WHERE LATENTS BELONG AND HOW THEY FLOW: JOINT IDENTIFICATION OF STRUCTURAL ROLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent factors in dynamical systems can play different roles across observations and time. For example, an idea in a group meeting may be expressed by only some participants, inherit information from earlier discussions, and shape what is said next. We formalize this notion as the *structural role* of a latent factor, describing where it is expressed, whether it receives influence from the past, and whether it transmits influence to the future. The main challenge is correspondence, since the three structures may be identified under different permutations, while a structural role is defined only when they are aligned to the same latent coordinate. We show that, under suitable structural and variability conditions, all three roles are jointly identifiable under a common permutation at each time point. Guided by this theory, we develop ROLETRACE, a framework that traces structural roles in a shared latent coordinate system. Synthetic experiments validate structural recovery, while experiments on multiparty meetings examine the identified roles in real sequential interactions.

1 INTRODUCTION

Understanding how hidden factors are organized and evolve is fundamental to explaining complex systems from their observations (Locatello et al., 2019; Schölkopf et al., 2021). In many dynamic systems, these hidden factors can be expressed in different parts of the system and can continue to influence what happens next. Consider a group meeting in Figure 1 (a), where participants discuss different aspects of a project. We observe their remarks over time, while the latent factors shaping the discussion remain hidden. An underlying standpoint may be expressed in some participants' remarks, inherit influence from earlier discussion, and shape later decisions. Understanding such a factor therefore involves where it is expressed, what influence it receives from the past, and whether its influence continues into the future. The same pattern appears beyond meetings, including multimodal streams (Tsai et al., 2019; Leglaive, 2022), distributed sensing (Olfati-Saber, 2007; Yang et al., 2018), biological systems (Velten et al., 2022; Li et al., 2024; Liu et al., 2025), and other settings (Wu et al., 2023; Li et al., 2026) where an evolving hidden process is observed through several meaningful parts of a system. This raises a general question:

Can we identify how hidden factors participate in an evolving system?

These systems contain two complementary forms of structure: observations are structured across groups, while latent states are dependent across time. The group structure characterizes where each latent factor is expressed. At each time point, the groups may correspond to participants in a meeting, modalities in a multimodal system or regions in a spatial process. For example, in Figure 1 (b), the latent $Z_{t,2}$ is expressed in the observations of the Manager $X_t^{(1)}$ and Designer $X_t^{(2)}$, but not the Engineer $X_t^{(3)}$, giving $\left(\mathbb{A}_{t,2}^{(1)}, \mathbb{A}_{t,2}^{(2)}, \mathbb{A}_{t,2}^{(3)}\right) = (1, 1, 0)$. The temporal structure characterizes its incoming and outgoing influence. $Z_{t,2}$ receives influence from the previous state and affects the next, corresponding to $\mathbb{B}_{t,2} = 1$ and $\mathbb{C}_{t,2} = 1$. Together, these attributes give $((1, 1, 0), 1, 1)$, which we call the *structural role* of a latent variable.

This formulation extends the usual target of latent variable identification. Independent component analysis (ICA) studies the recovery of latent sources from observed mixtures (Comon, 1994; Eriksson & Koivunen, 2004). Nonlinear ICA extends this setting, with identifiability established un-

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

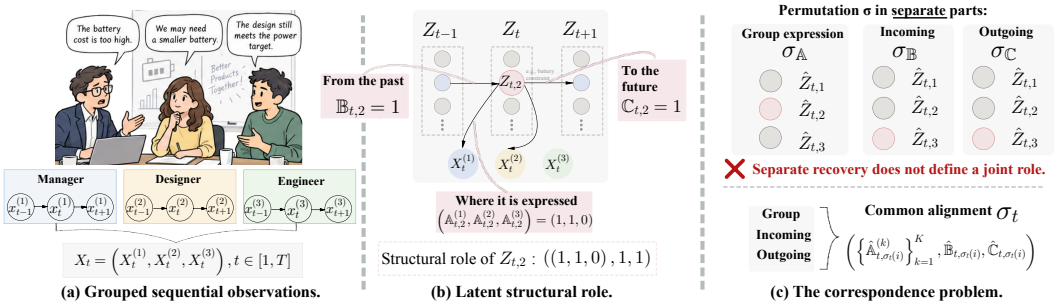


Figure 1: Example of latent structural roles and their alignment. For observation X_t and its latent $Z_{t,2}$: $\mathbb{A}_{t,2}^{(1)} = 1$ means that $Z_{t,2}$ is expressed in group 1, $\mathbb{B}_{t,2} = 1$ indicates incoming influence, and $\mathbb{C}_{t,2} = 1$ indicates outgoing influence. Joint identification requires a common permutation σ_t.

der additional structures such as auxiliary variables (Hyvarinen et al., 2019; Khemakhem et al., 2020), restricted function classes (Buchholz et al., 2022; Lachapelle et al., 2023), or structural sparsity (Zheng et al., 2022; Zheng & Zhang, 2023). More recent work (Zheng et al., 2025) further identifies the incidence structure between latent factors and observation groups. For sequential data, temporal causal representation learning exploits temporal dependence to identify latent processes and their causal relations under assumptions on temporal variation (Yao et al., 2022b;a; Song et al., 2023), historical context (Chen et al., 2024), or sparse latent dynamics (Li et al., 2025b). Specifically, (Li et al., 2025a) further allows the active latent support to vary over time, but the identified temporal support corresponds to our incoming role $\mathbb{B}$, with observation and outgoing activity coupled into the support definition. Existing work therefore identifies latent components, observation-side structure, or temporal causal structure, while the complete role of a latent remains unresolved.

The main challenge is establishing correspondence among these structural views. Latent coordinates have no intrinsic labels, so each structure is generally identified only up to permutation. As illustrated in Figure 1(c), separately recovering group expression, incoming influence, and outgoing influence under σ_A, σ_B, and σ_C may assign the same true latent to different learned coordinates. Joint identification therefore requires a common permutation σ_t that aligns all three structural views simultaneously: $\left(\{\widehat{\mathbb{A}}_{t,\sigma_t(i)}^{(k)}\}_{k=1}^K, \widehat{\mathbb{B}}_{t,\sigma_t(i)}, \widehat{\mathbb{C}}_{t,\sigma_t(i)}\right) = \left(\{\mathbb{A}_{t,i}^{(k)}\}_{k=1}^K, \mathbb{B}_{t,i}, \mathbb{C}_{t,i}\right)$. Under this common alignment, the recovered attributes define the structural role of the same latent coordinate.

Our theory provides guarantees for recovering the complete structural role of each latent coordinate from grouped sequential observations. Variation in the grouped observation map reveals which latent directions are expressed in each group, while the row and column supports of the transition cross Hessian reveal which coordinates receive and transmit delayed influence. Since all three views constrain the same latent reparameterization at time t, their correspondence is recovered through a common permutation σ_t, yielding joint identification of $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$. Guided by the theory, we develop RoleTrace, which learns a latent representation and its transition density in the same coordinate system. The decoder Jacobian estimates group expression, while derivatives of the transition density characterize temporal roles. Our contributions are threefold: **Conceptually,** we formulate structural role identification for grouped sequential observations, characterizing each latent coordinate by its group and temporal roles. **Theoretically,** we establish identifiability of all three structural roles under a common permutation, which enables recovery of their correspondence at each latent coordinate. **Practically,** we develop RoleTrace, a theory guided estimator of structural roles in a shared latent coordinate system, and evaluate it on both synthetic and real-world data.

2 Preliminaries

In this section, we formulate the temporal data-generating process (DGP), introduce the structural support objects used in our analysis, and define the latent roles that will be identified in Section 3.

Data-generating process. For a positive integer n, let $[n] := \{1, \ldots, n\}$. For time horizon T and latent dimension N, we consider a latent sequence $Z_t = (Z_{t,1}, \ldots, Z_{t,N}) \in \mathcal{Z}_t \subseteq \mathbb{R}^N$, $t \in [T]$. We

observe $X_{1:T} := (X_1, \dots, X_T)$, where $X_t = (X_t^{(1)}, \dots, X_t^{(K)}) = (X_{t,1}, \dots, X_{t,d_x}) \in \mathcal{X}_t \subseteq \mathbb{R}^{d_x}$, where d_x is the observation dimension. We partition the observation coordinates into K groups, where $X_t^{(k)} := X_{t,I_k}$, $I_k \subseteq [d_x]$, and $\bigsqcup_{k=1}^K I_k = [d_x]$. The number of latent coordinates expressed in group k may vary with t. The initial state follows $Z_1 \sim p_1$, where p_1 is supported on $\mathcal{Z}_1$. The first-order latent dynamics and observation model are:

$$Z_{t,i} = f_i\big(\mathrm{Pa}_d(Z_{t,i}), \epsilon_{t,i}\big), \quad i \in [N], \tag{1}$$

$$X_t = g_t(Z_t) \in \mathcal{X}_t \subseteq \mathbb{R}^{d_x}, \tag{2}$$

where f_i is a transition mechanism, and the noises $\{\epsilon_{t,i}\}_{t,i}$ are mutually independent with positive smooth densities p_{ϵ_i}. The induced transition density $p(z_t \mid z_{t-1})$ is positive and twice continuously differentiable on its support. $\mathrm{Pa}_d(Z_{t,i})$ denotes the set of delayed parents of $Z_{t,i}$. We assume a time lag of 1 in Eq. 1. Instantaneous dependencies and higher-order lags are discussed in Appendix B.

For $t \in \{2, \dots, T\}$, the delayed adjacency matrix $G_t^d \in \{0,1\}^{N \times N}$ is defined by $(G_t^d)_{i,j} = 1$ iff $Z_{t-1,j} \in \mathrm{Pa}_d(Z_{t,i})$. Its rows therefore index coordinates at time t, whereas its columns index coordinates at time $t-1$. For a differentiable map ϕ, J_ϕ denotes its Jacobian. In particular, $J_{g_t}(z_t) \in \mathbb{R}^{d_x \times N}$, whose (r,i)-th entry is $\partial g_{t,r}(z_t)/\partial z_{t,i}$. To formalize the identification problem, let $\mathcal{G}$ denote the DGP specified by p_1, $\{(f_i, p_{\epsilon_i})\}_{i=1}^N$, and $\{g_t\}_{t=1}^T$. Let $\widehat{\mathcal{G}}$ be a learned DGP of the same form, with latent variables $\widehat{Z}_t$, transition mechanisms $\widehat{f}_i$, noises $\widehat{\epsilon}_{t,i}$, observation maps $\widehat{g}_t$.

Definition 2.1 (Coordinate subspace)**.** For $S \subseteq [n]$, the coordinate subspace of $\mathbb{R}^n$ indexed by S is:

$$\mathbb{R}^n_S := \{v \in \mathbb{R}^n : v_i = 0 \text{ for } i \notin S\}.$$

Definition 2.2 (Support of a matrix-valued map)**.** For $F : \mathcal{U} \to \mathbb{R}^{m \times n}$, define:

$$\mathrm{supp}(F) := \big\{(r,i) \in [m] \times [n] : \exists u \in \mathcal{U} \text{ s.t. } F(u)_{r,i} \neq 0\big\}.$$

Definition 2.3 (Matrix support space)**.** For a binary pattern $G \in \{0,1\}^{m \times n}$, define:

$$\mathbb{R}^{m \times n}_G := \{Q \in \mathbb{R}^{m \times n} : G_{r,i} = 0 \Rightarrow Q_{r,i} = 0\}.$$

Definition 2.4 (Observational equivalence)**.** The models $\mathcal{G}$ and $\widehat{\mathcal{G}}$ are observationally equivalent if they induce the same distribution over observed sequences, $p_{\widehat{\mathcal{G}}}(X_{1:T}) = p_{\mathcal{G}}(X_{1:T})$.

Grouped observations and structural roles. The observation map and delayed graph assign three complementary roles to each latent coordinate. We define the observation roles for $t \in [T]$, the incoming roles for $t \in \{2, \dots, T\}$, and the outgoing roles for $t \in \{1, \dots, T-1\}$ as:

$$\mathcal{S}^{\mathbb{A}}_{t,k} := \big\{i \in [N] : \exists r \in I_k \text{ s.t. } (r,i) \in \mathrm{supp}(J_{g_t})\big\}, \qquad \mathbb{A}^{(k)}_{t,i} := \mathbf{1}\{i \in \mathcal{S}^{\mathbb{A}}_{t,k}\}, \tag{3}$$

$$\mathcal{S}^{\mathbb{B}}_t := \big\{i \in [N] : \exists j \in [N] \text{ s.t. } (G_t^d)_{i,j} = 1\big\}, \qquad \mathbb{B}_{t,i} := \mathbf{1}\{i \in \mathcal{S}^{\mathbb{B}}_t\}, \tag{4}$$

$$\mathcal{S}^{\mathbb{C}}_t := \big\{i \in [N] : \exists j \in [N] \text{ s.t. } (G_{t+1}^d)_{j,i} = 1\big\}, \qquad \mathbb{C}_{t,i} := \mathbf{1}\{i \in \mathcal{S}^{\mathbb{C}}_t\}. \tag{5}$$

Then $\mathbb{A}^{(k)}_{t,i}$ indicates expression in group k, while $\mathbb{B}_{t,i}$ and $\mathbb{C}_{t,i}$ indicate incoming and outgoing delayed influence. For $t \in \{2, \dots, T-1\}$, combining these roles gives $\Big(\{\mathbb{A}^{(k)}_{t,i}\}_{k=1}^K, \mathbb{B}_{t,i}, \mathbb{C}_{t,i}\Big)$. For $\boldsymbol{a} = (a_1, \dots, a_K) \in \{0,1\}^K$ and $b, c \in \{0,1\}$, the associated role class is:

$$\mathcal{R}_t^{\boldsymbol{a},b,c} := \big\{i \in [N] : \mathbb{A}^{(k)}_{t,i} = a_k \ \forall k \in [K],\ \mathbb{B}_{t,i} = b,\ \mathbb{C}_{t,i} = c\big\}. \tag{6}$$

Thus, the $2^{K+2} - 4$ possible signatures [1] retain the entire group-incidence pattern together with the incoming and outgoing temporal roles. Our identification target is a single permutation σ_t at each interior time s.t. $\Big(\{\widehat{\mathbb{A}}^{(k)}_{t,\sigma_t(i)}\}_{k=1}^K, \widehat{\mathbb{B}}_{t,\sigma_t(i)}, \widehat{\mathbb{C}}_{t,\sigma_t(i)}\Big) = \Big(\{\mathbb{A}^{(k)}_{t,i}\}_{k=1}^K, \mathbb{B}_{t,i}, \mathbb{C}_{t,i}\Big), \quad \forall i \in [N]$.

3 Identifiability Theory

In this section, we establish identifiability of the structural roles defined above. Theorem 3.1 first identifies all group-specific observation supports $\mathbb{A}$. Theorem 3.2 and Corollary 3.3 then identify the outgoing and incoming temporal supports $\mathbb{C}$ and $\mathbb{B}$. Lemma 3.4 aligns these three results, leading to joint identification of the complete signature in Theorem 3.5. Complete proofs are in Appendix A.

[1] We assume that a latent variable belongs to at least one group. Then we have $2^K - 1$ possibilities. Together with the four possible $(\mathbb{B}, \mathbb{C})$ patterns, this yields $4(2^K - 1) = 2^{K+2} - 4$ joint roles.

3.1 IDENTIFYING GROUPED OBSERVATION SUPPORT

We first identify which latent coordinates are expressed in each observation group. The collection $\{\mathcal{S}^{\mathbb{A}}_{t,k}\}_{k=1}^{K}$ characterizes the complete incidence structure between latent coordinates and observation groups at time t, including which latents are expressed jointly or exclusively across different groups.

Theorem 3.1 (Identifiability of grouped observation support)**.** *For $t \in [T]$, suppose that $\mathcal{G}$ and $\widehat{\mathcal{G}}$ are observationally equivalent and that:*

i. The maps $g_t : \mathcal{Z}_t \to \mathcal{X}_t$ and $\widehat{g}_t : \widehat{\mathcal{Z}}_t \to \mathcal{X}_t$ are C^1 embeddings onto the same observation manifold. Consequently, $h_t = \widehat{g}_t^{-1} \circ g_t$ is a C^1 diffeomorphism.

ii. There exists $z_t^\star \in \mathcal{Z}_t$ s.t., for every $k \in [K]$, $\operatorname{span}\big\{(J_{g_t}(z_t^\star))_{r,:} : r \in I_k\big\} = \mathbb{R}^{N}_{\mathcal{S}^{\mathbb{A}}_{t,k}}$.

iii. The learned grouped supports are no denser overall than the true grouped supports: $\sum_{k=1}^{K} \left|\widehat{\mathcal{S}}^{\mathbb{A}}_{t,k}\right| \le \sum_{k=1}^{K} \left|\mathcal{S}^{\mathbb{A}}_{t,k}\right|$.

Let $\widehat{z}_t^\star = h_t(z_t^\star)$ and $D_t^\star := J_{h_t^{-1}}(\widehat{z}_t^\star)$. Then $(D_t^\star)_{i,v} = 0, \quad \forall i \in \mathcal{S}^{\mathbb{A}}_{t,k},\ \forall v \notin \widehat{\mathcal{S}}^{\mathbb{A}}_{t,k},\ k \in [K]$. Moreover, every permutation σ_t s.t. $(D_t^\star)_{i,\sigma_t(i)} \neq 0 \forall i \in [N]$ aligns the group supports $\widehat{\mathcal{S}}^{\mathbb{A}}_{t,k} = \sigma_t\big(\mathcal{S}^{\mathbb{A}}_{t,k}\big),\ \forall k \in [K]$. At least one such permutation exists.

