.68	66.25±0.78	55.55±0.42	75.12±0.29	**82.71±0.17**
Contact	95.31±1.33	95.98±0.15	96.28±0.09	96.89±0.56	90.26±0.28	92.58±0.00	92.44±0.12	91.92±0.03	97.39±0.22	98.64±0.02	98.29±0.01	98.66±0.01	**99.11±0.04**
Avg. Rank	8.85	9.77	9.54	6.04	8.00	11.54	10.77	9.31	5.15	4.73	4.12	1.96	1.23

Table 14: DGB configurations of the full TSNN model, distinct from the intervention configurations in Table 21. T/F denotes enabled/disabled; a dash denotes an inapplicable or unrecorded setting. The EdgeBank count mixing of Eq. (19) and the CN channel are enabled in every column, with unlimited history (window ratio 0).

Symbol / Hyperparameter	Wiki	Reddit	MOOC	LastFM	Enron	S. Evo.	Can. Parl.	US Legis.	UN Vote	Contact	UN Trade	UCI	Flights
Model hyperparameters (architecture)													
d (latent width)	192	192	192	192	128	128	192	192	128	192	192	192	192
d_{cd} (coordinate-difference projection width)	16	16	16	16	-	-	16	-	-	16	-	16	32
Input edge-feature width	172	172	4	2	32	2	1	1	1	1	1	0	1
k (Householder reflectors)	2	2	2	2	2	2	2	2	2	2	2	2	2
p_s (residual-score hidden width)	192	192	192	192	192	192	192	192	192	192	192	192	192
$\mathbf{1}$[nbr. attn.]	F	F	T	F	F	F	T	T	F	F	F	T	F
nbr. attn. history cap B_{attn}	-	-	15	-	-	-	15	15	-	-	-	15	-
$\mathbf{1}[\phi(\Delta)$ in $\psi]$	F	T	T	T	F	F	T	F	F	T	F	T	T
Training hyperparameters													
K (diffusion steps)	2	1	2	2	2	1	2	2	1	2	1	2	1
Epochs	25	25	25	25	25	25	25	25	25	25	25	25	25
Early-stop patience	5	5	5	5	5	5	5	5	5	5	5	5	5
η (diffusion step size)	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
λ_{geo} (geo-branch reg. coeff.)	0.3	0.3	0.5	0.3	0.3	0.3	0.5	0.5	0.3	0.3	0.3	0.3	0.5
μ_{geo} (geo-param LR multiplier)	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1.5	2.0	1.5	2.0	2.0
α_{lr} (AdamW base LR)	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}	5×10^{-4}	5×10^{-4}	8×10^{-4}	5×10^{-4}
Recent-neighbor budget	15	18	10	15	12	18	10	10	22	18	22	10	22
$K^{\mathrm{train}}_{\mathrm{neg}}$ (neg. per pos. in $\mathcal{L}_{\mathrm{CE}}$)	1	1	3	1	1	1	3	3	1	1	1	3	1
L_c (BPTT chunk length)	2000	2000	2000	2000	2000	2000	2000	2000	2000	1800	2000	1400	1800
λ_{bias} (clipped residual-score penalty weight)	0.1	0.1	0.1	0.1	0.07	0.1	0.1	0.07	0.1	0.1	0.15	0.05	0.12
λ_s (smoothness weight)	10^{-4}	10^{-4}	10^{-5}	10^{-4}	10^{-4}	5×10^{-4}	10^{-5}	5×10^{-4}	5×10^{-4}	10^{-4}	10^{-3}	10^{-4}	10^{-3}
$\mathbf{1}$[EdgeBank] (mode)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)	T (unlim.)
EdgeBank window ratio	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathbf{1}$[CN feats]	T	T	T	T	T	T	T	T	T	T	T	T	T

Table 15: Transductive DGB test ROC AUC (%), same models as Table 13. Five-seed means ± standard deviations; baselines from Lu et al. (2024).

