BIXFORMER: A BIDIRECTIONAL CROSS ATTENTION TRANSFORMER FOR DISENTANGLING INTER-REGIONAL NEURAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in high-throughput neural recording technologies enable simultaneous measurement of activity across multiple brain regions in behaving animals, producing datasets of unprecedented scale and richness. Interpreting these data remains challenging due to the bidirectional and temporally offset nature of inter-regional communication, where feedforward and feedback signals are superimposed within neural populations. We introduce BiXformer, a bidirectional cross-attention transformer that disentangles these interactions by decomposing inter-regional communication into temporally leading and lagging streams using directionally masked attention. By enforcing temporal constraints within attention heads, BiXformer recovers low-dimensional, directed latent dynamics and estimates communication delays without relying on linearity or stationarity assumptions. Validated on synthetic data with known ground-truth delays, BiXformer accurately recovers both the latent structure and inter-regional timing, outperforming existing methods in both accuracy and speed. Applied to simultaneous neural-behavioral recordings and multi-region neural recordings during a movement task, BiXformer reveals interpretable, temporally structured components consistent with the coexistence of sensory feedback and motor-related signals. These results establish BiXformer as a flexible framework for uncovering dynamic, directed communication in complex neural circuits.

1 INTRODUCTION

Recent advances in recording techniques now enable simultaneous measurement of neural activity across multiple brain regions in a behaving animal, producing datasets of unprecedented scale and richness (Jun et al., 2017; Steinmetz et al., 2018; Bondy et al., 2024; Chen et al., 2024; Khilkevich et al., 2024). The same scale that renders these datasets powerful also makes them remarkably difficult to interpret. A key part of this difficulty lies in the brain's architecture: regions underlying different processes, such as cognition (Miller et al., 2018; Hasnain et al., 2025), sensation (Lamme et al., 1998; Angelucci & Bressloff, 2006; Gilbert & Li, 2013; Harris & Mrsic-Flogel, 2013), and movement (Shadmehr & Krakauer, 2008; Chen et al., 2024), are densely interconnected, with information relayed bi-directionally through many pathways and across multiple timescales. A central challenge in systems neuroscience is characterizing the information flow between brain regions – specifically, the content, timing, and directionality of inter-regional communication.

This challenge is acute in sensorimotor circuits (Mayrhofer et al., 2019; Xu et al., 2022) where neural responses often integrate signals that both drive and reflect movement. For instance, in a rhythmic licking task, neurons in the primary motor cortex can exhibit both feedforward motor commands that drive movement and feedback signals reflecting proprioceptive input arising from movement (Keemink & Machens, 2019; Dragoi et al., 2026). Crucially, these signals do not arrive simultaneously; communication between populations is subject to conduction delays that reflect the underlying anatomical organization and physiological properties of neurons (Schmolesky et al., 1998). The mixing of these temporally offset signals within a region's neural population creates a fundamental confound: without explicitly disentangling feedforward from feedback interactions, the direction of information flow – between neural populations or between neural activity and behavior – remains ambiguous.

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.

Addressing this central challenge has motivated a growing body of computational approaches aimed at modeling inter-regional interactions (Semedo et al., 2019; Kang & Druckmann, 2020; Keeley et al., 2020; Glaser et al., 2020; Perich & Rajan, 2020; Gokcen et al., 2022; Karniol-Tambour et al., 2023; Gokcen et al., 2023; Liu et al., 2025; Xin & Kass, 2025). While these approaches have made significant progress, particularly in identifying smooth, low-dimensional latent dynamics across neural populations, they either fail to model the temporal delays inherent to inter-regional signaling, or rely on linearity and stationarity assumptions that constrain them to a fixed mode of communication. Neural circuits are inherently nonlinear (Vyas et al., 2020) and non-stationary, re-configuring their dynamics across timescales and behavioral contexts (Mante et al., 2013; Driscoll et al., 2017). An ideal model of inter-regional communication should be capable of capturing nonlinear interactions between populations, time-dependent changes in communication, and temporal delays inherent to inter-regional signaling.

Transformers offer a principled framework for satisfying these criteria. While their ability to capture time-dependent interactions in high dimensional sequential data makes them particularly well-suited for modeling neural dynamics, they have largely been used as black-box predictive models rather than tools for mechanistic interpretation.

To address this need, we introduce the *Bidirectional Cross Attention Transformer* (BiXformer), a novel architecture designed to disentangle dynamic and directed information flow across neural populations. BiXformer decomposes inter-regional communication into two parallel streams — a leading stream and a lagging stream — by leveraging masked cross-attention heads to explicitly disentangle feedforward and feedback interactions. In this work, we demonstrate the model's ability to recover these temporally-asymmetric signals in synthetic data and show that it yields interpretable, directed latent dynamics when applied to real simultaneous neural-behavioral and multi-region recordings from mice performing a controlled motor task.

Our contributions can be summarized as follows:

- **We introduce BiXformer**, a transformer architecture that decomposes inter-regional communication into leading and lagging streams via masked cross-attention.
- **We validate BiXformer on synthetic data**, demonstrating that BiXformer accurately recovers leading and lagging latents and their associated inter-regional delays, outperfoming existing methods in both accuracy and speed.
- **We apply BiXformer to simultaneous neural-behavioral recordings** from the medulla (Md) during a rhythmic licking task, revealing interpretable latent components consistent with coexisting sensory and motor signals.
- **We use BiXformer to uncover interactions between primary motor cortex (M1) and primary somatosensory cortex (S1)** during a delayed reward licking task, confirming existing hypotheses about sensorimotor communication while shedding new light on the nature of S1–M1 communication.

2 Related Work

2.1 Latent Variable Models of Inter-Regional Communication

A common approach to understanding inter-regional communication has been to identify low-dimensional latent structure that mediates the relationship between neural populations. Methods such as reduced rank regression (RRR) (Semedo et al., 2019) and canonical correlation analysis (CCA) (Hotelling, 1936) identify shared and private dimensions of neural activity across regions, but leave the directionality of inter-regional communication unresolved. DLAG (Gokcen et al., 2022) and its multi-region extension mDLAG (Gokcen et al., 2023) extended this framework by modeling time-lagged latent trajectories with Gaussian process priors, enabling inference of bidirectional signal flow between neural populations. However, their linear-Gaussian observation model precludes nonlinear interactions between populations, and their reliance on stationary temporal kernels further constrains them to a single, time-invariant communication mode, preventing them from capturing within-trial reconfigurations in coupling strength or direction. Dynamical system approaches, such as LFADS/MR-LFADS (Pandarinath et al., 2018; Liu et al., 2025) and rsLDS (Linderman et al., 2016; Glaser et al., 2020), offer greater expressivity by capturing nonlinear interactions between neural

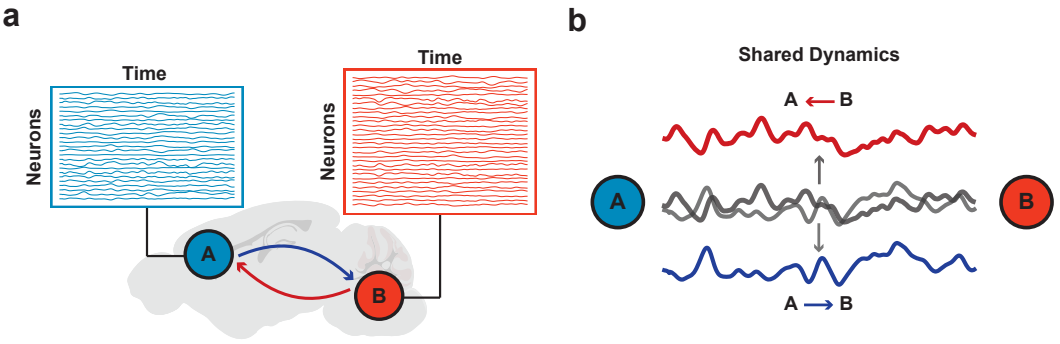

Figure 1: **Bidirectional, concurrent interactions between neural populations. (a)** Simultaneous recordings from two brain regions, A and B, reveal rich population activity that reflects a mixture of inter-regional signals relayed concurrently in both directions. **(b)** The shared dynamics between A and B are composed of two temporally asymmetric components: a feedforward signal ($A \to B$, blue) in which A leads B, and a feedback signal ($B \to A$, red) in which B leads A. These signals are superimposed in the recorded activity (gray), making their disentanglement a central challenge for inter-regional communication analysis.

populations. However, inter-regional delays must be absorbed into the latent dynamics, making directionality and lag structure difficult to disentangle from the underlying system state.