Theorem 3.1 recovers the full expression pattern under one permutation. For example, the coordinate corresponding to $Z_{t,2}$ in Figure 1 is identified as being expressed in Groups 1 and 2 and absent from Group 3. Assumption (i) gives a valid invertible change of latent coordinates. Assumption (ii) requires enough variation within each group to distinguish the latent directions it expresses. Conditions of this type have been widely adopted in the identifiability literature to exclude degenerate generating mechanisms and ensure that the relevant structural directions are sufficiently excited (Zheng & Zhang, 2023; Li et al., 2025b). Assumption (iii) expresses structural minimality by allowing no more links between groups and latents than the true structure. Together, these conditions rule out indistinguishable directions and extra links between groups and latents.

3.2 IDENTIFYING INCOMING AND OUTGOING TEMPORAL SUPPORT

Having identified the group-specific observation supports $\mathbb{A}$, we now turn to $\mathbb{B}$ and $\mathbb{C}$, which characterize how latent coordinates receive and transmit delayed influence across consecutive time steps.

Temporal propagation is encoded by the transition distribution between latent states. Let $q_{t+1}(z_{t+1} \mid z_t) := \log p(z_{t+1} \mid z_t)$ and define the transition cross-Hessian:

$$M_t(z_{t+1}, z_t) := \nabla_{z_{t+1}} \nabla_{z_t}^{\top} q_{t+1}(z_{t+1} \mid z_t). \tag{7}$$

The learned cross-Hessian $\widehat{M}_t(\widehat{z}_{t+1}, \widehat{z}_t)$ is defined analogously from the learned transition density $p_{\widehat{\mathcal{G}}}(\widehat{z}_{t+1} \mid \widehat{z}_t)$. Rows of M_t correspond to latent coordinates at time $t+1$, whereas columns correspond to latent coordinates at time t. Its row and column supports therefore characterize incoming and outgoing delayed influence. We first identify the outgoing support $\mathcal{S}_t^{\mathbb{C}}$.

Theorem 3.2 (Identifiability of outgoing temporal support)**.** *For $t \in \{1, \dots, T-1\}$, suppose that the true model $\mathcal{G}$ and learned model $\widehat{\mathcal{G}}$ are observationally equivalent and that:*

i. The transition densities $p(z_{t+1} \mid z_t), p(\widehat{z}_{t+1} \mid \widehat{z}_t)$ are positive and C^2, and the observation embeddings induce C^1 diffeomorphisms $h_t : \mathcal{Z}_t \to \widehat{\mathcal{Z}}_t$ and $h_{t+1} : \mathcal{Z}_{t+1} \to \widehat{\mathcal{Z}}_{t+1}$.

ii. The true and learned delayed graphs are faithful to the cross-Hessian. For every $i, j \in [N]$

$$(G^d_{t+1})_{j,i} = 1 \iff \exists (z_{t+1}, z_t) \in \mathcal{Z}_{t+1} \times \mathcal{Z}_t, \ \text{ s.t. } [M_t(z_{t+1}, z_t)]_{j,i} \neq 0,$$

with the analogous equivalence holding for $\widehat{G}^d_{t+1}$ and $\widehat{M}_t$.

iii. For $\forall\ z_t^\star \in \mathcal{Z}_t$, $\operatorname{span}\big\{M_t(z_{t+1}, z_t^\star) : z_{t+1} \in \mathcal{Z}_{t+1}\big\} = \mathbb{R}^{N \times N}_{G^d_{t+1}}$.

iv. The learned outgoing support satisfies $|\widehat{\mathcal{S}}_t^{\mathbb{C}}| \leq |\mathcal{S}_t^{\mathbb{C}}|$.

Then, for $\forall\ \widehat{z}_t \in \widehat{\mathcal{Z}}_t$, $\frac{\partial Z_{t,i}}{\partial \widehat{Z}_{t,v}} = 0 \quad$ *for* $i \in \mathcal{S}_t^{\mathbb{C}},\ v \notin \widehat{\mathcal{S}}_t^{\mathbb{C}}$. *For any* $\widehat{z}_t^0$, *let* $D_t^0 := J_{h_t^{-1}}(\widehat{z}_t^0)$. *At least one permutation* σ_t *exists s.t.* $(D_t^0)_{i,\sigma_t(i)} \neq 0$ *for* $\forall\ i \in [N]$ *satisfies* $\widehat{\mathcal{S}}_t^{\mathbb{C}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{C}}\big)$.

Theorem 3.2 identifies whether each latent has delayed influence on the next state. Thus, $\mathbb{C}_{t,2} = 1$ in Figure 1 is preserved for the corresponding learned coordinate. Assumption (ii) requires each delayed edge to leave a detectable cross derivative in the transition density. Assumption (iii) provides enough transition variation to reveal the supported directions, and Assumption (iv) imposes minimality on the number of outgoing-active latent coordinates. This is weaker than constraining every delayed edge and is sufficient for identifying the outgoing structural role considered here.

The column support of $M_t(z_{t+1}, z_t)$ is used in Theorem 3.2 to identify $\mathbb{C}$. For $M_{t-1}(z_t, z_{t-1})$, the row support reveals which coordinates receive delayed influence from $t-1$, corresponding to $\mathbb{B}$.

Corollary 3.3 (Identifiability of incoming temporal support)**.** *For* $t \in \{2, \ldots, T\}$, *suppose that the true model* $\mathcal{G}$ *and learned model* $\widehat{\mathcal{G}}$ *are observationally equivalent and that:*

i. The smoothness and faithfulness conditions in Theorem 3.2 hold for $t-1 \to t$.

ii. For $\forall\ z_t^\star \in \mathcal{Z}_t$, $\operatorname{span}\big\{M_{t-1}(z_t^\star, z_{t-1}) : z_{t-1} \in \mathcal{Z}_{t-1}\big\} = \mathbb{R}^{N \times N}_{G_t^d}$.

iii. The learned incoming support satisfies $|\widehat{\mathcal{S}}_t^{\mathbb{B}}| \leq |\mathcal{S}_t^{\mathbb{B}}|$.

Then, for $\forall\ \widehat{z}_t$, $\frac{\partial Z_{t,i}}{\partial \widehat{Z}_{t,v}} = 0 \quad$ *for* $i \in \mathcal{S}_t^{\mathbb{B}},\ v \notin \widehat{\mathcal{S}}_t^{\mathbb{B}}$. *For any* $\widehat{z}_t^0$, *At least one permutation* σ_t *exists, supported by the nonzero pattern of* $D_t^0 = J_{h_t^{-1}}(\widehat{z}_t^0)$ *satisfies* $\widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{B}}\big)$.

Corollary 3.3 gives the guarantee for incoming influence. In Figure 1, it preserves $\mathbb{B}_{t,2} = 1$, so the recovered coordinate for $Z_{t,2}$ receives influence from the past. Its assumptions play the same roles as those in Theorem 3.2, now applied to the row support of the transition cross Hessian.

3.3 Joint identification of structural roles

Theorem 3.1, Theorem 3.2, and Corollary 3.3 identify $\mathbb{A}$, $\mathbb{C}$, and $\mathbb{B}$ separately up to permutation. We next show that these structures can be aligned by one permutation at the same time point.

Lemma 3.4 (Common support alignment)**.** *For* $t \in \{2, \ldots, T-1\}$, *suppose that the assumptions of Theorem 3.1, Theorem 3.2, and Corollary 3.3 hold. Let* $z_t^\star \in \mathcal{Z}_t$ *be the point in assumption (ii) of Theorem 3.1, and define* $\widehat{z}_t^\star := h_t(z_t^\star)$ *and* $D_t^\star := J_{h_t^{-1}}(\widehat{z}_t^\star)$. *There exists at least one permutation* $\sigma_t : [N] \to [N]$ *s.t.* $(D_t^\star)_{i,\sigma_t(i)} \neq 0$ *for* $\forall\ i \in [N],\ \forall\ k \in [K]$ *simultaneously satisfies:*

$$\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}} = \sigma_t\big(\mathcal{S}_{t,k}^{\mathbb{A}}\big),\ \ \widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{B}}\big),\ \ \widehat{\mathcal{S}}_t^{\mathbb{C}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{C}}\big). \tag{8}$$

Lemma 3.4 resolves the permutation ambiguity in Figure 1 (c). Intuitively, the three preceding results all constrain the row support of the same time-t reparameterization Jacobian $J_{h_t^{-1}}$. Evaluating them at a common point therefore yields a single permutation that aligns all structures. Hence, the group, incoming, and outgoing roles attached to one coordinate correspond to the same true latent.

Theorem 3.5 (Joint identifiability of temporal latent roles)**.** *Under the conditions of Lemma 3.4, there exists a permutation* $\sigma_t : [N] \to [N]$ *s.t., for* $\forall\ i \in [N]$,

$$\left(\left\{\widehat{\mathbb{A}}_{t,\sigma_t(i)}^{(k)}\right\}_{k=1}^{K}, \widehat{\mathbb{B}}_{t,\sigma_t(i)}, \widehat{\mathbb{C}}_{t,\sigma_t(i)}\right) = \left(\left\{\mathbb{A}_{t,i}^{(k)}\right\}_{k=1}^{K}, \mathbb{B}_{t,i}, \mathbb{C}_{t,i}\right). \tag{9}$$

Consequently, for every $\boldsymbol{a} \in \{0,1\}^K$ *and* $b, c \in \{0,1\}$, $\widehat{\mathcal{R}}_t^{\boldsymbol{a},b,c} = \sigma_t\big(\mathcal{R}_t^{\boldsymbol{a},b,c}\big)$.

Discussion. Theorem 3.5 establishes a common correspondence between group expression and temporal influence. Lemma 3.4 obtains this correspondence from one nonsingular Jacobian $D_t^\star$, yielding joint support identification without a path-connectedness assumption. (Chen et al., 2024;

Lachapelle et al., 2023) use continuity and connectedness, together with local Jacobian structure, to keep a coordinate or block permutation fixed across latent states. Our result aligns supports defined over the entire latent domain while allowing ambiguity among coordinates with the same structural role. This suffices to identify all admissible role classes at time t. Appendix A.5 gives a geometric explanation and compares this guarantee with related temporal identification results.

4 ROLETRACE: ESTIMATION OF TEMPORAL LATENT ROLES

Guided by the identifiability results in Section 3, we propose ROLETRACE, a practical framework for estimating latent structural roles from finite data. ROLETRACE first learn a latent representation and its temporal transition density, estimate the three structural roles from the corresponding derivatives, and finally refine a common latent coordinate system in which the three roles are jointly sparse.

We use an encoder–decoder pair to estimate the latent representation underlying the grouped observations $X_t = (X_t^{(1)}, \dots, X_t^{(K)})$. Let e_ω denote the encoder and $\widetilde{g}_t$ the decoder. We define $\widetilde{Z}_t = e_\omega(X_t) \in \mathbb{R}^N$ and reconstruct $\widetilde{X}_t = \widetilde{g}_t(\widetilde{Z}_t)$. Given $\widetilde{Z}_{1:T}$, we model the transition with a differentiable conditional density $\widetilde{p}_\eta(\widetilde{z}_{t+1} \mid \widetilde{z}_t)$ whose log density has mixed second derivatives:

$$\mathcal{L}_{\text{rec}} = \frac{1}{T}\sum_{t=1}^{T} \ell_x\Big(X_t, \widetilde{g}_t(\widetilde{Z}_t)\Big), \quad \mathcal{L}_{\text{trans}} \quad = -\frac{1}{T-1}\sum_{t=1}^{T-1} \log \widetilde{p}_\eta\Big(\widetilde{Z}_{t+1} \mid \widetilde{Z}_t\Big). \tag{10}$$

Here, ℓ_x denotes the reconstruction loss. We first learn the latent representation and keep it fixed during temporal density estimation, which preserves the geometry of the learned latent space.

Structural Role Estimation. We estimate the three structural roles from the derivative objects appearing in Section 3. Define the decoder Jacobian $\widetilde{J}_t(\widetilde{z}_t) := J_{\widetilde{g}_t}(\widetilde{z}_t)$ and let $\widetilde{q}_{t+1}(\widetilde{z}_{t+1} \mid \widetilde{z}_t) := \log \widetilde{p}_\eta(\widetilde{z}_{t+1} \mid \widetilde{z}_t)$, with transition cross-Hessian $\widetilde{M}_t(\widetilde{z}_{t+1}, \widetilde{z}_t) := \nabla_{\widetilde{z}_{t+1}} \nabla_{\widetilde{z}_t}^\top \widetilde{q}_{t+1}(\widetilde{z}_{t+1} \mid \widetilde{z}_t)$. Its rows correspond to coordinates at time $t+1$, while its columns correspond to coordinates at time t.

For each role, we extract the derivative slice associated with latent coordinate i:

$$\begin{aligned} v_{t,k,i}^{\mathbb{A}} &:= [\widetilde{J}_t(\widetilde{Z}_t)]_{I_k,i}, & v_{t,i}^{\mathbb{B}} &:= [\widetilde{M}_{t-1}(\widetilde{Z}_t, \widetilde{Z}_{t-1})]_{i,:}, & v_{t,i}^{\mathbb{C}} &:= [\widetilde{M}_t(\widetilde{Z}_{t+1}, \widetilde{Z}_t)]_{:,i}, \\ d_k^{\mathbb{A}} &:= |I_k|, & d^{\mathbb{B}} &:= N, & d^{\mathbb{C}} &:= N. \end{aligned} \tag{11}$$

We then use a common root-mean-square score:

$$s^{\mathbb{X}} = \left[\mathbb{E}_{\text{emp}}\left(\frac{1}{d^{\mathbb{X}}}\left\|v^{\mathbb{X}}\right\|_2^2\right)\right]^{1/2}, \quad \mathbb{X} \in \{\mathbb{A}, \mathbb{B}, \mathbb{C}\}, \tag{12}$$

with the indices inherited from Eq. 11. Thus, $s_{t,k,i}^{\mathbb{A}}$ measures how strongly latent coordinate i is expressed in group k, $s_{t,i}^{\mathbb{B}}$ and $s_{t,i}^{\mathbb{C}}$ measures its incoming and outgoing temporal influence. These scores are continuous empirical surrogates for the support conditions defining $\mathcal{S}_{t,k}^{\mathbb{A}}$, $\mathcal{S}_t^{\mathbb{B}}$, and $\mathcal{S}_t^{\mathbb{C}}$. After structural refinement, we threshold the three score types to obtain $\widehat{\mathbb{A}}_{t,i}^{(k)}$, $\widehat{\mathbb{B}}_{t,i}$, and $\widehat{\mathbb{C}}_{t,i}$.

Common Coordinate Structural Refinement. The derivative scores above depend on the coordinates selected by the initial representation learner. Motivated by Theorem 3.5, we search for a single refined coordinate system at each time point that makes the three structural views jointly sparse.

For $t \in \{2, \dots, T-1\}$, let $Q_t \in \mathbb{R}^{N \times N}$ be an orthogonal transformation, $Q_t^\top Q_t = I$, and define the latent coordinates by $\widehat{Z}_t = Q_t^\top \widetilde{Z}_t$. Equivalently, $\widetilde{Z}_t = Q_t \widehat{Z}_t$. The observation decoder in the coordinates is therefore $\widehat{g}_t(\widehat{z}_t) = \widetilde{g}_t(Q_t \widehat{z}_t)$, which gives $J_{\widehat{g}_t}(\widehat{z}_t) = \widetilde{J}_t(\widetilde{z}_t) Q_t$. For the transitions $t-1 \to t$ and $t \to t+1$, the cross-Hessians transform on the time t coordinate as $\widehat{M}_{t-1} = Q_t^\top \widetilde{M}_{t-1}$, $\widehat{M}_t = \widetilde{M}_t Q_t$. Hence, the same Q_t simultaneously controls the group-incidence columns of the observation Jacobian, the incoming rows of $\widetilde{M}_{t-1}$, and the outgoing columns of $\widetilde{M}_t$. Let $s_{t,k,i}^{\mathbb{A}}(Q_t)$, $s_{t,i}^{\mathbb{B}}(Q_t)$, and $s_{t,i}^{\mathbb{C}}(Q_t)$ denote the resulting scores. We use the sparsity objective:

$$\Omega_t(Q_t) = \lambda_{\mathbb{A}} \sum_{k=1}^{K}\sum_{i=1}^{N} s_{t,k,i}^{\mathbb{A}}(Q_t) + \lambda_{\mathbb{B}} \sum_{i=1}^{N} s_{t,i}^{\mathbb{B}}(Q_t) + \lambda_{\mathbb{C}} \sum_{i=1}^{N} s_{t,i}^{\mathbb{C}}(Q_t), \tag{13}$$

This finite-sample sparsity objective is a practical surrogate for the minimality conditions in Section 3. We minimize Eq. 13 using deterministic Jacobi coordinate rotations. Let $G_{ij}(\theta)$ denote the identity matrix with its (i, j) block replaced by a rotation of angle θ. For each pair (i, j), we update $\theta^\star = \arg\min_\theta \Omega_t \left(Q_t G_{ij}(\theta)\right)$, $Q_t \leftarrow Q_t G_{ij}(\theta^\star)$, and sweep over coordinate pairs until the structural objective no longer decreases appreciably. Because the estimated transition density may change after the latent coordinates are refined, we perform one additional transition refit: after obtaining the initial Q_t, we refit the conditional density in the coordinates, recompute the cross-Hessians, and run a final Jacobi sweep. The resulting representation yields the joint estimated signature.

5 EXPERIMENTS

We conduct experiments in both synthetic and real-world settings, guided by four research questions: **RQ1 (Recovery).** Can ROLETRACE recover $\{\mathbb{A}^{(k)}\}_{k=1}^K$, $\mathbb{B}$ and $\mathbb{C}$? **RQ2 (Alignment).** Does common alignment enable more consistent joint recovery than aligning the three structural roles separately? **RQ3 (Sensitivity to Identifiability Conditions).** How does recovery change as the key conditions underlying our identifiability results are weakened? **RQ4 (Applicability).** Does ROLETRACE capture meaningful information flow in real-world scenario?

