COUPLED USAGE–SENSE PROCESSES: TEMPORAL AND ATTRIBUTABLE LEXICAL SEMANTIC CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Lexical semantic change is usually summarized by a scalar distance between independently sampled period distributions. This measures how much a word changed, but does not reveal when it changed, which mechanisms and component movements carried the change, or which usages support the attribution. We introduce Coupled Usage–Sense Processes (CUSP), which derives these answers from a single marginal preserving temporal process. A hierarchical coupling relates contextual distributions through latent usage components, while Markov composition makes adjacent and longer span correspondences compatible. Displacement operators quantify change magnitude and timing, split variation exactly between movement of component centers and reorganization within components, and attribute it to transported component pairs. Word-local modes resolve distinct directions of change and their activity over time, while representative passages from attributed components ground the analysis in text. Under a Gaussian mixture specialization, we prove parametric recovery of the operators and squared distances. Synthetic experiments support the predicted rate. CUSP remains competitive on English and German DWUG and recovers controlled Janus profiles while maintaining compositionally coherent transport. A large corpus of US court opinions demonstrates transition, mode, and passage attribution in unlabeled natural text. CUSP thus makes magnitude, timing, mechanism, movement, modes, and textual evidence compatible views of one lexical history.

1 INTRODUCTION

Computational lexical semantic change (LSC) studies how a word's contextual usages evolve. Standard evaluations condense this history to a scalar distance between periods. A score cannot reveal when change occurred, whether it reflects movement between usage components or reorganization within them, which transitions carried it, or which usages support the attribution. A multi-period analysis should answer these questions as compatible views of one lexical history.

Prior work addresses individual parts of this problem. Aligned and dynamic embeddings track changes in a word's distributional representation across periods (Hamilton et al., 2016; Bamler & Mandt, 2017). Contextual approaches compare individual usages directly or cluster them into usage types, then measure change with distances, divergences, or optimal transport (Giulianelli et al., 2020; Periti & Tahmasebi, 2024; Montariol & Allauzen, 2021). Sense based methods track prevalence, cluster continuity, and contributing senses (Frermann & Lapata, 2016; Hu et al., 2019; Periti et al., 2025; Kolli et al., 2026; Kokosinskii et al., 2026), while diachronic similarity matrices characterize temporal patterns (Kiyama et al., 2025). Unbalanced transport identifies gains and losses in usage mass (Kishino et al., 2025), and embedding axes aid interpretation (Aida & Bollegala, 2025a;b). Chen et al. (2026) identify interpretability as a major unresolved challenge, noting that current models often cannot explain how or why meanings change and explicitly interpretable approaches remain at an early stage. Prior methods already track senses and attribute change (Table 3). To our knowledge, no existing LSC method builds one process that preserves period marginals and makes correspondence consistent across time, exactly decomposes its change score by mechanism and transported component pair, and links its word-local modes and attributed movements to representative passages.

Diachronic corpora provide independent samples from each period's usage distribution, not correspondences between periods. A directly optimized $t \to t+2$ transport plan can disagree with the

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.

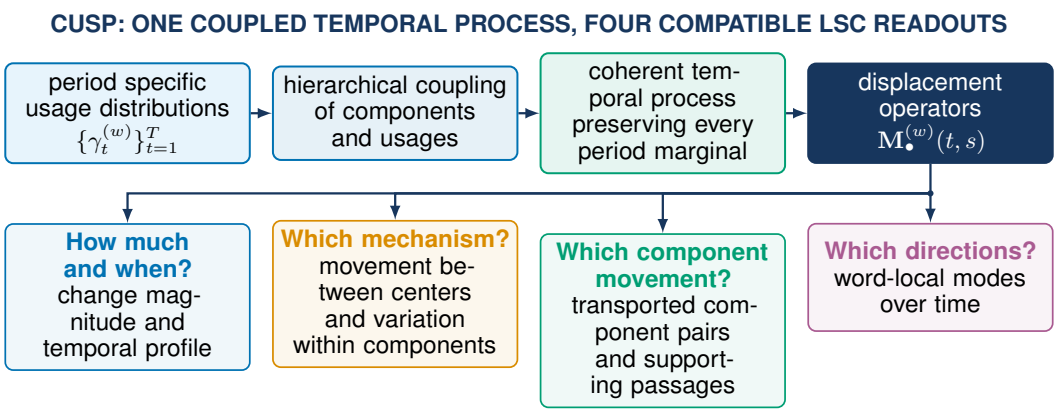

Figure 1: CUSP turns independent period distributions into one marginal preserving temporal process. Its displacement operators yield compatible LSC readouts of magnitude and timing, mechanism, component and passage attribution, and word-local temporal modes.

plan composed through $t{+}1$. Scores and transition explanations built from those separate plans can therefore describe different histories. What is needed is one process that preserves each period marginal and makes all readouts refer to the same temporal correspondence.

We introduce Coupled Usage–Sense Processes (CUSP) to construct exactly this object (Figure 1). CUSP makes three technical contributions. (i) A hierarchical coupling matches contextual distributions within component pairs, transports their prevalence mass with the prescribed marginals, and composes adjacent plans into coherent longer span correspondence. (ii) Displacement operators quantify magnitude and timing and decompose change exactly by center movement, within component variation, and transported component pair. The summed mean operator yields word-local modes, while representative passages make the attributed movements inspectable. (iii) Building on the second-moment geometry of MENT (Ezoe & Hisano, 2026), we prove marginal preservation, temporal compatibility, exact decompositions, and parametric recovery of the operators and squared distances for the trace and retained modes under identification and optimization stability conditions. These are compatible readouts of one fitted temporal process.

Empirically, we test each level of this construction in increasing proximity to natural LSC. Synthetic data support the predicted recovery rate. DWUG establishes that the operator trace retains standard scalar validity while its exact decomposition distinguishes mechanisms. Janus tests temporal profile recovery and compositional coherence across several periods. A large corpus of U.S. court opinions shows that the complete construction connects an unsupervised temporal signal to component movements, word-local modes, and directly inspectable passages. Together, these settings show that CUSP preserves the conventional answer to *how much* while adding compatible answers to *when*, *through which movement*, and *with what textual evidence*.

2 Coupled Usage–Sense Processes

A fixed encoder places independently sampled usages from each period in a common embedding space, but shared coordinates do not determine temporal correspondence. CUSP represents each period distribution as a mixture of latent usage components and couples their prevalences and contextual distributions into a joint process whose period marginals reproduce the observed distributions.

2.1 Usage distributions in each period

Let V be the vocabulary and $[T] = \{1, \dots, T\}$ the periods. For word $w \in V$ and period $t \in [T]$, let

$$X_{t,i}^{(w)} \overset{\text{i.i.d.}}{\sim} \gamma_t^{(w)}, \qquad \gamma_t^{(w)} = \sum_{k=1}^{K_t^{(w)}} \pi_{t,k}^{(w)} \nu_{t,k}^{(w)}, \qquad \nu_{t,k}^{(w)} \in \mathcal{P}_2(\mathbb{R}^d).$$

Here, $X_{t,i}^{(w)}$ is a contextual usage embedding, $K_t^{(w)}$ counts components, $\pi_{t,k}^{(w)}$ are positive prevalence weights summing to one, and $\nu_{t,k}^{(w)}$ is the contextual distribution of component k. Components are period local, so equal indices across periods do not imply correspondence. Mixtures are fitted independently by period, and their marginals do not identify temporal correspondence. After coupling, prevalence changes describe relative mass redistribution, while component changes describe movement or reshaping.

2.2 Hierarchical coupling and Markov composition

We use optimal transport hierarchically: contextual distributions are coupled within each candidate component pair, and their induced costs define transport between component prevalences, following the general structure of mixture and hierarchical OT (Chen et al., 2019; Delon & Desolneux, 2020; Yurochkin et al., 2019). For each adjacent pair $(t, t+1)$ and possible component transition $k \to \ell$, CUSP first selects a contextual coupling

$$\eta_{t,k,\ell}^{(w)} = \arg\min_{\eta \in \mathcal{A}_{t,k,\ell}^{(w)}} \int c(x,y)\, d\eta(x,y), \qquad \mathcal{A}_{t,k,\ell}^{(w)} \subseteq \Pi\Big(\nu_{t,k}^{(w)}, \nu_{t+1,\ell}^{(w)}\Big), \tag{1}$$

where $\mathcal{A}_{t,k,\ell}^{(w)}$ is the admissible class of contextual couplings for the component pair and equals $\Pi(\nu_{t,k}^{(w)}, \nu_{t+1,\ell}^{(w)})$ in the Gaussian specialization below. We denote the minimum cost by $C_{t,k,\ell}^{(w)}$. CUSP then transports the component prevalences using

$$\begin{aligned} Q_t^{(w)} = \arg\min_{Q} \quad & \sum_{k,\ell} Q(k,\ell) C_{t,k,\ell}^{(w)} \\ \text{s.t.} \quad & \sum_{\ell} Q(k,\ell) = \pi_{t,k}^{(w)}, \qquad \sum_{k} Q(k,\ell) = \pi_{t+1,\ell}^{(w)}, \qquad Q(k,\ell) \geq 0. \end{aligned} \tag{2}$$

The two transport levels have distinct roles. $C_{t,k,\ell}^{(w)}$ measures the contextual cost of a candidate component transition, while $Q_t^{(w)}$ allocates prevalence mass among those candidates subject to the two observed period marginals.

Assumption 2.1 (Uniqueness of the coupling construction)**.** For every adjacent period pair, each problem in equation 1 has a unique finite-cost minimizer, and equation 2 has a unique minimizer.

The adjacent plans share the prescribed intermediate marginals and can therefore be composed into a joint coupling (Peyré & Cuturi, 2019). CUSP uses Markov composition to induce non-adjacent correspondences from adjacent plans. Specifically, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which the latent component process $Z_{1:T}^{(w)}$ and the contextual state process $X_{1:T}^{(w)}$ are defined. The adjacent plans define a Markov chain over usage components local to each period:

$$\Pr(Z_1^{(w)} = k) = \pi_{1,k}^{(w)}, \qquad P_t^{(w)}(k,\ell) := \Pr(Z_{t+1}^{(w)} = \ell \mid Z_t^{(w)} = k) = \frac{Q_t^{(w)}(k,\ell)}{\pi_{t,k}^{(w)}}.$$

Disintegrating $\eta_{t,k,\ell}^{(w)}(dx, dy) = \nu_{t,k}^{(w)}(dx) \mathcal{K}_{t,k,\ell}^{(w)}(x, dy)$ defines the corresponding contextual process conditional on the entire latent component sequence $Z_{1:T}^{(w)}$:

$$X_1^{(w)} \mid Z_{1:T}^{(w)} \sim \nu_{1,Z_1^{(w)}}^{(w)}, \qquad X_{t+1}^{(w)} \mid (X_t^{(w)}, Z_{1:T}^{(w)}) \sim \mathcal{K}_{t,Z_t^{(w)},Z_{t+1}^{(w)}}^{(w)}(X_t^{(w)}, \cdot).$$

The construction preserves both levels of the observed mixture representation.