2.2 Transformers in Neuroscience

Transformers have been increasingly applied to neural population recordings, demonstrating advantages in generalization, scalability, and flexible mapping between neural and behavioral data (Ye et al., 2023; Azabou et al., 2023; 2025; Zhang et al., 2025). Their application to mechanistic interpretation, however, has remained limited until recently, with transformers now being leveraged toward understanding inter-regional communication. CTAE (Sristi et al., 2025) introduced a transformer autoencoder that disentangles shared and private latent structure across regions and captures nonlinear and non-stationary dynamics, though it offers no insight into the directionality or temporal lags between populations. The GLM-Transformer (Xin & Kass, 2025) captures within-area dynamics through a transformer-based variational autoencoder, but relies on a GLM coupling term for cross-area interactions, constraining communication to be stationary. Despite these advances, the directed, dynamic nature of inter-regional communication remains unresolved, demanding new approaches.

3 Methods

3.1 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{N \times T \times p}$ and $\mathbf{Y} \in \mathbb{R}^{N \times T \times q}$ denote simultaneously recorded spiking activity from two distinct neural populations (Fig. 1a) — population A with p units and population B with q units — measured across N trials of length T. (Note that for one application of our approach, one region is instead considered as behavioral measurements, but the above formalism still applies.) We assume that the activity of population B at any time t arises from private dynamics intrinsic to B and shared dynamics in the form of temporally asymmetric, bidirectional interactions with population A (Fig. 1b):

1. **Leading (feedforward) stream**: activity in population A at time t' leads activity in population B at time $t = t' + \tau_{\text{lead}}$, so that A leads B by delay $\tau_{\text{lead}} > 0$.
2. **Lagging (feedback) stream**: activity in population B at time t' leads activity in population A at time $t = t' + \tau_{\text{lag}}$, so that A lags B by delay $\tau_{\text{lag}} > 0$.

The delays τ_{lead} and τ_{lag} are unknown, potentially distinct across communication channels, and reflect the underlying anatomical and physiological organization of the inter-regional pathways. Our goal

is to jointly estimate these delays and recover the low-dimensional latent dynamics mediating each direction of communication from the simultaneously observed population recordings.

3.2 Model Architecture

BiXformer is a bidirectional cross-attention transformer that decomposes inter-regional communication into two parallel streams — a **leading stream** and a **lagging stream** — via directionally masked cross-attention with rotary positional encoding. The full architecture is illustrated in Figure 2.

Both populations are linearly embedded into d-dimensional spaces:

$$\mathbf{X}^e = \mathbf{X}\mathbf{W}^x + \mathbf{b}^x \in \mathbb{R}^{N\times T\times d}, \qquad \mathbf{Y}^e = \mathbf{Y}\mathbf{W}^y + \mathbf{b}^y \in \mathbb{R}^{N\times T\times d} \tag{1}$$

During training, the population activity of region $\mathbf{B}$ is set to zero ($\mathbf{Y} \leftarrow \mathbf{0}$), so the embedding reduces to the learned bias alone: $\mathbf{Y}^e = \mathbf{b}^y$. This bias is then projected into the query matrix $\mathbf{Q} = \mathbf{Y}^e\mathbf{W}^Q$, so that the reconstruction of $\mathbf{Y}$ arises entirely through cross-attention to $\mathbf{X}$. This prevents degenerate solutions in which the model bypasses the cross-attention mechanism by attending to its own input. Rotary positional embeddings (RoPE; (Su et al., 2023)) are applied to query and key vectors so that attention scores depend only on the relative temporal offset $t - t'$, enabling each head to learn a preference for a specific inter-regional lag.

3.2.1 Masked Cross-Attention Streams

BiXformer uses two parallel cross-attention modules, each with $H/2$ heads (H total). In both streams, queries are formed from population B's embedding, while keys and values are formed from population A's embedding:

$$\mathbf{Q} = \mathbf{Y}^e\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}^e\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}^e\mathbf{W}^V \tag{2}$$

The masked attention weight from population B at time t to population A at time i is:

$$A_{ti} = \mathrm{softmax}_i\left(\frac{\langle \mathbf{q}'_t, \mathbf{k}'_i\rangle}{\sqrt{d_h}} + m_{ti}\right), \quad m_{ti} = \begin{cases} -\infty & \text{if masked} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This directional mask distinguishes the two streams: a causal mask (positions $i > t$ masked) defines the **leading stream**, restricting B to past or concurrent activity in A ($i \le t$) so that A leads B, while an acausal mask (positions $i < t$ masked) defines the **lagging stream**, restricting B to future or concurrent activity in A ($i \ge t$) so that B leads A.

The inter-regional communication delay is read out from the **diagonal profile** of each head's $T \times T$ attention matrix — the mean attention weight as a function of lag $\delta = i - t$:

$$\phi(\delta) = \frac{1}{T - |\delta|}\sum_t A_{t,\,t+\delta} \tag{4}$$

The profile concentrates at $\delta^* = \arg\max_\delta \phi(\delta)$, giving a direct estimate of the lag preferred by that channel. After multi-head aggregation and a linear projection, each stream is compressed to an r-dimensional latent representation:

$$\mathbf{z}^{(\text{lead})} = \mathrm{CrossAttn}_{\text{lead}}(\mathbf{Y}^e, \mathbf{X}^e)\,\mathbf{P}^{(\text{lead})} \in \mathbb{R}^{N\times T\times r} \tag{5}$$

$$\mathbf{z}^{(\text{lag})} = \mathrm{CrossAttn}_{\text{lag}}(\mathbf{Y}^e, \mathbf{X}^e)\,\mathbf{P}^{(\text{lag})} \in \mathbb{R}^{N\times T\times r} \tag{6}$$

3.2.2 Stream-Orthogonal Stiefel Decoder

Within a neural population, feedforward and feedback signals should not share the same activity subspace, as overlapping representations would cause interference between the two communication channels. To enforce this, we impose a **hard orthogonality constraint** on the decoder weight matrices via QR reparameterization (Edelman et al., 1998). Given learnable raw parameters $\tilde{\mathbf{W}}_{\text{lead}}, \tilde{\mathbf{W}}_{\text{lag}} \in \mathbb{R}^{q\times r}$, we compute the thin QR decomposition of their column-concatenation at every forward pass:

$$\big[\mathbf{W}_{\text{lead}} \mid \mathbf{W}_{\text{lag}}\big] = \mathrm{QR}\big(\big[\tilde{\mathbf{W}}_{\text{lead}} \mid \tilde{\mathbf{W}}_{\text{lag}}\big]\big) \tag{7}$$

The resulting matrices $\mathbf{W}_{\text{lead}}, \mathbf{W}_{\text{lag}} \in \mathbb{R}^{q\times r}$ satisfy $\mathbf{W}_{\text{lead}}^\top\mathbf{W}_{\text{lead}} = \mathbf{I}_r$, $\mathbf{W}_{\text{lag}}^\top\mathbf{W}_{\text{lag}} = \mathbf{I}_r$, and $\mathbf{W}_{\text{lead}}^\top\mathbf{W}_{\text{lag}} = \mathbf{0}$, constraining the two decoders to lie on the Stiefel manifold and to span orthogonal subspaces of $\mathbb{R}^q$. Population B's activity is reconstructed as the sum of the two stream contributions:

$$\hat{\mathbf{Y}} = \mathbf{z}^{(\text{lead})}\mathbf{W}_{\text{lead}}^\top + \mathbf{z}^{(\text{lag})}\mathbf{W}_{\text{lag}}^\top \tag{8}$$

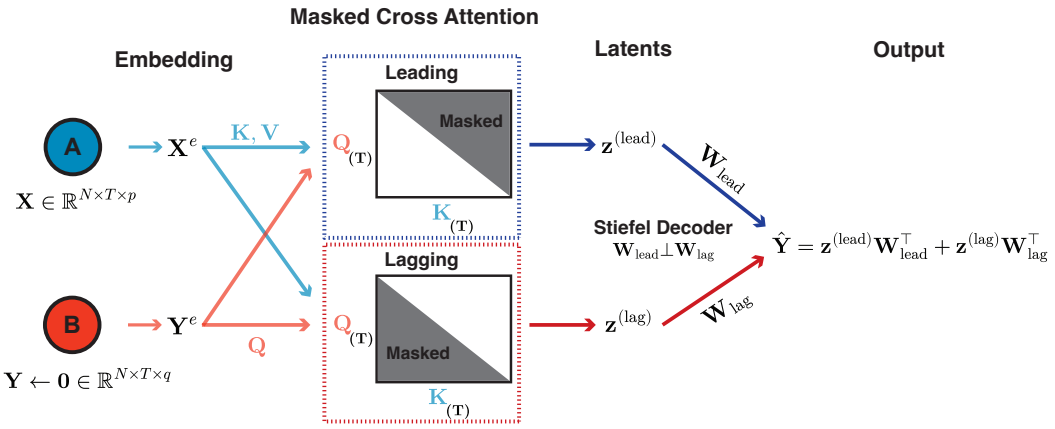

Figure 2: **BiXformer architecture.** Population A's activity $\mathbf{X}$ is embedded into $\mathbf{X}^e$, providing keys and values to both streams. Population B's input is zeroed during training so its embedding $\mathbf{Y}^e$ reduces to a learned bias; this bias is then linearly projected into the query matrix. Two parallel masked cross-attention modules enforce directional constraints: the leading stream attends only to past activity in A ($i \leq t$), and the lagging stream only to future activity ($i \geq t$). Each stream produces a latent trajectory decoded into orthogonal subspaces of B's activity space via a Stiefel-constrained decoder ($\mathbf{W}_{\text{lead}} \perp \mathbf{W}_{\text{lag}}$), and their contributions are summed to reconstruct $\hat{\mathbf{Y}}$.