Dataset	JODIE	DyRep	TGAT	TGN	CAWN	EdgeBank	TCL	GraphMixer	NAT	PINT	DyGFormer	TPNet	TSNN (EB+CN)
Wikipedia	96.33±0.07	94.37±0.09	96.67±0.07	98.37±0.07	98.54±0.04	90.78±0.00	95.84±0.18	96.92±0.03	97.75±0.11	98.16±0.06	98.91±0.02	99.30±0.02	**99.31±0.02**
Reddit	98.31±0.05	98.17±0.05	98.47±0.02	98.60±0.06	99.01±0.01	95.37±0.00	97.42±0.02	97.17±0.02	99.09±0.10	99.09±0.03	99.15±0.01	99.22±0.00	**99.68±0.02**
MOOC	83.81±2.09	85.03±0.58	87.11±0.19	91.21±1.15	80.38±0.26	60.86±0.00	83.12±0.18	84.01±0.17	87.42±0.58	90.55±0.43	87.91±0.58	97.17±0.08	**99.49±0.01**
LastFM	70.49±1.66	71.16±1.89	71.59±0.18	78.47±2.94	85.92±0.10	83.77±0.00	64.06±1.16	73.53±0.12	86.92±2.72	89.28±1.63	93.05±0.10	94.39±0.04	**95.04±0.13**
Enron	87.96±0.52	84.89±3.00	68.89±1.10	88.32±0.99	90.45±0.14	87.05±0.00	75.74±0.72	84.38±0.21	91.68±0.83	92.87±0.34	93.33±0.13	93.98±0.26	**97.88±0.01**
Social Evo.	92.05±0.46	90.76±0.21	94.76±0.16	95.39±0.17	87.34±0.08	81.60±0.00	94.84±0.17	95.23±0.07	90.84±3.72	96.16±0.02	96.30±0.01	96.43±0.02	**97.38±0.24**
UCI	90.44±0.49	68.77±2.34	78.53±0.74	92.03±1.13	93.87±0.08	77.30±0.00	87.82±1.36	91.81±0.67	92.31±0.37	95.57±0.16	94.49±0.26	**96.79±0.05**	96.23±0.05
Flights	96.21±1.42	95.95±0.62	94.13±0.17	98.22±0.13	98.45±0.01	90.23±0.00	91.21±0.02	91.13±0.01	98.69±0.10	98.89±0.02	98.93±0.01	**99.00±0.02**	98.91±0.03
Can. Parl.	78.21±0.23	73.35±3.67	75.69±0.78	76.99±1.80	75.70±3.27	64.14±0.00	72.46±3.23	83.17±0.53	84.04±1.13	77.96±1.46	97.76±0.41	92.05±0.34	**98.28±0.18**
US Legis.	82.85±1.07	82.28±0.32	75.84±1.99	83.34±0.43	77.16±0.39	62.57±0.00	76.27±0.63	76.96±0.79	85.36±0.52	82.10±0.85	77.90±0.58	86.49±0.18	**92.47±0.07**
UN Trade	69.62±0.44	67.44±0.83	64.01±0.12	69.10±1.67	68.54±0.18	66.75±0.00	64.72±0.05	65.52±0.51	77.61±1.36	74.87±0.53	70.20±1.44	89.17±0.46	**92.15±0.77**
UN Vote	68.53±0.95	67.18±1.04	52.83±1.12	69.71±2.65	53.09±0.22	62.97±0.00	51.88±0.36	52.46±0.27	75.32±0.63	70.69±1.02	57.12±0.62	79.88±0.30	**87.64±0.06**
Contact	96.66±0.89	96.48±0.14	96.95±0.08	97.54±0.35	89.99±0.34	94.34±0.00	94.15±0.09	93.94±0.02	97.79±0.16	98.90±0.02	98.53±0.01	98.91±0.01	**99.26±0.03**
Avg. Rank	8.15	9.69	9.92	6.08	8.15	11.31	11.15	9.62	5.12	4.50	4.15	1.92	1.23

Table 16: Inductive DGB test AP (%) across thirteen datasets and twelve models. Five-seed means ± standard deviations; baselines from Lu et al. (2024).

