DEMIXING TASK-CONDITIONED SHARED AND PRIVATE NEURAL SUBSPACES WITH TEMPORAL ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-region neural populations often mix sensory and decision-related signals, making it hard to tell which activity reflects sensory evidence, which reflects emerging choice, and when each becomes useful on a single trial. We introduce the *Time-varying Subspace Autoencoder* (TSA), a Poisson autoencoding model with subspace-specific temporal attention readouts. TSA organizes neural population activity into stimulus-private, choice-private, and shared latent subspaces, where "private" denotes preferential rather than exclusive task information, and uses attention gates to determine when each subspace contributes to stimulus and choice decoding. Its encoder, decoder, and attention scorer are all causal in time, so activity in a later time bin can never influence an earlier latent state or readout weight. On synthetic data, TSA recovers the ground-truth stimulus-before-choice ordering more often than matched controls without the task partition. On barrel-cortex recordings from a perceptual decision-making task, TSA gives the highest held-out stimulus decoding among the methods we compare; its stimulus readout tracks whisker contact, which is never used in training, and on average peaks before the choice readout. Controls show that this timing is not created by the separation penalties, and that at matched capacity the partition improves stimulus decoding and contact alignment. Together, these results position TSA as a descriptive tool for studying when sensory and decision signals are read out from shared and private neural subspaces.

1 INTRODUCTION

Large-scale recordings have shifted neural computation analysis from isolated neurons to population activity. A central lesson is that neural activity is often lower-dimensional than the number of recorded neurons, yet that low-dimensional structure is not monolithic. In the same recorded population, activity can reflect sensory evidence, latent task variables, movement-related activity, and trial-to-trial fluctuations (Cunningham & Yu, 2014; Mante et al., 2013; Churchland et al., 2012). Mixed selectivity makes this both powerful and difficult: heterogeneous responses can support flexible linear readouts, yet they obscure which population dimensions carry which task variables and when those variables are read out (Rigotti et al., 2013; Fusi et al., 2016).

This ambiguity is especially acute in perceptual decision-making tasks, where the main task variables are stimulus and choice. Because these variables are correlated, stimulus-related activity can help predict choice and choice-related activity can help predict stimulus, even though the two are not the same computation. Time provides one useful source of difference: stimulus-related and choice-related signals should emerge at different moments, but those moments are not directly observed and can vary from trial to trial. Contact timing, go-cue onset, and movement time do not by themselves identify when stimulus-conditioned or choice-conditioned neural subspaces are active, nor how long those signals persist. A useful model should therefore ask not only *whether* stimulus and choice are decodable, but also *from which latent subspace* and *when on each trial* they are read out.

Existing methods point in two complementary directions. One direction is *task-conditioned but not explicitly time-varying at the single-trial readout level*: methods such as demixed PCA (Kobak et al., 2016) can separate stimulus- and choice-related structure, but they do not provide explicit trial-specific readout weights over time, nor do they distinguish stimulus-private, choice-private, and shared subspaces within one population. The other direction is *time-varying but not explicitly task-*

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.

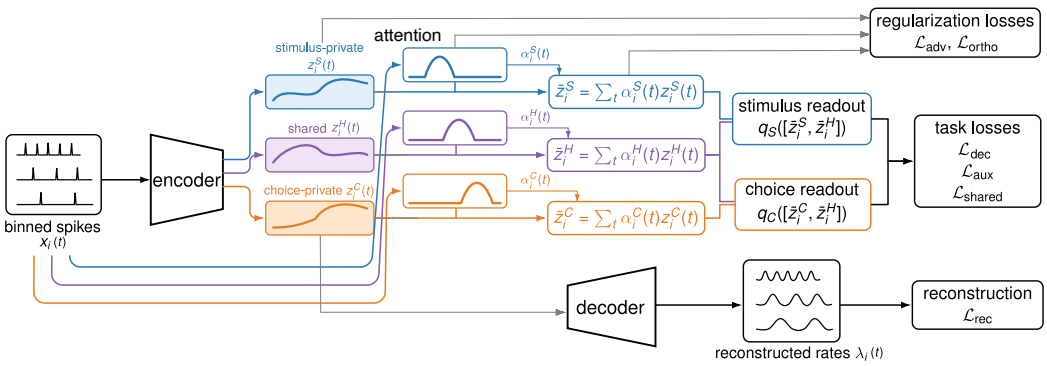

Figure 1: Overview of the Time-Varying Subspace Autoencoder. A causal convolutional encoder maps spikes to a time-resolved latent that is partitioned into stimulus-private (S), choice-private (C), and shared (H) subspaces, and a pointwise decoder reconstructs activity from the full latent. Attention gates, each a single causal convolution, pool the latent over time for the linear stimulus and choice readouts. The gates read only the spikes, with no input from the latent, so no component can see future time bins. Auxiliary, shared, cross-task, and covariance terms encourage the intended roles of the subspaces.

conditioned: latent dynamical models such as SLDS (Zoltowski et al., 2020) can recover time-varying structure, but they do not assign it to explicit task conditions, nor do they separate stimulus-related and choice-related signals into private and shared roles. Shared and private subspaces are usually discussed across brain regions, but the idea is also meaningful within one area: even in a single population, stimulus and choice should have some common and some variable-specific structure.

We propose the *Time-Varying Subspace Autoencoder* (TSA) for this setting. TSA uses a Poisson autoencoding model to keep the latent state tied to the recorded spikes, partitions that latent representation into stimulus-private, choice-private, and shared components, and uses attention gates to determine when each subspace is used for readout. Our main scientific readout is this *temporal attention*: when stimulus-related and choice-related information becomes prominent within a trial. The private/shared split serves that goal. Because stimulus and choice are behaviorally correlated, an exclusive split is neither achievable nor desirable, so "private" here means *preferential* task information. What the split provides is a *task-attributed* readout: TSA can say when the stimulus-private subspace is read out, which a single undivided latent cannot express at all. We evaluate TSA on synthetic data with a known latent organization and on barrel-cortex recordings. Although we focus on stimulus and choice, the framework applies whenever several task variables may share part of their neural representation while retaining variable-specific structure (Appendix G).

Scientifically, the value of TSA is not only that it recovers low-dimensional structure, but that it identifies when different task-conditioned subspaces become informative and when their contribution fades. This provides a temporal picture of computation that standard latent subspace methods usually do not offer: they can recover low-dimensional structure, but they do not typically distinguish private versus shared dimensions across task conditions, nor do they directly localize the time windows in which different task-conditioned subspaces drive decoding. In principle, this separation can generate new hypotheses about how a single neural population routes sensory and decision signals.

2 Related Work

Latent and sequence models. Many latent neural population models recover smooth low-dimensional activity, support denoising, or improve prediction, including GPFA, LFADS, and neural transformers (Yu et al., 2009; Pandarinath et al., 2018; Pei et al., 2021; Le & Shlizerman, 2022; Liu et al., 2022; Ye et al., 2023; Azabou et al., 2023). These methods capture single-trial population structure, but they do not assign distinct task roles to different latent dimensions.

Demixing, mixed selectivity, and task-aligned subspaces. Mixed selectivity motivates methods that explicitly demix task structure, including dPCA, tensor methods, and related decompositions (Rigotti et al., 2013; Fusi et al., 2016; Kobak et al., 2016; Williams et al., 2018; Pellegrino et al., 2024); Aoi et al. (2020) further showed that task variables are encoded along multiple, dynamically changing dimensions in prefrontal cortex. A related line studies preparatory/output, communication, or task-shared subspaces across populations or regions (Kaufman et al., 2014; Elsayed et al., 2016; Semedo et al., 2019; Gokcen et al., 2022; 2023; Takagi et al., 2020; Tafazoli et al., 2026), including multi-region deep models (Sristi et al., 2026; Li et al., 2026). TAME-GP comes especially close by introducing explicit private/shared structure in a probabilistic setting (Balzani et al., 2023). However, none of these jointly provides a within-region split into task-private and shared subspaces together with explicit single-trial time-varying readout weights for each subspace.

Behavior-guided and disentangled neural representations. Behavior-guided and disentangled latent models introduce supervision or structure to make representations interpretable (Kingma & Welling, 2014; Higgins et al., 2017; Kim & Mnih, 2018; Locatello et al., 2019; Zhou & Wei, 2020; Schneider et al., 2023). PSID, TNDM, DPAD, BRAID, and related approaches separate behaviorally relevant dynamics from other neural variation (Sani et al., 2021; Hurwitz et al., 2021; Sani et al., 2024; Vahidi et al., 2025; Wang et al., 2024), and multimodal models learn shared and modality-private latents between neural and behavioral data (Gondur et al., 2024; Yi et al., 2025). These models distinguish *neural versus behavioral* or *relevant versus irrelevant* structure, rather than partitioning two correlated task variables into task-private and shared roles within the neural latent.