5.1 SYNTHETIC EXPERIMENTS

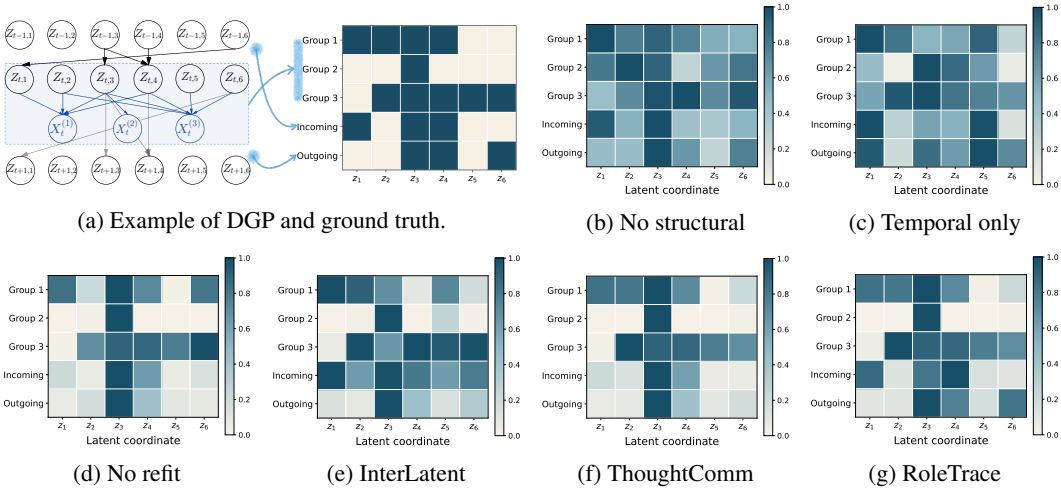
Setup. We evaluate structural role recovery across randomly generated mechanisms. The RQ1 datasets are generated by a random nonlinear latent process with Gaussian noise and a nonlinear grouped observation map. Both structural supports and numerical parameters are sampled for each mechanism; Figure 2a illustrates one example. We use 16 mechanisms with $N = 6$, $K = 3$, $T = 5$, and six observed dimensions per group. RQ2 and RQ3 use controlled mechanisms with $N = 4$. Details of DGP, training, and evaluation metrics are provided in Appendix D.1.

Baselines and evaluation. We compare two adapted methods and three ablations. THOUGHTCOMM uses the observation structure $\mathbb{A}$ for coordinate refinement (Zheng et al., 2025). INTERLATENT jointly trains the representation, temporal density, and structural penalties (Li et al., 2025a). NO STRUCTURAL omits coordinate refinement; TEMPORAL-ONLY uses only $\mathbb{B}$ and $\mathbb{C}$; and NO-REFIT omits the transition refit and final refinement. Methods share data and matched component initializations, with a common derivative readout for all three roles. We measure structural recovery under a common evaluation permutation. Role F1 scores assess $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$. Joint accuracy counts coordinates with a correct signature, while Hamming accuracy counts correct signature entries.

In Table 1, ROLETRACE achieves the highest mean joint recovery among the compared methods. At $n = 3200$, its Joint accuracy is 0.854, compared with 0.708 for both THOUGHTCOMM and NO-REFIT. It also reaches 0.973 F1 for both temporal roles, while THOUGHTCOMM has slightly higher $\mathbb{A}$ F1. The improvement over NO-REFIT supports refitting the transition after coordinate refinement. Figure 2 illustrates the recovered signatures in detail for the example mechanism **(Answering RQ1)**. Joint alignment reduces the correspondence ambiguity left by individual roles. In Figure 3, we hold the recovered scores fixed and select evaluation permutations using $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$, or all three jointly. Averaging over equally optimal permutations, the respective individual alignments give Joint accuracies of 0.700, 0.475, and 0.458, compared with 0.950 for joint alignment. Thus, accurate matching of one role can leave the complete signature unresolved **(Answering RQ2)**. Figure 4 examines recovery as the identifiability conditions weaken. Using true derivatives under orthogonal coordinate mixing, we vary group conditioning, temporal span, and edge faithfulness separately. Mean Joint accuracy decreases from 1 to 0.633, 0.750, and 0.500 at the respective endpoints. The first two paths show losses before the endpoint. In the cancellation experiment, the declared target remains fixed, and recovery stays perfect until the selected edge has zero net influence. These results characterize sensitivity along the tested degradation paths **(Answering RQ3)**.

5.2 REAL-WORLD CASE: MULTI-PARTY MEETINGS

Dataset and setup. We consider the AMI Meeting Corpus, a multimodal corpus containing 100 hours of recorded meetings with manual transcripts and linguistic annotations (Carletta et al., 2005). We focus on scenario-based meetings in which four participants collaborate on a shared design project over multiple project phases. Each participant is assigned one fixed functional role:



(a) Example of DGP and ground truth. (b) No structural (c) Temporal only

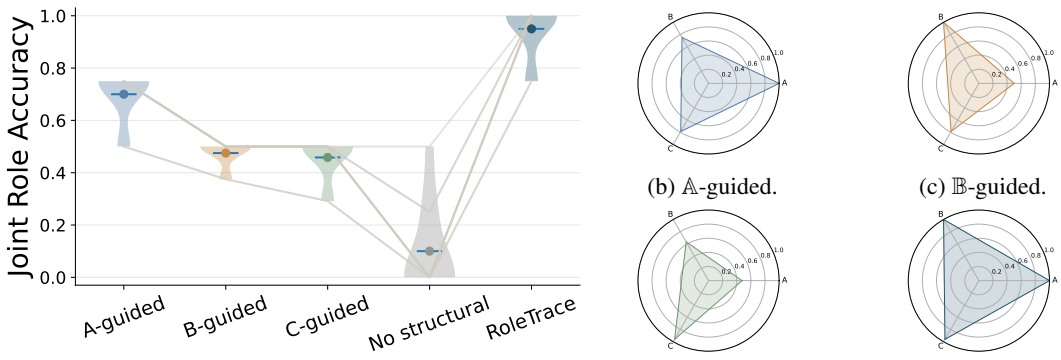
(d) No refit (e) InterLatent (f) ThoughtComm (g) RoleTrace

Figure 2: Heatmaps of recovered structural signatures at $n = 3200$.

Table 1: Structural recovery at $n = 3200$ (mean $\pm$ SE over 16 paired mechanisms).

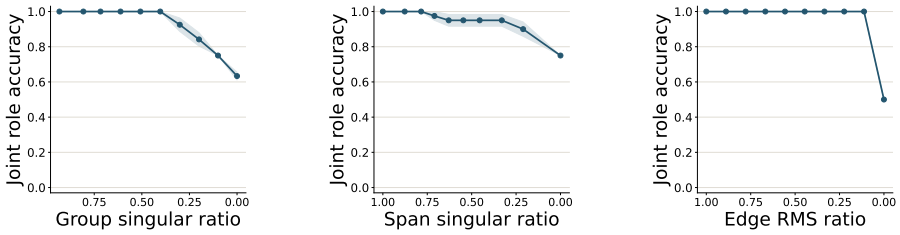
Method	$\mathbb{A}$ F1	$\mathbb{B}$ F1	$\mathbb{C}$ F1	Joint Acc.	Hamming
No structural	0.796 ± 0.027	0.823 ± 0.034	0.795 ± 0.043	0.281 ± 0.036	0.769 ± 0.015
Temporal-only	0.850 ± 0.025	0.783 ± 0.033	0.899 ± 0.040	0.469 ± 0.051	0.819 ± 0.021
No-refit	0.950 ± 0.020	0.880 ± 0.037	0.886 ± 0.023	0.708 ± 0.058	0.919 ± 0.020
InterLatent	0.847 ± 0.017	0.717 ± 0.048	0.833 ± 0.032	0.323 ± 0.042	0.790 ± 0.016
ThoughtComm	$\mathbf{0.964 \pm 0.013}$	0.859 ± 0.039	0.848 ± 0.027	0.708 ± 0.058	0.919 ± 0.017
RoleTrace	0.957 ± 0.017	$\mathbf{0.973 \pm 0.014}$	$\mathbf{0.973 \pm 0.014}$	$\mathbf{0.854 \pm 0.050}$	$\mathbf{0.954 \pm 0.017}$

Project Manager (PM), Industrial Designer (ID), User Interface Designer (UI), and Marketing Expert (ME). The four participant roles provide natural observation groups for an underlying conversational state that evolves throughout the meeting, which closely matches our grouped temporal setting. Let $\mathcal{K}_{\mathrm{AMI}} := \{\mathrm{PM}, \mathrm{ID}, \mathrm{UI}, \mathrm{ME}\}$, with $K = 4$. At each time step t, we write $X_t = \left(X_t^{(\mathrm{PM})}, X_t^{(\mathrm{ID})}, X_t^{(\mathrm{UI})}, X_t^{(\mathrm{ME})}\right) \in \mathbb{R}^{4\times d}$. We let $Z_t \in \mathbb{R}^N$ denote the latent conversational state underlying these grouped observations. After requiring the meetings to have the corresponding RoleTrace representations and downstream labels, our evaluation contains 118 meetings from 30 teams. We partition each meeting into 15-second windows and evaluate next-speaker prediction. Further preprocessing and evaluation details are provided in Appendix D.2.



(a) Joint role recovery. (b) $\mathbb{A}$-guided. (c) $\mathbb{B}$-guided. (d) $\mathbb{C}$-guided. (e) RoleTrace.

Figure 3: Structural role identification under different permutations.



(a) Group conditioning. (b) Temporal span. (c) Edge faithfulness.

Figure 4: Sensitivity to structural conditions using true derivatives.

Table 2: AMI next-speaker accuracy (%, mean $\pm$ standard deviation over ten seeds) at matched latent budgets. Additional comparisons are provided in Appendix D.2.

Selection rule	$k = 4$	$k = 8$	$k = 12$	$k = 16$	$k = 24$
Random	42.822 ± 0.588	47.528 ± 0.259	49.341 ± 0.154	50.190 ± 0.105	51.108 ± 0.088
Variance	42.442 ± 0.748	47.139 ± 0.584	49.222 ± 0.338	50.118 ± 0.310	51.077 ± 0.126
Activity	44.774 ± 0.936	48.031 ± 0.522	49.686 ± 0.450	50.438 ± 0.293	51.197 ± 0.289
Content	44.261 ± 0.875	49.016 ± 0.566	50.299 ± 0.256	50.841 ± 0.289	$\mathbf{51.475 \pm 0.119}$
Density	45.155 ± 1.397	49.430 ± 0.493	50.503 ± 0.266	$\mathbf{50.954 \pm 0.218}$	$\mathbf{51.475 \pm 0.119}$
Role $\mathbb{C}$	$\mathbf{45.854 \pm 1.413}$	$\mathbf{49.470 \pm 0.384}$	$\mathbf{50.546 \pm 0.317}$	50.906 ± 0.261	$\mathbf{51.475 \pm 0.119}$

We evaluate whether the identified outgoing role can retain information useful for next-speaker prediction under limited latent budgets. All methods select k coordinates from the same representation and use the same downstream classifier. RANDOM samples coordinates uniformly. VARIANCE keeps those with the largest variation on fit data. ACTIVITY ranks coordinates by their sensitivity to current speech activity. CONTENT uses the corresponding temporal readout for linguistic content, and DENSITY uses sensitivity from the learned transition density. Our Role $\mathbb{C}$ selector prioritizes coordinates whose current variation is more relevant to subsequent interaction. Table 2 shows that Role $\mathbb{C}$ performs best at the tighter budgets $k = 4, 8, 12$, consistently outperforming random, variance, and activity selection. At $k = 16$, it remains comparable to the strongest density readout, while at $k = 24$ the three temporal selectors retain the same coordinates and therefore achieve the same accuracy of $51.475 \pm 0.119\%$. These results show that the outgoing structural role learned by ROLETRACE can benefit the downstream next-speaker task by preserving more predictive information when the latent representation is compressed **(Answering RQ4)**.

6 CONCLUSION

We study the joint identification of latent structural roles from grouped sequential observations. Under suitable structural and variability conditions, we show that group expression $\mathbb{A}$, incoming influence $\mathbb{B}$, and outgoing influence $\mathbb{C}$ are jointly identifiable under a common permutation σ_t, recovering their correspondence at each latent coordinate. Guided by these results, we develop ROLETRACE to estimate the roles from finite data in a shared latent coordinate system. Synthetic experiments support structural recovery and highlight the importance of common alignment, while the AMI results suggest that our approach is beneficial for downstream tasks. Together, our theory and framework provide a foundation for understanding where hidden factors are expressed and how their influence propagates through evolving systems. Although we have not explored a broader range of real-world applications empirically, the proposed formulation applies to a broader class of grouped sequential systems than those evaluated in this work. We hope this work sheds light on the latent organization underlying sequential observations and opens new directions for understanding complex evolving systems through identifiable representations.

AI USE DISCLOSURE

In this work, generative AI tools were used to assist with checking theoretical proofs, implementing and debugging research code, and translating draft material. We also used generative AI for literature exploration and summarization, drafting and editing parts of the manuscript and formatting LaTeX content. The illustrative cartoon in the first panel of Figure 1 was generated with the assistance of a generative image model. All AI-assisted code and experimental procedures were verified and tested by the authors. All AI-assisted text and artifacts were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, March 2003. ISSN 1532-4435. doi: 10.1162/153244303768966085. URL https://doi.org/10.1162/153244303768966085.

Philippe Brouillard, Sebastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Learning a spatial partitioning and its causal relations from temporal data. In Bijan Mazaheri and Niels Richard Hanson (eds.), *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 547–592. PMLR, 06–08 Apr 2026. URL https://proceedings.mlr.press/v323/brouillard26a.html.

S. Buchholz, M. Besserve, and B. Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, pp. 16946–16961. Curran Associates, Inc., December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6c5da478b9d13f541993d67897a0bb30-Abstract-Conference.html.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: a pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pp. 28–39, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3540325492. doi: 10.1007/11677482_3. URL https://doi.org/10.1007/11677482_3.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: learning temporal causal representation under non-invertible generation process. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0165168494900299. Higher Order Statistics.

Zhongfen Deng, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Quan Hung Tran, Shuaiqi Liu, Wenting Zhao, Tao Zhang, Yibo Wang, and Philip S Yu. Aspect-based meeting transcript summarization: A two-stage approach with weak supervision on sentence classification. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 636–645. IEEE, 2023.

Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ica models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004. ISSN 1558-2361.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3808–3814. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/524. URL https://doi.org/10.24963/ijcai.2021/524. Main Track.

Matthew Frampton, Jia Huang, Trung Bui, and Stanley Peters. Real-time decision detection in multi-party dialogue. In Philipp Koehn and Rada Mihalcea (eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1133–1141, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1118/.

L. Gresele, P. K. Rubenstein*, A. Mehrjou, F. Locatello, and B. Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227. PMLR, July 2019. URL https://proceedings.mlr.press/v115/gresele20a.html. *equal contribution.

Elio Gruttadauria, Mathieu Fontaine, and Slim Essid. Online speaker diarization of meetings guided by speech separation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11356–11360. IEEE, 2024.

Lukas Hilgert and Jan Niehues. Next speaker prediction for multi-speaker dialogue with large language models. In Mourad Abbas, Tariq Yousef, and Lukas Galke (eds.), *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pp. 60–71, Southern Denmark University, Odense, Denmark, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.icnlsp-1.7/.

Zenan Huang, Haobo Wang, Junbo Zhao, and Nenggan Zheng. Latent processes identification from multi-view time series. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/428. URL https://doi.org/10.24963/ijcai.2023/428.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3772–3780, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/hyvarinen17a.html.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *ArXiv*, abs/1805.08651, 2018. URL https://api.semanticscholar.org/CorpusID:46897575.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*,

volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi LE PRIOL, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

Sebastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=R6KJN1AUAR.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27(71):1–90, 2026. URL http://jmlr.org/papers/v27/24-0771.html.

Simon Leglaive. A multimodal dynamical variational autoencoder for audiovisual speech representation learning. In *Proceedings of the 1st International Workshop on Methodologies for Multimedia*, M4MM '22, pp. 3, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450395052. doi: 10.1145/3552487.3556435. URL https://doi.org/10.1145/3552487.3556435.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2190–2196, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1210. URL https://aclanthology.org/P19-1210/.

Weihan Li, Chengrui Li, Yule Wang, and Anqi Wu. Multi-region markovian gaussian process: an efficient method to discover directional communications across multiple brain regions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Xintong Li, Haoran Zhang, and Xiao Zhou. Spatio-temporal hierarchical causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23230–23238, 2026.

Yihang Li and Chenhui Chu. Rethinking meeting effectiveness: A benchmark and framework for temporal fine-grained automatic meeting effectiveness evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5049–5066, 2026.

Yuke Li, Yujia Zheng, Guangyi Chen, Kun Zhang, and Heng Huang. Identification of intermittent temporal latent process. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=6Pz7afmsOp.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *International Conference on Learning Representations*, volume 2025, pp. 35203–35242, 2025b.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=itZ6ggvMnzS.

Belle Liu, Jacob Sacks, and Matthew D. Golub. Accurate identification of communication between multiple interacting neural populations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39381–39404. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25bh.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Hiroshi Morioka and Aapo Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36249–36293. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/morioka24a.html.

Hiroshi Morioka, Hermanni Hälvä, and Aapo Hyvarinen. Independent innovation analysis for nonlinear vector autoregressive process. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1549–1557. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/morioka21a.html.

R. Olfati-Saber. Distributed kalman filtering for sensor networks. In *2007 46th IEEE Conference on Decision and Control*, pp. 5492–5498, 2007. doi: 10.1109/CDC.2007.4434303.

A. Sapru, Sree Harsha Yella, and Hervé Bourlard. Improving speaker diarization using social role information. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

Ashtosh Sapru and Fabio Valente. Automatic speaker role labeling in ami meetings: Recognition of formal and social roles. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5057–5060, 2012. doi: 10.1109/ICASSP.2012.6289057.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Haochen Tan, Han Wu, Wei Shao, Xinyun Zhang, Mingjie Zhan, Zhaohui Hou, Ding Liang, and Linqi Song. Reconstruct before summarize: An efficient two-step framework for condensing and summarizing meeting transcripts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13128–13141, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.812. URL https://aclanthology.org/2023.emnlp-main.812/.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019.