Proposition 2.2 (Marginal preservation and temporal compatibility)**.** *For every $w \in V$, $t \in [T]$, and $k \in [K_t^{(w)}]$,*

$$\Pr(Z_t^{(w)} = k) = \pi_{t,k}^{(w)}, \qquad \mathcal{L}(X_t^{(w)} \mid Z_{1:T}^{(w)}) = \nu_{t,Z_t^{(w)}}^{(w)}.$$

Consequently,

$$\mathcal{L}(X_1^{(w)}, \ldots, X_T^{(w)}) \in \Pi(\gamma_1^{(w)}, \ldots, \gamma_T^{(w)}). \tag{3}$$

The correspondence from t to s is therefore induced through the intervening adjacent plans, rather than being optimized independently. All period pairs belong to one process with the observed marginals, and there is no assertion of trajectories for individual utterances.

2.3 FULL, MEAN, AND DEVIATION PROCESSES

Let $m_{t,k}^{(w)} = \int x \, d\nu_{t,k}^{(w)}(x)$. The coupled contextual state can be separated into a component center process and a centered deviation process within components:

$$\phi_{\text{full}}^{(w)}(t) = X_t^{(w)}, \qquad \phi_{\text{mean}}^{(w)}(t) = m_{t,Z_t^{(w)}}^{(w)}, \qquad \phi_{\text{dev}}^{(w)}(t) = X_t^{(w)} - m_{t,Z_t^{(w)}}^{(w)}.$$

These processes satisfy the following reconstruction and centering properties:

$$\phi_{\text{full}}^{(w)}(t) = \phi_{\text{mean}}^{(w)}(t) + \phi_{\text{dev}}^{(w)}(t), \qquad \mathbb{E}[\phi_{\text{dev}}^{(w)}(t) \mid Z_{1:T}^{(w)}] = 0. \tag{4}$$

A component may therefore move through the embedding space, or its usages may broaden, contract, or rotate around its center. The full process retains both mechanisms.

3 QUANTIFYING AND ATTRIBUTING SEMANTIC CHANGE

The coupled process provides compatible answers to how much a word has changed, when it changed, and which mechanisms and component movements carried the change. Second-moment operators retain directional information, their traces yield scalar change scores and temporal profiles, and their exact decompositions attribute change to movement between components, variation within components, and transported component pairs.

3.1 DISPLACEMENT OPERATORS AND TEMPORAL PROFILES

For $\bullet \in \{\text{full}, \text{mean}, \text{dev}\}$, define

$$\Delta\phi_{\bullet}^{(w)}(t,s) := \phi_{\bullet}^{(w)}(t) - \phi_{\bullet}^{(w)}(s), \qquad \mathbf{M}_{\bullet}^{(w)}(t,s) := \mathbb{E}[\Delta\phi_{\bullet}^{(w)}(t,s)\Delta\phi_{\bullet}^{(w)}(t,s)^{\top}].$$

The trace distance

$$d_{\bullet,\text{tr}}^{(w)}(t,s) := \{\text{tr}\, \mathbf{M}_{\bullet}^{(w)}(t,s)\}^{1/2}$$

measures the total coupled displacement between two periods. Evaluating this distance on adjacent pairs yields a temporal change profile. The trace reflects how much change occurred, while the operator retains the embedding directions carrying that displacement. Adjacent evaluations locate the change in time, and non-adjacent evaluations inherit the same composed temporal correspondence. For any fixed word-local orthonormal basis $\{\boldsymbol{u}_r^{(w)}\}_{r=1}^d$, define

$$d_{\bullet,r}^{(w)}(t,s) := \{(\boldsymbol{u}_r^{(w)})^{\top}\mathbf{M}_{\bullet}^{(w)}(t,s)\boldsymbol{u}_r^{(w)}\}^{1/2}.$$

The trace summarizes the total semantic variation, while the mode-wise quantities provide its orthogonal decomposition:

$$\left(d_{\bullet,\text{tr}}^{(w)}(t,s)\right)^2 = \sum_{r=1}^{d}\left(d_{\bullet,r}^{(w)}(t,s)\right)^2. \tag{5}$$

Remark. Because $\nu_{t,k}^{(w)} \in \mathcal{P}_2(\mathbb{R}^d)$, all $\phi_{\bullet}^{(w)}(t)$ belong to $L^2(\Omega;\mathbb{R}^d)$. Hence $\mathbf{M}_{\bullet}^{(w)}(t,s)$ and the resulting distances are finite. Moreover, the trace and mode-wise semantic distances define finite pseudometrics on $[T]$, with the zero-distance conditions given by $\phi_{\bullet}^{(w)}(t) = \phi_{\bullet}^{(w)}(s)$ almost surely, and $(\boldsymbol{u}_r^{(w)})^{\top}\phi_{\bullet}^{(w)}(t) = (\boldsymbol{u}_r^{(w)})^{\top}\phi_{\bullet}^{(w)}(s)$ almost surely, respectively. These properties are the same as those of the corresponding distances defined in Ezoe & Hisano (2026). Appendix E.1 gives the details.

3.2 EXACT MECHANISM AND COMPONENT TRANSITION ATTRIBUTION

Conditional centering removes the mean–deviation cross term.

Theorem 3.1 (Exact mean–deviation decomposition)**.** *For every $w \in V$ and $t, s \in [T]$,*

$$\mathbf{M}^{(w)}_{\mathrm{full}}(t,s) = \mathbf{M}^{(w)}_{\mathrm{mean}}(t,s) + \mathbf{M}^{(w)}_{\mathrm{dev}}(t,s). \tag{6}$$

Consequently, for $\rho \in \{\mathrm{tr}, 1, \ldots, d\}$,

$$\left(d^{(w)}_{\mathrm{full},\rho}(t,s)\right)^2 = \left(d^{(w)}_{\mathrm{mean},\rho}(t,s)\right)^2 + \left(d^{(w)}_{\mathrm{dev},\rho}(t,s)\right)^2. \tag{7}$$

Further conditioning on $(Z^{(w)}_t, Z^{(w)}_s)$ attributes each operator to transported component pairs. Define

$$\begin{aligned}
\mathbf{C}^{(w)}_{\mathrm{mean}}(t,s;k,\ell) &:= (m^{(w)}_{t,k} - m^{(w)}_{s,\ell})(m^{(w)}_{t,k} - m^{(w)}_{s,\ell})^\top, \\
\mathbf{C}^{(w)}_{\mathrm{dev}}(t,s;k,\ell) &:= \mathrm{Cov}(X^{(w)}_t - X^{(w)}_s \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell), \\
q^{(w)}(t,s;k,\ell) &:= \Pr\left(Z^{(w)}_t = k,\ Z^{(w)}_s = \ell\right).
\end{aligned}$$

The component-pair covariance $\mathbf{C}^{(w)}_{\mathrm{dev}}(t,s;k,\ell)$ is defined only when $q^{(w)}(t,s;k,\ell) > 0$. Zero-mass pairs contribute nothing and are omitted from the weighted sums.

Theorem 3.2 (Exact component transition attribution)**.** *For $\bullet \in \{\mathrm{mean}, \mathrm{dev}\}$,*

$$\mathbf{M}^{(w)}_{\bullet}(t,s) = \sum_{k=1}^{K^{(w)}_t} \sum_{\ell=1}^{K^{(w)}_s} q^{(w)}(t,s;k,\ell)\mathbf{C}^{(w)}_{\bullet}(t,s;k,\ell). \tag{8}$$

Consequently,

$$\left(d^{(w)}_{\bullet,\mathrm{tr}}(t,s)\right)^2 = \sum_{k,\ell} q^{(w)}(t,s;k,\ell)\,\mathrm{tr}\,\mathbf{C}^{(w)}_{\bullet}(t,s;k,\ell), \tag{9}$$

$$\left(d^{(w)}_{\bullet,r}(t,s)\right)^2 = \sum_{k,\ell} q^{(w)}(t,s;k,\ell)(\boldsymbol{u}^{(w)}_r)^\top \mathbf{C}^{(w)}_{\bullet}(t,s;k,\ell)\boldsymbol{u}^{(w)}_r. \tag{10}$$

Together, Theorems 3.1 and 3.2 form an exact accounting of change. Every unit of the squared trace or mode-wise variation is assigned to between center or within component change and to transported component pairs. This attribution decomposes the score itself rather than fitting a separate explanatory model. In the empirical analysis, passages assigned to the selected source and target components make these numerical attributions inspectable.

3.3 Word-local temporal modes

To adapt the basis to the history of word w, we use the aggregated mean operator

$$\mathcal{M}^{(w)}_{\mathrm{mean}} := \sum_{t=1}^{T-1} \mathbf{M}^{(w)}_{\mathrm{mean}}(t, t+1).$$

We take the orthonormal eigenvectors of $\mathcal{M}^{(w)}_{\mathrm{mean}}$ as the word specific basis $\{\boldsymbol{u}^{(w)}_r\}_{r=1}^d$, with eigenvalues ordered decreasingly:

$$\mathcal{M}^{(w)}_{\mathrm{mean}}\boldsymbol{u}^{(w)}_r = \lambda^{(w)}_r \boldsymbol{u}^{(w)}_r.$$

Let $R^{(w)} \in [d]$ denote the number of modes retained for individual analysis. We require their eigenvalues to be non-degenerate[1].

Assumption 3.3 (Non-degenerate retained modes)**.** There exists a constant $\delta > 0$ such that, for every $r \in [R^{(w)}]$,

$$\lambda^{(w)}_r - \lambda^{(w)}_{r+1} \geq \delta,$$

where we define $\lambda^{(w)}_{d+1} = -\infty$.

[1]For $r > R^{(w)}$, the eigenvectors need not be uniquely determined in the presence of eigenvalue degeneracy, but this freedom is immaterial because these modes are not used for individual analysis.

The eigenpairs admit the following spectral characterization.