Temporal alignment and attention. Time-warping and switching linear dynamical systems model single-trial timing variability (Williams et al., 2020; Zoltowski et al., 2020); they can indicate which discrete regime is active, but they do not tie regimes to task-specific private/shared subspaces. Attention builds readout-weighted temporal summaries (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), although attention weights should not be over-interpreted as causal explanations (Jain & Wallace, 2019; Serrano & Smith, 2019). TSA uses attention in a narrower sense: a continuous, task-attributed readout weight that asks when each subspace is useful for decoding its supervised variable, with several subspaces allowed to be active at the same time.

3 METHOD

We first define a latent-variable model tied to neural activity through reconstruction, then split the latent into task-conditioned subspaces, introduce temporal attention to determine when each subspace is informative, and finally define the losses that make these roles operational. Every component is designed so that information only flows forward in time.

3.1 NEURAL LATENT MODEL

For trial i, let $\mathbf{x}_i \in \mathbb{R}^{T\times N}$ denote binned activity from N neurons over T time bins, with binary stimulus $s_i \in \{0,1\}$ and choice $c_i \in \{0,1\}$. The encoder E_ϕ maps neural activity to a time-resolved latent, $\mathbf{z}_i(1:T) = E_\phi(\mathbf{x}_i(1:T))$, and the decoder maps it back to nonnegative rates, $\lambda_i(1:T) = \mathrm{softplus}(D_\theta(\mathbf{z}_i(1:T)))$.

The encoder consists of two causal temporal convolution layers, so the latent $\mathbf{z}_i(t)$ depends only on spikes at times $\leq t$. The decoder mirrors the encoder's depth but is pointwise (kernel size 1), so the reconstructed rate at time t is a function of the latent at time t alone. Together, these two choices remove every path by which activity in a later bin could influence an earlier latent state or an earlier reconstructed rate. This matters for our question: if the model could look ahead, a stimulus signal might appear to rise before whisker contact simply because later contact-driven activity leaked backwards.

We train the reconstruction pathway with the Poisson negative log likelihood

$$\mathcal{L}_{\mathrm{rec}} = \textstyle\sum_{i,t,n} \left[\lambda_{i,t,n} - x_{i,t,n} \log(\lambda_{i,t,n} + \epsilon)\right].$$

Without this term, a small supervised subspace could solve the tasks while discarding large amounts of neurally real structure.

3.2 TASK-CONDITIONED SUBSPACE SPLIT

We split the latent state at each time bin into $\mathbf{z}_i(t) = \left[\mathbf{z}_i^S(t), \mathbf{z}_i^C(t), \mathbf{z}_i^H(t)\right] \in \mathbb{R}^{d_z}$ with $d_z = d_S + d_C + d_H$, where $\mathbf{z}^S$ is stimulus-private, $\mathbf{z}^C$ is choice-private, and $\mathbf{z}^H$ is shared task-related activity. These labels describe *intended, preferential* roles induced by the objective, not guaranteed causal or identifiable quantities. The split reflects our working hypothesis that, in a single area such as barrel cortex (Nogueira et al., 2023), some dimensions are more closely tied to sensory evidence, some to emerging choice, and some to both. On real data we use a balanced layout of $23/23/24$ dimensions, selected by a search over 14 layouts (Appendix C.1).

3.3 SUBSPACE-SPECIFIC TEMPORAL ATTENTION

Once the latent representation has been split into task-conditioned subspaces, the next question is when each subspace becomes informative within a trial. In perceptual decision-making, sensory and choice-related activity may overlap in population space yet emerge at different moments, and those moments need not be fixed across trials. Standard subspace methods can recover low-dimensional structure, but they do not directly reveal when a task-conditioned subspace drives readout on a given trial. We therefore introduce explicit temporal attention to identify, for each trial, when stimulus-private, choice-private, and shared activity are most informative.

Each task-relevant subspace has a separate temporal attention network. In the implementation used here, for $r \in \{S, C, H\}$, the scorer g_r is a single causal convolutional layer. At each time bin, g_r takes only the observed activity in the current and two preceding bins, $\mathbf{x}_i(t-2:t)$, and produces a scalar score that is normalized over time as

$$\alpha_i^r(t) = \frac{\sigma\big(\beta\, g_r(\mathbf{x}_i(t-2:t))\big)}{\sum_{\tau=1}^{T} \sigma\big(\beta\, g_r(\mathbf{x}_i(\tau-2:\tau))\big) + \epsilon},$$

where β is a fixed temperature that controls how peaked the attention is. The resulting subspace-specific pooled representation is $\bar{\mathbf{z}}_i^r = \sum_{t=1}^{T} \alpha_i^r(t)\mathbf{z}_i^r(t)$. The stimulus classifier reads out from stimulus-private and shared pooled activity as $\hat{s}_i = q_S\left([\bar{\mathbf{z}}_i^S, \bar{\mathbf{z}}_i^H]\right)$, and the choice classifier reads out from choice-private and shared pooled activity as $\hat{c}_i = q_C\left([\bar{\mathbf{z}}_i^C, \bar{\mathbf{z}}_i^H]\right)$. Here q_S and q_C are linear classifiers on pooled latent features. The attention heads are supervised only through task losses and regularizers; they are not given stimulus or choice timing labels. The resulting trajectories $\alpha_i^S(t)$, $\alpha_i^C(t)$, and $\alpha_i^H(t)$ therefore provide a direct single-trial summary of when each task-conditioned subspace contributes to decoding.

Two properties of this scorer are deliberate. First, it is causal. We avoid centered convolutions, bidirectional RNNs, and Transformers over the whole trial in this scorer, because a model that can see future time bins could leak contact-timing information backward in time and create an artificial stimulus ramp before contact. Together with the causal encoder, this means that no attention weight depends on activity that occurs later in the trial. Second, the scorer reads only the observed spikes and not the latent state. The post hoc decoders we fit to the baselines read latent states inferred from the spikes, so TSA's timing cannot come from a richer input than the baselines receive.

3.4 LOSS CONSTRUCTION

Given the neural latent model, the subspace split, and the temporal attention mechanism, we now specify the losses that encourage the intended roles of the subspaces. All losses are averaged over the training batch, and BCE denotes binary cross-entropy. The decoding loss decodes stimulus and choice from their intended private-plus-shared features,

$$\mathcal{L}_{\text{dec}} = \text{BCE}(\hat{s}_i, s_i) + \text{BCE}(\hat{c}_i, c_i).$$

This is the minimum supervision that encourages task information to appear in the intended readout channels. By itself, however, it is not enough: both tasks could be solved by routing nearly all discriminative signal through the shared subspace, or by letting one private subspace carry the wrong variable.

Auxiliary losses decode each task from its private subspace alone, using separate heads,

$$\mathcal{L}_{\text{aux}} = \text{BCE}(q_S^{\text{aux}}(\bar{\mathbf{z}}_i^S), s_i) + \text{BCE}(q_C^{\text{aux}}(\bar{\mathbf{z}}_i^C), c_i),$$

so that $\mathbf{z}^S$ and $\mathbf{z}^C$ remain useful for their designated variables instead of becoming empty slots. Shared losses allow the shared subspace to carry task-positive signal,

$$\mathcal{L}_{\text{shared}} = \text{BCE}(q_S^H(\bar{\mathbf{z}}_i^H), s_i) + \text{BCE}(q_C^H(\bar{\mathbf{z}}_i^H), c_i).$$

Without $\mathcal{L}_{\text{shared}}$, the optimizer can push almost all discriminative content into the private subspaces, leaving the shared block underused even when stimulus and choice are correlated.

The auxiliary and shared terms only add positive pressure to carry desired signals; they do not remove undesired ones. We therefore discourage cross-task leakage with gradient-reversal adversaries (Ganin et al., 2016), where GRL denotes a gradient reversal layer:

$$\mathcal{L}_{\text{adv}} = \text{BCE}(a_C(\text{GRL}(\bar{\mathbf{z}}_i^S)), c_i) + \text{BCE}(a_S(\text{GRL}(\bar{\mathbf{z}}_i^C)), s_i).$$

Here a_C tries to decode choice from the stimulus-private pooled latent and a_S tries to decode stimulus from the choice-private pooled latent, while the encoder receives reversed gradients that make these predictions harder. In the real-data model, this role is filled by a cross-task probe penalty together with a directional HSIC dependence penalty (Gretton et al., 2005); adding gradient-reversal adversaries instead gives similar decoding (Appendix E.2).