3.3 Training Objective

BiXformer is trained to minimize the mean-squared reconstruction error of target population B:

$$\mathcal{L} = \frac{1}{NTq} \left\| \hat{\mathbf{Y}} - \mathbf{Y} \right\|_F^2 \tag{9}$$

The Stiefel orthogonality constraint on the decoder ($\mathbf{W}_{\text{lead}} \perp \mathbf{W}_{\text{lag}}$) is enforced structurally, ensuring that the leading and lagging streams write into orthogonal subspaces of the target population's activity. This means that the contribution of each stream to any given target variable can be read off directly and unambiguously, without mixing. Directional separation between streams is enforced by the causal and acausal masks, so no auxiliary losses are required. Parameters are optimized with Adam, with an 80–20 train-test split; training epochs vary over experiment; model performance is reported on the held-out fold of each split (see Appendix A1).

4 Experiments

4.1 Validating BiXformer on Synthetic Data

To verify that BiXformer correctly recovers inter-regional delays and latent communication structure, we generated synthetic datasets with known ground truth using Gaussian Process (GP)-driven latent dynamics (Fig. 3a; see Appendix B1). The dataset was designed to reflect the scale and structure of real neural recordings: two populations of $p = 80$ and $q = 50$ neurons recorded over 200 trials of $T = 200$ time bins (5 ms resolution), with both shared cross-area latent dimensions and population-specific private latents to model within-area dynamics not involved in inter-regional communication. Four cross-area latent dimensions were included — two leading ($A \to B$) and two lagging ($B \to A$) — each with GP timescales and communication delays independently sampled from a uniform distribution spanning a range of biologically plausible values. Population activity was constructed by multiplying the latent trajectories by random mixing matrices and adding i.i.d. Gaussian observation noise. To ensure clean separation between shared and private variance in the ground truth, the within-area loading vectors were projected onto the null space of the cross-area loading matrix (via SVD), guaranteeing that private latents explain variance exclusively in directions orthogonal to those used by cross-area communication.

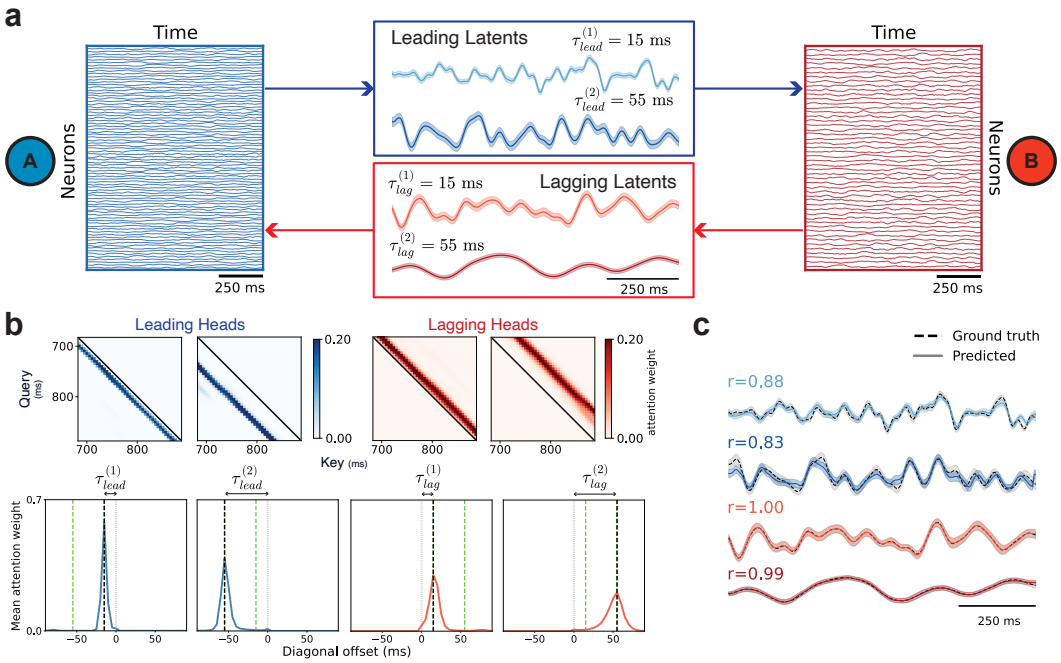

Figure 3: **BiXformer recovers directed latent dynamics and communication delays from synthetic data. (a)** Population activity from regions A (blue) and B (red) is decomposed into leading ($A \to B$) and lagging ($B \to A$) latents, each associated with a known ground-truth delay. **(b–c)** Results are shown for a representative held-out fold from 5-fold cross-validation. **(b)** Representative attention heads with the strongest diagonal profiles are shown (zoomed in for clarity); diagonal profiles peak at offsets closely matching the true delays (dashed lines). See Appendix D1 for all heads. **(c)** Predicted latent trajectories (solid) closely track ground-truth latents (dashed) across all four dimensions (Pearson r reported for single-trial predictions).

Following training, we evaluated BiXformer on held-out data. A subset of attention heads developed peaked diagonal profiles at offsets closely matching the true communication delays (Fig. 3b), while the remaining heads exhibited diffuse, non-selective attention patterns — consistent with the model allocating representational capacity selectively across channels (see Appendix D1). Latent trajectories recovered by each stream were well-correlated with the corresponding ground-truth GP signals on single trials (Fig. 3c).

On this same dataset, we benchmarked BiXformer against DLAG (Gokcen et al., 2022), an established model for inferring inter-regional communication with explicit delay modeling. BiXformer achieved substantially higher held-out R^2 (0.887 ± 0.018 vs. 0.590 ± 0.074), recovered ground-truth delays with a lower mean absolute error (1.0 ± 2.0 ms vs. 13.5 ± 8.9 ms), and trained over $200\times$ faster per fold (0.44 ± 0.01 min vs. 93.9 ± 3.9 min) (Appendix C). Together, these results establish that BiXformer reliably identifies the direction, magnitude, and latent structure of bidirectional inter-regional communication in synthetic data.

4.2 Disentangling Feedforward and Feedback Signals in Neural-Behavioral Recordings

As described in the Introduction, neural activity in sensorimotor regions simultaneously reflects feedforward motor commands and ascending sensory feedback, which are temporally offset yet superimposed within the same population — creating a confound that motivates an explicitly bidirectional decomposition.

We applied BiXformer to simultaneous electrophysiological recordings (118 units; spike counts binned at 2.5 ms) from the medulla, specifically the Intermediate and Parvocellular Reticular Nuclei (IRN/PARN), and high-speed videography (56 orofacial kinematic keypoint trajectories) in a head-

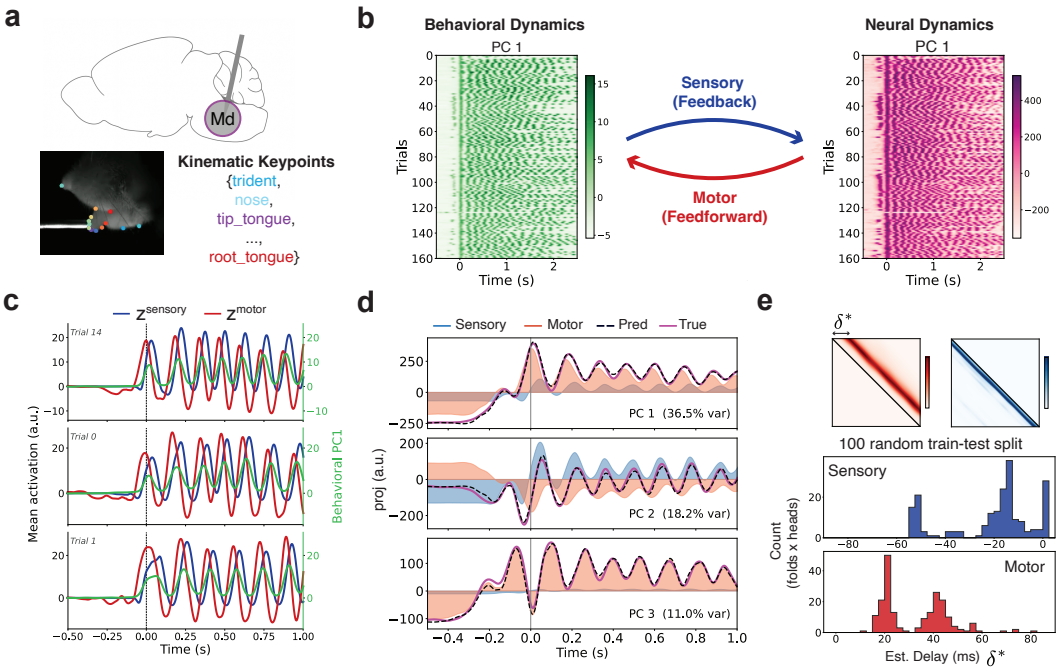

Figure 4: **BiXformer disentangles sensory and motor signals in neural-behavioral recordings from the medulla.** **(a)** Electrophysiological recordings from the medulla (Md) (sagittal view) during a rhythmic licking task, with orofacial kinematics tracked via high-speed videography. **(b)** The first principal component of behavioral keypoint features (left) and Md population activity (right) for multiple behavioral trials; time is aligned to first lickport contact. Bidirectional arrows indicate the sensory (feedback, blue) and motor (feedforward, red) pathways modeled by BiXformer. **(c)** Mean activation of sensory (z^{sensory}) and motor (z^{motor}) latents, overlaid with behavioral PC1 (40.4% variance explained) for multiple trials. **(d)** Contribution of each stream to reconstructing the top neural PCs (mean). **(e)** Estimated communication delays across 100 random train-test splits (2 heads per stream per run).