Britta Velten, Jana M Braunger, Ricard Argelaguet, Damien Arnol, Jakob Wirbel, Danila Bredikhin, Georg Zeller, and Oliver Stegle. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature methods*, 19(2):179–186, 2022.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/wu2023iclr-slotformer/.

Danru Xu, Dingling Yao, Sebastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55389–55433. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xu24ac.html.

Chao Yang, Jiangying Zheng, Xiaoqiang Ren, Wen Yang, Hongbo Shi, and Ling Shi. Multi-sensor kalman filtering with intermittent measurements. *IEEE Transactions on Automatic Control*, 63 (3):797–804, 2018. doi: 10.1109/TAC.2017.2734643.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OGtnhKQJms.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022a. Curran Associates Inc. ISBN 9781713871088.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *AAAI 2022*, September 2022. URL https://www.microsoft.com/en-us/research/publication/ts2vec-towards-universal-representation-of-time-series/.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: a general setting. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: sparsity and beyond. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Yujia Zheng, Zhuokai Zhao, Zijian Li, Yaqi Xie, Mingze Gao, Lizhu Zhang, and Kun Zhang. Thought communication in multiagent collaboration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=tq9lyV9Cml.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 194–203, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.19. URL https://aclanthology.org/2020.findings-emnlp.19/.

TABLE OF CONTENTS

A Proofs . . . 15

A.1 Proof of Theorem 3.1 . . . 15

A.2 Proofs of Theorem 3.2 and Corollary 3.3 . . . 17

A.3 Proof of Lemma 3.4 . . . 19

A.4 Proof of Theorem 3.5 . . . 21

A.5 Common structural alignment and global permutation invariance . . . 22

B Extensions of the Temporal Model . . . 23

B.1 Higher-order temporal dependencies . . . 23

B.2 Instantaneous dependencies . . . 25

C Related Work . . . 26

D Additional Experimental Details . . . 27

D.1 Additional Details for Synthetic Experiments . . . 27

D.2 Additional Details for AMI Multi-party Meeting Experiment . . . 31

A PROOFS

A.1 PROOF OF THEOREM 3.1

Theorem A.1 (Identifiability of grouped observation support)**.** *For $t \in [T]$, suppose that $\mathcal{G}$ and $\widehat{\mathcal{G}}$ are observationally equivalent and that:*

i. The maps $g_t : \mathcal{Z}_t \to \mathcal{X}_t$ and $\widehat{g}_t : \widehat{\mathcal{Z}}_t \to \mathcal{X}_t$ are C^1 embeddings onto the same observation manifold. Consequently, $h_t = \widehat{g}_t^{-1} \circ g_t$ is a C^1 diffeomorphism.

ii. There exists $z_t^\star \in \mathcal{Z}_t$ s.t., for every $k \in [K]$, $\operatorname{span}\big\{(J_{g_t}(z_t^\star))_{r,:} : r \in I_k\big\} = \mathbb{R}^N_{\mathcal{S}^{\mathbb{A}}_{t,k}}$.

iii. The learned grouped supports are no denser overall than the true grouped supports: $\sum_{k=1}^K \left|\widehat{\mathcal{S}}^{\mathbb{A}}_{t,k}\right| \le \sum_{k=1}^K \left|\mathcal{S}^{\mathbb{A}}_{t,k}\right|$.

Let $\widehat{z}_t^\star = h_t(z_t^\star)$ and $D_t^\star := J_{h_t^{-1}}(\widehat{z}_t^\star)$. Then $(D_t^\star)_{i,v} = 0, \quad \forall i \in \mathcal{S}^{\mathbb{A}}_{t,k},\ \forall v \notin \widehat{\mathcal{S}}^{\mathbb{A}}_{t,k},\ k \in [K]$. Moreover, every permutation σ_t s.t. $(D_t^\star)_{i,\sigma_t(i)} \neq 0 \forall i \in [N]$ aligns the group supports:

$$\widehat{\mathcal{S}}^{\mathbb{A}}_{t,k} = \sigma_t\big(\mathcal{S}^{\mathbb{A}}_{t,k}\big), \quad \forall k \in [K].$$

At least one such permutation exists.

Proof. By assumption (i), $h_t : \mathcal{Z}_t \to \widehat{\mathcal{Z}}_t$ is a C^1 diffeomorphism. Hence, for every $\widehat{z}_t \in \widehat{\mathcal{Z}}_t$, with $z_t = h_t^{-1}(\widehat{z}_t)$, $\widehat{g}_t(\widehat{z}_t) = g_t\left(h_t^{-1}(\widehat{z}_t)\right)$. Define $D_t(\widehat{z}_t) := J_{h_t^{-1}}(\widehat{z}_t)$. Applying the chain rule gives:

$$J_{\widehat{g}_t}(\widehat{z}_t) = J_{g_t}(z_t) D_t(\widehat{z}_t), \quad z_t = h_t^{-1}(\widehat{z}_t). \tag{14}$$

At the point in assumption (ii), let $\widehat{z}_t^\star = h_t(z_t^\star)$ and $D_t^\star = D_t(\widehat{z}_t^\star)$. Since h_t^{-1} is a diffeomorphism, $D_t^\star$ is nonsingular.

We first establish the support separation. Fix an arbitrary group $k \in [K]$, an index $i \in \mathcal{S}^{\mathbb{A}}_{t,k}$, and an index $v \notin \widehat{\mathcal{S}}^{\mathbb{A}}_{t,k}$. By the definition of the learned group support, $v \notin \widehat{\mathcal{S}}^{\mathbb{A}}_{t,k}$ means that no observation coordinate indexed by I_k depends on the learned latent coordinate v. Equivalently,

$$\left(J_{\widehat{g}_t}(\widehat{z}_t)\right)_{r,v} = 0, \quad \forall r \in I_k,\ \forall \widehat{z}_t \in \widehat{\mathcal{Z}}_t. \tag{15}$$

In particular, Eq. 15 holds at $\widehat{z}_t^\star$.

Since $i \in \mathcal{S}_{t,k}^{\mathbb{A}}$, the standard row vector $e_i^\top$ belongs to $\mathbb{R}^N_{\mathcal{S}_{t,k}^{\mathbb{A}}}$. Assumption (ii) therefore implies that there exist coefficients $\{\alpha_r\}_{r\in I_k}$ s.t.

$$e_i^\top = \sum_{r\in I_k} \alpha_r \left(J_{g_t}(z_t^\star)\right)_{r,:}. \tag{16}$$

Right-multiplying Eq. 16 by $D_t^\star$ yields:

$$\begin{aligned} e_i^\top D_t^\star &= \sum_{r\in I_k} \alpha_r \left(J_{g_t}(z_t^\star)\right)_{r,:} D_t^\star \\ &= \sum_{r\in I_k} \alpha_r \left(J_{\widehat{g}_t}(\widehat{z}_t^\star)\right)_{r,:}, \end{aligned} \tag{17}$$

where the second equality follows from Eq. 14.

Taking the v-th coordinate of Eq. 17 gives $(D_t^\star)_{i,v} = \sum_{r\in I_k} \alpha_r \left(J_{\widehat{g}_t}(\widehat{z}_t^\star)\right)_{r,v} = 0$. Since k, i, and v were arbitrary, $(D_t^\star)_{i,v} = 0, \quad \forall k \in [K],\ \forall i \in \mathcal{S}_{t,k}^{\mathbb{A}},\ \forall v \notin \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}$. Equivalently,

$$\operatorname{supp}\left((D_t^\star)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}, \quad \forall k \in [K],\ \forall i \in \mathcal{S}_{t,k}^{\mathbb{A}}. \tag{18}$$

We next show that at least one permutation is supported by the nonzero pattern of $D_t^\star$. Since $D_t^\star$ is nonsingular, $\det(D_t^\star) \neq 0$. By the Leibniz formula,

$$\det(D_t^\star) = \sum_{\sigma\in S_N} \operatorname{sgn}(\sigma) \prod_{i=1}^{N} (D_t^\star)_{i,\sigma(i)} \neq 0, \tag{19}$$

where S_N denotes the set of permutations of $[N]$. Therefore, at least one term in the sum is nonzero. Hence there exists a permutation $\sigma_t : [N] \to [N]$ s.t.

$$(D_t^\star)_{i,\sigma_t(i)} \neq 0, \quad \forall i \in [N]. \tag{20}$$

Now fix any permutation σ_t satisfying Eq. 20. We show that this permutation aligns every group support. Fix $k \in [K]$ and $i \in \mathcal{S}_{t,k}^{\mathbb{A}}$. By Eq. 18, every nonzero entry in row i of $D_t^\star$ has its column index in $\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}$. Since $(D_t^\star)_{i,\sigma_t(i)} \neq 0$, it follows that $\sigma_t(i) \in \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}$. Thus,

$$\sigma_t\left(\mathcal{S}_{t,k}^{\mathbb{A}}\right) \subseteq \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}, \quad \forall k \in [K]. \tag{21}$$

Since σ_t is a permutation and hence injective,

$$\left|\mathcal{S}_{t,k}^{\mathbb{A}}\right| = \left|\sigma_t\left(\mathcal{S}_{t,k}^{\mathbb{A}}\right)\right| \leq \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right|, \quad \forall k \in [K]. \tag{22}$$

Summing Eq. 22 over k gives:

$$\sum_{k=1}^{K} \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right| \leq \sum_{k=1}^{K} \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right|. \tag{23}$$

Therefore, by assumption (iii),

$$\sum_{k=1}^{K} \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right| = \sum_{k=1}^{K} \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right|. \tag{24}$$

For each $k \in [K]$, define $\Delta_k := \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right| - \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right|$.

By Eq. 22, $\Delta_k \geq 0$ for every k, while Eq. 24 implies $\sum_{k=1}^{K} \Delta_k = 0$. Hence $\Delta_k = 0$ for every $k \in [K]$, i.e.,

$$\left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right| = \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right|, \quad \forall k \in [K]. \tag{25}$$

Combining Eq. 21 with Eq. 25 yields:

$$\widehat{\mathcal{S}}^{\mathbb{A}}_{t,k} = \sigma_t\left(\mathcal{S}^{\mathbb{A}}_{t,k}\right), \quad \forall k \in [K].$$

The permutation σ_t was chosen once from the single nonsingular matrix $D^\star_t$, and the argument above holds simultaneously for every $k \in [K]$. Therefore, the same permutation aligns all group-specific observation supports. Since the argument applies to any permutation satisfying Eq. 20, every such permutation has the claimed property.

□

A.2 Proofs of Theorem 3.2 and Corollary 3.3

Theorem A.2 (Identifiability of outgoing temporal support)**.** *For $t \in \{1, \dots, T-1\}$, suppose that the true model $\mathcal{G}$ and learned model $\widehat{\mathcal{G}}$ are observationally equivalent and that:*

i. The transition densities $p(z_{t+1} \mid z_t), p(\widehat{z}_{t+1} \mid \widehat{z}_t)$ are positive and C^2, and the observation embeddings induce C^1 diffeomorphisms $h_t : \mathcal{Z}_t \to \widehat{\mathcal{Z}}_t$ and $h_{t+1} : \mathcal{Z}_{t+1} \to \widehat{\mathcal{Z}}_{t+1}$.

ii. The true and learned delayed graphs are faithful to the cross-Hessian. For every $i, j \in [N]$

$$(G^d_{t+1})_{j,i} = 1 \iff \exists (z_{t+1}, z_t) \in \mathcal{Z}_{t+1} \times \mathcal{Z}_t, \ \text{ s.t. } [M_t(z_{t+1}, z_t)]_{j,i} \neq 0,$$

with the analogous equivalence holding for $\widehat{G}^d_{t+1}$ and $\widehat{M}_t$.

iii. For $\forall\, z^\star_t \in \mathcal{Z}_t$, $\operatorname{span}\{M_t(z_{t+1}, z^\star_t) : z_{t+1} \in \mathcal{Z}_{t+1}\} = \mathbb{R}^{N \times N}_{G^d_{t+1}}$.

iv. The learned outgoing support satisfies $|\widehat{\mathcal{S}}^{\mathbb{C}}_t| \leq |\mathcal{S}^{\mathbb{C}}_t|$.

Then, for $\forall\, \widehat{z}_t \in \widehat{\mathcal{Z}}_t$, $\frac{\partial Z_{t,i}}{\partial \widehat{Z}_{t,v}} = 0$ for $i \in \mathcal{S}^{\mathbb{C}}_t$, $v \notin \widehat{\mathcal{S}}^{\mathbb{C}}_t$. For any $\widehat{z}^0_t$, let $D^0_t := J_{h_t^{-1}}(\widehat{z}^0_t)$. At least one permutation σ_t exists s.t. $(D^0_t)_{i,\sigma_t(i)} \neq 0$ for $\forall\, i \in [N]$ satisfies $\widehat{\mathcal{S}}^{\mathbb{C}}_t = \sigma_t\big(\mathcal{S}^{\mathbb{C}}_t\big)$.

Proof. For $\widehat{z}_t \in \widehat{\mathcal{Z}}_t$ and $\widehat{z}_{t+1} \in \widehat{\mathcal{Z}}_{t+1}$, let $z_t = h_t^{-1}(\widehat{z}_t), \quad z_{t+1} = h_{t+1}^{-1}(\widehat{z}_{t+1})$, and define:

$$A_t(\widehat{z}_{t+1}) := J_{h^{-1}_{t+1}}(\widehat{z}_{t+1}), \quad D_t(\widehat{z}_t) := J_{h_t^{-1}}(\widehat{z}_t).$$

By the change-of-variables formula for conditional densities,

$$\widehat{p}(\widehat{z}_{t+1} \mid \widehat{z}_t) = \frac{\widehat{p}(\widehat{z}_{t+1}, \widehat{z}_t)}{\widehat{p}(\widehat{z}_t)} = p(z_{t+1} \mid z_t)\,|\det A_t(\widehat{z}_{t+1})|\,. \tag{26}$$

Let $q_{t+1} := \log p$ and $\widehat{q}_{t+1} := \log \widehat{p}$. Taking logarithms in Eq. 26 gives

$$\widehat{q}_{t+1}(\widehat{z}_{t+1} \mid \widehat{z}_t) = q_{t+1}(z_{t+1} \mid z_t) + \log|\det A_t(\widehat{z}_{t+1})|\,. \tag{27}$$

The second term in Eq. 27 depends only on $\widehat{z}_{t+1}$ and therefore vanishes after taking a derivative with respect to $\widehat{z}_t$.

We now derive the transformation rule for the cross-Hessian. For $r, v \in [N]$, the chain rule gives:

$$\begin{aligned}[\widehat{M}_t(\widehat{z}_{t+1}, \widehat{z}_t)]_{r,v} &= \frac{\partial^2 \widehat{q}_{t+1}(\widehat{z}_{t+1} \mid \widehat{z}_t)}{\partial \widehat{z}_{t+1,r} \partial \widehat{z}_{t,v}} \\ &= \sum_{j=1}^{N}\sum_{i=1}^{N} \frac{\partial z_{t+1,j}}{\partial \widehat{z}_{t+1,r}} \frac{\partial^2 q_{t+1}(z_{t+1} \mid z_t)}{\partial z_{t+1,j} \partial z_{t,i}} \frac{\partial z_{t,i}}{\partial \widehat{z}_{t,v}}.\end{aligned}$$

Hence,

$$\widehat{M}_t(\widehat{z}_{t+1}, \widehat{z}_t) = A_t(\widehat{z}_{t+1})^\top M_t(z_{t+1}, z_t) D_t(\widehat{z}_t), \tag{28}$$

where both $A_t(\widehat{z}_{t+1})$ and $D_t(\widehat{z}_t)$ are nonsingular by assumption (i).

For any $i \in \mathcal{S}_t^{\mathbb{C}}$ and an arbitrary $z_t^\star \in \mathcal{Z}_t$, since i is outgoing-active, there exists $j \in [N]$ s.t. $(G_{t+1}^d)_{j,i} = 1$. Therefore, the coordinate matrix $E_{ji} := e_j e_i^\top$ belongs to $\mathbb{R}^{N\times N}_{G_{t+1}^d}$. By assumption (iii), there exist $L \in \mathbb{N}$, points $\{z_{t+1}^{(\ell)}\}_{\ell=1}^L \subseteq \mathcal{Z}_{t+1}$, and coefficients $\{\alpha_\ell\}_{\ell=1}^L$ such that:

$$E_{ji} = \sum_{\ell=1}^{L} \alpha_\ell M_t\left(z_{t+1}^{(\ell)}, z_t^\star\right). \tag{29}$$

Let $\widehat{z}_t^\star = h_t(z_t^\star)$ and $D_t^\star := D_t(\widehat{z}_t^\star)$. For every $\ell \in [L]$, define:

$$\widehat{z}_{t+1}^{(\ell)} := h_{t+1}\left(z_{t+1}^{(\ell)}\right), \quad A_\ell := A_t\left(\widehat{z}_{t+1}^{(\ell)}\right),$$

$$M_\ell := M_t\left(z_{t+1}^{(\ell)}, z_t^\star\right), \quad \widehat{M}_\ell := \widehat{M}_t\left(\widehat{z}_{t+1}^{(\ell)}, \widehat{z}_t^\star\right).$$

Eq. 28 gives $\widehat{M}_\ell = A_\ell^\top M_\ell D_t^\star$. Since A_ℓ is nonsingular,

$$M_\ell D_t^\star = A_\ell^{-\top} \widehat{M}_\ell. \tag{30}$$

Right-multiplying Eq. 29 by $D_t^\star$ and applying Eq. 30 yields:

$$E_{ji} D_t^\star = \sum_{\ell=1}^{L} \alpha_\ell A_\ell^{-\top} \widehat{M}_\ell. \tag{31}$$

Now take any $v \notin \widehat{\mathcal{S}}_t^{\mathbb{C}}$. We have $(\widehat{G}_{t+1}^d)_{r,v} = 0,\ \forall r \in [N]$. Assumption (ii) therefore implies that the v-th column of the learned cross-Hessian vanishes everywhere:

$$[\widehat{M}_t(\widehat{z}_{t+1}, \widehat{z}_t)]_{:,v} = 0, \quad \forall (\widehat{z}_{t+1}, \widehat{z}_t) \in \widehat{\mathcal{Z}}_{t+1} \times \widehat{\mathcal{Z}}_t. \tag{32}$$

In particular, $(\widehat{M}_\ell)_{:,v} = 0$ for every $\ell \in [L]$. Therefore,

$$(A_\ell^{-\top} \widehat{M}_\ell)_{:,v} = 0, \quad \forall \ell \in [L].$$

Taking the v-th column of Eq. 31 gives:

$$(E_{ji} D_t^\star)_{:,v} = 0. \tag{33}$$

Since $E_{ji} D_t^\star = e_j (e_i^\top D_t^\star)$, its v-th column is $(E_{ji} D_t^\star)_{:,v} = e_j (D_t^\star)_{i,v}$. Combining this identity with Eq. 33 yields $(D_t^\star)_{i,v} = 0$.