Proposition 3.4 (Spectral characterization of word-local modes)**.** *For every $r \in [d]$,*

$$\lambda_r^{(w)} = \sum_{t=1}^{T-1} \left(d_{\mathrm{mean},r}^{(w)}(t, t+1) \right)^2, \tag{11}$$

and

$$\boldsymbol{u}_r^{(w)} \in \arg \max_{\substack{\|\boldsymbol{u}\|=1 \\ \boldsymbol{u} \perp \boldsymbol{u}_1^{(w)}, \ldots, \boldsymbol{u}_{r-1}^{(w)}}} \sum_{t=1}^{T-1} \boldsymbol{u}^\top \mathbf{M}_{\mathrm{mean}}^{(w)}(t, t+1) \boldsymbol{u}. \tag{12}$$

The path sum identifies the directions that organize the word's complete history, while the adjacent distances locate each direction in time. Because the basis is word local, its modes describe distinct developments within a word rather than vocabulary wide axes. We use the mean basis to isolate center movement while keeping the full, mean, and deviation contributions exactly comparable.

3.4 Gaussian implementation and statistical recovery

For a tractable estimator, we specialize to Gaussian usage components, quadratic cost, and unrestricted component couplings using the established Gaussian and Gaussian-mixture Wasserstein geometry (Chen et al., 2019; Delon & Desolneux, 2020):

$$\nu_{t,k}^{(w)} = \mathcal{N}(\mu_{t,k}^{(w)}, \Sigma_{t,k}^{(w)}), \qquad \mathcal{A}_{t,k,\ell}^{(w)} = \Pi(\nu_{t,k}^{(w)}, \nu_{t+1,\ell}^{(w)}), \qquad c(x,y) = \|x - y\|^2.$$

For completeness, the resulting component-level optimal coupling has the standard Gaussian form.

Proposition 3.5 (Gaussian optimal coupling)**.** *If $\Sigma_{t,k}^{(w)} \succ 0$, the optimizer in equation 1 is unique and equals*

$$\eta_{t,k,\ell}^{(w)} = (\mathrm{id}, T_{t,k,\ell}^{(w)})_{\#} \nu_{t,k}^{(w)}, \qquad T_{t,k,\ell}^{(w)}(x) = \mu_{t+1,\ell}^{(w)} + A_{t,k,\ell}^{(w)}(x - \mu_{t,k}^{(w)}), \tag{13}$$

where

$$A_{t,k,\ell}^{(w)} = \left(\Sigma_{t,k}^{(w)}\right)^{-1/2} \left[\left(\Sigma_{t,k}^{(w)}\right)^{1/2} \Sigma_{t+1,\ell}^{(w)} \left(\Sigma_{t,k}^{(w)}\right)^{1/2} \right]^{1/2} \left(\Sigma_{t,k}^{(w)}\right)^{-1/2}.$$

Under this specialization, the component transition terms and induced distances admit closed form expressions and can be computed by dynamic programming for every period pair (t, s). Details are deferred to Appendices F and G.

Finally, we have the statistical recovery results. Let hats denote estimated quantities.

Assumption 3.6 (Estimation and optimization conditions)**.** The Gaussian mixtures have known, correctly specified component counts, positive weights, positive definite covariances and pairwise distinct component parameters (see Definition B.2). Their parameters are estimated by a penalized maximum likelihood estimator satisfying the conditions of Chen & Tan (2009). Each adjacent optimal component plan is unique and has support size $K_t^{(w)} + K_{t+1}^{(w)} - 1$.

Theorem 3.7 (Statistical recovery)**.** *Under the Gaussian specialization and Assumptions 3.3 and 3.6, the following holds using the word-local basis induced by $\mathcal{M}_{\mathrm{mean}}^{(w)}$. As $n^{(w)} := \min_t n_t^{(w)} \to \infty$, for $\bullet \in \{\mathrm{full}, \mathrm{mean}, \mathrm{dev}\}$ and $\rho \in \{\mathrm{tr}, 1, \ldots, R^{(w)}\}$,*

$$\max_{t,s} \left\| \widehat{\mathbf{M}}_\bullet^{(w)}(t,s) - \mathbf{M}_\bullet^{(w)}(t,s) \right\|_2 = O_p\left((n^{(w)})^{-1/2} \right),$$

and

$$\max_{t,s} \left| \widehat{d}_{\bullet,\rho}^{(w)}(t,s)^2 - d_{\bullet,\rho}^{(w)}(t,s)^2 \right| = O_p\left((n^{(w)})^{-1/2} \right).$$

Thus, the operators, squared trace and mode distances used below stabilize jointly as the period sample sizes increase. The guarantee covers the CUSP quantities derived from the fitted process, as well as the underlying Gaussian mixture parameters.

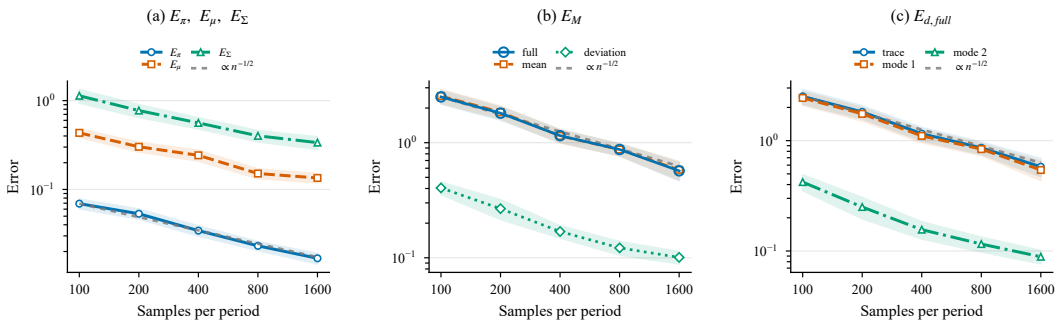

Figure 2: Finite-sample recovery of (a) component parameters, (b) full, mean, and deviation operators, and (c) squared full-trace and word-local mode distances. Curves show means and 95% confidence intervals over 50 repetitions. Gray dashed lines show the $n^{-1/2}$ rate.

4 Experiments

The experiments follow Figure 1: synthetic data test operator and distance recovery, DWUG (Schlechtweg et al., 2021) tests scalar validity and decomposition, Janus (Cassotti & Tahmasebi, 2025) tests multi-period recovery and coherence, and Court opinions test passage-grounded attribution. The synthetic experiment uses Gaussian samples. The lexical experiments use target-token XL-LEXEME embeddings (Cassotti et al., 2023) and Gaussian CUSP, with component couplings given by Proposition 3.5. The dominant directions shared across contextual embeddings can distort the distances (Timkey & van Schijndel, 2021). For each dataset, we therefore center its pooled usage panel, remove the leading principal direction ($k = 1$), and apply ℓ_2-normalization (Mu & Viswanath, 2018). The transform is fixed across words and periods within each dataset. DWUG uses separate English and German panels, Janus uses its original released pool before schedule construction, and Court uses its balanced analysis panel.

4.1 Synthetic mixture processes: finite-sample recovery

We test Theorem 3.7 on a three-period Gaussian process in $d = 2$ with $K = 2$, where the component prevalences, centers, and covariances all change. For 50 repetitions at $n \in \{100, 200, 400, 800, 1600\}$ usages per period, we fit independent full-covariance Gaussian mixture models (GMMs) with K fixed and withhold the planted component labels.

Figure 2 shows that the parameter errors decrease with increasing sample size. The full operator and squared full-trace errors both have a log–log slope of -0.532, while the two squared mode distance errors have slopes of -0.540 and -0.561. Their confidence intervals contain $-1/2$, agreeing with the rate predicted by Theorem 3.7. Appendix I.1 reports the mean and deviation distances and pure-mechanism checks.

4.2 DWUG: scalar validity and decomposition

The English and German DWUG datasets contain 46 and 50 two-period targets with human graded-change scores (Schlechtweg et al., 2021). We compare the CUSP full trace with APD, prototype distance, and AP-JSD (Giulianelli et al., 2020; Periti & Tahmasebi, 2024), balanced usage-level OT (Montariol & Allauzen, 2021), and fSUS (Kishino et al., 2025). We also evaluate the exact mean–deviation split (Theorem 3.1). Tables 1 and 6 report the full/mean and complete results, respectively.

Table 1 places CUSP first in English and within 0.012 of the strongest German result. The overlapping intervals support competitive scalar validity rather than uniform superiority. Sensitivity at $k = 0$ and $k = 3$ preserves the same conclusion, as reported in Appendix I.2. Center movement supplies median shares of 0.868 and 0.906 of the full-trace variation in English and German. The deviation trace remains correlated with human judgments at 0.681 and 0.674. The deviation distance is the second largest in each dataset for English *bar* (0.323) and German *Rezeption* (0.326), both split across

Table 1: Spearman correlation with DWUG judgments under pooled panel $k = 1$. Brackets give 95% paired word-bootstrap intervals.

Method	English ρ [95% CI]	German ρ [95% CI]
CUSP full trace	0.746 **[0.536, 0.883]**	0.815 [0.664, 0.903]
CUSP mean trace	0.742 [0.531, 0.881]	0.809 [0.649, 0.902]
APD	0.706 [0.482, 0.853]	0.815 [0.669, 0.894]
Prototype	0.636 [0.375, 0.817]	0.801 [0.641, 0.896]
AP-JSD	0.531 [0.242, 0.737]	0.578 [0.344, 0.761]
Usage-level OT	0.718 [0.501, 0.865]	0.827 **[0.675, 0.915]**
fSUS	0.726 [0.516, 0.866]	0.782 [0.621, 0.881]

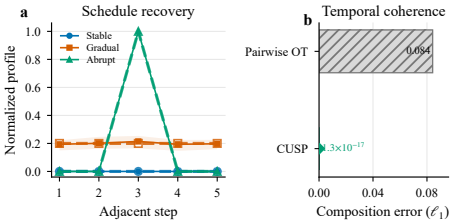

Figure 3: Janus at $k = 1$. (a) Planted and recovered profiles. Bands span the 2.5–97.5 percentiles across 16 lemmas per schedule. (b) Mean ℓ_1 composition error. CUSP is consistent by construction.

multiple DWUG usage clusters. The decomposition therefore preserves a useful within-component reorganization signal even though center displacement drives most graded change.