Finally, we penalize cross-subspace covariance. Let $\tilde{Z}^r \in \mathbb{R}^{BT \times d_r}$ be the batch-time matrix of subspace r, stacking all B trials and T time bins in the batch, with each dimension standardized across its BT rows. The orthogonality penalty is

$$\mathcal{L}_{\text{ortho}} = \sum_{r<r'} \left\| \frac{(\tilde{Z}^r)^\top \tilde{Z}^{r'}}{BT} \right\|_F^2.$$

This term addresses a different failure mode from the cross-task terms: even if wrong-label decoding is weak, two subspaces can still carry nearly the same directions up to a linear reparameterization. The covariance penalty discourages that redundancy and makes the decomposition easier to interpret.

The full objective is

$$\mathcal{L} = w_{\text{rec}}\mathcal{L}_{\text{rec}} + w_{\text{dec}}\mathcal{L}_{\text{dec}} + w_{\text{aux}}\mathcal{L}_{\text{aux}} + w_{\text{shared}}\mathcal{L}_{\text{shared}} + w_{\text{adv}}\mathcal{L}_{\text{adv}} + w_{\text{ortho}}\mathcal{L}_{\text{ortho}},$$

together with a small smoothness penalty on the attention weights (Appendix A). Each loss targets a specific failure mode of an underdetermined decomposition: $\mathcal{L}_{\text{rec}}$ preserves neural fidelity, $\mathcal{L}_{\text{dec}}$ encourages task decodability, $\mathcal{L}_{\text{aux}}$ keeps the private subspaces informative, $\mathcal{L}_{\text{shared}}$ prevents the shared subspace from collapsing, $\mathcal{L}_{\text{adv}}$ suppresses wrong-variable leakage, and $\mathcal{L}_{\text{ortho}}$ reduces geometric redundancy. We therefore view the multiple losses not as overengineering, but as complementary inductive biases required by the scientific objective.

Optimizing all of these losses at once from random initialization is unstable, so we use a staged strategy within a single training run. We first optimize $\mathcal{L}_{\text{rec}}$ alone, then turn on the decoding loss, then the auxiliary and shared losses, and only in the final stage the adversarial and orthogonality penalties. The stage boundaries are fixed fractions of training rather than being chosen by validation performance. Staging mainly improves reconstruction and the cross-seed stability of the ordering (Appendix E.2).

4 Synthetic Validation Experiment

We first validate TSA in a synthetic perceptual decision-making task with early stimulus-related and later choice-related information, where we can ask whether the model recovers the known timing and latent structure. The simulator contains known stimulus-private, choice-private, and shared factors, and produces noisy Poisson observations through mixed neural readouts with contact-time jitter, amplitude variability, nuisance activity, and baseline drift. We use 2048 trials, 14 observed neurons, and 25 time bins per trial (Appendix B).

We generated three dataset seeds and trained three model seeds per dataset seed, yielding nine runs in total. On this benchmark, TSA achieved high balanced accuracy for both labels (0.956 ± 0.004 for stimulus and 0.992 ± 0.002 for choice), showing that the model reliably recovers the supervised variables from noisy population activity. More importantly, stimulus attention peaked before choice

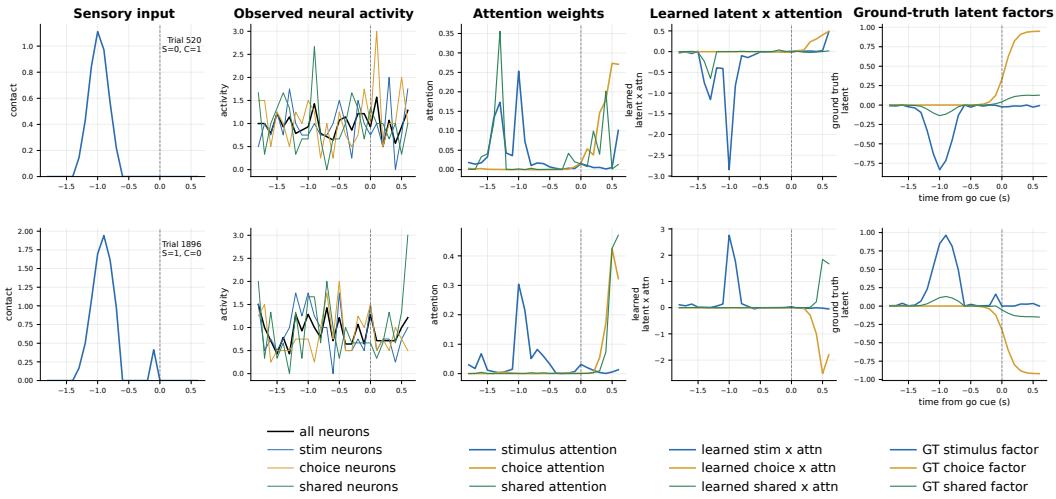

Figure 2: Representative synthetic single-trial examples.

attention on 0.949 ± 0.006 of trials, indicating that the model recovered the intended stimulus-before-choice ordering rather than collapsing both variables into one generic readout.

We also quantified latent recovery using held-out R^2 of the best linear map from each *unweighted* learned latent block (the latents that reconstruct the spikes, before attention weighting) to its ground-truth factor. Recovery was partial: $R^2 = 0.41$ for the stimulus-private factor, 0.52 for the choice-private factor, and 0.13 for the shared factor. The model without separation penalties reached similar values, so recovery of the factors themselves does not depend on those penalties; what the penalties change is the recovered timing. We read this as partial recovery of task-related structure, not near-perfect identification of the true latents.

Figure 2 shows two representative single-trial examples from the synthetic dataset. Across columns, the figure moves from stimulus presentation and observed neural activity to learned attention, attention-weighted latent signals, and the corresponding ground-truth factors. These examples show that TSA does more than decode stimulus and choice correctly: it captures temporally organized single-trial readout dynamics in which stimulus-related activity tends to emerge earlier, choice-related activity peaks later, and the learned latent structure qualitatively follows the ground-truth factors.

5 Real Neural Data Experiment

5.1 Dataset, trial structure, and baselines

We apply TSA to barrel-cortex recordings from a mouse whisker-based shape-discrimination task (Rodgers, 2022) (Figure 3A). On each trial, the animal samples the object with its whiskers, accumulates contact-driven sensory evidence, and then reports a left or right choice about whether the object is convex or concave. Barrel cortex carries strong contact-related signals while the task still exhibits later choice-related variability, so stimulus and choice are correlated but not identical. The analyzed session contains 14 simultaneously recorded neurons and roughly 159 valid trials, split 80/20 (stratified by stimulus) into about 127 training and 32 test trials. Each trial is aligned to response-window opening at time 0 and represented as a 25×14 spike-count matrix from -2.0 to $+0.5$ s in 100 ms bins. Whisker-contact records serve as a timing reference for interpretation only and are never used as model inputs or targets. Unless stated otherwise, we report the mean and standard deviation over 10 independently trained seeds (Appendix C). We repeat the main analyses on a second session from a different animal in Appendix D.

Because there is no ground-truth decomposition, we treat TSA as a *descriptive* model and ask whether it decodes held-out labels, reconstructs held-out spikes, and produces timing that relates sensibly to variables it was never trained on. We compare TSA with dPCA (Kobak et al., 2016), a trial-averaged task-conditioned method; PSID and DPAD (Sani et al., 2021; 2024), which prioritize behaviorally relevant dynamics and receive stimulus and choice as trial-level behavioral variables; sticky and