The point $z_t^\star$ was arbitrary in $\mathcal{Z}_t$. Since h_t is a bijection between $\mathcal{Z}_t$ and $\widehat{\mathcal{Z}}_t$, we obtain:

$$[J_{h_t^{-1}}(\widehat{z}_t)]_{i,v} = 0, \quad \forall \widehat{z}_t \in \widehat{\mathcal{Z}}_t,\ \forall i \in \mathcal{S}_t^{\mathbb{C}},\ \forall v \notin \widehat{\mathcal{S}}_t^{\mathbb{C}}. \tag{34}$$

Equivalently,

$$\operatorname{supp}\left(J_{h_t^{-1}}(\widehat{z}_t)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{C}}, \quad \forall \widehat{z}_t \in \widehat{\mathcal{Z}}_t,\ \forall i \in \mathcal{S}_t^{\mathbb{C}}. \tag{35}$$

We next identify the outgoing support up to permutation. For $\widehat{z}_t^0 \in \widehat{\mathcal{Z}}_t$ and let $D_t^0 := J_{h_t^{-1}}(\widehat{z}_t^0)$. Since h_t^{-1} is a diffeomorphism, D_t^0 is nonsingular and $\det(D_t^0) \neq 0$. By the Leibniz formula,

$$\det(D_t^0) = \sum_{\sigma \in S_N} \operatorname{sgn}(\sigma) \prod_{i=1}^{N} (D_t^0)_{i,\sigma(i)} \neq 0. \tag{36}$$

Therefore, at least one permutation $\sigma_t \in S_N$ satisfies:

$$(D_t^0)_{i,\sigma_t(i)} \neq 0, \quad \forall i \in [N]. \tag{37}$$

Now fix any permutation σ_t satisfying Eq. 37. For every $i \in \mathcal{S}_t^{\mathbb{C}}$, Eq. 35 evaluated at $\widehat{z}_t^0$ implies that every nonzero entry in row i of D_t^0 has its column index in $\widehat{\mathcal{S}}_t^{\mathbb{C}}$. Since $(D_t^0)_{i,\sigma_t(i)} \neq 0$, we have $\sigma_t(i) \in \widehat{\mathcal{S}}_t^{\mathbb{C}}$. Thus,

$$\sigma_t\left(\mathcal{S}_t^{\mathbb{C}}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{C}}. \tag{38}$$

Since σ_t is a permutation, $\left|\sigma_t\left(\mathcal{S}_t^{\mathbb{C}}\right)\right| = \left|\mathcal{S}_t^{\mathbb{C}}\right|$. Eq. 38 therefore gives:

$$|\mathcal{S}_t^{\mathbb{C}}| \leq |\widehat{\mathcal{S}}_t^{\mathbb{C}}|.$$

Combining this inequality with assumption (iv) yields:

$$\widehat{\mathcal{S}}_t^{\mathbb{C}} = \sigma_t\left(\mathcal{S}_t^{\mathbb{C}}\right).$$

The argument applies to every permutation supported by the nonzero pattern of D_t^0, and Eq. 36 guarantees that at least one such permutation exists. □

Corollary A.3 (Identifiability of incoming temporal support)**.** *For $t \in \{2, \ldots, T\}$, suppose that the true model $\mathcal{G}$ and learned model $\widehat{\mathcal{G}}$ are observationally equivalent and that:*

i. The smoothness and faithfulness conditions in Theorem 3.2 hold for $t-1 \to t$.

ii. For $\forall\ z_t^\star \in \mathcal{Z}_t$,$\mathrm{span}\big\{M_{t-1}(z_t^\star, z_{t-1}) : z_{t-1} \in \mathcal{Z}_{t-1}\big\} = \mathbb{R}^{N\times N}_{G_t^d}$.

iii. The learned incoming support satisfies $|\widehat{\mathcal{S}}_t^{\mathbb{B}}| \leq |\mathcal{S}_t^{\mathbb{B}}|$.

Then, for $\forall\ \widehat{z}_t$, $\frac{\partial Z_{t,i}}{\partial \widehat{Z}_{t,v}} = 0 \quad$ for $i \in \mathcal{S}_t^{\mathbb{B}},\ v \notin \widehat{\mathcal{S}}_t^{\mathbb{B}}$. For any $\widehat{z}_t^0$, At least one permutation σ_t exists, supported by the nonzero pattern of $D_t^0 = J_{h_t^{-1}}(\widehat{z}_t^0)$ satisfies $\widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{B}}\big)$.

Proof. The result follows from the row-side counterpart of the argument for Theorem 3.2. For the transition from $t-1$ to t, the cross-Hessian transforms as:

$$\widehat{M}_{t-1}(\widehat{z}_t, \widehat{z}_{t-1}) = J_{h_t^{-1}}(\widehat{z}_t)^\top M_{t-1}(z_t, z_{t-1}) J_{h_{t-1}^{-1}}(\widehat{z}_{t-1}).$$

Theorem 3.2 uses the column support of the cross-Hessian and right multiplication by $J_{h_t^{-1}}$ to identify outgoing-active coordinates. Here, we exchange the row and column roles.

For $z_t^\star$ and vary z_{t-1}, Assumption (ii) allows every coordinate matrix $E_{ij} = e_i e_j^\top$ associated with an edge $(G_t^d)_{i,j} = 1$ to be expressed as a linear combination of $M_{t-1}(z_t^\star, z_{t-1})$. Since $J_{h_t^{-1}}(h_t(z_t^\star))$ is fixed while z_{t-1} varies, multiplying $\widehat{M}_{t-1}(\widehat{z}_t, \widehat{z}_{t-1})$ on the right by the inverse of $J_{h_{t-1}^{-1}}(\widehat{z}_{t-1})$ shows that

$$[J_{h_t^{-1}}(\widehat{z}_t)]_{i,v} = 0, \quad \forall i \in \mathcal{S}_t^{\mathbb{B}},\ \forall v \notin \widehat{\mathcal{S}}_t^{\mathbb{B}}.$$

The same supported-permutation argument as in Theorem 3.2, followed by the cardinality condition in assumption (iii), gives $\widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t\left(\mathcal{S}_t^{\mathbb{B}}\right)$. □

A.3 Proof of Lemma 3.4

Lemma A.4 (Common support alignment)**.** *For $t \in \{2, \ldots, T-1\}$, suppose that the assumptions of Theorem 3.1, Theorem 3.2, and Corollary 3.3 hold. Let $z_t^\star \in \mathcal{Z}_t$ be the point in assumption (ii) of Theorem 3.1, and define $\widehat{z}_t^\star := h_t(z_t^\star)$ and $D_t^\star := J_{h_t^{-1}}(\widehat{z}_t^\star)$. There exists at least one permutation $\sigma_t : [N] \to [N]$ s.t. $(D_t^\star)_{i,\sigma_t(i)} \neq 0$ for $\forall\ i \in [N],\ \forall\ k \in [K]$ simultaneously satisfies:*

$$\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}} = \sigma_t\big(\mathcal{S}_{t,k}^{\mathbb{A}}\big),\ \ \widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{B}}\big),\ \ \widehat{\mathcal{S}}_t^{\mathbb{C}} = \sigma_t\big(\mathcal{S}_t^{\mathbb{C}}\big). \tag{39}$$

Proof. Since h_t is a diffeomorphism, $D_t^\star = J_{h_t^{-1}}(\widehat{z}_t^\star)$ is nonsingular. Hence $\det(D_t^\star) \neq 0$. By the Leibniz formula,

$$\det(D_t^\star) = \sum_{\sigma \in S_N} \mathrm{sgn}(\sigma) \prod_{i=1}^{N} (D_t^\star)_{i,\sigma(i)} \neq 0,$$

where S_N denotes the set of permutations of $[N]$. Therefore, at least one term in the sum is nonzero, and there exists a permutation $\sigma_t : [N] \to [N]$ s.t.

$$(D_t^\star)_{i,\sigma_t(i)} \neq 0, \quad \forall i \in [N]. \tag{40}$$

We show that every permutation satisfying Eq. 40 simultaneously aligns the three support structures. First, consider the grouped observation supports. By Theorem 3.1,

$$\operatorname{supp}\left((D_t^\star)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}, \quad \forall k \in [K],\ \forall i \in \mathcal{S}_{t,k}^{\mathbb{A}}. \tag{41}$$

Fix $k \in [K]$ and $i \in \mathcal{S}_{t,k}^{\mathbb{A}}$. Eqs. 40 and 41 imply $\sigma_t(i) \in \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}$. Hence,

$$\sigma_t\left(\mathcal{S}_{t,k}^{\mathbb{A}}\right) \subseteq \widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}, \quad \forall k \in [K]. \tag{42}$$

Since σ_t is injective,

$$\left|\mathcal{S}_{t,k}^{\mathbb{A}}\right| = \left|\sigma_t\left(\mathcal{S}_{t,k}^{\mathbb{A}}\right)\right| \leq \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right|, \quad \forall k \in [K]. \tag{43}$$

$$\sum_{k=1}^{K} \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right| \leq \sum_{k=1}^{K} \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right|. \tag{44}$$

The total group-incidence sparsity assumption in Theorem 3.1 gives the reverse inequality, so:

$$\sum_{k=1}^{K} \left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right| = \sum_{k=1}^{K} \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right|. \tag{45}$$

Every difference $\left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right| - \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right|$ is nonnegative by Eq. 43, while their sum is zero by Eq. 45. Therefore, $\left|\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}}\right| = \left|\mathcal{S}_{t,k}^{\mathbb{A}}\right|,\ \forall k \in [K]$. Combining this equality with Eq. 42 yields:

$$\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}} = \sigma_t\left(\mathcal{S}_{t,k}^{\mathbb{A}}\right), \quad \forall k \in [K]. \tag{46}$$

Next, we consider the incoming temporal support. By Corollary 3.3, for every $\widehat{z}_t \in \widehat{\mathcal{Z}}_t$, $\operatorname{supp}\left(J_{h_t^{-1}}(\widehat{z}_t)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{B}},\ \forall i \in \mathcal{S}_t^{\mathbb{B}}$. Evaluating this relation at $\widehat{z}_t^\star$ gives

$$\operatorname{supp}\left((D_t^\star)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{B}}, \quad \forall i \in \mathcal{S}_t^{\mathbb{B}}. \tag{47}$$

For every $i \in \mathcal{S}_t^{\mathbb{B}}$, Eqs. 40 and 47 imply $\sigma_t(i) \in \widehat{\mathcal{S}}_t^{\mathbb{B}}$. Thus,

$$\sigma_t\left(\mathcal{S}_t^{\mathbb{B}}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{B}}. \tag{48}$$

Since σ_t is injective, $\left|\mathcal{S}_t^{\mathbb{B}}\right| = \left|\sigma_t\left(\mathcal{S}_t^{\mathbb{B}}\right)\right| \leq \left|\widehat{\mathcal{S}}_t^{\mathbb{B}}\right|$. The sparsity condition in Corollary 3.3 gives $\left|\widehat{\mathcal{S}}_t^{\mathbb{B}}\right| \leq \left|\mathcal{S}_t^{\mathbb{B}}\right|$. Therefore the two cardinalities are equal, and Eq. 48 gives:

$$\widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t\left(\mathcal{S}_t^{\mathbb{B}}\right). \tag{49}$$

Finally, Theorem 3.2 establishes that, for every $\widehat{z}_t \in \widehat{\mathcal{Z}}_t$, $\operatorname{supp}\left(J_{h_t^{-1}}(\widehat{z}_t)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{C}},\ \forall i \in \mathcal{S}_t^{\mathbb{C}}$. Evaluating at $\widehat{z}_t^\star$ gives

$$\operatorname{supp}\left((D_t^\star)_{i,:}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{C}}, \quad \forall i \in \mathcal{S}_t^{\mathbb{C}}. \tag{50}$$

Hence, for every $i \in \mathcal{S}_t^{\mathbb{C}}$, Eq. 40 implies $\sigma_t(i) \in \widehat{\mathcal{S}}_t^{\mathbb{C}}$, and therefore:

$$\sigma_t\left(\mathcal{S}_t^{\mathbb{C}}\right) \subseteq \widehat{\mathcal{S}}_t^{\mathbb{C}}. \tag{51}$$

Injectivity of σ_t and the sparsity condition in Theorem 3.2 give $\left|\mathcal{S}_t^{\mathbb{C}}\right| = \left|\sigma_t\left(\mathcal{S}_t^{\mathbb{C}}\right)\right| \leq \left|\widehat{\mathcal{S}}_t^{\mathbb{C}}\right| \leq \left|\mathcal{S}_t^{\mathbb{C}}\right|$. All inequalities are therefore equalities, and Eq. 51 yields:

$$\widehat{\mathcal{S}}_t^{\mathbb{C}} = \sigma_t\left(\mathcal{S}_t^{\mathbb{C}}\right). \tag{52}$$

Eqs. 46, 49, and 52 are obtained from the same permutation σ_t selected from the nonzero pattern of the single matrix $D_t^\star$. Hence σ_t simultaneously aligns the grouped observation, incoming, and outgoing supports. The argument applies to every permutation satisfying Eq. 40. □

A.4 Proof of Theorem 3.5

Theorem A.5 (Joint identifiability of temporal latent roles)**.** *Under the conditions of Lemma 3.4, there exists a permutation $\sigma_t : [N] \to [N]$ s.t., for $\forall\, i \in [N],\ \forall\, k \in [K]$,*

$$\left(\widehat{\mathbb{A}}^{(k)}_{t,\sigma_t(i)}, \widehat{\mathbb{B}}_{t,\sigma_t(i)}, \widehat{\mathbb{C}}_{t,\sigma_t(i)}\right) = \left(\mathbb{A}^{(k)}_{t,i}, \mathbb{B}_{t,i}, \mathbb{C}_{t,i}\right). \tag{53}$$

Consequently, for every $\boldsymbol{a} \in \{0,1\}^K$ and $b, c \in \{0,1\}$,

$$\widehat{\mathcal{R}}^{\boldsymbol{a},b,c}_t = \sigma_t\big(\mathcal{R}^{\boldsymbol{a},b,c}_t\big). \tag{54}$$

Proof. By Lemma 3.4, there exists a single permutation $\sigma_t : [N] \to [N]$ s.t. for $\forall k \in [K]$:

$$\widehat{\mathcal{S}}^{\mathbb{A}}_{t,k} = \sigma_t\big(\mathcal{S}^{\mathbb{A}}_{t,k}\big),\ \ \widehat{\mathcal{S}}^{\mathbb{B}}_t = \sigma_t\big(\mathcal{S}^{\mathbb{B}}_t\big),\ \ \widehat{\mathcal{S}}^{\mathbb{C}}_t = \sigma_t\big(\mathcal{S}^{\mathbb{C}}_t\big).$$

For every $i \in [N]$ and $k \in [K]$,

$$\begin{aligned}
\widehat{\mathbb{A}}^{(k)}_{t,\sigma_t(i)} = 1 &\iff \sigma_t(i) \in \widehat{\mathcal{S}}^{\mathbb{A}}_{t,k} \\
&\iff \sigma_t(i) \in \sigma_t\left(\mathcal{S}^{\mathbb{A}}_{t,k}\right) \\
&\iff i \in \mathcal{S}^{\mathbb{A}}_{t,k} \\
&\iff \mathbb{A}^{(k)}_{t,i} = 1.
\end{aligned}$$

Since both quantities are binary,

$$\widehat{\mathbb{A}}^{(k)}_{t,\sigma_t(i)} = \mathbb{A}^{(k)}_{t,i}, \quad \forall i \in [N],\ \forall k \in [K]. \tag{55}$$

For the incoming indicator, for every $i \in [N]$,

$$\begin{aligned}
\widehat{\mathbb{B}}_{t,\sigma_t(i)} = 1 &\iff \sigma_t(i) \in \widehat{\mathcal{S}}^{\mathbb{B}}_t \\
&\iff \sigma_t(i) \in \sigma_t\left(\mathcal{S}^{\mathbb{B}}_t\right) \\
&\iff i \in \mathcal{S}^{\mathbb{B}}_t \\
&\iff \mathbb{B}_{t,i} = 1.
\end{aligned}$$

Hence,

$$\widehat{\mathbb{B}}_{t,\sigma_t(i)} = \mathbb{B}_{t,i}, \quad \forall i \in [N]. \tag{56}$$

Similarly, for the outgoing indicator,

$$\begin{aligned}
\widehat{\mathbb{C}}_{t,\sigma_t(i)} = 1 &\iff \sigma_t(i) \in \widehat{\mathcal{S}}^{\mathbb{C}}_t \\
&\iff \sigma_t(i) \in \sigma_t\left(\mathcal{S}^{\mathbb{C}}_t\right) \\
&\iff i \in \mathcal{S}^{\mathbb{C}}_t \\
&\iff \mathbb{C}_{t,i} = 1.
\end{aligned}$$

Therefore,

$$\widehat{\mathbb{C}}_{t,\sigma_t(i)} = \mathbb{C}_{t,i}, \quad \forall i \in [N]. \tag{57}$$

Eqs. 55, 56, and 57 establish the joint signature equality. For $\mathbf{a} = (a_1, \ldots, a_K) \in \{0,1\}^K$ and $b, c \in \{0,1\}$. For every $i \in [N]$,

$$\begin{aligned}
i \in \mathcal{R}^{\mathbf{a},b,c}_t &\iff \mathbb{A}^{(k)}_{t,i} = a_k \text{ for all } k \in [K], \quad \mathbb{B}_{t,i} = b, \quad \mathbb{C}_{t,i} = c \\
&\iff \widehat{\mathbb{A}}^{(k)}_{t,\sigma_t(i)} = a_k \text{ for all } k \in [K], \quad \widehat{\mathbb{B}}_{t,\sigma_t(i)} = b, \quad \widehat{\mathbb{C}}_{t,\sigma_t(i)} = c \\
&\iff \sigma_t(i) \in \widehat{\mathcal{R}}^{\mathbf{a},b,c}_t.
\end{aligned}$$

Thus,

$$i \in \mathcal{R}^{\mathbf{a},b,c}_t \iff \sigma_t(i) \in \widehat{\mathcal{R}}^{\mathbf{a},b,c}_t.$$

Since σ_t is bijective, the pointwise membership equivalence implies:

$$\widehat{\mathcal{R}}^{\mathbf{a},b,c}_t = \sigma_t\left(\mathcal{R}^{\mathbf{a},b,c}_t\right).$$

□

A.5 COMMON STRUCTURAL ALIGNMENT AND GLOBAL PERMUTATION INVARIANCE

Lemma 3.4 concerns the correspondence among the three structural roles at the same time point. Recall that $D_t^\star = J_{h_t^{-1}}(\widehat{z}_t^\star)$ is nonsingular. The preceding results show that the group, incoming, and outgoing supports all restrict the rows of this same matrix. Hence, a permutation supported by the nonzero pattern of $D_t^\star$ can align $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$ simultaneously. Figure 5 illustrates how this common alignment differs from the global permutation invariance considered in related identifiability results.