fixed mouse performing a delayed licking task (Fig. 4a). Although our approach is designed primarily for inter-regional communication (as we do below), application to a neural-behavioral dataset enables greater interpretability owing to the direct interpretation that behavior affords. The medulla exhibits highly stereotyped, rhythmic activity tightly coupled to orofacial movement, reflecting its prominent role in motor execution — making it a compelling model system for evaluating bidirectional decomposition (Fig. 4b). To aid interpretability and visualization, PCA was applied independently to both behavioral and neural datasets; all principal components were retained, preserving 100% of the variance. In this application, behavioral kinematics serve as the source population (X) and medulla neural activity as the target population (Y), so that BiXformer predicts neural dynamics from behavior.

Evaluated on held-out data, BiXformer successfully disentangles feedforward and feedback signals within medulla dynamics. The sensory (z^{sensory}) and motor (z^{motor}) latents exhibit distinct, interpretable temporal profiles: the motor stream precedes the dominant behavioral PC (40.4% of variance explained) consistent with a feedforward motor command, while the sensory stream follows it, consistent with ascending sensory feedback (Fig. 4c). Despite this explicit disentanglement, the model retains strong predictive power, achieving $R^2 = 0.638 \pm 0.032$ across 5-fold cross-validation (see Appendix D2).

The Stiefel orthogonality constraint on the decoder (Fig. 2) ensures that the two streams write into orthogonal subspaces of the medulla's neural population activity, allowing us to directly attribute each neural PC's variance to either the sensory or motor stream (Fig. 4d). We find that the motor stream strongly dominates the first PC of medulla activity, consistent with the IRN/PARN being a major motor hub (Takatoh et al., 2021; Kaku et al., 2025). Finally, to characterize the temporal structure of

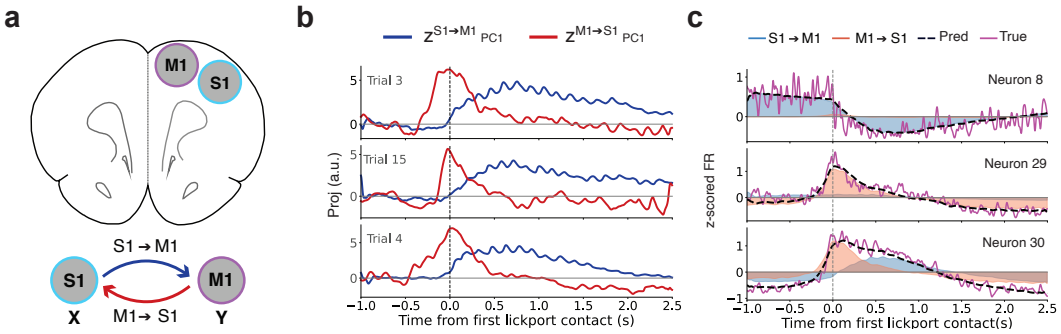

Figure 5: **BiXformer disentangles bi-directional signals in simultaneous S1–M1 recordings during a delayed reward task. (a)** BiXformer is applied to simultaneous recordings in S1 (X) and M1 (Y), where the leading stream models S1→M1 communication and the lagging stream models M1→S1 communication. **(b)** PC1 of the S1→M1 ($z^{\mathrm{S1}\to\mathrm{M1}}$) and M1→S1 ($z^{\mathrm{M1}\to\mathrm{S1}}$) latent spaces, capturing 51.0% and 89.9% of their respective variances, shown for three example trials. **(c)** Contribution of each stream to reconstructing the mean activity of individual neurons; three example neurons shown.

communication within each stream, we estimated neural-behavioral lags from the attention heads across 100 random train-test splits (2 heads per stream per run; Appendix D3). The resulting delay distributions reveal distinct characteristic lags within each stream (Fig. 4e), suggesting that multiple signals are transmitted at different timescales within both the sensory and motor pathways (GMM fit: sensory $\mu = \{16.0, 50.5\}$ ms, $\sigma = \{5.2, 6.2\}$ ms; motor $\mu = \{19.4, 40.4\}$ ms, $\sigma = \{2.6, 3.6\}$ ms; see Appendix D4). Across sessions from two different animals, the estimated delay distributions remain distinct but shifted in their characteristic lags (see Appendix D5), reflecting differences in recording sites and the specific neural populations sampled within the IRN/PARN. Together, these results demonstrate that BiXformer can recover interpretable feedforward and feedback components from mixed neural-behavioral recordings, revealing the temporal structure of communication within each pathway, while preserving strong predictive performance.

4.3 Uncovering Directed Interactions Between S1 and M1 During a Delayed Reward Task

Next, we used BiXformer to disentangle sensory- and motor-related signals from multi-site neural recordings during movement. Recent experimental measurements from primary motor and sensory areas in mice executing controlled movements have demonstrated a rich interplay of sensory- and motor-related signals across the cortex (Dragoi et al., 2026). Activity in primary motor cortex (M1) was shown to contain temporally multiplexed sensory- and motor-related signals, each of which was differentially represented within specific temporal epochs during a licking task. Motor signals were transiently present during periods when the consequences of movements were uncertain, while "residual" low-amplitude sensory signals were present during all periods of the task. Primary somatosensory cortex (S1) was active throughout the movement period, suggesting that residual M1 activity may reflect ongoing sensory input from S1 rather than intrinsic motor dynamics.

To test this hypothesis, we applied BiXformer to simultaneous S1 and M1 recordings, treating S1 as the source population (X) from which to predict M1 activity (Y), thereby decomposing M1 activity into components attributable to S1 input (S1 → M1) and components that precede S1 activity and are thus not attributable to it (M1 → S1) (Fig. 5a).

The latent dynamics recovered by each stream show distinct signatures of S1–M1 communication. The dominant mode of the S1 → M1 stream (PC1, capturing 51.0% of the stream's variance) exhibits a sustained signal that rises around the first lickport contact and persists throughout the lick bout, consistent with an ongoing sensory signal transmitted from S1 to M1 (Fig. 5b). The dominant mode of the M1 → S1 stream (PC1, capturing 89.9% of the stream's variance) resembles the identified engagement-disengagement profile that peaks sharply at first lickport contact and rapidly dissipates (Dragoi et al., 2026). These results demonstrate the ability of BiXformer to partition

neural responses into motor- and sensory-related signals originating within interconnected neural populations. Additional model fits, including attention maps and diagonal profiles for both streams, are provided in Appendix D6.

Assessing each stream's contribution to reconstructing individual M1 neuron activity revealed heterogeneity across the population, reflecting the diversity of functional signals present in M1 during movement (Fig. 5c). Some neurons were well-described by the M1 $\rightarrow$ S1 stream alone. For example, Neuron 29 exclusively exhibited a strong motor drive that disengaged rapidly, and reflected virtually no sensory-related signals from S1. Other neurons show mixed contributions from both streams (e.g., Neuron 30), consistent with the highly mixed responses observed across sensorimotor circuits and which are necessary for sensory-guided motor control. Somewhat surprisingly, a subset of neurons were predominantly and almost exclusively reconstructed by the S1 $\rightarrow$ M1 stream (e.g., Neuron 8), indicating they were largely and almost exclusively driven by S1 input. These neurons did not reflect strong motor-related signals around movement initiation, suggesting a different functional contribution to motor control than neurons primarily reflecting M1 $\rightarrow$ S1 signals. In addition, these neurons were active prior to movement onset, suggesting the S1 $\rightarrow$ M1 pathway may carry more than purely sensory-driven signals, such as anticipatory or top-down influences during motor planning. Collectively, these findings are consistent with the hypothesis that S1 input contributes to residual M1 activity following cortical disengagement, while also revealing intriguing top-down sensory signals that provide deeper insight into the nature of S1–M1 communication.