Dataset	JODIE	DyRep	TGAT	TGN	CAWN	TCL	GraphMixer	NAT	PINT	DyGFormer	TPNet	TSNN (EB+CN)
Wikipedia	94.82±0.20	92.43±0.37	96.22±0.07	97.83±0.04	98.24±0.03	96.22±0.17	96.65±0.02	96.30±0.08	98.03±0.04	98.59±0.03	98.91±0.01	**99.23±0.03**
Reddit	96.50±0.13	96.09±0.11	97.09±0.04	97.50±0.19	98.62±0.01	94.09±0.07	95.26±0.02	98.24±0.04	98.56±0.05	98.84±0.02	98.86±0.01	**99.44±0.07**
MOOC	79.63±1.92	81.07±0.44	85.50±0.19	89.04±1.17	81.42±0.24	80.60±0.22	81.41±0.21	83.62±1.19	87.90±0.98	86.96±0.43	95.07±0.26	**99.45±0.03**
LastFM	81.61±3.82	83.02±1.48	78.63±0.31	81.45±4.29	89.42±0.07	73.53±1.66	82.11±0.42	92.24±0.93	92.42±0.64	94.23±0.09	95.36±0.11	**95.89±0.24**
Enron	80.72±1.39	74.55±3.95	67.05±1.51	77.94±1.02	86.35±0.51	76.14±0.79	75.88±0.48	87.18±1.24	88.12±0.30	89.76±0.34	90.34±0.28	**96.27±0.27**
Social Evo.	91.96±0.48	90.04±0.47	91.41±0.16	90.77±0.86	79.94±0.18	91.55±0.09	91.86±0.06	87.44±5.48	92.40±0.60	93.14±0.04	93.24±0.07	**97.23±0.07**
UCI	79.86±1.48	57.48±1.87	79.54±0.48	88.12±2.05	92.73±0.06	87.36±2.03	91.19±0.42	87.31±0.28	94.72±0.15	94.54±0.12	95.74±0.05	**96.67±0.04**
Flights	94.74±0.37	92.88±0.73	88.73±0.33	95.03±0.60	97.06±0.02	83.41±0.07	83.03±0.05	96.74±0.22	97.54±0.06	97.79±0.02	**97.97±0.04**	97.68±0.03
Can. Parl.	53.92±0.94	54.02±0.76	55.18±0.79	54.10±0.93	55.80±0.69	54.30±0.66	55.91±0.82	61.90±2.52	50.32±0.86	87.74±0.71	68.09±1.55	**96.75±0.04**
US Legis.	54.93±2.29	57.28±0.71	51.00±3.11	58.63±0.37	53.17±1.20	52.59±0.97	50.71±0.76	60.41±0.74	59.71±1.36	54.28±2.87	61.71±0.84	**73.10±0.06**
UN Trade	59.65±0.77	57.02±0.69	61.03±0.18	58.31±3.15	65.24±0.21	62.21±0.12	62.17±0.31	69.57±1.45	60.37±0.78	64.55±0.62	86.53±0.29	**91.44±0.73**
UN Vote	56.64±0.96	54.62±2.22	52.24±1.46	58.85±2.51	49.94±0.45	51.60±0.97	50.68±0.44	66.60±0.98	57.43±1.24	55.93±0.39	58.00±3.21	**82.67±0.15**
Contact	94.34±1.45	92.18±0.41	95.87±0.11	93.82±0.99	89.55±0.30	91.11±0.12	90.59±0.05	96.12±0.08	97.41±0.14	98.03±0.02	98.39±0.02	**98.85±0.08**
Avg. Rank	8.62	9.69	8.96	7.15	7.15	9.50	8.85	5.69	5.15	3.92	2.15	1.15