Relation to global permutation invariance. Figure 5 (a) illustrates why a locally valid permutation need not remain the same over the whole latent domain. The two disconnected regions $\widehat{\mathcal{U}}_1$ and $\widehat{\mathcal{U}}_2$ admit different local Jacobian support patterns: one corresponds to the identity assignment and the other to a coordinate swap. Since there is no continuous path connecting the two regions, these local assignments can differ without contradiction. This type of example appears in the discussion of local and global disentanglement in (Lachapelle et al., 2023).

Figure 5(b) gives the geometric intuition for how path connectedness and continuity can prevent such a switch. Consider one row of $J_{h_t^{-1}}(\widehat{z}_t)$ in two dimensions. Its two entries are:

$$\left(\frac{\partial Z_{t,i}}{\partial \widehat{Z}_{t,1}}, \frac{\partial Z_{t,i}}{\partial \widehat{Z}_{t,2}}\right).$$

If this row has only one nonzero entry, it lies on one of the two coordinate axes. Along the continuous path $\gamma(s)$ from $\widehat{z}_t^{(a)}$ to $\widehat{z}_t^{(b)}$, switching from one active coordinate to the other would require the row to pass through either the origin or a point with two nonzero entries. The origin gives a zero Jacobian row and violates nonsingularity, while the second case violates the assumed local permutation structure. The active coordinate therefore stays fixed along the path. This is the continuity argument used in (Chen et al., 2024) to establish permutation invariance, with a related block-level argument developed by (Lachapelle et al., 2023).

Our common alignment result addresses a different source of ambiguity, as shown in Figure 5(c). At the reference point, consider:

$$D_t^\star = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & 0 \end{pmatrix}.$$

The true coordinates $Z_{t,1}$ and $Z_{t,2}$ have the same structural signature $((1,0),1,0)$, while $Z_{t,3}$ has signature $((0,1),0,1)$. In the learned representation, $\widehat{Z}_{t,2}$ and $\widehat{Z}_{t,3}$ share the first signature, and $\widehat{Z}_{t,1}$ has the second. The nonzero pattern of $D_t^\star$ supports both:

$$\sigma_t: \quad Z_{t,1} \mapsto \widehat{Z}_{t,2}, \quad Z_{t,2} \mapsto \widehat{Z}_{t,3}, \quad Z_{t,3} \mapsto \widehat{Z}_{t,1},$$

$$\widetilde{\sigma}_t: \quad Z_{t,1} \mapsto \widehat{Z}_{t,3}, \quad Z_{t,2} \mapsto \widehat{Z}_{t,2}, \quad Z_{t,3} \mapsto \widehat{Z}_{t,1}.$$

Both permutations preserve the group, incoming, and outgoing roles. Their difference only exchanges coordinates with the same structural signature. Thus, Lemma 3.4 requires a common permutation for the three structural views and does not require this permutation to be unique.

This distinction also clarifies the role of path connectedness in related temporal identification results. (Li et al., 2025a) uses transition cross Hessians and structural sparsity to identify a temporal active support and includes path connectedness among its assumptions. Its active support combines observation and temporal participation. Our analysis keeps $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$ as separate structural objects and then aligns them through Lemma 3.4. The resulting guarantee is sufficient for identifying the structural roles studied here, while global component-wise disentanglement requires a stronger form of permutation consistency.

Temporal coordinate consistency. The global permutation discussion above concerns consistency over the latent domain at a fixed time point. A stronger temporal requirement would use one permutation σ over the full sequence, so that $\sigma_t = \sigma$ for all t. Our main results allow σ_t to depend on t: Lemma 3.4 aligns $\mathbb{A}_t$, $\mathbb{B}_t$, and $\mathbb{C}_t$ under the same permutation at each time point, while the argument is applied separately to each $D_t = J_{h_t^{-1}}$. The relation $\widehat{M}_t = D_{t+1}^\top M_t D_t$ is also

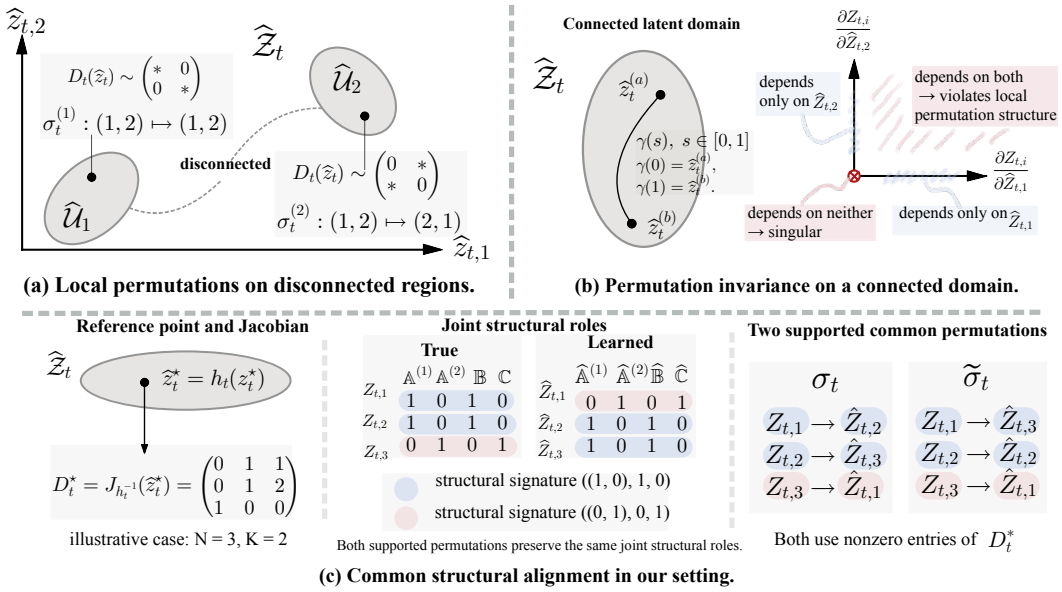


Figure 5: Common structural alignment and global permutation invariance. (a) Disconnected regions may admit different local permutations; (b) continuity on a connected domain prevents permutation switching under the local permutation constraint; (c) in our setting, a supported permutation of $D_t^\star$ aligns the group, incoming, and outgoing structural roles.

compatible with distinct coordinate relabelings at adjacent times, since its rows and columns follow the coordinates at $t+1$ and t. The present guarantee therefore identifies the structural role of each latent coordinate at every time point. Tracking a persistent latent identity over the full sequence requires additional conditions that link the reparameterizations across time, such as a shared latent reparameterization or suitable continuity conditions over a common temporal context.

B Extensions of the Temporal Model

The main analysis uses a first-order delayed process to isolate the core structural argument. This section describes how the same objects can be adapted to richer temporal models.

B.1 Higher-order temporal dependencies

Higher-order dynamics allow latent coordinates to interact across several time lags. Let L be a positive integer denoting the maximum lag and write $z_{a:b} := (z_a, \dots, z_b)$ for $a \le b$. We retain the observation model $X_t = g_t(Z_t)$ and extend the dynamics of each coordinate $i \in [N]$ to:

$$Z_{t,i} = f_i\big(\mathrm{Pa}_d(Z_{t,i}), \epsilon_{t,i}\big), \quad \mathrm{Pa}_d(Z_{t,i}) \subseteq \bigcup_{\ell=1}^{L} Z_{t-\ell}, \quad t = L+1, \dots, T. \tag{58}$$

The initial block $Z_{1:L}$ follows a joint initial distribution, and the subsequent noises remain mutually independent and independent of this block. We assume no instantaneous dependencies in this subsection. The transition density is $p(z_t \mid z_{t-L:t-1})$. For clarity, take $T > 2L$ and consider times $L+1 \le t \le T-L$, where all required transitions are available. Related temporal identification frameworks also accommodate longer conditioning histories.

The temporal roles can summarize participation across all lags. Let $G_t^{d,[\ell]} \in \{0,1\}^{N \times N}$ satisfy $(G_t^{d,[\ell]})_{i,j} = 1$ iff $Z_{t-\ell,j} \in \mathrm{Pa}_d(Z_{t,i})$. We extend the incoming and outgoing supports to:

$$\begin{aligned} \mathcal{S}_t^{\mathbb{B}} &:= \big\{ i \in [N] : \exists \ell \in [L], j \in [N] \text{ s.t. } (G_t^{d,[\ell]})_{i,j} = 1 \big\}, \\ \mathcal{S}_t^{\mathbb{C}} &:= \big\{ i \in [N] : \exists \ell \in [L], j \in [N] \text{ s.t. } (G_{t+\ell}^{d,[\ell]})_{j,i} = 1 \big\}. \end{aligned} \tag{59}$$

As before, $\mathbb{B}_{t,i} = \mathbf{1}\{i \in \mathcal{S}_t^{\mathbb{B}}\}$ and $\mathbb{C}_{t,i} = \mathbf{1}\{i \in \mathcal{S}_t^{\mathbb{C}}\}$. These roles record whether a coordinate has any delayed parent or child within the lag range. The group roles $\{\mathbb{A}_{t,i}^{(k)}\}_{k=1}^{K}$ remain unchanged. The derivative objects must condition on the full history. For each valid transition and $\ell \in [L]$:

$$M_t^{[\ell]} := \nabla_{z_{t+\ell}} \nabla_{z_t}^{\top} \log p\big(z_{t+\ell} \mid z_{t+\ell-L:t+\ell-1}\big). \tag{60}$$

All other history coordinates are held fixed when taking these derivatives. This conditioning is important for preserving the interpretation of a direct delayed edge. For example, a path through Z_{t+1} can make Z_{t+2} depend on Z_t after Z_{t+1} is marginalized out, even when the graph has no direct edge from Z_t to Z_{t+2}. We therefore use the conditional density in Eq. 60 to characterize influence at lag ℓ. The learned blocks $\widehat{M}_t^{[\ell]}$ are defined analogously.

The same change of latent coordinates applies to every lag. Suppose that the true and learned models are observationally equivalent, their transition densities are positive and C^2, and their observation embeddings induce the C^1 diffeomorphisms $h_s = \widehat{g}_s^{-1} \circ g_s$. Write $D_s := J_{h_s^{-1}}(\widehat{z}_s)$. The conditional density transformation gives:

$$\widehat{M}_t^{[\ell]} = D_{t+\ell}^{\top} M_t^{[\ell]} D_t. \tag{61}$$

The log Jacobian term depends only on the target state and disappears after differentiation with respect to the history coordinate. Thus, the calculation in Appendix A.2 extends to every lag.

The incoming and outgoing blocks can be collected around the same time point. Suppressing their state arguments, define:

$$\mathcal{M}_t^{\mathbb{B}} := \begin{bmatrix} M_{t-1}^{[1]} & \cdots & M_{t-L}^{[L]} \end{bmatrix}, \quad \mathcal{M}_t^{\mathbb{C}} := \begin{bmatrix} (M_t^{[1]})^{\top} & \cdots & (M_t^{[L]})^{\top} \end{bmatrix}^{\top}. \tag{62}$$

These matrices have dimensions $N \times LN$ and $LN \times N$. Assume the counterpart of cross Hessian faithfulness for each lag, $\operatorname{supp}(M_t^{[\ell]}) = \operatorname{supp}(G_{t+\ell}^{d,[\ell]})$, in both models. The row support of $\mathcal{M}_t^{\mathbb{B}}$ and the column support of $\mathcal{M}_t^{\mathbb{C}}$ then coincide with the corresponding temporal role supports. Let $\Delta_t^{-} := \operatorname{diag}(D_{t-1}, \ldots, D_{t-L})$ and $\Delta_t^{+} := \operatorname{diag}(D_{t+1}, \ldots, D_{t+L})$. Eq. 61 yields:

$$\widehat{\mathcal{M}}_t^{\mathbb{B}} = D_t^{\top} \mathcal{M}_t^{\mathbb{B}} \Delta_t^{-}, \quad \widehat{\mathcal{M}}_t^{\mathbb{C}} = (\Delta_t^{+})^{\top} \mathcal{M}_t^{\mathbb{C}} D_t. \tag{63}$$

Both temporal views therefore constrain the same matrix D_t. A sufficient variation condition can be stated directly for the role directions. For every fixed $z_t^{\star} \in \mathcal{Z}_t$, suppose that:

$$\begin{aligned} \operatorname{span}\big\{[\mathcal{M}_t^{\mathbb{B}}]_{:,j} : j \in [LN],\ z_t = z_t^{\star}\big\} &= \mathbb{R}^N_{\mathcal{S}_t^{\mathbb{B}}}, \\ \operatorname{span}\big\{[\mathcal{M}_t^{\mathbb{C}}]_{j,:}^{\top} : j \in [LN],\ z_t = z_t^{\star}\big\} &= \mathbb{R}^N_{\mathcal{S}_t^{\mathbb{C}}}. \end{aligned} \tag{64}$$

Here, all state arguments other than z_t range over the admissible support. The first condition separates the incoming directions using the history, while the second separates the outgoing directions using the remaining temporal context. When $L = 1$, the matrix spanning assumptions in Theorem 3.2 and Corollary 3.3 imply these conditions.

These conditions recover the support separation used in the main analysis. If $v \notin \widehat{\mathcal{S}}_t^{\mathbb{B}}$, the vth row of $\widehat{\mathcal{M}}_t^{\mathbb{B}}$ vanishes throughout its domain. Since Δ_t^{-} is nonsingular, Eq. 63 gives $(D_t e_v)^{\top} \mathcal{M}_t^{\mathbb{B}} = 0$. Similarly, $v \notin \widehat{\mathcal{S}}_t^{\mathbb{C}}$ implies $\mathcal{M}_t^{\mathbb{C}} D_t e_v = 0$. At fixed z_t, D_t is fixed, so Eq. 64 forces $(D_t)_{i,v} = 0$ for the true active indices i. With $|\widehat{\mathcal{S}}_t^{\mathbb{B}}| \le |\mathcal{S}_t^{\mathbb{B}}|$ and $|\widehat{\mathcal{S}}_t^{\mathbb{C}}| \le |\mathcal{S}_t^{\mathbb{C}}|$, the supported permutation argument applies. Together with the assumptions of Theorem 3.1, it gives one permutation σ_t s.t.

$$\widehat{\mathcal{S}}_{t,k}^{\mathbb{A}} = \sigma_t(\mathcal{S}_{t,k}^{\mathbb{A}}), \quad \widehat{\mathcal{S}}_t^{\mathbb{B}} = \sigma_t(\mathcal{S}_t^{\mathbb{B}}), \quad \widehat{\mathcal{S}}_t^{\mathbb{C}} = \sigma_t(\mathcal{S}_t^{\mathbb{C}}). \tag{65}$$

Consequently, the joint signature and its role classes are preserved under the same relabeling at time t. This extension retains the original granularity of the temporal roles. Identifying the incoming and outgoing supports separately for each lag would require variation and minimality conditions for those separate supports. For estimation, an extension of RoleTrace can condition its transition model on L previous states and compute role scores from the blocks in Eq. 62, using one common coordinate transformation for all blocks involving Z_t.

B.2 INSTANTANEOUS DEPENDENCIES

Instantaneous dependencies allow latent coordinates to influence one another within the same time step. We return to a temporal lag of one and consider:

$$Z_{t,i} = f_i\big(\mathrm{Pa}_d(Z_{t,i}), \mathrm{Pa}_e(Z_{t,i}), \epsilon_{t,i}\big), \quad \mathrm{Pa}_d(Z_{t,i}) \subseteq Z_{t-1}, \tag{66}$$

for $t = 2, \dots, T$ and $i \in [N]$, where $\mathrm{Pa}_e(Z_{t,i}) \subseteq Z_t \setminus \{Z_{t,i}\}$ denotes the instantaneous parents. We assume that these edges form a directed acyclic graph within each time step and retain mutually independent innovations that are independent of the past. The observation model remains $X_t = g_t(Z_t)$. The supports $\mathcal{S}_t^{\mathbb{B}}$ and $\mathcal{S}_t^{\mathbb{C}}$ again refer to direct delayed edges between adjacent times.