4.3 Janus: controlled multi-period recovery

Beyond two-period DWUG, Janus (Cassotti & Tahmasebi, 2025) provides controlled six-period histories: 48 lemmas, each with four ten-sentence pools resampled into 40 usages per period. Stable, gradual, and abrupt prevalence schedules each receive 16 lemmas. Released sense tags define schedules but are withheld during fitting. CUSP receives no cross-period usage links. A BIC-selected diagonal GMM is fitted independently to every lemma and period. Figure 3(a) shows that CUSP achieves exact recovery of the stable profile, a TV error of 0.044 for the gradual profile, and the correct boundary for all 16 abrupt lemmas. The recovered and planted adjacent changes correlate at $\rho = 0.954$. The path length separates the changed and the stable schedules with an AUROC value of 1.000. Pairwise OT fits balanced usage-level transport independently for each period pair. Its direct and composed plans have mean ℓ_1 disagreement 0.084, concentrated in gradual histories (Figure 3(b)). CUSP composes non-adjacent plans by construction, giving numerical zero disagreement while preserving endpoint marginals (Proposition 2.2). Appendix I.3 gives the protocol and schedule-specific results.

4.4 Court opinions: passage-grounded attribution

Court opinions provide a natural, unlabeled setting: 258,480 balanced usages of 60 legal lemmas from 100,473 opinions in eight decade bins (1950–2020), drawn from the May 6, 2024 CourtListener release (Free Law Project, 2024). The vocabulary is fixed before fitting. A fixed LLM procedure retains legal-reasoning usages and extracts verbatim passages. Each excerpt remains traceable through its CourtListener opinion ID. Counts are equalized across decades within each lemma, and BIC selects period-local diagonal GMMs with $K \leq 5$. Appendix I.4 details construction and fitting.

Without gold change points or sense labels, we select transitions numerically before passage inspection. For each lemma's seven adjacent full-trace distances d_t, we compute $z_t = (d_t - \bar{d})/\operatorname{sd}(d)$ and select the maximum. The descriptive threshold $\max_t z_t \geq 1.2$ retains 53 lemmas. At the selected transition, Theorems 3.1 and 3.2 give

$$a_{k\ell} = q_{k\ell} \operatorname{tr}(\mathbf{C}_{\text{mean}}(k,\ell) + \mathbf{C}_{\text{dev}}(k,\ell)), \qquad d_t^2 = \sum_{k,\ell} a_{k\ell}.$$

With $\alpha_{k\ell} = a_{k\ell}/d_t^2$, the ratio $\alpha_{k\ell}/q_{k\ell}$ compares the share of change in the pair with its share of transported mass. We report $E_{\max}$, the largest such ratio at the selected transition, to identify a movement that is unusually consequential for its size. Component labels are local to each period, so their numerical indices have no semantic meaning. Proposition 3.4 supplies the complementary temporal analysis. Its word-local mean modes distinguish directions that peak together from directions active in different decades. Appendix I.5 presents the fixed thresholds, the top ten transition and temporal-separation rankings, and the passage analyses.

After numerical selection, usages are assigned by maximum posterior probability. A fixed retrieval score favors proximity to component centers and clear target-lemma matches (Appendix I.5). Source

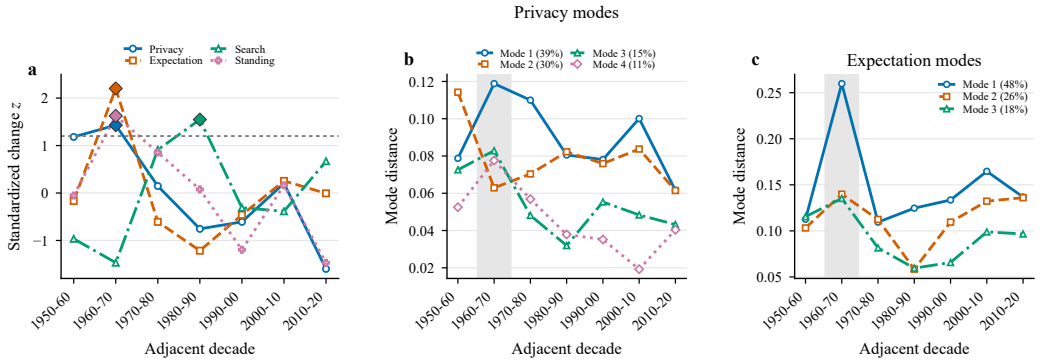

Figure 4: Court localization and word-local modes. (a) Standardized adjacent change profiles for the three highest-enrichment lemmas and *expectation*. Diamonds mark maxima and the line is $z = 1.2$. (b) *Privacy* modes peak in different decades. (c) The three leading *expectation* modes peak together. Legends give the retained mean-path energy shares. Markers and line styles distinguish series.

and target passages are representative endpoints, not observed sentence pairs. Figure 4(a) shows *privacy*, *search*, and *standing*, the three lemmas with the highest $E_{\max}$ scores, alongside *expectation*, whose modes peak together. Cited decisions situate the attributed usage movements in their doctrinal context.

Privacy. *Privacy* ranks first in both Court rankings. Its standardized profile peaks at 1960→1970 ($z = 1.43$), where center displacement supplies 93.3% of the variation. The selected pair carries 2.46% of the transported mass, but 37.1% of the contribution ($E_{\max} = 15.10$). Its source invokes Prosser's "four categories" (opinion `1179312`) and its target says that warrantless inspection poses "only a minimal threat to justifiable expectations of privacy" (opinion `1602980`). At the preceding step, the leading pair moves from privacy as "a direct wrong of a personal character" (opinion `2604478`) toward the same Prosser component. The successive pairs place Prosser's tort classification (Prosser, 1960) between an earlier privacy tort rationale and post-*Katz* expectation-of-privacy language (Supreme Court of the United States, 1967a). The modes of this lemma separately distinguish tort, constitutional, informational, search, and family privacy across several decades, consistent with legal scholarship treating privacy as context-dependent (Solove, 2006). The underlying passages and case anchors appear in Appendix I.5.

Expectation. Its standardized profile peaks at 1960→1970 ($z = 2.20$), where center displacement supplies 86.8% of the variation. Sources concern "an expectation of pay or compensation" and contributions made "without expectation of services or direct benefits" (opinions `2226138`, `1636296`). Targets ask whether the accused showed "a reasonable expectation of privacy infringed by the search and seizure" and describe an area "protected by an expectation of privacy" (opinions `1365600`, `1311660`). The transition moves from ordinary anticipation toward the post-*Katz* Fourth Amendment standard (Supreme Court of the United States, 1967a; 1978). The three leading modes of this lemma, carrying 48%, 26%, and 18% of retained mean-path energy, all peak at this transition. Unlike *privacy*, the change is synchronized rather than staggered. The Court dataset thus demonstrates the complete empirical chain from independently sampled period distributions to a coupled process, exact contributions, temporal modes, and passages that can be checked independently.

5 Conclusion

Diachronic corpora provide period specific usage distributions without observed trajectories. CUSP turns them into one marginal preserving process whose adjacent couplings determine longer span correspondence. Its central contribution is a common process yielding magnitude and timing, exact accounting by mechanism and transported component pair, word-local modes, and representative passages. Gaussian specialization makes the operators and squared distances tractable and recoverable. Synthetic, DWUG, Janus, and Court experiments support recovery, scalar validity, multi-period coherence, and passage-grounded attribution. CUSP thus makes scalar measurement, coherent temporal structure, and inspectable explanation compatible views of one lexical history.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval and discovery, including identifying related work, and with designing and implementing the code used to generate the synthetic datasets described in Section 4.1 and Appendix I.1. They also assisted with experimental analysis, interpretation of results, review of mathematical arguments and proofs, and manuscript organization and editing. The authors reviewed the AI-assisted work and take responsibility for the final text, code, results, and claims.

REPRODUCIBILITY STATEMENT

Appendix A guides readers to the assumptions, formulas, and proofs. Section 4 and Appendix I describe the experimental protocols, including preprocessing, fitting, evaluation, and passage retrieval. The anonymous supplementary package provides the code, configurations, prompts, fitted outputs, and scripts needed to regenerate the reported tables and figures. As detailed in Appendix I.4, rebuilding the Court panel from raw opinions requires the specified CourtListener snapshot and LLM-based collection procedure. The complete processed datasets exceed the submission size limit and will be released after acceptance.

REFERENCES

Taichi Aida and Danushka Bollegala. Investigating the contextualised word embedding dimensions specified for contextual and temporal semantic changes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1413–1437, Abu Dhabi, UAE, 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.95/.

Taichi Aida and Danushka Bollegala. SCDTour: Embedding axis ordering and merging for interpretable semantic change detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14775–14785, Suzhou, China, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.797. URL https://aclanthology.org/2025.findings-emnlp.797/.

Clare Arrington, Mauricio Gruppi, and Sibel Adali. ConShift: Sense-based language variation analysis using flexible alignment. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 167–181, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.9. URL https://aclanthology.org/2025.findings-naacl.9/.

Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 380–389. PMLR, 2017. URL https://proceedings.mlr.press/v70/bamler17a.html.

Pierluigi Cassotti and Nina Tahmasebi. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708, 2025. doi: 10.1162/tacl_a_00761. URL https://aclanthology.org/2025.tacl-1.32/.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1577–1585, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.135. URL https://aclanthology.org/2023.acl-short.135/.

Jiahua Chen and Xianming Tan. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100(7):1367–1383, 2009. doi: 10.1016/j.jmva.2008.12.005.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, and Chu-Ren Huang. Lexical semantic change detection: A survey of tasks, benchmarks, models, and potential impacts in digital humanities and social sciences. *Natural Language Processing*, 32(4):391–429, 2026. doi: 10.1017/nlp.2026.10032.

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019. doi: 10.1109/ACCESS.2018.2889838.

Julie Delon and Agnès Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020. doi: 10.1137/19M1301047.

Haruka Ezoe and Ryohei Hisano. Multiscale euclidean network trajectories: Second-moment geometry, attribution, and change points. *arXiv preprint arXiv:2605.04589*, 2026. URL https://arxiv.org/abs/2605.04589.

Free Law Project. CourtListener bulk legal data, 2024. URL https://wiki.free.law/c/courtlistener/help/api/bulk-data/bulk-legal-data. Opinion-data snapshot dated 2024-05-06.

Lea Frermann and Mirella Lapata. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016. doi: 10.1162/tacl_a_00081. URL https://aclanthology.org/Q16-1003/.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3960–3973, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.365. URL https://aclanthology.org/2020.acl-main.365/.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1501, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL https://aclanthology.org/P16-1141/.

Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3899–3908, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1379. URL https://aclanthology.org/P19-1379/.

Tao Huang, Heng Peng, and Kun Zhang. Model selection for gaussian mixture models. *Statistica Sinica*, 27(1):147–169, 2017. doi: 10.5705/ss.2014.105.