Table 17: Inductive DGB test ROC AUC (%) across thirteen datasets and twelve models. Five-seed means ± standard deviations; baselines from Lu et al. (2024).

Dataset	JODIE	DyRep	TGAT	TGN	CAWN	TCL	GraphMixer	NAT	PINT	DyGFormer	TPNet	TSNN (EB+CN)
Wikipedia	94.33±0.27	91.49±0.45	95.90±0.09	97.72±0.03	98.03±0.04	95.57±0.20	96.30±0.04	95.82±0.18	97.76±0.05	98.48±0.03	98.90±0.01	**99.12±0.04**
Reddit	96.52±0.13	96.05±0.12	96.98±0.04	97.39±0.07	98.42±0.02	93.80±0.07	94.97±0.05	98.00±0.04	98.38±0.07	98.71±0.01	98.73±0.02	**99.32±0.16**
MOOC	83.16±1.30	84.03±0.19	86.84±0.17	91.24±0.99	81.86±0.25	81.43±0.19	82.77±0.24	84.72±1.31	90.27±0.96	87.62±0.51	95.55±0.25	**99.55±0.02**
LastFM	81.13±3.39	82.24±1.51	76.99±0.29	82.61±3.15	87.82±0.12	70.84±0.85	80.37±0.18	91.60±1.31	92.15±0.68	94.08±0.08	95.36±0.06	**95.90±0.24**
Enron	81.96±1.34	76.34±4.20	64.63±1.74	78.83±1.11	87.02±0.50	72.33±0.99	76.51±0.71	87.95±0.58	87.97±0.61	90.69±0.26	90.21±0.49	**96.75±0.30**
Social Evo.	93.70±0.29	91.18±0.49	93.41±0.19	93.43±0.59	84.73±0.27	93.71±0.18	94.09±0.07	88.15±6.36	94.78±0.37	95.29±0.03	95.47±0.04	**97.50±0.13**
UCI	78.80±0.94	58.08±1.81	77.64±0.38	86.68±2.29	90.40±0.11	84.49±1.82	89.30±0.57	83.78±0.37	93.18±0.17	92.63±0.13	94.40±0.03	**95.75±0.03**
Flights	95.21±0.32	93.56±0.70	88.64±0.35	95.92±0.43	96.86±0.02	82.48±0.01	82.27±0.06	96.97±0.20	97.69±0.05	97.80±0.02	**98.05±0.04**	97.84±0.04
Can. Parl.	53.81±1.14	55.27±0.49	56.51±0.75	55.86±0.75	58.83±1.13	55.83±1.07	58.32±1.08	62.70±2.91	49.64±0.69	89.33±0.48	69.21±1.31	**97.23±0.09**
US Legis.	58.12±2.35	61.07±0.56	48.27±3.50	62.38±0.48	51.49±1.13	50.43±1.48	47.20±0.89	64.22±0.65	61.89±1.52	53.21±3.04	65.29±0.61	**84.18±0.02**
UN Trade	62.28±0.50	58.82±0.98	62.72±0.12	59.99±3.50	67.05±0.21	63.76±0.07	63.48±0.37	69.15±2.33	64.05±0.72	67.25±1.05	86.88±0.23	**91.92±0.88**
UN Vote	58.13±1.43	55.13±3.46	51.83±1.35	61.23±2.71	48.34±0.76	50.51±1.05	50.04±0.86	68.55±0.90	59.01±1.88	56.73±0.69	54.82±4.04	**87.42±0.16**
Contact	95.37±0.92	91.89±0.38	96.53±0.10	94.84±0.75	89.07±0.34	93.05±0.09	92.83±0.05	95.70±0.06	97.66±0.12	98.30±0.02	98.51±0.01	**99.04±0.07**
Avg. Rank	8.46	9.62	8.92	6.69	7.46	9.77	9.00	5.77	4.92	3.77	2.54	1.08