5 Discussion

We introduced BiXformer, a transformer that disentangles directed, time-delayed communication from simultaneous neural population recordings. Masked cross-attention enforces directionality and a Stiefel-constrained decoder enforces subspace separation, yielding latent dynamics that are both expressive and structurally interpretable, without linearity or stationarity assumptions. On synthetic data, BiXformer recovers latent trajectories and inter-regional delays, outperforming existing methods in both predictive performance and delay estimation accuracy while training orders of magnitude faster. Applied to medulla recordings during rhythmic licking, it cleanly separates feedforward motor commands from ascending sensory feedback. In simultaneous S1–M1 recordings during a delayed reward task, it supports the experimentally motivated hypothesis that residual M1 activity reflects ongoing S1 input, while revealing a heterogeneous M1 population differentially driven by S1 input versus intrinsic motor dynamics, including a subset whose pre-movement activity is explained by the S1$\rightarrow$M1 stream.

Existing approaches to inter-regional communication face a recurring tradeoff between expressivity and interpretability. Linear-Gaussian methods (Gokcen et al., 2022; 2023) model delays explicitly through time-shifted latent Gaussian processes, but their linear observation model precludes nonlinear interactions between populations. Dynamical-system approaches (Liu et al., 2025; Glaser et al., 2020) capture nonlinear dynamics but represent delays only implicitly through the latent state, obscuring the direction and timing of communication. Recent transformer-based methods either omit directionality (Sristi et al., 2025) or constrain communication to be stationary (Xin & Kass, 2025). BiXformer resolves this tradeoff: masked cross-attention enforces directionality architecturally, per-head attention exposes lag structure interpretably, and the underlying mapping remains nonlinear.

Limitations & Future Directions. Our experiments involve only two populations at a time. Extending BiXformer to a multi-region setting — with directed communication estimated across many populations — is conceptually straightforward via additional cross-attention streams, but raises nontrivial questions about identifiability, scaling, and interpretation of inter-regional interactions. Beyond scaling, the attention heads offer several avenues for further improving the model's interpretability. In our experiments, only a subset of heads developed selective diagonal profiles, suggesting that head pruning or a sparsity-inducing prior over heads could yield more compact models and a sharper estimate of the number of distinct communication channels. The heads could also be parameterized to learn unimodal lag kernels with explicit centers and widths, providing a more directly interpretable representation of communication delays. Finally, imposing temporal or population-level sparsity on the latents would likely improve the identifiability of individual channels within each stream, particularly when they operate at different timescales, as observed in the neural-behavioral experiment.

AI USE STATEMENT

In this work, we used generative AI to reformat datasets into the input required by the models evaluated in Appendix C. We also used generative AI to polish written text for grammar and clarity, refactor code, and support data and results visualizations. We have reviewed all AI-assisted work: reformatted datasets were checked against the original data for consistency; refactored code was verified through testing to confirm it preserved the original functionality; and visualization code was checked against the underlying data for correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. All experimental data were obtained under protocols approved by the institution's Institutional Animal Care and Use Committee (IACUC).

REPRODUCIBILITY STATEMENT

The model's repository is available at `https://anonymous.4open.science/r/BiXformer-EB1B/`. The model and its training procedure are described in sufficient detail to reproduce both the implementation and the results reported in the paper. The experimental data analyzed here are not released with the submission. However, they can be made available to reviewers upon request, and according to the data's stewards, they will be made publicly available upon final publication of the cited work.

REFERENCES

Alessandra Angelucci and Paul C. Bressloff. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Progress in Brain Research*, 154:93–120, 2006.

Mehdi Azabou et al. A unified, scalable framework for neural population decoding. *arXiv*, 2023.

Mehdi Azabou et al. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *International Conference on Learning Representations*, 2025.

Adrian G. Bondy et al. Brain-wide coordination of internal signals during decision-making. *bioRxiv*, 2024. doi: 10.1101/2024.08.21.609044.

Susu Chen et al. Brain-wide neural activity underlying memory-guided movement. *Cell*, 187(3): 676–691, 2024.

Tudor Dragoi et al. Dynamic engagement of the motor cortex in controlling movement. *bioRxiv*, 2026. doi: 10.64898/2026.01.13.699314.

Laura N. Driscoll et al. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, 170:986–999, 2017.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. *The Geometry of Algorithms with Orthogonality Constraints*, volume 20. 1998.

Charles D. Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14:350–363, 2013.

Joshua Glaser, Matthew Whiteway, John P. Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in Neural Information Processing Systems*, 33:14867–14878, 2020.

Evren Gokcen et al. Disentangling the flow of signals between populations of neurons. *Nature Computational Science*, 2:512–525, 2022.

Evren Gokcen et al. Uncovering motifs of concurrent signaling across multiple neuronal populations. *Advances in Neural Information Processing Systems*, 2023.

Kenneth D. Harris and Thomas D. Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503:51–58, 2013.

Munib A. Hasnain, Jaclyn E Birnbaum, et al. Separating cognitive and motor processes in the behaving mouse. *Nature Neuroscience*, 28:640–653, 2025.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

James J. Jun et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017. doi: 10.1038/nature24636.

Heet Kaku, Liu D. Liu, Runbo Gao, Steven West, Song-Mao Liao, Arseny Finkelstein, David Kleinfeld, Alyse Thomas, Sri Laasya Tipparaju, Karel Svoboda, and Nuo Li. A brainstem map of orofacial rhythms. *bioRxiv*, pp. 2025.01.27.635041, 2025. doi: 10.1101/2025.01.27.635041.

Byungwoo Kang and Shaul Druckmann. Approaches to inferring multi-regional interactions from simultaneous population recordings. *Current Opinion in Neurobiology*, 65:108–119, 2020.

O. Karniol-Tambour et al. Modeling state-dependent communication between brain regions with switching nonlinear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2023.

Stephen L. Keeley, David M. Zoltowski, Mikio C. Aoi, and Jonathan W. Pillow. Modeling statistical dependencies in multi-region spike train data. *Current Opinion in Neurobiology*, 65:194–202, 2020.

Sander W. Keemink and Christian K. Machens. Decoding and encoding (de)mixed population responses. *Current Opinion in Neurobiology*, 58:112–121, 2019.

Andrei Khilkevich et al. Brain-wide dynamics linking sensation to action during decision-making. *Nature*, 634(8035):890–900, 2024. doi: 10.1038/s41586-024-07908-w.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Victor A. Lamme, Hans Supèr, and Henk Spekreijse. Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, 8:529–535, 1998.

Scott W. Linderman et al. Recurrent switching linear dynamical systems. *arXiv*, 2016. doi: 10.48550/ARXIV.1610.08466.

B. Liu, J. Sacks, and Matthew D. Golub. Accurate identification of communication between multiple interacting neural populations. *arXiv*, 2025.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503:78–84, 2013.

Johannes M. Mayrhofer et al. Distinct contributions of whisker sensory cortex and tongue-jaw motor cortex in a goal-directed sensorimotor transformation. *Neuron*, 103:1034–1043, 2019.

Earl K. Miller, Mikael Lundqvist, and Andre M. Bastos. Working Memory 2.0. *Neuron*, 100:463–475, 2018.

Chethan Pandarinath et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15:805–815, 2018.

Matthew G. Perich and Kanaka Rajan. Rethinking brain-wide interactions through multi-region 'network of networks' models. *Current Opinion in Neurobiology*, 65:146–151, 2020.

Matthew T. Schmolesky et al. Signal timing across the macaque visual system. *Journal of Neurophysiology*, 79:3272–3278, 1998.

Joao D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102:249–259, 2019.

Reza Shadmehr and John W. Krakauer. A computational neuroanatomy for motor control. *Experimental Brain Research*, 185:359–381, 2008.

Ram Dyuthi Sristi et al. Coupled transformer autoencoder for disentangling multi-region neural latent dynamics. *arXiv*, 2025.

Nicholas A. Steinmetz, Christof Koch, Kenneth D. Harris, and Matteo Carandini. Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Current Opinion in Neurobiology*, 50:92–100, 2018.

Jianlin Su et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2023.

Jun Takatoh et al. Constructing an adult orofacial premotor atlas in Allen mouse CCF. *eLife*, 10: e67291, 2021. doi: 10.7554/eLife.67291.

Saurabh Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

Qi Xin and Robert E. Kass. Identifying interactions across brain areas while accounting for individual-neuron dynamics with a transformer-based variational autoencoder. *arXiv*, 2025.

Da Xu et al. Cortical processing of flexible and context-dependent sensorimotor sequences. *Nature*, 603:464–469, 2022.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. *bioRxiv*, 2023. doi: 10.1101/2023.09.18.558113.

Yizi Zhang et al. Neural encoding and decoding at scale. *arXiv*, 2025.