Ryo Kishino, Hiroaki Yamagiwa, Ryo Nagata, Sho Yokoi, and Hidetoshi Shimodaira. Quantifying lexical semantic shift via unbalanced optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15913–15933, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.774. URL https://aclanthology.org/2025.acl-long.774/.

Hajime Kiyama, Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. Analyzing continuous semantic shifts with diachronic word similarity matrices. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1613–1631, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.109/.

Denis V. Kokosinskii, Dominik Schlechtweg, and Nikolay V. Arefyev. Selecting and controlling sense granularity for lexical semantic change detection. *Program Systems: Theory and Applications*, 17(2):103–146, 2026. doi: 10.25209/2079-3316-2026-17-2-103-146. URL https://psta.psiras.ru/en/2026/2_103-146.

Imene Kolli, Kai-Robin Lange, Jonas Rieger, and Carsten Jentsch. Word-centered semantic graphs for interpretable diachronic sense tracking, 2026. URL https://arxiv.org/abs/2601.22410.

Syrielle Montariol and Alexandre Allauzen. Transport optimal pour le changement sémantique à partir de plongements contextualisés (optimal transport for semantic change detection using contextualised embeddings). In *Actes de la 28e Conférence sur le Traitement Automatique des*

Langues Naturelles. Volume 1 : conférence principale, pp. 81–90, Lille, France, 2021. ATALA. URL https://aclanthology.org/2021.jeptalnrecital-taln.7/.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Ryo Okano and Masaaki Imaizumi. Distribution-on-distribution regression with wasserstein metric: Multivariate gaussian case. *Journal of Multivariate Analysis*, 203:105334, 2024. doi: 10.1016/j.jmva.2024.105334.

Francesco Periti and Nina Tahmasebi. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4262–4282. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.240. URL https://aclanthology.org/2024.naacl-long.240/.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. What is done is done: an incremental approach to semantic shift detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pp. 33–43, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.lchange-1.4. URL https://aclanthology.org/2022.lchange-1.4/.

Francesco Periti, Sergio Picascia, Stefano Montanelli, Alfio Ferrara, and Nina Tahmasebi. Studying word meaning evolution through incremental semantic shift detection. *Language Resources and Evaluation*, 59:1363–1399, 2025. doi: 10.1007/s10579-024-09769-1. URL https://link.springer.com/article/10.1007/s10579-024-09769-1.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Bach Phan-Tat, Kris Heylen, Dirk Geeraerts, Stefano De Pascale, and Dirk Speelman. SynFlow: A multidimensional diachronic semantic analysis toolkit, 2026. URL https://arxiv.org/abs/2608.19472.

William L. Prosser. Privacy. *California Law Review*, 48(3):383–423, 1960. doi: 10.2307/3478805.

Filippo Santambrogio. Euclidean, Metric, and Wasserstein gradient flows: An overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017. doi: 10.1007/s13373-017-0101-1.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7079–7091, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.567. URL https://aclanthology.org/2021.emnlp-main.567/.

Daniel J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–564, 2006. doi: 10.2307/40041279. URL https://scholarship.law.gwu.edu/faculty_publications/921/.

Supreme Court of the United States. Griswold v. Connecticut, 381 u.s. 479, 1965.

Supreme Court of the United States. Katz v. United States, 389 U.S. 347, 1967a.

Supreme Court of the United States. Warden v. Hayden, 387 u.s. 294, 1967b.

Supreme Court of the United States. Roe v. Wade, 410 u.s. 113, 1973.

Supreme Court of the United States. Rakas v. Illinois, 439 u.s. 128, 1978.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M. Solomon. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1601–1611, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8b5040a8a5baf3e0e67386c2e3a9b903-Abstract.html.

A Appendix guide and notation

The appendix follows the construction in the main text. Appendix B formalizes identifiability of the period specific mixture representation. Appendix C proves marginal preservation and the basic properties of the coupled process. Appendices D and E establish the exact mean–deviation and sense-transition decompositions and the trace and mode-wise geometry, respectively. Appendix F gives explicit Gaussian formulas, and Appendix G shows how to compute non-adjacent transition quantities without enumerating latent paths. Appendix H proves the finite-sample recovery result. Appendix I then provides the complete experimental constructions, robustness checks, rankings, and passage analyses.

Throughout the paper, w denotes a target word, t, s denote periods, k, ℓ denote period local components, and r denotes a word-local mode. Component numbers are local to each period and do not identify the same sense across time. Their relation across periods is determined by the fitted coupling. We use $\bullet \in \{\mathrm{full}, \mathrm{mean}, \mathrm{dev}\}$ for statements that apply to any of the three processes and $\rho \in \{\mathrm{tr}, 1, \ldots, d\}$ for either the trace or a mode-wise readout. Table 2 collects the notation used throughout the paper and appendix.

A.1 Structural comparison with related approaches

Table 3 compares the structures provided by closely related approaches. A checkmark denotes a directly provided capability, $\triangle$ a related but non-equivalent object, and a dash a capability that is not directly provided. The comparison concerns the joint set of capabilities rather than any single column. Compositional coupling means that adjacent and non-adjacent usage–component correspondences arise from one multi-period process rather than independent fits. An exact mechanism split and transition attribution decompose the same change quantity by mechanism and transported component pair, respectively. Temporal modes are word-local directions with period local activity profiles. For SynFlow, $\triangle$ denotes related filler cluster continuity and value contributions rather than transported usage–component accounting.

B Identifiability of the sense mixture representation

The mixture representation of a period specific word distribution is not unique in general, since different parameterizations may induce the same distribution. The following assumption formalizes the identifiability condition stated in the main text: the mixture representation is uniquely determined up to permutation of the sense labels.

Assumption B.1 (Identifiability up to sense-label permutation)**.** Suppose that two mixture representations

$$\sum_{k=1}^{K_t^{(w)}} \pi_{t,k}^{(w)} \nu_{t,k}^{(w)} \qquad \text{and} \qquad \sum_{k=1}^{\tilde{K}_t^{(w)}} \tilde{\pi}_{t,k}^{(w)} \tilde{\nu}_{t,k}^{(w)}$$

represent the same period specific word distribution. Then

$$K_t^{(w)} = \tilde{K}_t^{(w)},$$

Table 2: Notation used in the CUSP construction, theoretical results, and experiments.

Notation	Meaning
Indices, spaces, and period distributions	
$V, [T], d$	Vocabulary, period index set $[T] = \{1, \dots, T\}$, and contextual-embedding dimension.
w, t, s, k, ℓ, r	Word, periods, period local source and target components, and word-local mode.
$\mathcal{P}_2(\mathbb{R}^d), \Pi(\mu, \nu)$	Probability measures with finite second moment, and couplings with marginals μ and ν.
$X_{t,i}^{(w)} \sim \gamma_t^{(w)}$	Contextual embedding of usage i, sampled independently from the distribution of word w in period t.
$\gamma_t^{(w)} = \sum_k \pi_{t,k}^{(w)} \nu_{t,k}^{(w)}$	period specific usage distribution represented as a finite mixture.
$K_t^{(w)}, \pi_{t,k}^{(w)}, \nu_{t,k}^{(w)}$	Number, prevalence, and contextual distribution of the period local sense components.
$m_{t,k}^{(w)}$	Center $\int x \, d\nu_{t,k}^{(w)}(x)$ of a general sense component.
Hierarchical coupling and temporal process	
$c, \mathcal{A}_{t,k,\ell}^{(w)}$	Contextual transport cost and admissible couplings for the candidate transition $k \to \ell$.
$\eta_{t,k,\ell}^{(w)}, C_{t,k,\ell}^{(w)}$	Optimal adjacent contextual coupling and its scalar transport cost.
$Q_t^{(w)}(k, \ell)$	Adjacent sense-level transport mass from component k at t to component ℓ at $t+1$.
$P_t^{(w)}(k, \ell)$	Sense transition probability $Q_t^{(w)}(k, \ell)/\pi_{t,k}^{(w)}$.
$\mathcal{K}_{t,k,\ell}^{(w)}(x, dy)$	Contextual transition kernel obtained by disintegrating $\eta_{t,k,\ell}^{(w)}$.
$Z_{1:T}^{(w)}, X_{1:T}^{(w)}$	Coupled latent sense process and contextual state process.
$R_{t,s}^{(w)}(k, \ell)$	Non-adjacent conditional probability $\Pr(Z_s^{(w)} = \ell \mid Z_t^{(w)} = k)$.
$q^{(w)}(t, s; k, \ell)$	Joint transported mass $\Pr(Z_t^{(w)} = k, Z_s^{(w)} = \ell) = \pi_{t,k}^{(w)} R_{t,s}^{(w)}(k, \ell)$.
Processes, operators, decompositions, and modes	
$\phi_{\text{full}}^{(w)}, \phi_{\text{mean}}^{(w)}, \phi_{\text{dev}}^{(w)}$	Contextual state, transported sense center, and centered within-sense deviation processes.
$\Delta\phi_\bullet^{(w)}(t, s)$	Coupled displacement $\phi_\bullet^{(w)}(t) - \phi_\bullet^{(w)}(s)$.
$\mathbf{M}_\bullet^{(w)}(t, s)$	Second-moment operator of $\Delta\phi_\bullet^{(w)}(t, s)$.
$\mathbf{C}_\bullet^{(w)}$	For $\bullet \in \{\text{mean}, \text{dev}\}$, $\mathbf{C}_\bullet^{(w)}(t, s; k, \ell)$ is the operator contribution conditional on transported component pair (k, ℓ).
$\bullet \in \{\text{full}, \text{mean}, \text{dev}\}$	Choice of the full contextual, between center, or within component process.
$\rho \in \{\text{tr}, 1, \dots, d\}$	Trace readout or word-local mode index.
$d_{\bullet,\text{tr}}^{(w)}(t, s), d_{\bullet,r}^{(w)}(t, s)$	Trace distance and distance along word-local mode r.
$\mathcal{M}_{\text{mean}}^{(w)}$	Word-local path operator $\sum_{t=1}^{T-1} \mathbf{M}_{\text{mean}}^{(w)}(t, t+1)$.
$(\lambda_r^{(w)}, \boldsymbol{u}_r^{(w)})$	Eigenvalue and eigenvector of $\mathcal{M}_{\text{mean}}^{(w)}$. The eigenvector defines word-local mode r.
Gaussian specialization and recovery	
$\mu_{t,k}^{(w)}, \Sigma_{t,k}^{(w)}$	Mean and covariance of Gaussian sense component $\nu_{t,k}^{(w)}$.
$A_{t,k,\ell}^{(w)}, T_{t,k,\ell}^{(w)}$	Linear part and affine map of the adjacent optimal Gaussian coupling.
$M_{t,s}^{(w)}$	Product of adjacent Gaussian linear maps along a latent sense path from t to s, transposed according to Appendix F.
$\Xi^{(w)}(t, s; k, \ell)$	Endpoint conditioned cross-period covariance used in $\mathbf{C}_{\text{dev}}^{(w)}(t, s; k, \ell)$.
$n_t^{(w)}, n^{(w)}$	Period sample size and minimum sample size $\min_t n_t^{(w)}$ used in the recovery theorem.
$\widehat{\theta}$	Estimated version of a population quantity θ.
Court readouts	
d_t, z_t	Adjacent full-trace distance and its within-lemma standardized value.
$a_{k\ell}, \alpha_{k\ell}$	Squared full-trace contribution of pair (k, ℓ) at the selected peak and its share of total peak variation.
$E_{\max}^{(w)}$	Maximum transition enrichment $\max_{q_{k\ell}>0} \alpha_{k\ell}/q_{k\ell}$.
e_r, τ_r	Share of mean-path energy carried by mode r and the adjacent period at which that mode peaks.
S_2	Secondary-mode temporal-separation score used when the two leading retained modes peak in different periods.