F.5 Component and scoring controls

F.5.1 DGB average-precision interventions

The retrained DGB component interventions use EB and CN off by default, the historical TSNN-core reference, union-of-endpoint negatives, and global AP. Panel A of Table 18 reports the architectural interventions and matched ungated side-state control; Panel B reports width sensitivity and Panel C additional controls. Panel A uses the full-budget reference; Panel B is a separate width sweep. Panel C includes results near the budgeted rerun (USLegis 69.76, CanParl 69.56), but proximity of scores alone does not establish matched training budgets. We therefore state which reference is used for each numerical comparison below and do not interpret all Panel C differences as isolated component effects. The historical reference differs from the full TSNN of Table 13; rows enabling CN are marked in the table.

Against the full-budget TSNN-core reference, the compound no-transport intervention reduces AP by up to 7.04 points, redrawing fixed frames each epoch and removing smoothness each reduce AP by less than 3 points, and the ungated side-state control stays within 2.24 AP points.

Table 18: DGB component ablations, width sensitivity, and additional controls: test AP (%). Retrained rows report three-seed mean±s.d. over seeds 42–44. The budgeted rerun is identified in its row. EB and CN are off unless a row states otherwise. The compound no-transport intervention also freezes frames and removes carry-over, smoothness, and attention.

Intervention	USLegis	CanParl	UCI	MOOC	Enron	Wikipedia
TSNN-core reference	81.90 ± 0.246	97.44 ± 0.293	95.05 ± 0.135	99.15 ± 0.280	94.12 ± 0.383	98.45 ± 0.325
TSNN-core reference, budgeted rerun	69.76 ± 0.264 †	69.56 ± 0.384 †	93.01 ± 0.163 †	98.80 ± 0.093 †	90.95 ± 0.035 †	94.47 ± 0.127 †
Panel A: retrained architectural interventions; reference is full-budget						
Without transport (compound)	74.86 ± 0.167	95.25 ± 0.062	94.45 ± 0.196	98.60 ± 0.023	92.60 ± 0.179	96.84 ± 0.095
Fixed frames, redrawn per epoch	78.92 ± 0.368	95.19 ± 0.161	93.84 ± 0.313	98.24 ± 0.072	91.88 ± 0.237	97.16 ± 0.151
Without smoothness (λ_s=0)	80.21 ± 0.183	94.70 ± 0.046	94.78 ± 0.036	98.79 ± 0.349	92.53 ± 0.286	97.45 ± 0.292
State, matched ungated side state	79.66 ± 0.119	96.18 ± 0.071	94.43 ± 0.335	98.95 ± 0.240	93.87 ± 0.271	97.87 ± 0.332
Panel B: capacity sensitivity, latent width d (width-study reference d=256)						
$d = 64$	80.10 ± 0.388	92.30 ± 0.049	93.92 ± 0.201	98.82 ± 0.275	91.30 ± 0.272	97.32 ± 0.018
$d = 128$	81.23 ± 0.187	95.56 ± 0.257	94.63 ± 0.252	98.83 ± 0.054	90.95 ± 0.015	97.70 ± 0.349
$d = 256$	81.86 ± 0.383	96.96 ± 0.086	95.05 ± 0.273	98.84 ± 0.257	92.00 ± 0.360	98.07 ± 0.254
$d = 384$	81.39 ± 0.089	97.22 ± 0.326	95.07 ± 0.274	98.80 ± 0.224	89.04 ± 0.354	96.79 ± 0.254
Panel C: additional controls						
Frames held fixed for the whole run	69.89 ± 0.315	71.50 ± 0.215	92.87 ± 0.123	98.78 ± 0.127	91.39 ± 0.201	93.21 ± 0.127
Transport-only removal, frames retained	69.23 ± 0.308	69.28 ± 0.138	92.94 ± 0.357	98.75 ± 0.264	88.81 ± 0.306	95.83 ± 0.099
Common-neighbour features enabled (CN on)	71.79 ± 0.036	97.70 ± 0.085	94.12 ± 0.124	99.28 ± 0.138	95.08 ± 0.180	99.13 ± 0.053
Attention enabled (separate fits)	69.76 ± 0.194	69.56 ± 0.046	93.01 ± 0.059	98.81 ± 0.154	91.96 ± 0.306	97.63 ± 0.083
Connected-gradient path, auxiliary losses off	72.58 ± 0.225	72.44 ± 0.073	94.20 ± 0.156	98.79 ± 0.335	91.93 ± 0.180	97.14 ± 0.070
Reversed event orientation	73.00 ± 0.336	70.43 ± 0.095	97.05 ± 0.081	97.25 ± 0.373	95.80 ± 0.174	91.69 ± 0.111

† Budgeted reference rerun (epochs: USLegis 4, CanParl 3, UCI 4, MOOC 1, Enron 2, Wikipedia 2). Panel A uses the full-budget reference; the budgeted row is shown to identify the separate rerun.

In Panel B, $d = 256$ is strongest on USLegis (81.86), Enron (92.00), and Wikipedia (98.07), while $d = 384$ is strongest on CanParl (97.22) and UCI (95.07). MOOC varies by only 0.04 AP points across widths. Increasing width from 256 to 384 lowers AP by 2.96 points on Enron and 1.28 on Wikipedia.

The Panel C CN-on row exceeds the budgeted reference on all six datasets, by 0.48–28.14 AP points. Against the full-budget reference, the same row is higher on CanParl, MOOC, Enron, and Wikipedia, but lower on USLegis and UCI; the two comparisons are reported separately. The attention row contains separate fits: its displayed means match the budgeted rerun on USLegis, CanParl, and UCI

and differ by 0.01 on MOOC (98.81 vs. 98.80), but its standard deviations differ, so these are not repeated reference summaries. On Enron and Wikipedia its means exceed the budgeted reference by 1.01 and 3.16 points. Reversing event orientation also has mixed effects relative to the budgeted reference: AP rises on USLegis, CanParl, UCI, and Enron, but falls on MOOC and Wikipedia.

F.5.2 TEMPORAL-LINK CONTROLS AND PAIRED COMPARISONS

Table 19: Additional ablations and controls not shown in Table 3: test MRR (higher is better). MRR and its standard deviation are reported as fractions, not percentages. Entries report three-seed mean±s.d. over seeds 42–44. The full-model reference is shown; rows repeated from the main table are omitted.