APPENDIX

A MODEL DETAILS

A.1 TRAINING

All models were trained using the Adam optimizer (Kingma & Ba, 2015). For the synthetic data experiments, BiXformer was trained for 500 epochs per fold under 5-fold cross-validation. For the neural-behavioral data experiments, training ran for 400 epochs per fold under 5-fold cross-validation. For the multi-region (S1–M1) experiments, training ran for 400 epochs per fold under standard 5-fold cross-validation. For the 100-run random train–test split evaluation used to assess delay distribution robustness (Figures 4e and 12), each independent run was trained for 400 epochs on an $80\%/20\%$ random train–test split of the data, with the split seed set equal to the run index for reproducibility.

A.2 HYPERPARAMETERS

No hyperparameter search was conducted; all values were set based on initial exploratory runs and held fixed across experiments. Table 1 reports the hyperparameters used across all experiments. The synthetic and real-data experiments share the same model dimension, batch size, and learning rate, with minor differences in the number of attention heads and latent dimensionality.

Table 1: Hyperparameters used across all experiments.

Hyperparameter	**Synthetic Data**	**Real Data & 100-Split CV**
Model dimension d	128	128
Attention heads H	8	4
Latent dimension r	16	20
Batch size	32	32
Learning rate	10^{-3}	10^{-3}
Weight decay	10^{-4}	10^{-3}
Gradient clip (max norm)	–	0.5
Training epochs	500	400
Evaluation strategy		5-fold CV
Optimizer		Adam

A.3 COMPUTE

BiXformer was trained on NVIDIA L40S and A100 GPUs for the synthetic-data experiments and on an Apple Silicon GPU (via MPS) for the real-data experiments. The 100-run random train–test split evaluation was conducted across three sessions from three animals, distributed across NVIDIA A100, A40, and L40S GPUs.

B DATASETS

B.1 SYNTHETIC DATA GENERATION

We generated synthetic paired population recordings $\mathbf{X}^{(n)} \in \mathbb{R}^{T\times p}$ and $\mathbf{Y}^{(n)} \in \mathbb{R}^{T\times q}$ for $n = 1, \ldots, N$ trials, where T is the number of time bins and p, q are the dimensionalities of the two populations.

Latent processes. Each latent z_k is an independent draw from a Gaussian Process (GP) with a squared-exponential covariance kernel,

$$\mathrm{Cov}\big(z_k(t),\, z_k(t')\big) = \exp\!\left(-\frac{(t-t')^2}{2\ell_k^2}\right), \tag{10}$$

where $\ell_k > 0$ is the GP timescale of latent k, controlling the smoothness of the trajectory over time.

Cross-area latents and delays. There are n_{lead} leading latents ($\mathbf{X}$ leads $\mathbf{Y}$, delay $\tau_k > 0$) and n_{lag} lagging latents ($\mathbf{Y}$ leads $\mathbf{X}$, delay $\tau_k < 0$), where τ_k denotes the inter-regional communication delay for latent k.

Private latents. In addition, n_x within-$\mathbf{X}$ private latents $\{z_{x,j}^{(n)}\}_{j=1}^{n_x}$ and n_y within-$\mathbf{Y}$ private latents $\{z_{y,j}^{(n)}\}_{j=1}^{n_y}$ are drawn from independent GPs and contribute to only one population with no delay.

Observations. Observations are formed by linearly mixing the latents and adding i.i.d. Gaussian observation noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$:

$$\mathbf{X}^{(n)}(t) = \sum_{k=1}^{n_{\text{lead}}+n_{\text{lag}}} z_k^{(n)}(t-\tau_k^X)\,\mathbf{c}_{X,k}^\top + \sum_{j=1}^{n_x} z_{x,j}^{(n)}(t)\,\mathbf{d}_{X,j}^\top + \boldsymbol{\varepsilon}_X^{(n)}(t), \tag{11}$$

$$\mathbf{Y}^{(n)}(t) = \sum_{k=1}^{n_{\text{lead}}+n_{\text{lag}}} z_k^{(n)}(t-\tau_k^Y)\,\mathbf{c}_{Y,k}^\top + \sum_{j=1}^{n_y} z_{y,j}^{(n)}(t)\,\mathbf{d}_{Y,j}^\top + \boldsymbol{\varepsilon}_Y^{(n)}(t), \tag{12}$$

where $\tau_k^X = 0$, $\tau_k^Y = \tau_k$ for leading latents (so $\mathbf{Y}$ is shifted by delay $\tau_k > 0$ relative to $\mathbf{X}$) and $\tau_k^X = |\tau_k|$, $\tau_k^Y = 0$ for lagging latents (so $\mathbf{X}$ is shifted relative to $\mathbf{Y}$). The vectors $\mathbf{c}_{X,k} \in \mathbb{R}^p$ and $\mathbf{c}_{Y,k} \in \mathbb{R}^q$ are fixed random cross-area loading vectors. The private loading vectors $\mathbf{d}_{X,j} \in \mathbb{R}^p$ and $\mathbf{d}_{Y,j} \in \mathbb{R}^q$ are projected onto the null space of the cross-area loading matrix, ensuring that private variance is orthogonal to the shared cross-area subspace.

Demo dataset. The dataset used in Section 4.1 was generated with $N = 200$ trials, $T = 200$ bins ($\Delta t = 5$ ms), $p = 80$, $q = 50$, $n_{\text{lead}} = n_{\text{lag}} = 2$, $n_x = n_y = 2$. True communication delays were $\boldsymbol{\tau} = [+15, +55, -15, -55]$ ms (positive = $\mathbf{X}$ leads $\mathbf{Y}$; negative = $\mathbf{Y}$ leads $\mathbf{X}$) and cross-area GP timescales were $\boldsymbol{\ell}_{\text{cross}} = [15, 25, 30, 60]$ ms (leading: $\ell = 15, 25$ ms; lagging: $\ell = 30, 60$ ms), private-$\mathbf{X}$ timescales $\boldsymbol{\ell}_x = [20, 55]$ ms, and private-$\mathbf{Y}$ timescales $\boldsymbol{\ell}_y = [20, 55]$ ms.

B.2 NEURAL-BEHAVIORAL RECORDINGS

Adult C57BL/6J mice were implanted with a custom metal headbar, water-restricted, and trained on a head-fixed licking task in which the animal produced tongue movements across a range of angles toward a lickport to obtain water reward, with both go-cue and self-initiated trial types interleaved within each session. After training, a small craniotomy was made over the medullary reticular formation, and acute extracellular recordings were performed using 4-shank Neuropixels 2.0 probes (IMEC) targeting the intermediate reticular nucleus and adjacent areas in which putative licking oscillator neurons have been described (Takatoh et al., 2021; Kaku et al., 2025). Voltage signals were acquired at 30 kHz using SpikeGLX, band-pass filtered between 300 Hz and 6 kHz, spike-sorted with Kilosort 2.5, and automatically curated based on standard quality metrics. Spike times were binned at 2.5 ms resolution and the resulting firing rates were smoothed with a Gaussian kernel ($\sigma = 20$ ms) prior to analysis. During recording, high-speed video of the orofacial region was

acquired at 400 Hz using bottom-view and side-view cameras, and tongue and jaw keypoints were tracked with DeepLabCut.

Three recording sessions from three animals were analyzed, yielding 118, 162, and 181 isolated units respectively. Orofacial kinematics were tracked at 56 keypoints across all sessions.

B.3 Multi-Region Neural Recordings (S1–M1)

Data were obtained from (Dragoi et al., 2026); full surgical and experimental details are provided therein. Briefly, adult C57BL/6J mice (8–13 weeks) were implanted with a headbar and trained on a Delayed Reward licking task. A craniotomy was made over the tongue region of primary motor cortex (tjM1; AP: $+2.2$ mm, ML: ± 2.5 mm) and tongue-jaw primary somatosensory cortex (tjS1; AP: $+0.5$ mm, ML: ± 3.5 mm), and simultaneous extracellular recordings were performed in both areas using Neuropixels 1.0 probes. Spike times were binned at 2.5 ms resolution and firing rates smoothed with a Gaussian kernel ($\sigma = 10$ ms), yielding 47 units in tjM1 and 18 units in tjS1.

In the Delayed Reward task, each trial began with a 2 s LED cue followed by an auditory Go Cue, after which the animal licked a motorized port to receive a water reward. On each trial, reward was delivered on either the first (R1) or fourth (R4) port contact with equal probability (50%), with the lickport cycling among three lateral positions in blocks of 10 trials. Only R1 trials were used in the analysis presented here.

C MODEL COMPARISON

We compared BiXformer to two established inter-regional communication models (DLAG and MR-LFADS) and a linear baseline (RRR) on the synthetic dataset described in Section 4.1 (200 trials, T=200 bins at $\Delta t = 5$ ms, p=80, q=50, four cross-area latents with delays $\boldsymbol{\tau} = [+15, +55, -15, -55]$ ms and GP timescales $\boldsymbol{\ell} = [15, 25, 30, 60]$ ms). All methods were evaluated using the same 5-fold cross-validation split; held-out R^2 was computed on population $\mathbf{Y}$ predicted from population $\mathbf{X}$.