Table 3: Structural comparison with closely related approaches.

	Temporal representation			Attribution from that representation		
Approach	**Multi-period**	**Usage/component correspondence**	**Compositional coupling**	**Exact mechanism split**	**Transition attribution**	**Temporal modes**
Dynamic and evolving sense models (Frermann & Lapata, 2016; Hu et al., 2019)	✓	✓	–	△	–	–
WiDiD (Periti et al., 2022; 2025)	✓	✓	–	–	△	–
Multi-period similarity analysis (Kiyama et al., 2025)	✓	–	–	–	–	–
SynFlow (Phan-Tat et al., 2026)	✓	△	–	△	△	–
Pairwise OT (Montariol & Allauzen, 2021)	△	✓	–	–	–	–
ConShift (Arrington et al., 2025)	△	✓	–	△	△	–
UOT / SUS (Kishino et al., 2025)	△	✓	–	–	△	–
Directional methods (Aida & Bollegala, 2025a;b)	–	–	–	–	–	△
MENT (Ezoe & Hisano, 2026)	✓	–	–	△	–	✓
CUSP (ours)	✓	✓	✓	✓	✓	✓

and there exists a permutation σ of $\{1, \dots, K_t^{(w)}\}$ such that

$$\tilde{\pi}_{t,k}^{(w)} = \pi_{t,\sigma(k)}^{(w)}, \qquad \tilde{\nu}_{t,k}^{(w)} = \nu_{t,\sigma(k)}^{(w)}, \qquad k = 1, \dots, K_t^{(w)}.$$

Consequently, all label-dependent quantities are defined up to the corresponding label permutation. In the statistical recovery analysis, whenever population and estimated quantities are compared, we implicitly choose the permutation that aligns the corresponding sense labels.

The Gaussian-mixture specialization used in the statistical recovery analysis provides a concrete admissible class of mixture distributions and their parameterizations satisfying the identifiability condition above. The restrictions defining this class correspond to the first part of Assumption 3.6.

Definition B.2 (Admissible Gaussian mixture class)**.** The admissible Gaussian mixture class consists of Gaussian mixture distributions

$$\gamma = \sum_{k=1}^{K} \pi_k \nu_k, \qquad \nu_k = \mathcal{N}(m_k, \Sigma_k),$$

together with parameterizations satisfying

- $\pi_k > 0, \qquad \sum_{k=1}^{K} \pi_k = 1;$
- $\Sigma_k \succ 0, \qquad k = 1, \dots, K;$
- the component parameters are pairwise distinct:

$$(m_k, \Sigma_k) \neq (m_\ell, \Sigma_\ell), \qquad k \neq \ell.$$

C PROOFS OF BASIC PROPERTIES OF THE COUPLED USAGE-SENSE PROCESSES

We prove Proposition 2.2, which establishes preservation of the prescribed sense prevalences and contextual component distributions at each period, and consequently of the observed usage distributions as the period-wise marginals of the resulting joint process.

Proof. Fix $w \in V$, and suppress the superscript (w) throughout the proof. We first prove the preservation of the sense marginals by induction. By construction,

$$\Pr(Z_1 = k) = \pi_{1,k}$$

for every $k \in [K_1]$, so the claim holds at $t = 1$. If $\Pr(Z_t = k) = \pi_{t,k}$ for every $k \in [K_t]$, then

$$\begin{aligned}\Pr(Z_{t+1} = \ell) &= \sum_{k \in [K_t]} \Pr(Z_t = k) P_t(k, \ell) \\ &= \sum_{k \in [K_t]} \pi_{t,k} \frac{Q_t(k, \ell)}{\pi_{t,k}} = \sum_{k \in [K_t]} Q_t(k, \ell) = \pi_{t+1,\ell},\end{aligned}$$

where $P_t(k, \ell) = Q_t(k, \ell)/\pi_{t,k}$ is the transition probability, and the last equality follows from the second marginal constraint in 2, namely $\sum_{k \in [K_t]} Q_t(k, \ell) = \pi_{t+1,\ell}$. Thus, by induction,

$$\Pr(Z_t = k) = \pi_{t,k}$$

for all $t \in [T]$ and $k \in [K_t]$.

We next prove the preservation of the conditional contextual distributions by induction. Fix a sense path $z_{1:T}$ with positive probability. By construction,

$$\mathcal{L}(X_1 \mid Z_{1:T} = z_{1:T}) = \nu_{1,z_1},$$

so the claim holds at $t = 1$. Suppose that

$$\mathcal{L}(X_t \mid Z_{1:T} = z_{1:T}) = \nu_{t,z_t}.$$

Then, for any measurable set $A \subseteq \mathbb{R}^d$,

$$\begin{aligned}\Pr(X_{t+1} \in A \mid Z_{1:T} = z_{1:T}) &= \int \mathcal{K}_{t,z_t,z_{t+1}}(x, A)\, \nu_{t,z_t}(dx) \\ &= \int \mathbf{1}_A(y)\, \eta_{t,z_t,z_{t+1}}(dx, dy) \\ &= \nu_{t+1,z_{t+1}}(A),\end{aligned}$$

where we used the transition kernel defined by

$$\eta_{t,z_t,z_{t+1}}(dx, dy) = \nu_{t,z_t}(dx) \mathcal{K}_{t,z_t,z_{t+1}}(x, dy),$$

and the last equality follows from the marginal constraint in 1, namely $\eta_{t,z_t,z_{t+1}} \in \Pi(\nu_{t,z_t}, \nu_{t+1,z_{t+1}})$. Thus, by induction,

$$\mathcal{L}(X_t \mid Z_{1:T} = z_{1:T}) = \nu_{t,z_t}$$

for all $t \in [T]$, establishing the desired preservation of the conditional contextual distributions at each period.

Finally, using the sense marginal preservation and the conditional distribution established above, the marginal law of X_t is

$$\begin{aligned}\mathcal{L}(X_t) &= \sum_{z_{1:T}} \Pr(Z_{1:T} = z_{1:T})\, \mathcal{L}(X_t \mid Z_{1:T} = z_{1:T}) \\ &= \sum_{k} \left(\sum_{z_{1:T} : z_t = k} \Pr(Z_{1:T} = z_{1:T}) \right) \nu_{t,k} \\ &= \sum_{k \in [K_t]} \pi_{t,k} \nu_{t,k} = \gamma_t.\end{aligned}$$

Consequently,

$$\mathcal{L}(X_1, \ldots, X_T) \in \Pi(\gamma_1, \ldots, \gamma_T),$$

which proves the proposition. □

We prove the reconstruction and conditional centering properties stated in equation 4. The reconstruction identity follows directly from the definitions, while the conditional centering property follows from the preservation of the conditional marginals established in Proposition 2.2.

Proof. By definition,

$$\phi^{(w)}_{\mathrm{full}}(t) = \phi^{(w)}_{\mathrm{mean}}(t) + \phi^{(w)}_{\mathrm{dev}}(t),$$

so the reconstruction identity holds.

Moreover, by Proposition 2.2, for any sense path $z_{1:T}$ with positive probability,

$$\mathbb{E}\left[X^{(w)}_t \mid Z^{(w)}_{1:T} = z_{1:T}\right] = \int x \, d\nu^{(w)}_{t,z_t}(x) = m^{(w)}_{t,z_t}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\phi^{(w)}_{\mathrm{dev}}(t) \mid Z^{(w)}_{1:T}\right] &= \mathbb{E}\left[X^{(w)}_t - m^{(w)}_{t,Z^{(w)}_t} \mid Z^{(w)}_{1:T}\right] \\ &= m^{(w)}_{t,Z^{(w)}_t} - m^{(w)}_{t,Z^{(w)}_t} = 0.\end{aligned}$$

Thus, both identities in equation 4 hold. □

D Proofs of exact mechanistic and sense-transition attribution

We first prove Theorem 3.1. The key step is that the deviation process is conditionally centered given the full latent sense sequence, so the cross terms between the mean and deviation processes vanish.

Proof. Fix $w \in V$ and $t, s \in [T]$. By the reconstruction identity in equation 4,

$$\Delta\phi^{(w)}_{\mathrm{full}}(t,s) = \Delta\phi^{(w)}_{\mathrm{mean}}(t,s) + \Delta\phi^{(w)}_{\mathrm{dev}}(t,s).$$

Therefore, expanding the second-moment operator gives

$$\begin{aligned}\mathbf{M}^{(w)}_{\mathrm{full}}(t,s) = \mathbf{M}^{(w)}_{\mathrm{mean}}(t,s) &+ \mathbf{M}^{(w)}_{\mathrm{dev}}(t,s) \\ &+ \mathbb{E}\left[\Delta\phi^{(w)}_{\mathrm{mean}}(t,s)\Delta\phi^{(w)}_{\mathrm{dev}}(t,s)^\top\right] \\ &+ \mathbb{E}\left[\Delta\phi^{(w)}_{\mathrm{dev}}(t,s)\Delta\phi^{(w)}_{\mathrm{mean}}(t,s)^\top\right].\end{aligned}$$

We first consider the first cross term.

$$\begin{aligned}&\mathbb{E}\left[\Delta\phi^{(w)}_{\mathrm{mean}}(t,s)\Delta\phi^{(w)}_{\mathrm{dev}}(t,s)^\top\right] \\ &= \mathbb{E}\left[\left(m^{(w)}_{t,Z^{(w)}_t} - m^{(w)}_{s,Z^{(w)}_s}\right)\mathbb{E}\left[\Delta\phi^{(w)}_{\mathrm{dev}}(t,s)^\top \mid Z^{(w)}_{1:T}\right]\right] \\ &= 0,\end{aligned}$$

where the last equality follows from equation 4. The second cross term vanishes analogously. Consequently,

$$\mathbf{M}^{(w)}_{\mathrm{full}}(t,s) = \mathbf{M}^{(w)}_{\mathrm{mean}}(t,s) + \mathbf{M}^{(w)}_{\mathrm{dev}}(t,s),$$

which proves equation 6.