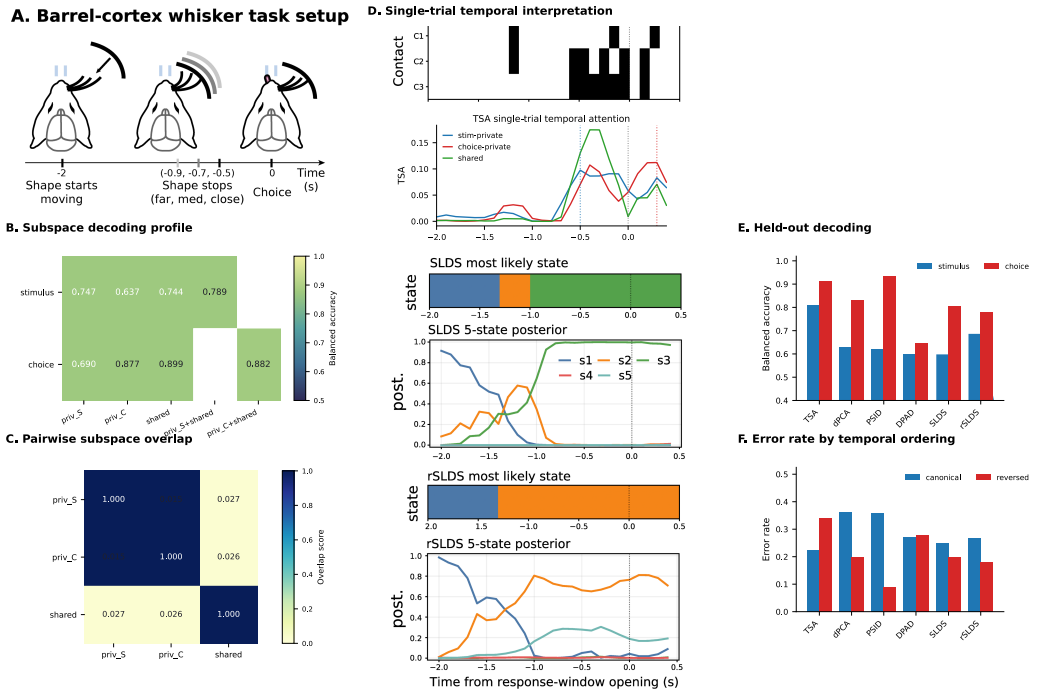

Figure 3: Real-data results. A. Task and trial timing (time 0: response-window opening). B. Held-out balanced accuracy of probes read from each subspace and their combinations. C. Pairwise subspace overlap (0 means orthogonal). D. Held-out trial 165: whisker contact (black), TSA attention for each subspace (dotted lines mark peaks), and SLDS/rSLDS states. E. Held-out decoding. F. Behavioral error rate on canonical and reversed trials.

recurrent SLDS (Fox et al., 2011; Zoltowski et al., 2020), which capture discrete state switches without task roles. All models are fit on the same training trials and scored on the same held-out trials. None of the baselines provides task-specific readout weights over time, so we derive post hoc timing signals from per-time-bin decoders fit to their latents (Appendix F.1).

5.2 Decoding and preferential role separation

We first ask how much task information TSA extracts and how that information is distributed across the learned subspaces. At the whole-model level (Figure 3E), TSA gives the highest held-out stimulus decoding of any method we compare and, among the models scored on the same held-out trials, the second-highest choice decoding after PSID. Stimulus is the harder variable on this dataset, so this is a gain exactly where the task is most difficult. The causal design also does not cost accuracy: compared with a non-causal variant of the same model, TSA improves held-out stimulus decoding by 12 BA points while matching choice decoding (Appendix A).

Figure 3B shows how this information is spread across the blocks, using attention-pooled probes on held-out trials. Each private block decodes its own label better than the other private block does: the stimulus-private block is more informative about stimulus than the choice-private block, and the choice-private block is more informative about choice than the stimulus-private block. The shared block carries both variables, and adding it to either private block improves the corresponding readout. The stimulus-private block nevertheless carries substantial choice information, and its selectivity is weaker than that of the choice-private block. This is expected on this session, where choice is the more decodable variable and stimulus and choice are correlated through behavior, so stimulus-related activity also carries choice information (Appendix D). We therefore read the roles as preferential rather than exclusive.

The overlap matrix in Figure 3C shows that the learned subspaces are close to orthogonal, so the decomposition is not achieved by relabeling one redundant latent cloud. Finally, TSA still reconstructs

Table 1: Timing summary (time 0: response-window opening). Peak times of the contact envelope and of the stimulus- and choice-related signals (TSA: mean attention; baselines: post hoc decoding curves), and the correlation $\rho_{\mathrm{stim,contact}}$ between the stimulus-related signal and the contact envelope.

Model	$t_{\mathrm{peak}}^{\mathrm{contact}}$ (s)	$t_{\mathrm{peak}}^{\mathrm{stim}}$ (s)	$t_{\mathrm{peak}}^{\mathrm{choice}}$ (s)	$\rho_{\mathrm{stim,contact}}$
TSA	-0.4	-0.3	-0.2	0.763
dPCA	-0.4	-0.9	-1.2	0.062
PSID	-0.4	0.0	0.3	0.795
DPAD	-0.4	-1.8	-1.3	-0.202
SLDS	-0.4	-1.6	0.0	-0.538
rSLDS	-0.4	-1.6	-0.1	-0.435

held-out spikes reasonably. The reconstruction cost comes from the demixing regularizers rather than the causal backbone, which reconstructs better when trained on reconstruction alone, and it can be traded back through the reconstruction weight (Appendix E.1).

5.3 When does stimulus-related readout emerge?

Because shape information arrives through whisker contact, a stimulus-related signal should arise after contact begins. Table 1 summarizes each model's timing and the correlation between its stimulus-related signal and the contact envelope.

TSA's stimulus attention is closely aligned with whisker contact ($\rho = 0.749$), comparable to PSID and far above the other baselines, even though contact is never used in training. This alignment is therefore a property that TSA finds in the neural data rather than one imposed by the objective. On average, the stimulus attention peaks just after the contact envelope and before the choice attention (Table 1; Appendix F), which is the order expected if sensory evidence is acquired first and the choice forms later.

Figure 3D shows a single held-out trial. All attention stays near zero before dense contact begins at about -0.6 s. Stimulus-private attention then peaks at about -0.5 s, which is biologically sensible because a strong stimulus signal should not precede sensory input. Shared attention peaks shortly afterwards, consistent with activity that is driven by the same sensory event but also participates in downstream processing. Choice-private attention ramps more gradually and peaks after response-window opening, consistent with a later commitment to the report. SLDS and rSLDS instead switch states well before the dense contacts on this trial, and their states carry no task roles, so their interpretation remains indirect. Shuffling the stimulus labels before training sharply reduces both the contact alignment and the cross-seed reproducibility of the stimulus attention, while shuffling the choice labels does not (Appendix E.3), so the stimulus timing is driven by the stimulus labels.

5.4 Does temporal order relate to behavioral error?

The intuition behind this question is that the animal should first acquire sensory evidence and then use it to form a choice. If, on some trial, the choice-related signal instead leads the stimulus-related signal, the decision may have formed before the stimulus was used normally, and such trials might be more error-prone. On each held-out trial, we compare the peak times of the stimulus and choice readout gates. We call a trial *canonical* if the stimulus peak comes first and *reversed* otherwise; ties are excluded. We then compare *behavioral* error rates, meaning that the animal's choice was inconsistent with the stimulus, between the two groups. The loss never sees behavioral correctness, so any association would be a property of the data rather than of the objective.

On individual trials, the ordering is variable. With all models scored on the same held-out trials (Figure 3F), reversed trials are more error-prone than canonical trials for TSA, whereas none of the refit baselines shows this pattern. However, with only 32 held-out trials, few of which are reversed, this difference is not statistically established. We therefore treat the ordering–error association as a descriptive trend in one session, not as an established effect.

Table 2: Controls at matched latent size (20 seeds each). Stim-first: fraction of held-out trials whose stimulus readout peaks first. Order κ: chance-corrected cross-seed agreement on each trial's ordering. NLL: held-out Poisson NLL.

Model	Stim. BA	Choice BA	$\rho_{\mathrm{stim,contact}}$	Stim-first	Order κ	NLL
TSA (full)	**0.805**	**0.937**	**0.685**	0.567	0.355	0.346
No separation	0.757	0.934	0.597	0.522	0.509	0.258
Monolithic	0.739	0.931	0.567	0.534	0.460	0.204
No private	0.725	0.928	0.594	0.486	0.459	0.242
No shared	0.789	0.913	0.634	0.645	0.473	0.376

5.5 What do the partition and the separation terms contribute?

A natural concern is that the separation penalties could create the timing we report, for example by pushing early, correlated information into one block. A related question is what the partition adds beyond a single latent. To answer both, we trained four controls at the same latent size as the full model (Table 2): the full model without separation penalties, a single undivided (monolithic) latent, a model without private blocks, and a model without a shared block.

Three conclusions follow. First, the timing structure is not created by the separation penalties: without them, the stimulus readout still aligns with contact, and the fraction of stimulus-first trials is essentially unchanged. Second, the partition and the separation terms do buy something concrete. At matched capacity, the full model decodes stimulus better than the monolithic latent, and its stimulus readout is the most closely aligned with contact. Most importantly, only the partitioned model can express *which* subspace is read out when. Third, the partition does not make single-trial timing more reproducible: the controls agree across seeds more often than the full model does. We therefore present the partition as what makes the timing readout task-attributed and stimulus-sensitive, not as what makes it more stable. Ablations of individual loss terms, the latent layout, and the scorer are in Appendices E.2, C.1, and A.1.

6 Discussion and Limitations

We introduced TSA to study how task information is divided into task-private and shared components and when each becomes prominent within a trial. On synthetic data, TSA recovered the intended ordering far more often than controls without the partition. On barrel-cortex data, it gave the strongest held-out stimulus decoding among the methods we compared, with all models scored on the same held-out trials, and its stimulus readout closely followed whisker contact even though contact is never used in training. Across two sessions from two animals, the decomposition remained close to orthogonal and combining a private block with the shared block improved decoding.