DLAG. DLAG (Gokcen et al., 2022) is a Gaussian-process-based latent variable model that estimates directed inter-regional communication delays from paired population recordings. It was fit with 4 across-area latents (matching ground truth) and 2 within-area latents per region. Delays were read directly from the fitted DLAG delay parameters.

MR-LFADS. MR-LFADS (Liu et al., 2025) is a multi-region sequential variational autoencoder that infers communication signals alongside local latent dynamics. It was fit with the communication dimension set to 4 and trained for 500 epochs per fold; at test time, the $\mathbf{Y}$ encoder input was zeroed so predictions depend only on communication from $\mathbf{X}$. In our evaluation, MR-LFADS underperformed RRR ($R^2 = 0.210$ vs. 0.341). The original authors acknowledge that MR-LFADS can be sensitive to hyperparameter settings, and we did not conduct a dedicated hyperparameter search; this likely accounts for the gap. We report these results transparently rather than omitting them.

Table 2: **Model comparison on synthetic dataset.** Held-out R^2 (variance-weighted, 5-fold CV), delay estimation accuracy (mean absolute error after Hungarian matching to ground truth, mean $\pm$ std across folds), and per-fold training time (mean $\pm$ std across folds).

Method	R^2 (mean $\pm$ std)	Delay MAE (ms)	Train time / fold (min)
RRR (r=4)	0.341 ± 0.005	N/A	N/A
DLAG	0.590 ± 0.074	13.5 ± 8.9	93.9 ± 3.9
MR-LFADS*	0.210 ± 0.011	N/A	81.6 ± 2.1
BiXformer	$\mathbf{0.887 \pm 0.018}$	$\mathbf{1.0 \pm 2.0}$	$\mathbf{0.44 \pm 0.01}$

*No hyperparameter search was conducted.

C.1 COMPUTE

BiXformer and MR-LFADS were trained using L40S and A100 GPUs, whereas DLAG was run on CPU, as its EM-based MATLAB implementation was not designed for GPU acceleration. These runtimes therefore reflect the practical training overhead of each approach on the hardware it was designed to use.

D Additional Results

D.1 Synthetic Data: Attention Maps and Diagonal Profiles Across All Heads

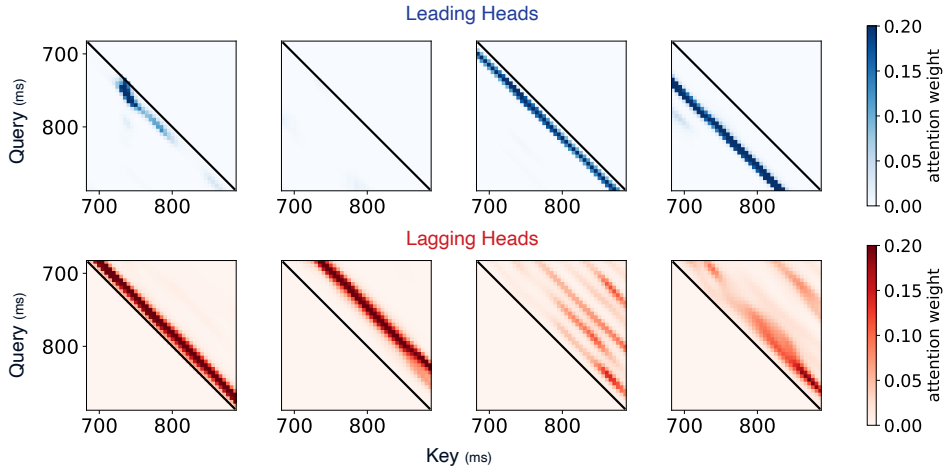

Figure 6: **Cross-attention weight maps for all attention heads.** Trial-averaged attention weights for all eight heads — four leading (blue) and four lagging (red) — from a representative held-out fold of the synthetic experiment, zoomed in on a narrow time window for clarity. Two heads per stream show sharp off-diagonal bands consistent with the ground-truth delays, while the remaining heads exhibit diffuse, non-selective attention.

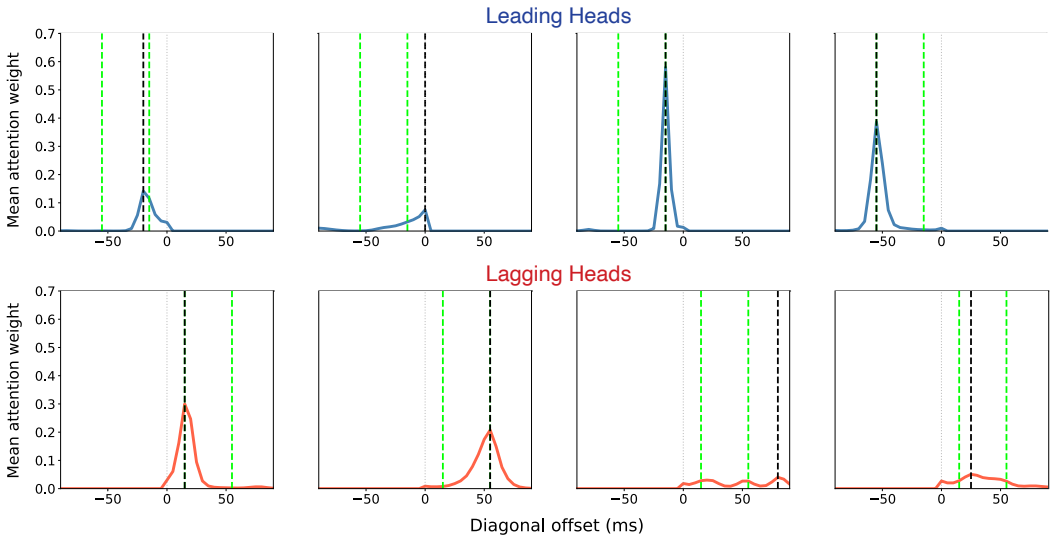

Figure 7: **Diagonal profiles for all attention heads.** Trial-averaged diagonal profiles for the same fold and heads as Figure 6. Leading heads 3 and 4 accurately recover the ground-truth delays of 15 and 55 ms, respectively, while leading heads 1–2 yield broad, uninformative profiles; lagging heads 1 and 2 recover $+15$ and $+55$ ms, while heads 3–4 remain diffuse. Green dashed lines mark ground-truth delays; black dashed lines indicate the per-head peak estimate.

D.2 Constrained vs. Unconstrained BiXformer: Held-Out Reconstruction Performance

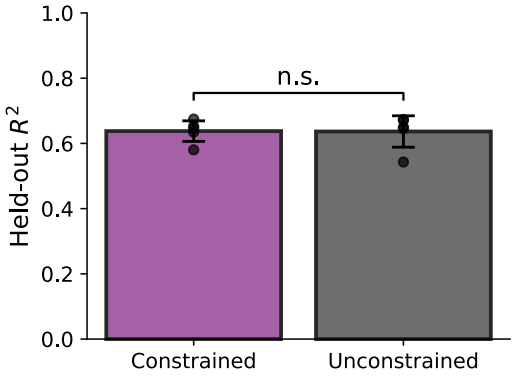

Figure 8: **Held-out R^2 for constrained and unconstrained BiXformer across 5-fold CV on the neural-behavioral dataset.** The unconstrained variant removes the causal/acausal attention masking and Stiefel decoder constraint, with all other hyperparameters and parameter counts held fixed. Bars show mean $\pm$ std across folds (constrained: 0.638 ± 0.032; unconstrained: 0.637 ± 0.048), with individual fold values overlaid. Differences are not statistically significant (paired t-test, $p = 0.932$).

D.3 Neural-Behavioral Data: Attention Maps and Diagonal Profiles for Sensory and Motor Heads

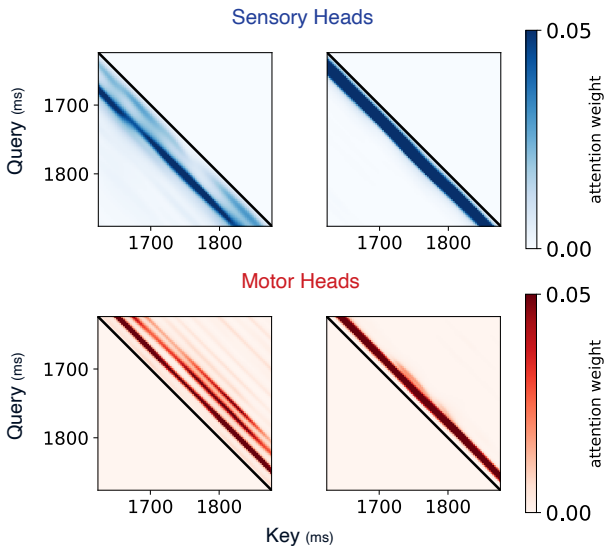

Figure 9: **Cross-attention weight maps for sensory and motor heads.** Trial-averaged attention weights for the two sensory heads (blue, leading stream) and two motor heads (red, lagging stream) from a representative held-out fold of the neural-behavioral experiment. Both head types show sharp off-diagonal bands, indicating that each consistently attends to a specific temporal offset between medullary neural activity and orofacial kinematics.