Intervention	tgbl-wiki	tkgl-smallpedia	YouTube	tkgl-polecat
Full TSNN, evolving frames (reference)	0.8223±0.00136	0.6221±0.00053	0.5207±0.00228	0.1849±0.00131
Panel A: additional components, gains, and capacity; matched settings as main-table Block C				
Frame-free, without pair recency	0.4170±0.0018	0.0132±0.0014	0.5170±0.0037	0.0379±0.0014
Frame-free, $K{=}0$	0.8220±0.0003	0.6204±0.0020	0.3298±0.0012	0.1820±0.0029
Isotropic gain, learned $c\mathbf{I}_d$	0.8216±0.0009	0.6219±0.0004	0.5183±0.0013	0.1854±0.0018
Frame-free, $\mathbf{D}_\theta{=}\mathbf{I}_d$	0.8196±0.0027	0.6144±0.0013	0.5173±0.0034	0.1846±0.0021
Frame-free, learned $c\mathbf{I}_d$	0.8192±0.0031	0.6152±0.0029	0.5204±0.0003	0.1837±0.0013
Roles removed from frame and message contexts	0.8215±0.0008	0.6141±0.0029	0.5176±0.0031	0.1810±0.0032
Detached recurrence	0.8201±0.0022	0.6074±0.0027	0.5169±0.0032	0.1763±0.0010
Reflectors $k{=}4$	0.8111±0.0028	0.6205±0.0016	0.5187±0.0020	0.1810±0.0036
Reflectors $k{=}8$	0.8221±0.0002	0.6193±0.0028	0.5023±0.0007	0.1824±0.0025
Width $d{=}192$	0.8212±0.0011	0.6218±0.0028	0.5014±0.0018	0.1538±0.0016
Width $d{=}256$	0.8112±0.0022	0.6140±0.0008	0.5103±0.0011	0.1504±0.0006
Panel B: transport-site and decoder controls; matched settings as main-table Block C				
Transport off in frame proposal	0.8140±0.0009	0.6212±0.0031	0.5192±0.0015	0.1842±0.0007
Transport off in messages	0.8144±0.0016	0.6208±0.0003	0.5201±0.0006	0.1518±0.0037
Carry-over and message transport off	0.8212±0.0011	0.6141±0.0032	0.5053±0.0004	0.1557±0.0033
Frame-free, ambient-vector MLP decoder	0.8221±0.0015	0.6009±0.0035	0.4335±0.0032	0.1661±0.0035
FiLM, ambient-vector MLP decoder	0.8217±0.0027	0.6044±0.0004	0.4503±0.0003	0.1523±0.0023

In Table 19, both panels use the full-model reference and the matched settings of Table 3, Block C. Frame-free variants use the same frame-free reference as Block C of the main table; $d = 128$ unless a row changes it. In Table 3, the * on the elementwise-GRU Wiki entry marks a run whose seed-42 validation MRR was 0.3002.

Among the additional controls, removing pair recency from the frame-free model gives 0.4170/0.0132 on Wiki/Smallpedia; for comparison, the frame-free reference in Table 3 is 0.8094/0.6089. These results show the importance of pair history in this setting. The other component and capacity controls lower MRR by dataset-dependent amounts: every entry of Table 19 is below the full-model reference except the learned scalar gain on Polecat (0.1854 vs. 0.1849). In Panel B, individual transport removals keep Smallpedia MRR near 0.621 (full-model reference 0.6221), and removing carry-over plus message transport gives 0.6141. The full-frame ambient-vector decoder is in Table 3; Panel B adds its frame-free and FiLM versions. The full-frame minus same-scorer frame-free difference in Table 3 is positive on all five datasets: $+0.0245$ on ICEWS and $+0.0091$–$+0.0132$ on the four datasets of Table 19, significant on every dataset at $p < 0.005$ (Welch t-test; Appendix F.5.2). The separate synthetic drift study's 24 paired intervals all contain zero. Constructed mechanism checks are in Appendix F.6.

Table 20 reports, for each dataset, the difference between the full-frame and same-scorer frame-free arm means of Table 3, Block C, together with a two-sided Welch t-test computed from the three-seed means and sample standard deviations of the two arms. Every difference is positive and significant at $p < 0.005$, and each exceeds four times the larger of the two arm standard deviations. The Welch test treats the arms as independent; since both arms share seeds, a paired test on per-seed outputs typically narrows the intervals further. In a synthetic drift study with the main wave-aggregated, roles-on, cross-entropy-only training, events carry features observed through node-specific rotating frames (null, global, and four node-specific drift regimes); full frames, frame-free, shared-frame, side-state, and FiLM arms were each fitted with three seeds per regime (90 fits). All 24 paired 95% intervals for full minus control contain zero; at the largest drift the full-frame mean test MRR is 0.3587 against

0.3595–0.3627 for the controls. In this synthetic setting the arms are not separated, and the study does not establish equivalence either. It rotates observed event features, whereas the cancellation result concerns latent coordinate changes, so it probes a different perturbation from the one the theory addresses.

Table 20: Full-frame minus same-scorer frame-free test MRR (Table 3, Block C), with two-sided Welch t-tests computed from the three-seed means and sample standard deviations of the two arms.