Applying the trace and the quadratic form $\boldsymbol{u}_r^\top(\cdot)\boldsymbol{u}_r$ gives equation 7 for every $\rho \in \{\mathrm{tr}, 1, \ldots, d\}$. □

We next prove Theorem 3.2. We first decompose each second-moment operator by conditioning on the latent sense path and then group the resulting terms according to the pair $(Z^{(w)}_t, Z^{(w)}_s)$.

Proof. Fix $w \in V$, $t, s \in [T]$, and $\bullet \in \{\mathrm{mean}, \mathrm{dev}\}$.

For the mean process,

$$\begin{aligned}\mathbf{M}^{(w)}_{\text{mean}}(t,s) &= \sum_{z_{1:T}} \Pr(Z^{(w)}_{1:T} = z_{1:T}) \left(m^{(w)}_{t,z_t} - m^{(w)}_{s,z_s}\right)\left(m^{(w)}_{t,z_t} - m^{(w)}_{s,z_s}\right)^\top \\ &= \sum_{k=1}^{K^{(w)}_t}\sum_{\ell=1}^{K^{(w)}_s}\left(\sum_{z_{1:T}: z_t=k,\ z_s=\ell} \Pr(Z^{(w)}_{1:T} = z_{1:T})\right) \\ &\qquad \cdot \left(m^{(w)}_{t,k} - m^{(w)}_{s,\ell}\right)\left(m^{(w)}_{t,k} - m^{(w)}_{s,\ell}\right)^\top \\ &= \sum_{k=1}^{K^{(w)}_t}\sum_{\ell=1}^{K^{(w)}_s} q^{(w)}(t,s;k,\ell)\mathbf{C}^{(w)}_{\text{mean}}(t,s;k,\ell).\end{aligned}$$

For the deviation process, first note that, for any $k \in [K^{(w)}_t]$ and $\ell \in [K^{(w)}_s]$,

$$\begin{aligned}&\mathbb{E}\left[X^{(w)}_t - X^{(w)}_s \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell\right] \\ &= \sum_{z_{1:T}: z_t=k,\ z_s=\ell} \Pr\left(Z^{(w)}_{1:T} = z_{1:T} \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell\right) \\ &\qquad \cdot \mathbb{E}\left[X^{(w)}_t - X^{(w)}_s \mid Z^{(w)}_{1:T} = z_{1:T}\right] \\ &= \sum_{z_{1:T}: z_t=k,\ z_s=\ell} \Pr\left(Z^{(w)}_{1:T} = z_{1:T} \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell\right) \\ &\qquad \cdot \left(m^{(w)}_{t,z_t} - m^{(w)}_{s,z_s}\right) \\ &= m^{(w)}_{t,k} - m^{(w)}_{s,\ell},\end{aligned}$$

where the second equality follows from the conditional marginal preservation in Proposition 2.2. Hence,

$$\begin{aligned}&\mathbb{E}\left[\Delta\phi^{(w)}_{\text{dev}}(t,s)\Delta\phi^{(w)}_{\text{dev}}(t,s)^\top \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell\right] \\ &= \mathbb{E}\left[\left\{X^{(w)}_t - X^{(w)}_s - \left(m^{(w)}_{t,k} - m^{(w)}_{s,\ell}\right)\right\}\left\{X^{(w)}_t - X^{(w)}_s - \left(m^{(w)}_{t,k} - m^{(w)}_{s,\ell}\right)\right\}^\top \middle| Z^{(w)}_t = k, Z^{(w)}_s = \ell\right] \\ &= \text{Cov}\left(X^{(w)}_t - X^{(w)}_s \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell\right) \\ &= \mathbf{C}^{(w)}_{\text{dev}}(t,s;k,\ell).\end{aligned}$$

Therefore, conditioning on the pair $(Z^{(w)}_t, Z^{(w)}_s)$ gives

$$\begin{aligned}\mathbf{M}^{(w)}_{\text{dev}}(t,s) &= \sum_{k=1}^{K^{(w)}_t}\sum_{\ell=1}^{K^{(w)}_s} q^{(w)}(t,s;k,\ell) \\ &\qquad \cdot \mathbb{E}\left[\Delta\phi^{(w)}_{\text{dev}}(t,s)\Delta\phi^{(w)}_{\text{dev}}(t,s)^\top \mid Z^{(w)}_t = k, Z^{(w)}_s = \ell\right] \\ &= \sum_{k=1}^{K^{(w)}_t}\sum_{\ell=1}^{K^{(w)}_s} q^{(w)}(t,s;k,\ell)\mathbf{C}^{(w)}_{\text{dev}}(t,s;k,\ell).\end{aligned}$$

Thus, equation 8 follows for both $\bullet \in \{\text{mean}, \text{dev}\}$.

Applying the trace and the quadratic form $\boldsymbol{u}^{(w)\top}_r(\cdot)\boldsymbol{u}^{(w)}_r$ gives equation 9 and equation 10. □

E Proofs of Properties of Trace and Mode-Wise Geometry

E.1 Validity and properties of the induced semantic distances

The semantic distances introduced in the main text define the geometry of semantic change on the period index set. This subsection establishes the pseudometric properties of the semantic distances induced by the corresponding L^2 random-vector representations.

The first property concerns the pseudo-metric structure induced by the random-vector representations.

Proposition E.1 (Pseudo-metric properties of trace and mode-wise semantic distances)**.** *For every $w \in V$ and $\bullet \in \{\mathrm{full}, \mathrm{mean}, \mathrm{dev}\}$, the quantities*

$$d^{(w)}_{\bullet,\mathrm{tr}}(t,s), \qquad d^{(w)}_{\bullet,r}(t,s), \quad r \in [d],$$

define finite pseudo-metrics on $[T]$.

Moreover, the zero-distance conditions are given by

$$d^{(w)}_{\bullet,\mathrm{tr}}(t,s) = 0 \iff \phi^{(w)}_{\bullet}(t) = \phi^{(w)}_{\bullet}(s) \quad \textit{almost surely},$$

and

$$d^{(w)}_{\bullet,r}(t,s) = 0 \iff \boldsymbol{u}_r^\top \phi^{(w)}_{\bullet}(t) = \boldsymbol{u}_r^\top \phi^{(w)}_{\bullet}(s) \quad \textit{almost surely}.$$

Remark. Zero distance implies equality only in the corresponding random-vector representation and does not necessarily imply $t = s$. For the mode-wise distances, the zero-distance condition only concerns the projection onto the corresponding direction $\boldsymbol{u}_r$.

Proof. Since $\nu^{(w)}_{t,k} \in \mathcal{P}_2(\mathbb{R}^d)$, the mixture distribution satisfies $\gamma^{(w)}_t \in \mathcal{P}_2(\mathbb{R}^d)$. By the marginal consistency of the full process in equation 3, $\phi^{(w)}_{\mathrm{full}}(t) \in L^2(\Omega;\mathbb{R}^d)$. Moreover, $\phi^{(w)}_{\mathrm{mean}}(t) \in L^2(\Omega;\mathbb{R}^d)$, since it takes values in the finite set of sense centers $\{m^{(w)}_{t,k}\}_{k=1}^{K^{(w)}_t}$. Therefore,

$$\phi^{(w)}_{\mathrm{dev}}(t) = \phi^{(w)}_{\mathrm{full}}(t) - \phi^{(w)}_{\mathrm{mean}}(t) \in L^2(\Omega;\mathbb{R}^d).$$

Hence, all the following L^2 distances are well-defined.

For the trace distance, the definition gives

$$d^{(w)}_{\bullet,\mathrm{tr}}(t,s) = \left(\mathbb{E}\left[\left\|\phi^{(w)}_{\bullet}(t) - \phi^{(w)}_{\bullet}(s)\right\|^2\right]\right)^{1/2} = \left\|\phi^{(w)}_{\bullet}(t) - \phi^{(w)}_{\bullet}(s)\right\|_{L^2(\Omega;\mathbb{R}^d)}.$$

Hence, $d^{(w)}_{\bullet,\mathrm{tr}}$ is the L^2 distance between the random vectors $\phi^{(w)}_{\bullet}(t)$ and $\phi^{(w)}_{\bullet}(s)$. Nonnegativity and symmetry follow directly from the properties of the norm, and the triangle inequality follows from Minkowski's inequality.

Similarly,

$$d^{(w)}_{\bullet,r}(t,s) = \left\|\boldsymbol{u}_r^\top \phi^{(w)}_{\bullet}(t) - \boldsymbol{u}_r^\top \phi^{(w)}_{\bullet}(s)\right\|_{L^2(\Omega;\mathbb{R})},$$

which is the L^2 distance between the scalar random variables obtained by projecting the process onto the direction $\boldsymbol{u}_r$. Therefore, the same arguments establish nonnegativity, symmetry, and the triangle inequality for $d^{(w)}_{\bullet,r}$.

Finally, an L^2 norm is zero if and only if the corresponding random elements agree almost surely. Applying this property to the vector-valued and projected representations gives the stated zero-distance conditions. □

E.2 Proofs of the decomposition and characterization of the semantic geometry

This section provides proofs of the characterization and decomposition results for the semantic geometry introduced in the main text. The proofs follow the same arguments as those in Ezoe & Hisano (2026).

The following proof establishes the decomposition of the total semantic change into orthogonal mode-wise contributions.

Proof of equation 5. Since $\{\boldsymbol{u}_r\}_{r=1}^d$ is an orthonormal basis,

$$\operatorname{tr} \mathbf{M}_\bullet^{(w)}(t,s) = \sum_{r=1}^{d} (\boldsymbol{u}_r^{(w)})^\top \mathbf{M}_\bullet^{(w)}(t,s) \boldsymbol{u}_r^{(w)}.$$

Using the definitions of $d_{\bullet,\mathrm{tr}}^{(w)}(t,s)$ and $d_{\bullet,r}^{(w)}(t,s)$ yields the claim. □

The following proof establishes the spectral characterization of word-local modes induced by each word's aggregated mean operator.