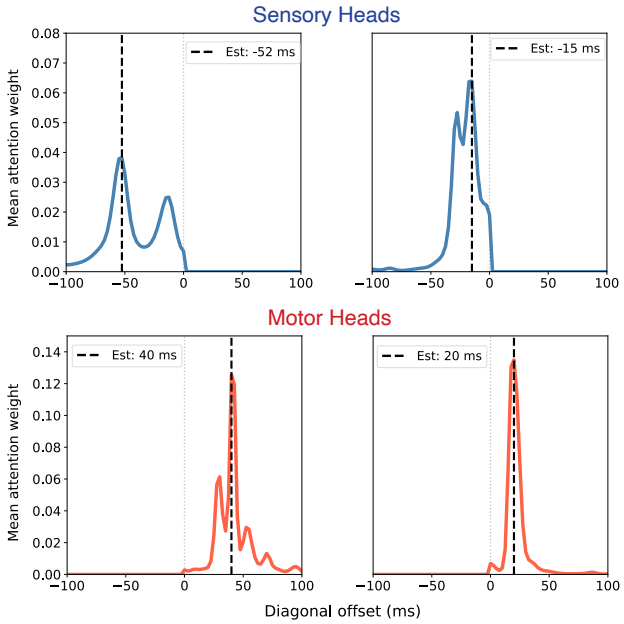

Figure 10: **Diagonal profiles for sensory and motor heads.** Diagonal profiles for the same fold and heads as Figure 9. Sensory heads identify negative diagonal offsets (-52 and -15 ms), indicating that behavior leads neural activity, consistent with ascending sensory feedback; motor heads identify positive offsets ($+40$ and $+20$ ms), indicating that neural activity leads behavior, consistent with a feedforward motor command. The sharp, well-localized peaks in all four profiles demonstrate that the model consistently latches onto specific temporal lags between medullary neural activity and orofacial kinematics.

D.4 GMM Model Fit

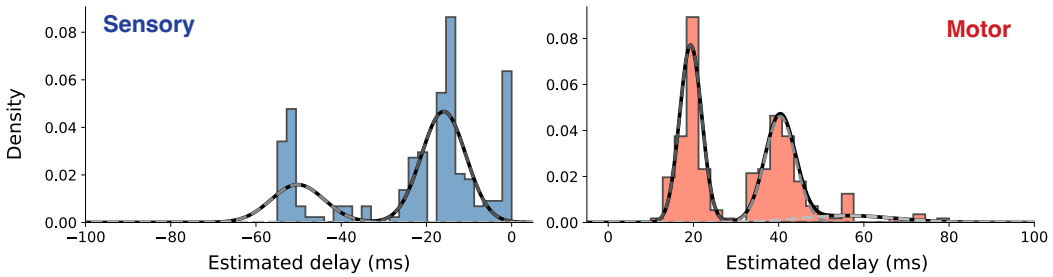

Figure 11: **Gaussian mixture model fit to the bootstrapped delay estimates shown in Fig. 4e.** Delays estimated across 100 random 80/20 train–test splits were pooled across all attention heads within each stream and fit with a Gaussian mixture model (`sklearn.mixture.GaussianMixture`, full covariance); the solid black curve shows the total mixture density, and dashed gray curves show the individual components. The sensory stream (X→Y) resolves into two dominant components at -16.0 ms and -50.2 ms (weights 0.61 and 0.25), consistent with behavior leading neural activity, while the motor stream (Y→X) resolves into components at $+19.4$ ms and $+40.4$ ms (weights 0.50 and 0.42), consistent with neural activity leading behavior.

D.5 Delay Distributions Across Additional Animals

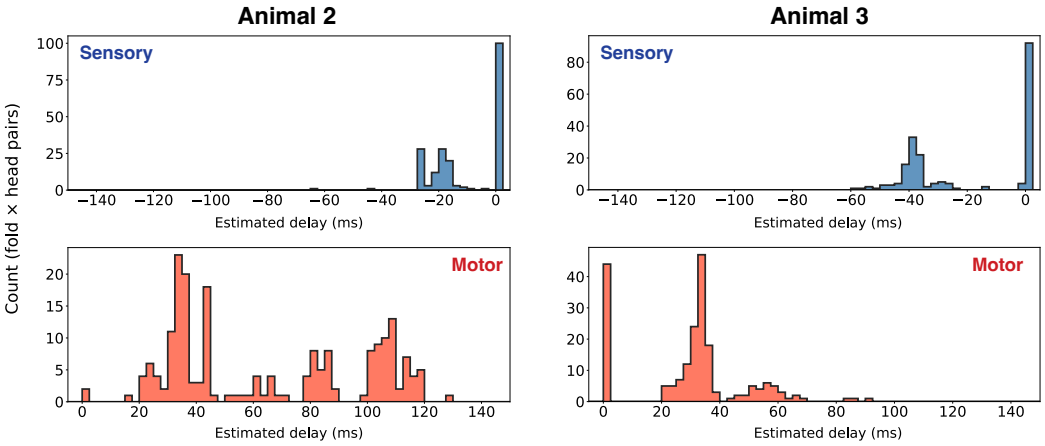

Figure 12: **Estimated delay distributions for different animals**. We performed the same random train-test split discussed in Section 4.2 for two different sessions from two different animals and we found distinct characteristic lags across animals, reflecting differences in recording sites and the specific neural populations sampled within the IRN/PARN regions of the medulla.

D.6 Multi-Region (S1–M1) Data: Attention Maps and Diagonal Profiles for S1→M1 and M1→S1 Heads

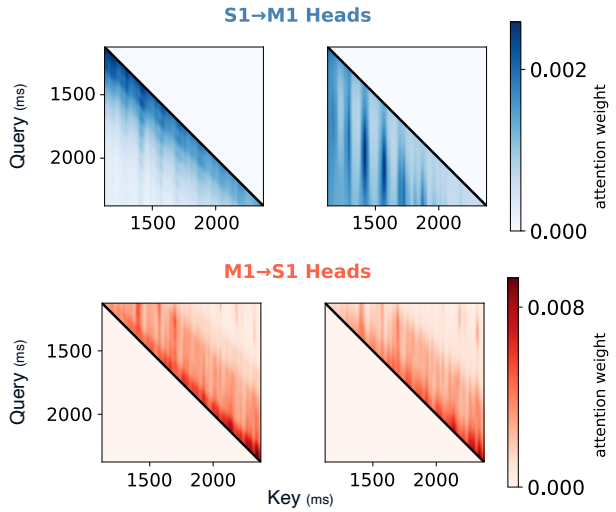

Figure 13: **Cross-attention weight maps for S1→M1 and M1→S1 heads.** Trial-averaged attention weights for the two S1→M1 heads (blue, leading stream) and two M1→S1 heads (red, lagging stream) from a representative held-out fold of the multi-region (S1–M1) experiment. Unlike the synthetic and neural-behavioral experiments, attention weights decay monotonically away from the main diagonal rather than concentrating into a sharp off-diagonal band, suggesting no single dominant communication delay between S1 and M1 in this session.

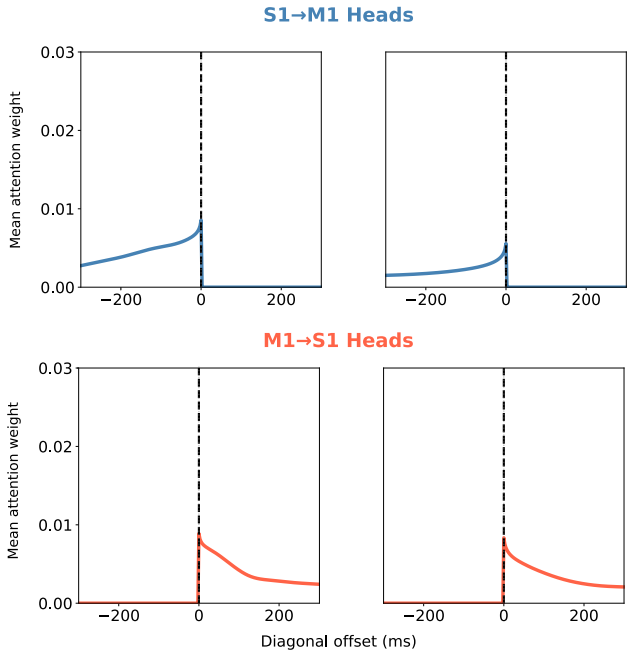

Figure 14: **Diagonal profiles for S1→M1 and M1→S1 heads.** Diagonal profiles for the same fold and heads as Figure 13. All four heads show attention weight decaying monotonically from zero offset, with no peak at a non-zero lag.