Dataset	Difference of means	Welch t	p	Seeds
tkgl-icews	$+0.0245$	15.06	0.0009	3
tkgl-smallpedia	$+0.0132$	23.82	0.0001	3
YouTube	$+0.0111$	6.34	0.0034	3
tkgl-polecat	$+0.0091$	6.83	0.0037	3
tgbl-wiki	$+0.0129$	14.16	0.0005	3

The ICEWS reference in Table 3 (0.2748) belongs to the separate three-seed ablation configuration; the benchmark result in Table 1 (0.2871) is from its five-seed benchmark run. These values describe different experiment configurations and should not be compared as replicate estimates.

The frozen-generator control in Table 3 still has evolving frames. It slightly exceeds the full-model reference on ICEWS (0.2750 vs. 0.2748), Wiki (0.8233 vs. 0.8223), and YouTube (0.5230 vs. 0.5207), but falls below it on Smallpedia (0.6130 vs. 0.6221) and Polecat (0.1841 vs. 0.1849). This distinguishes freezing the frame generator from holding the realized frames fixed for the run.

The matched frame controls in Block C of Table 3 and the additional controls in Panels A–B of Table 19 support a frame-specific comparison with the training of Section 3. They include a frame-free same-scorer arm with identical data, history, and budget; an elementwise-GRU side state and a FiLM arm of the same state size; frames held fixed from one initialisation; a shared frame; and isotropic versus diagonal gains. Their reference keeps pair recency on; it was selected after a first cohort with pair recency off had been evaluated on test (tgbl-wiki 0.45–0.53, tkgl-icews 0.02–0.03 test MRR), and all arms are compared within this fixed configuration.

F.5.3 DGB CONFIGURATIONS AND SENSITIVITY

The DGB component controls use global AP, negatives from the union of endpoint identifiers, 50 training negatives (the benchmark runs of Table 14 use 1 or 3), batches of 200 events, and chunk length 2,000; Table 21 lists the per-dataset settings. The Nbrs. column is a diffusion-neighbour budget, not the event-buffer cap $B_{\max}$.

Table 21: DGB control configurations by dataset. d: latent width; k: reflector count; Attn.: attention-history cap B_{attn}, with zero meaning disabled; K: diffusion rounds; η: step; Nbrs.: diffusion-neighbour budget. Epochs gives training budgets; ranges list separate configurations.

Dataset	d	k	Attn.	K	η	Nbrs.	Epochs
USLegis	192	2	12	3	0.02	20	5–10
CanParl	192	2	15	3	0.02	15	6–10
UCI	192	2	15	3	0.01	10	10–15
MOOC	192	2	15	3	0.01	10	5
Enron	128	2	0	3	0.01	15	5
Wikipedia	192	4	0	2	0.01	30	30

The TSNN-core reference of Table 18 reports UCI test AP $95.05 \pm 0.135\%$ at $d = 192$ (Table 21). Table 22 reports UCI AP of 95.05% at both $d = 192$ and $d = 256$; the $d = 256$ mean agrees with Panel B of Table 18. The full-budget reference and the width sweep retain their separately reported standard deviations.

In the Panel A intervention of Table 18, fixed random frames were held within an epoch and redrawn between epochs, whereas the Panel C control holds one draw for the whole run; the no-transport intervention set $\mathbf{Q} = \mathbf{I}$, froze frames, disabled carry-over and the smoothness penalty, and bypassed neighbour attention; the smoothness intervention sets $\lambda_s = 0$; and the ungated side-state control added a kd-dimensional recurrent side state per node. The score-time identity and frame-permutation columns of Table 24 rescore a fixed reference checkpoint instead of retraining. The main-training

controls of Table 3, Blocks D–E, and Table 19, Panel A, instead remove carry-over, set $K = 0$, or replace the learned diagonal gain with identity or scalar gains, one at a time.

Table 22 normalizes cost to $d = 256$; UCI AP peaks at $d = 384$ (95.07), 0.02 points above $d = 256$ at $1.64\times$ the time per event.

Table 22: Latent-width sensitivity (DGB setup). UCI AP in percent; parameter count and time per event are ratios to $d = 256$.

d	UCI AP	Params	Time/event
64	93.92	0.15×	0.34×
128	94.63	0.36×	0.57×
192	95.05	0.65×	0.79×
256	95.05	1.00×	1.00×
384	95.07	2.05×	1.64×

Table 23 compares identity, learned scalar, and learned diagonal gains; the learned-diagonal column equals the TSNN-core reference of Table 18 on the three DGB datasets. Its reported AP is higher than identity gain by 0.37, 0.29, and 0.23 points on USLegis, CanParl, and UCI, respectively. It exceeds learned scalar gain by 0.18 and 0.13 points on USLegis and CanParl, and ties it at 95.05 on UCI. This table reports point estimates without uncertainty for the gain comparisons.