Scope and identifiability. Like dPCA and SLDS, TSA is a descriptive method. Its blocks are task-conditioned summaries of population activity, not uniquely identified biological entities, and attention is a learned readout weighting rather than a causal explanation. We therefore interpret TSA timing only insofar as it is stable across seeds and relates to variables not used in training. The partition does not make single-trial ordering more reproducible, and the association between single-trial ordering and behavioral error remains a descriptive trend.

Data scale and generality. Our real-data results come from two sessions from two animals with 14 to 17 neurons each. The label-selectivity of the private blocks and the sharpness of the contact alignment differ between the two recordings. Replication in larger populations and in a different stimulus–choice paradigm is the most important next step, and larger datasets would also be needed to test whether single-trial ordering relates to behavior. The current model handles two binary variables in one region; multi-variable and multi-region extensions are discussed in Appendix G. Within these limits, TSA offers a task-attributed, single-trial timing signal for studying when sensory and decision information is most prominent in a population.

AI USE STATEMENT

We used generative AI tools for editing (grammar, spelling, and word choice) and for reformatting and condensing the manuscript for this venue, including reorganizing previously obtained results into the main text and appendix. We did not use generative AI tools to generate research ideas, design experiments, write analysis code, or produce experimental results. All AI-assisted text was reviewed and verified against our experiments by the authors, and all references were checked manually. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work analyzes a previously published, publicly available dataset of mouse barrel-cortex recordings (Rodgers, 2022); all animal procedures were performed and approved under the original study. No new animal experiments were conducted. TSA is a descriptive analysis tool, and we caution against interpreting its latent blocks or attention weights as causal biological mechanisms without experimental manipulation.

REPRODUCIBILITY STATEMENT

The model and losses are specified in Section 3, and the architecture, training schedule, hyperparameters, parameter count, and training time are given in Appendix A. Synthetic benchmark construction is described in Appendix B; real-data session identifiers, trial filtering, and splits in Appendix C; post hoc baseline timing constructions in Appendix F.1; and the statistical procedures in Appendix E.3. All TSA results were produced from a single code base, which will be released upon publication.

REFERENCES

Mikio C. Aoi, Valerio Mante, and Jonathan W. Pillow. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience*, 23(11):1410–1420, 2020.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh B. Nachimuthu, Michael J. Mendelson, Blake Aaron Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=sw2Y0sirtM.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.0473.

Edoardo Balzani, Jean-Paul Noël, Pedro Herrero-Vidal, Dora E. Angelaki, and Cristina Savin. A probabilistic framework for task-aligned intra- and inter-area neural manifold estimation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kt-dcBQcSA.

Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. *Nature*, 487: 51–56, 2012. doi: 10.1038/nature11129. URL https://www.nature.com/articles/nature11129.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17:1500–1509, 2014. doi: 10.1038/nn.3776. URL https://www.nature.com/articles/nn.3776.

Gamaleldin F. Elsayed, Antonio H. Lara, Matthew T. Kaufman, Mark M. Churchland, and John P. Cunningham. Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications*, 7:13239, 2016. doi: 10.1038/ncomms13239. URL https://www.nature.com/articles/ncomms13239.

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4): 1569–1585, 2011. doi: 10.1109/TSP.2010.2102756.

Stefano Fusi, Earl K. Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, 2016. doi: 10.1016/j.conb.2016.01.010. URL https://doi.org/10.1016/j.conb.2016.01.010.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Evren Gokcen, Anna I. Jasper, João D. Semedo, Amin Zandvakili, Adam Kohn, Christian K. Machens, and Byron M. Yu. Disentangling the flow of signals between populations of neurons. *Nature Computational Science*, 2(8):512–525, 2022. doi: 10.1038/s43588-022-00282-5. URL https://www.nature.com/articles/s43588-022-00282-5.

Evren Gokcen, Anna Jasper, Alison Xu, Adam Kohn, Christian K. Machens, and Byron M. Yu. Uncovering motifs of concurrent signaling across multiple neuronal populations. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rabia Gondur, Usama Bin Sikandar, Evan Schaffer, Mikio Christian Aoi, and Stephen L. Keeley. Multi-modal Gaussian process variational autoencoders for neural and behavioral data. In *The Twelfth International Conference on Learning Representations*, 2024.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pp. 63–77, 2005.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Cole Hurwitz, Akash Srivastava, Kai Xu, Justin Jude, Matthew G. Perich, Lee E. Miller, and Matthias H. Hennig. Targeted neural dynamical modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5cfbc876972bd0d031c8abc37344c28-Abstract.html.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Matthew T. Kaufman, Mark M. Churchland, Stephen I. Ryu, and Krishna V. Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, 17(3):440–448, 2014. doi: 10.1038/nn.3643. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3955357/.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658, 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E. Feierstein, Adam Kepecs, Zachary F. Mainen, Ranulfo Romo, Xue-Lian Qi, Naoshige Uchida, and Christian K. Machens. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, 2016. doi: 10.7554/eLife.10989. URL https://elifesciences.org/articles/10989.

Trung Le and Eli Shlizerman. STNDT: Modeling neural population activity with spatiotemporal transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/72163d1c3c1726f1c29157d06e9e93c1-Abstract-Conference.html.

Chengrui Li, Yunmiao Wang, Yule Wang, Weihan Li, Dieter Jaeger, and Anqi Wu. A factorized low-rank RNN framework for uncovering independent neural latent dynamics and connectivity. In *Forty-third International Conference on Machine Learning*, 2026.

Ran Liu, Mehdi Azabou, Max Dabagia, Jingyun Xiao, and Eva L. Dyer. Seeing the forest and the tree: Building representations of both individual and collective dynamics with transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2377–2391, 2022.

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

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166/.

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. doi: 10.1038/nature12742. URL https://www.nature.com/articles/nature12742.

Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26(2):239–250, 2023. doi: 10.1038/s41593-022-01237-9. URL https://www.nature.com/articles/s41593-022-01237-9. Original preprint: bioRxiv (2021).

Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15:805–815, 2018. doi: 10.1038/s41592-018-0109-9. URL https://www.nature.com/articles/s41592-018-0109-9.

Felix Pei, Joel Ye, David M. Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark M. Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan W. Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 34, 2021. URL https://arxiv.org/abs/2109.04463.

Arthur Pellegrino, Heike Stein, and N. Alex Cayco-Gajic. Dimensionality reduction beyond neural subspaces with slice tensor component analysis. *Nature Neuroscience*, 27:1199–1210, 2024. doi: 10.1038/s41593-024-01626-2. URL https://www.nature.com/articles/s41593-024-01626-2.

Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497:585–590, 2013. doi: 10.1038/nature12160. URL https://www.nature.com/articles/nature12160.

Chris C. Rodgers. A detailed behavioral, videographic, and neural dataset on object recognition in mice. *Scientific Data*, 9:620, 2022. doi: 10.1038/s41597-022-01728-1. URL https://www.nature.com/articles/s41597-022-01728-1. Original preprint: bioRxiv (2021).

Omid G. Sani, Hamidreza Abbaspourazad, Yan T. Wong, Bijan Pesaran, and Maryam M. Shanechi. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149, 2021. doi: 10.1038/s41593-020-00733-0. URL https://www.nature.com/articles/s41593-020-00733-0.

Omid G. Sani, Bijan Pesaran, and Maryam M. Shanechi. Dissociative and prioritized modeling of behaviorally relevant neural dynamics using recurrent neural networks. *Nature Neuroscience*, 27(10):2033–2045, 2024. doi: 10.1038/s41593-024-01731-2. URL https://www.nature.com/articles/s41593-024-01731-2.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617:360–368, 2023. doi: 10.1038/s41586-023-06031-6. URL https://www.nature.com/articles/s41586-023-06031-6.

João D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259.e4, 2019. doi: 10.1016/j.neuron.2019.01.026. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6449210/.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Ram Dyuthi Sristi, Sowmya Manojna Narasimha, Jingya Huang, Alice Despatin, Simon Musall, Vikash Gilja, and Gal Mishne. Coupled transformer autoencoder for disentangling multi-region neural latent dynamics. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sina Tafazoli, Flora M. Bouchacourt, Adel Ardalan, Nikola T. Markov, Motoaki Uchimura, Marcelo G. Mattar, Nathaniel D. Daw, and Timothy J. Buschman. Building compositional tasks with shared neural subspaces. *Nature*, 650:164–172, 2026. doi: 10.1038/s41586-025-09805-2. URL https://www.nature.com/articles/s41586-025-09805-2.