The observation argument and the temporal transformation rule remain available. Under the assumptions of Theorem 3.1, group support identification follows from the same decoder Jacobian argument. Positive C^2 transition densities and the induced diffeomorphisms also give $\widehat{M}_{t-1} = D_t^\top M_{t-1} D_{t-1}$. This identity follows the change-of-variables formula for the joint conditional density and does not require the coordinates of Z_t to be conditionally independent given Z_{t-1}.

The interpretation of incoming support requires additional effort. Let $P_t^e(i) := \{j \in [N] : Z_{t,j} \in \mathrm{Pa}_e(Z_{t,i})\}$. The transition density factorizes as:

$$p(z_t \mid z_{t-1}) = \prod_{i=1}^{N} p\big(z_{t,i} \mid z_{t,P_t^e(i)}, z_{t-1}\big), \tag{67}$$

where each factor depends on the past only through its delayed parents. An instantaneous parent appears in the factor of its child, so differentiating this factor can create a nonzero incoming row for a coordinate without a delayed parent. For example, consider:

$$Z_{t,1} = \epsilon_{t,1}, \quad Z_{t,2} = aZ_{t-1,1} + bZ_{t,1} + \epsilon_{t,2}, \quad ab \neq 0, \tag{68}$$

with independent standard Gaussian noises. Here, $(\mathbb{B}_{t,1}, \mathbb{B}_{t,2}) = (0, 1)$, while direct differentiation of the conditional density gives:

$$[M_{t-1}]_{:,1} = \begin{pmatrix} -ab \\ a \end{pmatrix}. \tag{69}$$

Both rows are nonzero. The additional row arises because $Z_{t,1}$ and $Z_{t-1,1}$ enter the same conditional mechanism for $Z_{t,2}$. Thus, the row support can differ from the set of direct delayed receivers. Therefore, a restriction on instantaneous parents preserves the zero rows needed by our argument. Specifically, suppose that, at each relevant time, both models satisfy:

$$P_t^e(i) \subseteq \mathcal{S}_t^{\mathbb{B}} \quad \text{for every } i \in \mathcal{S}_t^{\mathbb{B}}, \tag{70}$$

with the corresponding hatted quantities in the learned model. This condition requires every instantaneous parent of a delayed receiver to have a delayed parent as well. It allows instantaneous edges among delayed receivers and allows them to influence other coordinates within the same time step. Under Eq. 70, every factor in Eq. 67 that depends on the past involves current coordinates only from $\mathcal{S}_t^{\mathbb{B}}$. Hence, $i \notin \mathcal{S}_t^{\mathbb{B}} \implies [M_{t-1}]_{i,:} \equiv 0$. The same implication holds in the learned model. On the outgoing side, a coordinate without any delayed child appears in none of the conditional factors for the next state, which already gives $i \notin \mathcal{S}_t^{\mathbb{C}} \Rightarrow [M_t]_{:,i} \equiv 0$.

Joint role identification then follows under variation and minimality at the level of these supports. In addition to the smoothness and observation assumptions above, impose Eq. 64 with $L = 1$, $\mathcal{M}_t^{\mathbb{B}} = M_{t-1}$, and $\mathcal{M}_t^{\mathbb{C}} = M_t$, together with $|\widehat{\mathcal{S}}_t^{\mathbb{B}}| \leq |\mathcal{S}_t^{\mathbb{B}}|$ and $|\widehat{\mathcal{S}}_t^{\mathbb{C}}| \leq |\mathcal{S}_t^{\mathbb{C}}|$. The zero rows and columns established above yield:

$$(D_t e_v)^\top M_{t-1} = 0 \quad \text{if } v \notin \widehat{\mathcal{S}}_t^{\mathbb{B}}, \quad M_t D_t e_v = 0 \quad \text{if } v \notin \widehat{\mathcal{S}}_t^{\mathbb{C}}.$$

The spanning conditions again force the corresponding entries of D_t to vanish. Evaluating at the observation reference point and using the supported permutation argument in the proof of Lemma 3.4 gives the common alignment in Eq. 65.

More general instantaneous models can be approached through richer causal identification conditions. Existing work uses temporal context, sufficient variability, and sparse influence constraints to identify latent components in the presence of instantaneous dependence (Li et al., 2025b). Under their respective assumptions, these results provide a complementary route in which temporal roles can be read from a recovered delayed graph. The additional restriction in Eq. 70 specifies the class considered by the support argument above.

C Related Work

In this section, we review identification results for latent variable models, temporal processes, and partially observed systems, and explain how they relate to the structural roles studied in this work.

Latent Variable Identification and Observation Structure. Latent variable identification studies which properties of a hidden generating process can be recovered from observations. Independent component analysis (ICA) provides a classical formulation, in which statistically independent sources are recovered from their linear mixtures under suitable distributional conditions (Comon, 1994; Bach & Jordan, 2003; Eriksson & Koivunen, 2004). With general nonlinear mixing, independence alone leaves substantial ambiguity in the recovered sources (Hyvärinen & Pajunen, 1999). Auxiliary variable approaches obtain identification from conditional independence and sufficient variation across an observed variable (Hyvärinen et al., 2018; Khemakhem et al., 2020), while other results restrict the generating function (Buchholz et al., 2022; Lachapelle et al., 2023). Structural assumptions on the observation map provide another route. Structural sparsity constrains the dependency pattern between sources and observed variables, enabling identification under suitable conditions on the support of the mixing Jacobian Li et al. (2025b); Zheng & Zhang (2023). G-CaRL uses a known grouping of observational variables to identify latent causal variables and their relations, with observed and latent variables partitioned into corresponding groups (Morioka & Hyvarinen, 2024). Multiple views provide another source of identification information (Gresele et al., 2019; Zhang et al., 2024), while (Zheng et al., 2025) identifies shared and private latent information together with the incidence structure between latent factors and observation groups. Our setting allows one latent coordinate to be expressed in several groups and retains its complete incidence profile $\{\mathbb{A}_{t,i}^{(k)}\}_{k=1}^{K}$. We study this observation structure jointly with incoming and outgoing temporal participation, preserving the correspondence among all three attributes.

Temporal Structure for Latent Identification. Temporal observations provide identification information through changes in source distributions and dependence across time. One line of work exploits observed segments, latent regimes, or multiple distributions to identify nonlinear representations from nonstationary data (Hyvärinen & Morioka, 2016; Yao et al., 2022b; Song et al., 2023; Brouillard et al., 2026). Others use dependence across consecutive states or independent innovations as the identifying signal, covering stationary nonlinear ICA and more general autoregressive processes (Hyvarinen & Morioka, 2017; Morioka et al., 2021; Yao et al., 2022a). Historical context can recover latent states when a single observation loses information (Chen et al., 2024). Sparse temporal mechanisms support identification of latent components and dependency structure (Lachapelle et al., 2022; Li et al., 2025b; Lachapelle et al., 2026). Intervention-based approaches provide another source of temporal structure and establish identification of causal factors and their instantaneous or delayed relations (Lippe et al., 2022; 2023). These methods characterize latent components, mechanisms, or causal edges at different levels of granularity. Our target is complementary: we characterize each latent coordinate by whether it receives delayed influence and whether it transmits delayed influence, represented by $\mathbb{B}_{t,i}$ and $\mathbb{C}_{t,i}$. This factorized temporal role preserves distinctions between coordinates that receive, transmit, or locally terminate temporal information, and it can be combined with observation structure without requiring recovery of every delayed causal edge.

Partial Observation and Latent Participation. Several recent formulations allow latent information to be exposed only through selected views, instances, or time steps. MuLTI studies overlapping latent processes across multiple time series views and learns permutation-based correspondence between view specific estimates (Huang et al., 2023). Multi-view causal representation learning identifies information shared across subsets of simultaneously observed views under partial observability (Yao et al., 2024), while sparsity based approaches allow the visible subset to vary across unpaired observations (Xu et al., 2024). InterLatent considers a time varying active support defined jointly through participation in the observation and adjacent transition mechanisms (Li et al., 2025a). These works target view correspondence, partially visible latent subsets, or a single active support. Our formulation keeps group expression $\mathbb{A}$, incoming influence $\mathbb{B}$, and outgoing influence $\mathbb{C}$ as separate structural objects and identifies their correspondence under one common permutation. This separation distinguishes latent coordinates with the same observation profile but different temporal roles, including factors that only receive or only transmit delayed influence.

D Additional Experimental Details

D.1 Additional Details for Synthetic Experiments

This section specifies the synthetic mechanisms, estimator, and evaluation, followed by additional results for RQ1 and the controlled designs for RQ2 and RQ3.

Synthetic Data Generation. The RQ1 datasets are generated from independently sampled random mechanisms. Within each mechanism, the observation map and transition law are fixed across time, so its group structure and delayed graph remain constant, and we suppress the time index on the roles below. This construction is time homogeneous; the initial distribution need not be the stationary distribution. Figure 2a illustrates one sampled mechanism. RQ2 uses different graphs on its two transitions, as specified separately below.

We set $N = 6$, $K = 3$, and $T = 5$, with six observed dimensions in each group. The initial latent state is sampled as $Z_1 \sim \mathcal{N}(0, I_N)$. Let $\mathbf{A} \in \{0,1\}^{K\times N}$ denote the group–latent incidence matrix. Each entry is sampled with probability 0.55, subject to the conditions that every observation group contains at least one active latent coordinate and every latent coordinate is expressed in at least one group. For each $k \in [K]$, let $W^{(k)} \in \mathbb{R}^{6\times N}$. When $\mathbf{A}_{k,i} = 1$, the corresponding column of $W^{(k)}$ is sampled from a scaled orthonormal basis. Columns associated with $\mathbf{A}_{k,i} = 0$ are set to zero. The observation mechanism is:

$$a_t^{(k)} = W^{(k)} Z_t, \quad X_t^{(k)} = a_t^{(k)} + 0.12 \tanh\Big(a_t^{(k)}\Big). \tag{71}$$

Let $G^d \in \{0,1\}^{N\times N}$ denote the delayed graph. For every mechanism, three latent coordinates are sampled as candidate senders and three as candidate receivers. Additional admissible edges are sampled independently with probability 0.35. The matrices $W^\mu, W^\sigma \in \mathbb{R}^{N\times N}$ share the support G^d. For every active entry, the sign is sampled independently and the magnitude is sampled uniformly from $[0.45, 0.90]$. For $t \geq 2$, define:

$$\mu_i(Z_{t-1}) = 0.75 \tanh\left(\sum_{j=1}^{N} W_{ij}^{\mu} \left[Z_{t-1,j} + 0.15 \sin(Z_{t-1,j})\right]\right), \tag{72}$$

$$\log \sigma_i(Z_{t-1}) = -0.65 + 0.30 \tanh\left(\sum_{j=1}^{N} W_{ij}^{\sigma} \left[Z_{t-1,j} + 0.10 \tanh(2Z_{t-1,j})\right]\right), \tag{73}$$

$$Z_{t,i} = \mu_i(Z_{t-1}) + \sigma_i(Z_{t-1})\epsilon_{t,i}, \quad \epsilon_{t,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{74}$$

The state-dependent mean and variance make the transition cross-Hessian vary with the latent state. The temporal roles are:

$$\mathbb{B}_i = \mathbf{1}\left\{\sum_{j=1}^{N} G_{ij}^d > 0\right\}, \quad \mathbb{C}_i = \mathbf{1}\left\{\sum_{j=1}^{N} G_{ji}^d > 0\right\}. \tag{75}$$

Thus, $\mathbb{B}_i = 1$ indicates that coordinate i receives delayed influence, and $\mathbb{C}_i = 1$ indicates that coordinate i transmits delayed influence. We retain 16 independently sampled mechanisms. For every mechanism, we generate one master training set and use nested prefixes of sizes $n \in \{50, 100, 150, 200, 400, 800, 1600, 3200\}$. The validation and test sets contain 512 and 1024 additional sequences, respectively. They remain fixed across sample sizes and methods.

Estimator and Optimization. The representation model follows the procedure described in Section 4. The transition density is parameterized as a conditional diagonal Gaussian. Its mean and log variance networks each use two $\tanh$ hidden layers of width 32. The representation and transition stages are trained for 150 epochs using AdamW with batch size 32, initial learning rate 10^{-3}, and weight decay 10^{-5}. After epoch 15, a cosine schedule decreases the learning rate to 0.05 of its initial value. Before aggregation, derivative magnitudes are normalized by the empirical standard deviation of the corresponding latent coordinates. For the common coordinate objective, we use:

$$(\lambda_{\mathbb{A}}, \lambda_{\mathbb{B}}, \lambda_{\mathbb{C}}) = (0.004, 0.0015, 0.0015). \tag{76}$$

The common coordinate transformation is optimized using the Jacobi procedure described in Section 4. After the structural scores are obtained, each score vector is transformed to log space and partitioned by two means clustering. The procedure is applied separately to each $\mathbb{A}$ group and separately to $\mathbb{B}$ and $\mathbb{C}$. The resulting cluster assignment determines the estimated binary support.

For ROLETRACE, the first common coordinate update is followed by one transition refit in the updated latent coordinates. We then recompute the temporal derivatives and perform one final Jacobi update. NO-REFIT uses the first common coordinate update as its final estimate. The remaining baselines isolate different sources of structural information. NO STRUCTURAL retains the initial representation and transition, while TEMPORAL-ONLY uses $\lambda_{\mathbb{A}} = 0$ for coordinate selection. The THOUGHTCOMM adaptation selects coordinates using the observation term alone and evaluates the temporal roles in those same coordinates. The INTERLATENT adaptation trains reconstruction, transition likelihood, and the three role penalties jointly in the grouped observation setting.

Evaluation and Additional RQ1 Results. Evaluation compares the complete roles under a single relabeling of the learned coordinates. Let $\widehat{r}_i$ and r_i be the estimated and true $(K+2)$ bit signatures. Hungarian matching finds:

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^{N} \|\widehat{r}_{\sigma(i)} - r_i\|_1. \tag{77}$$

This permutation only relabels an already fitted model for evaluation. The orthogonal coordinate transformation Q_t is learned from the structural objective without ground truth, and all RQ1 metrics use the same σ^* for the three roles. The metrics distinguish recovery of individual supports from recovery of the complete signature. For each role, $\mathrm{F1} = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, with the observation entries pooled across groups. For aligned signatures, we define:

$$\text{Joint} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\widehat{r}_{\sigma^*(i)} = r_i\}, \qquad \text{Hamming} = 1 - \frac{\sum_i \|\widehat{r}_{\sigma^*(i)} - r_i\|_1}{N(K+2)}. \tag{78}$$

Joint accuracy counts coordinates whose entire signature is correct, while Hamming accuracy counts correct signature entries. When reported, Common F1 is the arithmetic mean of the three role F1 values under the same alignment. AUPRC uses continuous scores and is computed as average precision, $\sum_\ell (R_\ell - R_{\ell-1})P_\ell$, grouping equal scores at each threshold. Means and standard errors use mechanisms as the independent units, with $\mathrm{SE} = s/\sqrt{m}$ for m mechanisms.

The sample size study shows that recovery improves most clearly in the small data regime. In Figure 6, OURS increases from 0.396 Joint accuracy at $n = 50$ to 0.708 at 100 and 0.844 at 150. For $n \geq 150$, Joint accuracy remains high and ranges from 0.802 to 0.875 across the tested sizes. The incoming role reaches high F1 early, while the outgoing role improves from 0.781 at $n = 50$ to 0.954 at 150. OURS exceeds TEMPORAL-ONLY in Joint accuracy at every displayed size. The curves therefore show a clear gain from additional data at small sample sizes, followed by a high but fluctuating recovery level.

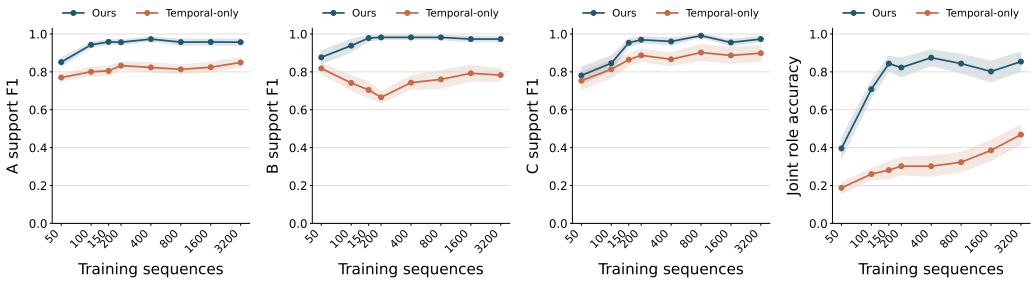


Figure 6: Sample-size scaling of structural recovery.

Here we provide complementary examples of the continuous structural scores. Each row is divided by its maximum, so color represents relative strength within a role or observation group and should not be interpreted as a probability or binary support. Figure 8 shows an additional mechanism, whose generating structure appears in Figure 7.