Table 23: Gain choices evaluated with the DGB setup: AP (%); $c > 0$ is a learned scalar and "diag" a learned positive diagonal. Theorem 5 covers the positive diagonal under its step bound.

Dataset	$\mathbf{D}_\theta{=}\mathbf{I}$	learned $c\mathbf{I}$	learned diag
US Legis.	81.53	81.72	81.90
Can. Parl.	97.15	97.31	97.44
UCI	94.82	95.05	95.05

F.5.4 SCORE-TIME GEOMETRIC DIAGNOSTICS

Table 24 rescores fixed reference checkpoints; its AP differences are fractions, not percentage points. With CN off, identity transport increases reported AP by 1.09, 0.37, and 0.04 points on USLegis, CanParl, and UCI. Frame permutation instead changes AP by $+0.27$, -5.46, and -0.06 points, respectively; the largest new-node change is also on CanParl (-3.03 points). With CN on, the absolute identity and permutation changes are at most 0.01 and 0.06 points, respectively, and new-node permutation changes are at most 0.11 points. The reported distance reduction is negative in five of six rows. These checkpoint diagnostics therefore do not show a uniform AP benefit from score-time alignment and are distinct from retrained interventions.

Table 24: Score-time alignment diagnostics. Drift is the mean of $\|I - (\mathbf{U}^{\mathrm{new}})^\top \mathbf{U}^{\mathrm{old}}\|_F/\sqrt{d}$, where old and new refer to the endpoint frame before and after its event update. Reduction is the reported relative decrease in summed positive squared distances after transport; a negative value denotes an increase. ΔAP is intervention minus reference, in fractional AP units, not percentage points.

Dataset	CN	Drift (test)	Reduction, permuted (test)	ΔAP identity	ΔAP permuted	New-node ΔAP permuted
USLegis	off	0.0156	-0.0024	+0.0109	+0.0027	+0.0064
CanParl	off	0.0422	+0.0006	+0.0037	-0.0546	-0.0303
UCI	off	0.1099	-0.0143	+0.0004	-0.0006	-0.0002
USLegis	on	0.0028	-0.0020	+0.0001	+0.0003	+0.0008
CanParl	on	0.0515	-0.0019	+0.0001	-0.0006	-0.0011
UCI	on	0.0915	-0.0018	+0.0000	-0.0002	-0.0002

F.6 CONSTRUCTED MECHANISM CHECKS

An independent rational-arithmetic reconstruction recovers the witness's exact gap $137950/104223$ (Appendix C). A related numerical experiment uses $d = 4$, $k = 1$, and 500 parameter draws, each sampling $\mathbf{h}_0 \sim \mathcal{N}(0, I_4)$ and a matrix $W \in \mathbb{R}^{4\times4}$ with independent $\mathcal{N}(0, 0.3^2)$ entries. All nodes start with that $\mathbf{h}_0$ and reflector $\mathbf{e}_1$. On $u \to v$ only the receiver updates: its raw frame increment is $W\mathbf{h}_u^{\mathrm{old}}$, followed by exact carry-over and a message $\mathbf{m}_v$ equal to the elementwise tanh of the

transported sender. The update is $\mathbf{h}_v^{\mathrm{new}} = \overline{\mathbf{h}}_v + g(\overline{\mathbf{h}}_v)\mathbf{m}_v$ with $g(\mathbf{h}) = \tanh(\|\mathbf{h} - \mathbf{h}_0\|_2)$. The shared-coordinate control uses the same draws, gate, and message without frames; the homogeneous control retains frame updates and carry-over but uses Proposition 10's ungated difference message with $\psi = \tanh$ on the first 200 draws. Parameters are sampled, not trained.

For the streams in Eq. (13), measure $D_{\mathrm{sep}} = \|\mathbf{g}_v(\mathcal{G}_1) - \mathbf{g}_v(\mathcal{G}_2)\|_2$ and count separation above 10^{-9}. Table 25 gives 500/500 framed separations and none for either control. A separate check on 60 small frame fields over repeated-edge, star, and isolate graphs confirms the operator factorization, spectra, carry-over, anisotropic descent, firm non-expansiveness, and perturbation bounds.

Table 25: Constructed frame-state mechanism check: number of separated stream pairs and the ambient separation D_{sep}. The framed and shared-coordinate gated updates use 500 parameter draws; the homogeneous zero-message control uses the first 200 draws.

Update	Separated/draws	Median D_{sep}	Maximum D_{sep}
Framed gated	500/500	0.362243	1.568899
Shared-coordinate gated	0/500	0	0
Homogeneous zero-message	0/200	3.140×10^{-16}	2.275×10^{-15}