Yu Takagi, Steven W. Kennerley, Jun-ichiro Hirayama, and Laurence T. Hunt. Demixed shared component analysis of neural population data from multiple brain areas. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.neurips.cc/paper/2020/hash/44ece762ae7e41e3a0b1301488907eaa-Abstract.html.

Parsa Vahidi, Omid G. Sani, and Maryam M. Shanechi. BRAID: Input-driven nonlinear dynamical modeling of neural-behavioral data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.neurips.cc/paper/7181-attention-is-all-you-need.

Yule Wang, Chengrui Li, Weihan Li, and Anqi Wu. Exploring behavior-relevant and disentangled neural dynamics with generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3d55170799265c03b37993e02b71b2cc-Abstract-Conference.html.

Alex H. Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I. Ryu, Krishna V. Shenoy, Mark Schnitzer, David Kahn, and Surya Ganguli. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98(6):1099–1115.e8, 2018. doi: 10.1016/j.neuron.2018.05.015. URL https://www.cell.com/neuron/fulltext/S0896-6273(18)30387-8.

Alex H. Williams, Ben Poole, Niru Maheswaranathan, Ashesh K. Dhawale, Tucker Fisher, Christopher D. Wilson, David H. Brann, Eric M. Trautmann, Stephen Ryu, Roman Shusterman,

Dmitry Rinberg, Bence P. Ölveczky, Krishna V. Shenoy, and Surya Ganguli. Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping. *Neuron*, 105(2):246–259.e8, 2020. doi: 10.1016/j.neuron.2019.10.020. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7336835/.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=CBBtMnlTGq.

Daiyao Yi, Hao Dong, Michael J. Higley, Anne Churchland, and Shreya Saxena. Shared-AE: Automatic identification of shared subspaces in high-dimensional neural and behavioral activity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009. doi: 10.1152/jn.90941.2008. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2712272/.

Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/510f2318f324cf07fce24c3a4b89c771-Abstract.html.

David M. Zoltowski, Jonathan W. Pillow, and Scott W. Linderman. A general recurrent state space framework for modeling neural dynamics during decision-making. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11680–11691. PMLR, 2020. URL https://proceedings.mlr.press/v119/zoltowski20a.html.

A ARCHITECTURE, TRAINING, AND ATTENTION SCORER

Architecture. Table 3 compares TSA with a non-causal variant that shares the same task partition, objective, and data pipeline. Only the network components and schedule differ: the non-causal variant uses a centered encoder, a transposed-convolution decoder (both can see future bins), and a scorer that reads latents as well as spikes.

Table 3: TSA compared with a non-causal variant of the same model.

Component	Non-causal variant	TSA (used throughout)
Encoder	2 × Conv1d, kernel 3, centered padding	2 × *causal* Conv1d, kernel 3, left padding
Decoder	2 × ConvTranspose1d, kernel 3	2 × Conv1d, kernel 1 (pointwise)
Attention scorer	3-layer per-bin MLP on $[\mathbf{x}_t, \mathbf{z}_t]$	one causal Conv1d, kernel 3, on $\mathbf{x}_t$ only
Gate / normalization	sigmoid, normalized over time	identical
Readout heads	3-layer MLP, width 64	linear
Latent layout	$23/23/24$	$23/23/24$
Schedule	warm start, then 80 epochs at 10^{-4}	one phase from random init, 500 epochs at 10^{-3}
Parameters	130,021	27,357

On the same real-data split (Table 4), TSA improves held-out stimulus decoding by 12 BA points while matching choice decoding, with about a fifth of the parameters.

Table 4: TSA versus the non-causal variant on the real-data session (same code base and split; 10 seeds, except the reproducibility and peak-time rows, which use 25 seeds).

Quantity	Non-causal variant	TSA
Held-out stimulus BA	0.674 ± 0.047	$\mathbf{0.795 \pm 0.036}$
Held-out choice BA	0.948 ± 0.037	0.935 ± 0.020
$\rho_{\mathrm{stim,contact}}$	0.559 ± 0.068	$\mathbf{0.632 \pm 0.116}$
Cross-seed single-trial readout reproducibility	0.529	**0.707**
Across-seed s.d. of mean stimulus peak time	0.142 s	**0.108 s**
Parameters	130,021	**27,357**
Training time per fit	21.4 s	21.8 s

Readout reproducibility. We train independently seeded models and compare the stimulus attention curves they assign to the same held-out trial. The reproducibility score is the mean pairwise agreement across seeds; shuffling trials between seeds gives the chance level. Normalized attention entropy (1 = uniform) measures how peaked the readout is.

Training. We train from random initialization for 500 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-2}, batch size 128). Reconstruction and attention smoothness are active throughout; decoding, auxiliary/shared losses, and separation penalties are switched on at fixed fractions of training. Loss weights: $w_{\mathrm{recon}} = 1$, $w_{\mathrm{aux}} = 5$, $w_{\mathrm{shared}} = 4$, $w_{\mathrm{probe}} = 32$, $w_{\mathrm{hsic}} = 4$ (directional $w_{\mathrm{hsic,S}\to\mathrm{choice}} = 10$, $w_{\mathrm{hsic,C}\to\mathrm{stim}} = 48$), $w_{\mathrm{role}} = 2$, $w_{\mathrm{ortho}} = 16$. The model has 27,357 parameters and takes about 22 s per fit on two CPU cores. Cross-seed readout reproducibility is the mean pairwise agreement of stimulus attention curves on the same held-out trial (chance: shuffle trials between seeds).

A.1 ATTENTION SCORER

Table 5 compares scorers on the same TSA backbone (10 seeds). Adding the latent to the scorer lowers stimulus decoding and contact alignment; a per-bin MLP on spikes alone yields nearly uniform attention. The reported scorer, a causal convolution with kernel 3 on spikes, best balances decoding, contact alignment, and reproducibility.

Table 5: Attention-scorer variants on the TSA backbone (10 seeds; the reported scorer is shown for reference from the 25-seed main run). $\mathbf{x}$: spikes; $\mathbf{z}$: latent. Repro.: cross-seed single-trial readout reproducibility; Entropy: normalized attention entropy (1 is uniform).

Scorer	Stim. BA	Choice BA	$\rho_{\text{stim,contact}}$	Repro.	Entropy	Params
Causal conv k=3, $\mathbf{x}$ (reported)	0.802	0.942	0.691	0.707	0.900	27,357
Per-bin MLP, $\mathbf{x}$	0.766	0.889	0.631	0.602	0.979	27,217
Per-bin MLP, $[\mathbf{x}, \mathbf{z}]$	0.749	0.945	0.563	0.600	0.741	27,567
Per-bin MLP, $[\mathbf{x}, \mathbf{z}]$, warm start	0.739	0.954	0.419	0.400	0.806	27,567
Per-bin MLP, $\mathbf{z}$	0.723	0.950	0.335	0.481	0.642	27,497
Causal conv, $[\mathbf{x}, \mathbf{z}]$	0.737	0.965	0.298	0.234	0.550	116,747
Conv, $\mathbf{x}$ (alternative)	0.751	0.915	0.242	0.373	0.694	43,247

B Synthetic benchmark

The synthetic data include contact-time jitter, amplitude variability, dropout, nuisance activity, baseline drift, and Poisson noise; ground-truth factors are mixed through random signed weights and a nonlinear rate. Stimulus and choice are correlated by design. We use 2048 trials, 14 neurons, 25 bins, an 80/20 split, three dataset seeds $\times$ three model seeds, and latent layout $23/23/24$. Architecture and optimizer match Appendix A.

B.1 Controls

Table 6 applies the same four controls as on real data. Decoding is near ceiling for every model; timing separates them. The full model peaks stimulus-first on 69% of trials versus 25–35% for monolithic, no-private, and no-separation controls. Latent recovery is similar with and without separation terms; the cost is reconstruction (R^2 0.45 versus 0.97).

Table 6: Synthetic controls (mean over 3 dataset seeds $\times$ 3 model seeds). Stim-first: fraction of held-out trials whose stimulus readout peaks first (ground truth: stimulus first). R^2_S, R^2_C, R^2_H: held-out latent recovery from the unweighted blocks. Recon. R^2: spike reconstruction.

Model	Stim. BA	Choice BA	Stim-first	R^2_S	R^2_C	R^2_H	Recon. R^2
TSA (full)	0.984	0.992	**0.694**	0.322	0.483	0.533	0.449
No separation terms	0.988	0.992	0.354	0.322	0.535	0.616	0.969
Monolithic latent	0.989	0.992	0.252	–	–	–	0.991
No private blocks	0.988	0.992	0.341	–	–	0.607	0.642
No shared block	0.980	0.988	0.624	0.361	0.531	–	0.577

C Real-data session and latent layout

The analyzed session is `sub-221CR_ses-20190515T160400` (Rodgers, 2022). We use 14 simultaneously recorded neurons. Trials are excluded if marked `ignore_trial`, if the stimulus is `nothing`, if the choice is not left/right, or if the choice occurs more than 0.5 s after response-window opening. After filtering, roughly 159 valid trials are split 80/20 (stimulus-stratified) into about 127 training and 32 test trials. Whisker-contact records from C1–C3 are timing references only.