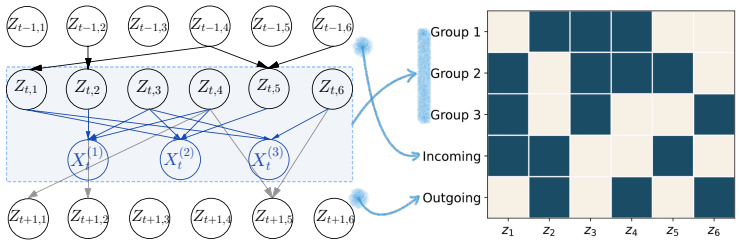
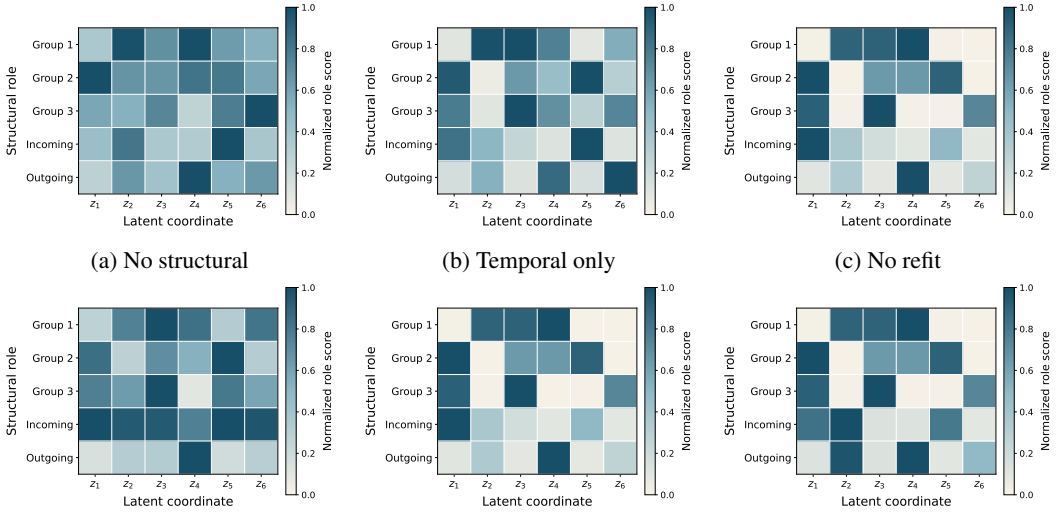


Figure 7: Data generation process and target identification of the simulation experiment.



(a) No structural (b) Temporal only (c) No refit

(d) InterLatent (e) ThoughtComm (f) RoleTrace

Figure 8: Heatmaps of recovered structural signatures.

Common Alignment (RQ2) RQ2 uses a controlled construction in which the three structural views leave different coordinate ambiguities. We use ten mechanisms with $N = 4$, $K = 3$, $T = 3$, and four observed dimensions per group, and evaluate the middle time point $t = 2$. The supports are:

$$(\mathbb{A}_{2,i}^{(k)})_{k,i} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbb{B}_2 = (1, 0, 1, 0), \qquad \mathbb{C}_2 = (0, 1, 1, 1). \tag{79}$$

Using indices starting at one, the first transition has edges $Z_{1,2} \to Z_{2,1}$ and $Z_{1,4} \to Z_{2,3}$, while the second has $Z_{2,2} \to Z_{3,2}$, $Z_{2,3} \to Z_{3,3}$, and $Z_{2,4} \to Z_{3,4}$. Both transitions use the nonlinear Gaussian construction above with separately sampled weights. Training uses 3200 sequences and the same validation, test, and optimization settings as RQ1. Separate temporal networks fit the two transitions. Refining the middle coordinate system transforms the derivative objects as $J_{\widehat{g}_2} Q_2$, $Q_2^\top \widehat{M}_1$, and $\widehat{M}_2 Q_2$. For $\mathbb{X} \in \{\mathbb{A}, \mathbb{B}, \mathbb{C}\}$, let $c_{\mathbb{X}}(\sigma)$ count mismatched entries in that role and define:

$$\Pi_{\mathbb{X}} = \arg\min_{\sigma \in \mathfrak{S}_4} c_{\mathbb{X}}(\sigma), \qquad \Pi_{\text{joint}} = \arg\min_{\sigma \in \mathfrak{S}_4} \sum_{\mathbb{X} \in \{\mathbb{A}, \mathbb{B}, \mathbb{C}\}} c_{\mathbb{X}}(\sigma). \tag{80}$$

We enumerate all 24 permutations and average each metric over the corresponding optimal set. The $\mathbb{A}$ guided, $\mathbb{B}$ guided, and $\mathbb{C}$ guided conditions use $\Pi_{\mathbb{A}}$, $\Pi_{\mathbb{B}}$, and $\Pi_{\mathbb{C}}$, while common alignment uses Π_{joint}. NO STRUCTURAL evaluates the initial scores with common matching. The results show how a single structural view can leave the complete coordinate correspondence unresolved. With learned scores, common alignment gives mean role F1 values $(1.000, 1.000, 0.960)$ and Joint accuracy 0.950 ± 0.033. The respective single-role alignments give Joint accuracies of 0.700 ± 0.033, 0.475 ± 0.017, and 0.458 ± 0.028, while NO STRUCTURAL gives 0.100.

Table 3: Coordinate refinement with fixed representations and pre-refit derivatives. Values are mean ± SE across ten mechanisms. Bold denotes the largest mean in each column, including ties.

Refinement	A F1	B F1	C F1	Joint Acc.	Hamming
None	0.694 ± 0.025	0.777 ± 0.070	0.770 ± 0.060	0.100 ± 0.055	0.725 ± 0.019
$\mathbb{A}$ only	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.967 \pm 0.033}$	0.910 ± 0.053	0.875 ± 0.056	0.975 ± 0.011
$\mathbb{B}$ only	0.792 ± 0.029	0.867 ± 0.069	0.920 ± 0.033	0.450 ± 0.082	0.850 ± 0.032
$\mathbb{C}$ only	0.822 ± 0.022	0.947 ± 0.037	$\mathbf{0.960 \pm 0.027}$	0.575 ± 0.065	0.890 ± 0.024
RoleTrace	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.967 \pm 0.033}$	0.930 ± 0.052	$\mathbf{0.900 \pm 0.055}$	$\mathbf{0.980 \pm 0.011}$

Common Alignment under Different Structural Views. To complement the permutation analysis from the perspective of the learned coordinate transformation, we compare which structural views are used to refine the common coordinate system Q_t. Starting from the same representation and derivatives, we optimize Q_t using $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$, $\{\mathbb{B}, \mathbb{C}\}$, or all three views, while keeping $Q_t = I$ as a reference. In particular, the full objective takes the form:

$$\Omega_t(Q_t) = \lambda_A \sum_{k,i} s^A_{t,k,i}(Q_t) + \lambda_B \sum_i s^B_{t,i}(Q_t) + \lambda_C \sum_i s^C_{t,i}(Q_t),$$

with the corresponding terms removed for the partial objectives. Transition refitting is disabled so that the comparison isolates the effect of coordinate refinement.

Table 3 shows that joint refinement achieves the highest Joint accuracy of 0.900, compared with 0.875, 0.450 and 0.575 using $\mathbb{A}$, $\mathbb{B}$ or $\mathbb{C}$ alone. These results indicate that combining observation and temporal structure can further resolve the coordinate correspondence, with the observation structure providing a particularly informative constraint in this controlled setting.

Sensitivity to Identifiability Conditions (RQ3). RQ3 isolates the structural recovery step by applying the same coordinate selection procedure directly to derivatives of the generating mechanisms. We use $N = 4$, $K = 3$, $T = 5$, four observed dimensions per group, and 512 latent sequences evaluated at $t = 3$. Each of the three degradation axes contains ten mechanisms, ten levels, and three orthogonal coordinate mixings, giving 900 evaluations in total. No neural model is trained. The transition mean scale and log standard deviation amplitude are 0.55 and 0.20, with intercept -0.65.

In Figure 4, the first path weakens the group span condition by making two active directions in one observation group increasingly dependent. For selected original columns u, v, we set:

$$w_p(\gamma) = u, \qquad w_q(\gamma) = \sqrt{1-\gamma^2}\, u + \gamma v, \tag{81}$$

while leaving the remaining columns unchanged. Let $J^{(s)}_{k_*,\gamma}$ denote the affected group's Jacobian restricted to its declared active columns. The diagnostic is

$$\rho_{\text{group}}(\gamma) = \min_{s \le S} \frac{\sigma_{\min}(J^{(s)}_{k_*,\gamma})}{\sigma_{\max}(J^{(s)}_{k_*,\gamma})}. \tag{82}$$

At $\gamma = 0$, those active columns lose rank while the complete observation map remains full rank, directly weakening the group variability condition in Theorem 3.1.

The second path weakens temporal variability by approaching a linear Gaussian transition. With $\ell = \log \sigma$ and $\ell_0 = -0.65$, define

$$\mu_\gamma(z) = (1-\gamma) W_0 z + \gamma \mu_1(z), \qquad \ell_\gamma(z) = \ell_0 + \gamma\big[\ell_1(z) - \ell_0\big]. \tag{83}$$

The delayed graph has edge set $E = \{(1,3),(2,3),(4,1),(4,2)\}$, where the first index denotes the receiver, and W_0 preserves this support. Writing vec_E for vectorization restricted to these edges:

$$(V^+_\gamma)_{s,:} = \text{vec}_E\, M_{t-1,\gamma}(z^{(s)}_t, z^{(1)}_{t-1}), \tag{84}$$

$$(V^-_\gamma)_{s,:} = \text{vec}_E\, M_{t,\gamma}(z^{(1)}_{t+1}, z^{(s)}_t). \tag{85}$$

For $\kappa(V) = \sigma_{\min}(V)/\sigma_{\max}(V)$, the displayed ratio is:

$$\rho_{\text{span}}(\gamma) = \frac{\min\{\kappa(V^+_\gamma), \kappa(V^-_\gamma)\}}{\min\{\kappa(V^+_1), \kappa(V^-_1)\}}. \tag{86}$$

At $\gamma = 0$, the cross Hessian is constant, so its span has dimension one within the four-dimensional edge space.

The third path weakens faithfulness through exact cancellation on one declared edge. The graph contains edges $(1, 2)$ and $(3, 4)$, with the first affected. For each mean or log scale coefficient w on this edge, two pathways have coefficients $(1 + c)w$ and $-(c + 1 - \gamma)w$, where $c = 0.35$, so their combined coefficient is γw. The second edge remains fixed. For the selected edge (i_*, j_*), define:

$$e_\gamma = \left[\frac{1}{S} \sum_{s=1}^{S} \left(M^{(s)}_{t-1,\gamma} \right)^2_{i_* j_*} \right]^{1/2}, \qquad \rho_{\text{edge}}(\gamma) = e_\gamma / e_1. \tag{87}$$

The target graph remains fixed at $\gamma = 0$, although the selected edge then has zero net influence on the transition distribution. Recovery at this endpoint is therefore evaluated against a declared role contribution that has become unobservable.

The three paths produce distinct degradation patterns. Mean Joint accuracy falls from one to 0.633, 0.750, and 0.500 at the group, temporal span, and cancellation endpoints, respectively. Group degradation primarily affects $\mathbb{A}$, while the temporal paths affect $\mathbb{B}$ and $\mathbb{C}$. The cancellation curve remains perfect at every tested positive ratio and drops only at zero. Plateaus are expected because the target is a discrete support: derivative magnitudes may change while the binary role remains the same. Losses before the other endpoints can also reflect the fixed structural objective and support threshold. These controlled experiments characterize the sensitivity of the specified recovery procedure along the tested paths; they do not establish that each condition is necessary for every estimator.

D.2 Additional Details for AMI Multi-party Meeting Experiment

Data Preprocessing and Split. We use the public manual annotations of the AMI Meeting Corpus (Carletta et al., 2005) and focus on scenario meetings, where four participants collaborate on a shared design project. The four functional roles, Project Manager (PM), Industrial Designer (ID), User Interface Designer (UI), and Marketing Expert (ME), define the observation groups used throughout the experiment. Speaker identifiers are mapped to these roles using the meeting metadata, so the group ordering is fixed as PM, ID, UI, and ME across meetings. Our evaluation uses 118 meetings from 30 teams. These meetings correspond to the 98 training and 20 development meetings in our original metadata split; the remaining 20 meetings from five held-out teams are not used in this experiment. The 118 meetings contain 28 teams with four project meetings and two teams with three, due to two meetings being absent from the scenario metadata. No meeting is filtered according to meeting effectiveness, speaking activity, or downstream prediction performance.

We construct five folds at the team level so that all meetings from the same team remain in the same partition. In each fold, 20 teams are used for model fitting, four for model selection, and six for evaluation. The fit partition determines all preprocessing statistics, representation parameters, and structural rankings. The selection partition is used for checkpoint selection and downstream regularization, while the evaluation partition is used only after these choices are fixed. Each team appears in the evaluation partition exactly once across the five folds. This separation prevents windows from different meetings of the same project team from appearing in both model fitting and evaluation.

We divide each meeting into non-overlapping 15-second windows, producing 14,513 windows across the 118 meetings. The 15-second interval provides a regular temporal resolution while retaining enough local conversational context for each participant. We use it as a fixed observation interval throughout the experiment. For a window ending at time T, the input contains valid words whose ending times fall in $(T - 15, T]$, ensuring that speech extending beyond the current boundary is not included prematurely.

Each participant contributes one grouped observation. For participant g, the raw observation is:

$$X^{(g)}_t = \left[u^{(g)}_t, \log(1 + n^{(g)}_t), r^{(g)}_t, q^{(g)}_t \right] \in \mathbb{R}^{35}, \tag{88}$$

where $u^{(g)}_t \in \mathbb{R}^{32}$ represents the linguistic content in the current window, $n^{(g)}_t$ is the number of completed words, $r^{(g)}_t$ records the normalized time since the participant's most recent completed word, and $q^{(g)}_t$ indicates whether the participant speaks in the current window. A participant with no

Table 4: Fixed-input reference results on AMI. Dimensions refer to the inputs supplied to the respective predictors. These rows are separate from the matched latent-budget comparisons.

Method / Representation	Input dimension	Accuracy (%)
PCA observations	32	50.969 ± 0.000
Full observations	140	50.657 ± 0.166
SVM	188	51.240 ± 0.143
XGBoost	188	50.783 ± 0.117
TS2Vec	32	50.623 ± 0.096
Role $\mathbb{C}$	32	$\mathbf{51.500 \pm 0.135}$

valid word receives a zero text representation and zero activity indicator, while the recency feature retains information about earlier speech. We keep silent windows in the sequence. Among all windows, 542 contain no active participant, while 3,033, 3,846, 4,281, and 2,811 windows contain one, two, three, and four active participants, respectively. The complete observation is therefore: $X_t = (X_t^{(\mathrm{PM})}, X_t^{(\mathrm{ID})}, X_t^{(\mathrm{UI})}, X_t^{(\mathrm{ME})}) \in \mathbb{R}^{4\times 35}$.

Feature construction, training and evaluation. For AMI, we use a practical adaptation of RoleTrace. We construct the text component $u_t^{(g)}$ from the manual transcript. Participant text within each window is first embedded into 384 dimensions and then projected to 32 dimensions by PCA. The PCA projection is fitted separately within each fold using only the fit partition and is shared across the four participant groups. The three activity features in Eq. 88 are appended to the projected text representation, and each feature is standardized using statistics computed from the fit partition. Thus, the four groups provide a 140-dimensional observation at each time point. The representation model maps the grouped observation to a 32-dimensional latent state Z_t. The encoder uses hidden dimensions 128 and 64, followed by a 32-dimensional output. We additionally provide one observed indicator for each participant, allowing the model to distinguish training-time input masking from naturally silent participants. Temporal context is summarized by a 32-dimensional GRU over the preceding three latent states. Training combines reconstruction of the current grouped observation with prediction of future latent states and linguistic content. No next-speaker labels are used in representation training. We train with Adam for at most 60 epochs with learning rate 10^{-3} and batch size 256.

We define the downstream task at the same 15-second boundaries used to construct the observations. At a boundary T, the target is the participant associated with the first valid dialogue act whose starting time is at or after T. We retain a sample when both this target and a previously completed dialogue act are available. The target may therefore coincide with the previous speaker, and its starting time is not restricted to the immediately following 15-second window. This construction yields 14,028 prediction samples from the 14,513 observation windows. The four target roles, PM, ID, UI, and ME account for 4,755, 3,118, 2,947, and 3,208 samples, respectively. We use a four-class logistic regression classifier on the selected latent coordinates. Its regularization is chosen on the selection partition, and final performance is evaluated on the held-out teams.

Additional comparisons. Table 4 collects the fixed-input representation controls and other baselines from the experimental records. Their input dimensions differ from the selected latent budgets, which means they do not isolate the effect of coordinate selection at a matched budget. PCA Observations applies PCA directly to the original observations and retains 32 dimensions, while Full Observations uses the complete 140-dimensional input without latent compression. We also include SVM and XGBoost (Chen & Guestrin, 2016) as conventional supervised predictors, together with TS2Vec (Yue et al., 2022), which learns a general-purpose temporal representation through contrastive learning. These methods use different fixed input dimensions and therefore serve as supplementary references to the matched-budget comparisons in Table 2. Over five seeds, the full 32-dimensional **Role** $\mathbb{C}$ representation achieves $\mathbf{51.500 \pm 0.135}\%$ accuracy, exceeding the strongest of these references.

AMI related work. It is noteworthy that AMI has supported a broad range of multi-party interaction tasks. Earlier work used linguistic and temporal context features for decision detection and so-

cial role recognition (Frampton et al., 2009; Sapru & Valente, 2012; Sapru et al., 2014). Subsequent methods for meeting summarization increasingly incorporated richer interaction structure, including multimodal hierarchical attention (Li et al., 2019), hierarchical speaker-aware representations (Zhu et al., 2020) and discourse graphs (Feng et al., 2021). More recent studies further emphasize selective and structured use of meeting information, including extraction of essential content (Tan et al., 2023), content selection (Deng et al., 2023) and the modeling of speaker dynamics (Gruttadauria et al., 2024; Hilgert & Niehues, 2025; Li & Chu, 2026). However, the basic structural units remain speakers, utterances, discourse relations, or content spans. Our formulation targets a different structural object. We treat participant roles as observation groups and identify the group-specific expression and temporal propagation roles of the latent information factors underlying the interaction.