C.1 Latent layout

To choose how the latent dimensions are divided among the stimulus-private, choice-private, and shared subspaces, we compared layouts that assign every latent dimension to one of the three subspaces. The total width is held at 70, and all other hyperparameters are those of the selected configuration. Table 7 reports eight seeds per layout, with all quantities computed on held-out trials.

Reproducibility is the cross-seed reproducibility of the single-trial readout, with its trial-shuffled chance level in parentheses.

Table 7: Latent layouts (total width 70, 8 seeds each). Layouts are $d_S/d_C/d_H$. NLL: held-out Poisson negative log-likelihood (lower is better). Repro.: cross-seed single-trial readout reproducibility (chance level in parentheses).

Layout	Stim. BA	Choice BA	NLL	Repro. (chance)	$\rho_{\text{stim,contact}}$	Params
23/23/24 (selected)	0.814	0.912	0.360	**0.693** (0.273)	**0.754**	27,357
12/46/12	0.818	0.895	0.317	0.550 (0.188)	0.621	28,101
2/2/66	0.813	0.906	0.266	0.510 (0.241)	0.598	24,753
18/34/18	0.809	0.887	0.309	0.599 (0.233)	0.616	27,729
34/18/18	0.807	0.918	0.373	0.495 (0.159)	0.684	27,729
12/12/46	0.806	0.922	0.305	0.624 (0.271)	0.686	25,993
30/30/10	0.800	0.905	0.356	0.460 (0.190)	0.676	28,225
10/10/50	0.798	0.940	0.291	0.518 (0.209)	0.598	25,745
18/18/34	0.795	0.945	0.325	0.493 (0.236)	0.640	26,737
6/6/58	0.793	0.950	0.278	0.424 (0.188)	0.607	25,249
35/35/0 (no shared)	0.790	0.908	0.400	0.450 (0.103)	0.725	28,800
46/12/12	0.784	0.935	0.373	0.479 (0.129)	0.703	28,101
66/2/2	0.779	0.906	0.385	0.376 (0.179)	0.611	28,721
2/66/2	0.722	0.873	0.278	0.098 (0.049)	0.323	28,721

Decoding does not select the layout. Stimulus accuracy spans roughly 0.78 to 0.82 and choice accuracy 0.87 to 0.95 across all allocations. The nominally highest stimulus accuracy ($12/46/12$) falls inside the seed spread of the selected layout, so the layout cannot be chosen from accuracy, and we do not claim otherwise.

Interpretability selects the balanced layout. The selected $23/23/24$ layout has the highest cross-seed readout reproducibility of any layout, about 2.5 times its shuffled chance level, and the highest correlation between the stimulus readout and whisker contact. Starving either private block degrades both quantities together. Removing the shared block entirely ($35/35/0$) gives the worst reconstruction of any layout and loses about a fifth of the reproducibility. The balanced allocation is therefore justified on the grounds the paper argues from, a readout that is stable across initializations and aligned to the sensory event, rather than on decoding accuracy, which is flat across layouts.

D A SECOND SESSION

To test whether the decomposition is specific to one recording, we repeated the analyses of Figure 3 on a second session from the same dataset (Rodgers, 2022), recorded from a different animal (`sub-231CR_ses-20190921T144923`). Preprocessing is identical to the main session: the same trial filters, the same window from -2.0 to $+0.5$ s in 100 ms bins, the same label map, and the same whisker-contact binning. The model, losses, and hyperparameters are unchanged, and we report ten seeds with all quantities computed on held-out trials.

The session yields 165 usable trials out of 246 (132 training and 33 held-out trials), with whisker contact on 99% of trials, a mean firing rate of 6.75 Hz, and an animal that is correct on 86% of trials, against 62% on the main session. It contains only 17 units in total. We therefore report two variants: the 14 highest-firing units, which matches the input dimension of the main session, and all 17 units. Units are selected by their firing rate within the analysis window, a criterion that never inspects stimulus or choice and so cannot inflate decodability.[1]

What transfers. The three subspaces remain close to orthogonal on the second session: every pairwise overlap is at most 0.09, against at most 0.03 on the main session. The geometry the model

[1] On both sessions we verified the behavioural labels against the `outcome` field of the NWB trials table: the convention `choice` = `stimulus` reproduces the recorded outcome on every trial of both sessions, and the `rewarded_side` field confirms the concave → left and convex → right mapping without exception.

Table 8: Main session versus a second session from a different animal (10 seeds each, held-out trials, mean ± s.d.). Off-diagonal probes are included so that the selectivity of each private block can be read directly.

	221CR (main)	231CR, 14 units	231CR, all 17
Stimulus BA	0.817 ± 0.028	0.899 ± 0.032	0.908 ± 0.033
Choice BA	0.917 ± 0.026	0.775 ± 0.041	0.833 ± 0.049
Stimulus-private → stimulus	0.727 ± 0.058	0.747 ± 0.126	0.776 ± 0.086
Choice-private → stimulus	0.642 ± 0.090	0.756 ± 0.095	0.777 ± 0.073
Choice-private → choice	0.880 ± 0.038	0.668 ± 0.057	0.716 ± 0.072
Stimulus-private → choice	0.725 ± 0.085	0.722 ± 0.043	0.793 ± 0.077
Shared → stimulus	0.725 ± 0.068	0.867 ± 0.026	0.851 ± 0.037
Shared → choice	0.894 ± 0.029	0.723 ± 0.046	0.801 ± 0.023
Stimulus-private + shared → stimulus	0.791 ± 0.021	0.870 ± 0.038	0.874 ± 0.033
Choice-private + shared → choice	0.887 ± 0.047	0.711 ± 0.046	0.787 ± 0.051
Overlap, stimulus-private / choice-private	0.015 ± 0.011	0.076 ± 0.115	0.051 ± 0.085
Overlap, stimulus-private / shared	0.025 ± 0.026	0.086 ± 0.136	0.050 ± 0.100
Overlap, choice-private / shared	0.025 ± 0.030	0.090 ± 0.153	0.052 ± 0.105
Poisson NLL	0.368 ± 0.032	0.445 ± 0.093	0.403 ± 0.052
$\rho_{\text{stim,contact}}$	0.763 ± 0.092	0.382 ± 0.106	0.376 ± 0.098

is constructed to produce is therefore not session-specific. Stimulus decoding is higher on the new session than on the main one (0.899 and 0.908 against 0.817), and reconstruction is comparable. The combination effect also reproduces, and is the clearest transfer in the table: adding the shared block to the stimulus-private block improves stimulus decoding on both sessions, from 0.747 to 0.870 here and from 0.727 to 0.791 on the main session. This is the property the shared block is meant to have, and it holds in a population where the private blocks behave quite differently. Results are also stable to where the unit cut falls, with all 17 units giving slightly better numbers than the top 14 throughout.

What does not transfer. The selectivity of the private blocks does not survive the change of recording. On the main session each private block decodes its own label better than the other, by $+0.085$ for stimulus-private and $+0.155$ for choice-private. On the second session both margins are approximately zero or slightly negative: -0.009 and -0.054 with 14 units, and -0.001 and -0.077 with all 17. The two private blocks are close to interchangeable as decoders here, even though they remain geometrically near-orthogonal. Orthogonality of the subspaces and label-selectivity of their contents are thus separate properties, and only the first is enforced by the model; the second depends on how strongly the recorded population distinguishes the two variables. The alignment between the stimulus readout and whisker contact is also weaker on this session, $\rho \approx 0.38$ against 0.763. Contact onset occurs earlier (-1.13 s against -0.79 s) and both readout ramps move earlier with it, so the readout still tracks contact well above zero, but the tight alignment reported in Table 1 is specific to the main session.

A caution on comparing the two task variables. Table 8 appears to show the stimulus/choice asymmetry reversing between sessions, with choice better decoded on the main session and stimulus better decoded here. That reading is not supported. The $80/20$ split is stratified on stimulus only, and on the second session it leaves the held-out set at 23 left against 10 right on choice while the training set is balanced at $62/70$, so a single split estimates choice accuracy from ten minority-class trials. Repeating the fit over ten independent splits with three seeds each (Table 9) shrinks the apparent gap on the second session from $+0.124$ to $+0.020$: the two variables decode about equally well there, while on the main session choice leads stimulus by 0.094. We therefore draw no conclusion about which variable a population represents more strongly, and in particular we do not read the second session as evidence about the origin of the stimulus-private block's weak selectivity on the main session.

Across both sessions the claims that hold are the near-orthogonality of the decomposition and the decoding gain from combining a private block with the shared block. The label-selectivity of the

Table 9: Decoding over ten independent stimulus-stratified splits, three seeds each (mean $\pm$ s.d. across all 30 fits). The last column gives the mean size of the smaller choice class in the held-out set, which the split does not control.

	Stimulus BA	Choice BA	Gap	Minority choice class
221CR (main)	0.791 ± 0.080	0.886 ± 0.067	-0.094	12 of 32
231CR, 14 units	0.863 ± 0.067	0.843 ± 0.086	$+0.020$	14 of 33
231CR, all 17 units	0.900 ± 0.061	0.876 ± 0.067	$+0.024$	14 of 33

private blocks, the decodability of each individual variable, and the sharpness of the contact alignment all vary from recording to recording.

E Ablations, reconstruction, and label controls

E.1 Reconstruction trade-off

Raising w_{rec} recovers reconstruction while preserving most of TSA's stimulus-decoding advantage over the non-causal variant (Table 10). Zeroing any one block with the decoder fixed worsens held-out NLL by 0.098 ± 0.026 (stimulus-private), 0.125 ± 0.060 (choice-private), and 0.176 ± 0.042 (shared), so the task blocks participate in reconstruction.

Table 10: Reconstruction-weight sweep (5 seeds), relative to the default w_{rec}.

	TSA			Non-causal variant		
w_{rec}	NLL	Recon. R^2	Stim. BA	NLL	Recon. R^2	Stim. BA
$\times 0.04$	0.461	0.651	0.796	0.289	0.873	0.693
$\times 0.2$	0.441	0.683	0.790	0.275	0.883	0.706
$\times 1$	0.339	0.797	0.776	0.248	0.896	0.661
$\times 5$	0.269	0.863	0.833	0.226	0.900	0.732
$\times 25$	0.200	0.928	0.782	0.210	0.899	0.674

E.2 Loss terms

Table 11 removes or modifies one part of the objective at a time (8 seeds; default $w_{\mathrm{ortho}} = 8$). Removing all separation terms lowers stimulus BA from 0.787 to 0.747; most of that drop comes from the cross-task probe. Without orthogonality, private-block overlap triples; doubling to 16 lowers it to 0.017 with no decoding cost. Turning all losses on from epoch one worsens reconstruction and ordering consistency, so staging mainly helps optimization.

E.3 Label shuffles

Table 12 retrains after shuffling stimulus labels, choice labels, or both (25 seeds for the main model, 12 per shuffle). Shuffling stimulus labels collapses stimulus-attention reproducibility (0.707 to 0.175) and reduces contact alignment (0.691 to 0.508); shuffling only choice leaves both largely intact. The ordering–error gap over 10 seeds is 0.051 ± 0.148 (mean permutation $p = 0.51$) and is reported only as a descriptive trend.

F Timing and post hoc baselines

Mean attention. Figure 4 shows trial-averaged TSA readout attention with the contact envelope. The stimulus readout is near chance before contact, rises afterward, peaks at response-window opening (single-bin BA 0.724), and stays elevated through $+0.4$ s. We use peak times rather than onset: a 30%-of-peak onset fires near trial start because attention never falls far below uniform.

Table 11: Loss ablations on TSA (8 seeds each, $w_{\text{ortho}} = 8$ unless varied). Overlap: stimulus-private versus choice-private overlap. Order κ: chance-corrected cross-seed agreement on each trial's ordering.

Variant	Stim. BA	Choice BA	NLL	Overlap	Order κ	$\rho_{\text{stim,cont.}}$
Full	0.787	0.939	0.343	0.052	0.309	0.657
No separation terms	0.747	0.931	0.258	0.095	0.508	0.571
No cross-task probe	0.754	0.934	0.260	0.036	0.447	0.606
No HSIC	0.799	0.948	0.334	0.066	0.356	0.619
No role penalty	0.787	0.935	0.347	0.049	0.309	0.656
No auxiliary losses	0.781	0.937	0.321	0.027	0.365	0.616
No shared losses	0.807	0.917	0.355	0.038	0.387	0.625
No attention smoothness	0.771	0.935	0.352	0.040	0.146	0.353
Probe via gradient reversal	0.781	0.944	0.321	0.236	0.407	0.568
GRL adversary added	0.802	0.925	0.337	0.078	0.419	0.638
Reconstruction only	0.501	0.503	0.180	0.099	-0.061	0.436
$w_{\text{ortho}} \times 0$	0.800	0.921	0.351	0.155	0.386	0.627
$w_{\text{ortho}} \times 0.5$	0.797	0.940	0.360	0.056	0.308	0.668
$w_{\text{ortho}} \times 2$ (reported, 16)	0.801	0.937	0.356	0.017	0.406	0.667
$w_{\text{ortho}} \times 4$	0.793	0.940	0.322	0.035	0.363	0.607
No staging (all losses on)	0.781	0.924	0.424	0.017	0.118	0.554
Cold-start variant	0.772	0.945	0.376	0.044	0.354	0.501

Table 12: Label-shuffled controls for TSA. Repro. (chance): cross-seed single-trial reproducibility of the stimulus attention, with its trial-shuffled chance level in parentheses.

Training labels	Stim. BA	Choice BA	Stim-first	Repro. (chance)	$\rho_{\text{stim,contact}}$
True labels	0.802	0.942	0.574 ± 0.120	0.707 (0.246)	0.691
Stimulus shuffled	0.586	0.924	0.607 ± 0.144	0.175 (0.110)	0.508
Choice shuffled	0.756	0.564	0.323 ± 0.132	0.689 (0.197)	0.559
Both shuffled	0.574	0.572	0.409 ± 0.128	0.128 (0.067)	0.480

F.1 Post hoc baseline timing

For baselines without task-specific readout weights, we fit per-bin decoders from $z_i(t)$ to the trial labels on the training set. Held-out balanced-accuracy curves yield contact alignment with the mean contact envelope; for TSA we use mean stimulus readout attention. Single-trial ordering uses peaks of positive-class probabilities (TSA: peaks of α_i^S and α_i^C). Whole-dataset peaks in Table 1 use $\arg\max_t s_t$. A raw-spike per-bin decoder already shows a stimulus-before-choice offset (peaks at -0.20 s and $+0.10$ s), so the ordering is present in the data independently of TSA.

G Extensions to multiple task variables and multiple regions

More than two task variables. The current implementation handles binary stimulus and choice. The likelihoods extend directly to multiclass cross-entropy, ordinal, continuous, or count objectives. The main design issue is the partition: allocating a block to every subset of K variables requires $2^K - 1$ blocks and does not scale. Practical alternatives are K variable-private blocks plus one global shared block, or a learned sparse variable-to-block membership matrix that retains only supported pairwise or higher-order shared blocks. Each would need new constraints and evaluation.

Multiple regions. A natural multi-region extension defines, for each region, private blocks, a block shared across task variables within the region, and blocks shared across regions, connecting TSA to multi-region latent models (Gokcen et al., 2022; 2023; Sristi et al., 2026; Li et al., 2026). This would let one ask which task variables are represented locally and which are coordinated across regions, and when. More blocks imply more constraints, and the staged training would need careful validation to

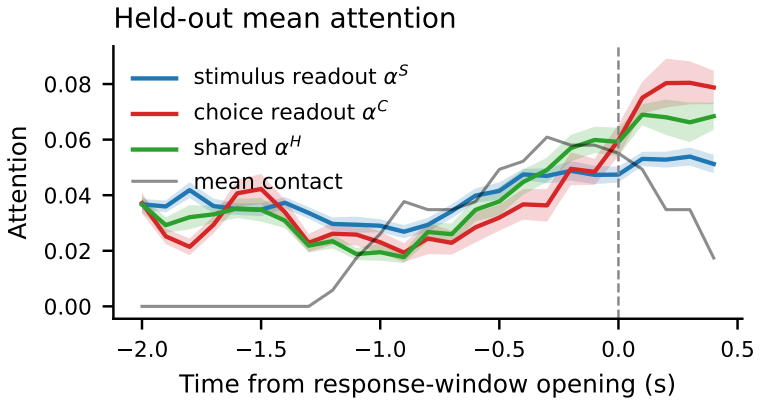

Figure 4: Trial-averaged TSA readout attention ($\pm$ s.e.m. across held-out trials) with the mean whisker-contact envelope, aligned to response-window opening at time 0.

avoid unstable optimization or arbitrary allocations across regions. We therefore view this as future work rather than a capability of the current model.