Proof of Proposition 3.4. By definition,

$$\mathcal{M}_{\mathrm{mean}}^{(w)} = \sum_{t=1}^{T-1} \mathbf{M}_{\mathrm{mean}}^{(w)}(t,t+1).$$

Since

$$\mathcal{M}_{\mathrm{mean}}^{(w)} \boldsymbol{u}_r^{(w)} = \lambda_r^{(w)} \boldsymbol{u}_r^{(w)} \quad \text{and} \quad \|\boldsymbol{u}_r^{(w)}\| = 1,$$

we have

$$\begin{aligned}
\lambda_r^{(w)} &= (\boldsymbol{u}_r^{(w)})^\top \mathcal{M}_{\mathrm{mean}}^{(w)} \boldsymbol{u}_r^{(w)} \\
&= \sum_{t=1}^{T-1} (\boldsymbol{u}_r^{(w)})^\top \mathbf{M}_{\mathrm{mean}}^{(w)}(t,t+1) \boldsymbol{u}_r^{(w)} \\
&= \sum_{t=1}^{T-1} \left(d_{\mathrm{mean},r}^{(w)}(t,t+1) \right)^2,
\end{aligned}$$

which proves equation 11.

Moreover, since $\mathcal{M}_{\mathrm{mean}}^{(w)}$ is symmetric, the Courant–Fischer variational characterization gives

$$\boldsymbol{u}_r^{(w)} \in \arg \max_{\substack{\|\boldsymbol{u}\|=1 \\ \boldsymbol{u} \perp \boldsymbol{u}_1^{(w)}, \ldots, \boldsymbol{u}_{r-1}^{(w)}}} \boldsymbol{u}^\top \mathcal{M}_{\mathrm{mean}}^{(w)} \boldsymbol{u}.$$

Substituting the definition of $\mathcal{M}_{\mathrm{mean}}^{(w)}$ yields

$$\boldsymbol{u}_r^{(w)} \in \arg \max_{\substack{\|\boldsymbol{u}\|=1 \\ \boldsymbol{u} \perp \boldsymbol{u}_1^{(w)}, \ldots, \boldsymbol{u}_{r-1}^{(w)}}} \sum_{t=1}^{T-1} \boldsymbol{u}^\top \mathbf{M}_{\mathrm{mean}}^{(w)}(t,t+1) \boldsymbol{u},$$

which proves equation 12. □

F Explicit expressions under the Gaussian specialization

Under the Gaussian and quadratic-cost specialization in Section 3, both the contextual coupling cost and the sense-pair second-moment components admit explicit expressions.

First, the contextual coupling problem in equation 1 reduces to the quadratic 2-Wasserstein problem between Gaussian measures. As established, e.g., in Okano & Imaizumi (2024); Santambrogio (2017), the optimal coupling exists and is unique under positive-definite covariance matrices, and is induced by the affine optimal transport map given in Proposition 3.5. Its minimum cost is therefore

$$\begin{aligned}
C_{t,k,\ell}^{(w)} &= W_2^2 \left(\nu_{t,k}^{(w)}, \nu_{t+1,\ell}^{(w)} \right) \\
&= \left\| \mu_{t,k}^{(w)} - \mu_{t+1,\ell}^{(w)} \right\|^2 + \operatorname{Tr} \left(\Sigma_{t,k}^{(w)} + \Sigma_{t+1,\ell}^{(w)} - 2 \left[(\Sigma_{t,k}^{(w)})^{1/2} \Sigma_{t+1,\ell}^{(w)} (\Sigma_{t,k}^{(w)})^{1/2} \right]^{1/2} \right). \qquad (14)
\end{aligned}$$

We next give explicit expressions for the sense-pair second-moment components used in Theorem 3.2.

Proposition F.1 (Explicit expressions for sense-pair second moments)**.** *Under the Gaussian specialization, for every $w \in V$, $t < s$, and pair (k, ℓ) with $q^{(w)}(t, s; k, \ell) > 0$, the mean component is*

$$\mathbf{C}^{(w)}_{\text{mean}}(t, s; k, \ell) = \left(\mu^{(w)}_{t,k} - \mu^{(w)}_{s,\ell}\right)\left(\mu^{(w)}_{t,k} - \mu^{(w)}_{s,\ell}\right)^{\top}, \tag{15}$$

and the deviation component is

$$\mathbf{C}^{(w)}_{\text{dev}}(t, s; k, \ell) = \Sigma^{(w)}_{t,k} + \Sigma^{(w)}_{s,\ell} - \Xi^{(w)}(t, s; k, \ell) - \Xi^{(w)}(t, s; k, \ell)^{\top}, \tag{16}$$

where

$$M^{(w)}_{t,s} := \left(A^{(w)}_{s-1,Z^{(w)}_{s-1},Z^{(w)}_{s}} \cdots A^{(w)}_{t,Z^{(w)}_{t},Z^{(w)}_{t+1}}\right)^{\top}, \tag{17}$$

$$\Xi^{(w)}(t, s; k, \ell) := \Sigma^{(w)}_{t,k}\,\mathbb{E}\left[M^{(w)}_{t,s}\,\middle|\,Z^{(w)}_{t} = k, Z^{(w)}_{s} = \ell\right] \tag{18}$$

Proof. The expression for $\mathbf{C}^{(w)}_{\text{mean}}(t, s; k, \ell)$ follows directly from the means of the Gaussian sense components.

For the deviation component, apply the law of total covariance by conditioning further on the complete latent sense sequence:

$$\begin{aligned}\mathbf{C}^{(w)}_{\text{dev}}(t, s; k, \ell) = \mathbb{E}\Big[\mathrm{Cov}\Big(X^{(w)}_{t} - X^{(w)}_{s} \mid Z^{(w)}_{1:T}\Big)\,\Big|\,Z^{(w)}_{t} = k, Z^{(w)}_{s} = \ell\Big] \\ + \mathrm{Cov}\Big(\mathbb{E}\Big[X^{(w)}_{t} - X^{(w)}_{s} \mid Z^{(w)}_{1:T}\Big]\,\Big|\,Z^{(w)}_{t} = k, Z^{(w)}_{s} = \ell\Big).\end{aligned} \tag{19}$$

By the conditional marginals established in Proposition 2.2,

$$\mathrm{Cov}\Big(\mathbb{E}\Big[X^{(w)}_{t} - X^{(w)}_{s} \mid Z^{(w)}_{1:T}\Big]\,\Big|\,Z^{(w)}_{t} = k, Z^{(w)}_{s} = \ell\Big) \tag{20}$$

$$= \mathrm{Cov}\Big(\mu^{(w)}_{t,k} - \mu^{(w)}_{s,\ell}\,\Big|\,Z^{(w)}_{t} = k, Z^{(w)}_{s} = \ell\Big) = 0, \tag{21}$$

thus the second term vanishes.

Now fix a latent sense sequence $z_{1:T}$ satisfying $z_t = k$ and $z_s = \ell$. The conditional covariance of the displacement decomposes as

$$\begin{aligned}&\mathrm{Cov}\Big(X^{(w)}_{t} - X^{(w)}_{s} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) \\ &= \mathrm{Var}\Big(X^{(w)}_{t} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) + \mathrm{Var}\Big(X^{(w)}_{s} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) \\ &\quad - \mathrm{Cov}\Big(X^{(w)}_{t}, X^{(w)}_{s} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) - \mathrm{Cov}\Big(X^{(w)}_{s}, X^{(w)}_{t} \mid Z^{(w)}_{1:T} = z_{1:T}\Big).\end{aligned} \tag{22}$$

By Proposition 2.2,

$$\mathrm{Var}\Big(X^{(w)}_{t} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) = \Sigma^{(w)}_{t,k}, \qquad \mathrm{Var}\Big(X^{(w)}_{s} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) = \Sigma^{(w)}_{s,\ell}. \tag{23}$$

Moreover, Proposition 3.5 gives the affine optimal transport map for each adjacent sense transition. Consequently, the Markov composition along the fixed sense path gives

$$X^{(w)}_{s} = \mu^{(w)}_{s,\ell} + A^{(w)}_{s-1,z_{s-1},z_s} \cdots A^{(w)}_{t,z_t,z_{t+1}}\left(X^{(w)}_{t} - \mu^{(w)}_{t,k}\right). \tag{24}$$

Taking the conditional covariance gives

$$\begin{aligned}\mathrm{Cov}\Big(X^{(w)}_{t}, X^{(w)}_{s} \mid Z^{(w)}_{1:T} = z_{1:T}\Big) &= \mathrm{Cov}\Big(X^{(w)}_{s}, X^{(w)}_{t} \mid Z^{(w)}_{1:T} = z_{1:T}\Big)^{\top} \\ &= \Sigma^{(w)}_{t,k}\left(A^{(w)}_{s-1,z_{s-1},z_s} \cdots A^{(w)}_{t,z_t,z_{t+1}}\right)^{\top},\end{aligned} \tag{25}$$

Substituting these expressions into equation 22, we obtain

$$\begin{aligned}
&\mathrm{Cov}\Big(X_t^{(w)} - X_s^{(w)} \mid Z_{1:T}^{(w)} = z_{1:T}\Big) \\
&= \Sigma_{t,k}^{(w)} + \Sigma_{s,\ell}^{(w)} - \Sigma_{t,k}^{(w)} \left(A_{s-1,z_{s-1},z_s}^{(w)} \cdots A_{t,z_t,z_{t+1}}^{(w)}\right)^\top \\
&\quad - \left[\Sigma_{t,k}^{(w)} \left(A_{s-1,z_{s-1},z_s}^{(w)} \cdots A_{t,z_t,z_{t+1}}^{(w)}\right)^\top\right]^\top . \qquad (26)
\end{aligned}$$

Combining the above expressions with equation 19, we obtain

$$\begin{aligned}
\mathbf{C}_{\mathrm{dev}}^{(w)}(t,s;k,\ell) &= \Sigma_{t,k}^{(w)} + \Sigma_{s,\ell}^{(w)} - \Sigma_{t,k}^{(w)} \mathbb{E}\Big[M_{t,s}^{(w)} \Big| Z_t^{(w)} = k, Z_s^{(w)} = \ell\Big] \\
&\quad - \mathbb{E}\Big[M_{t,s}^{(w)\top} \Big| Z_t^{(w)} = k, Z_s^{(w)} = \ell\Big] \Sigma_{t,k}^{(w)} \qquad (27) \\
&= \Sigma_{t,k}^{(w)} + \Sigma_{s,\ell}^{(w)} - \Xi^{(w)}(t,s;k,\ell) - \Xi^{(w)}(t,s;k,\ell)^\top, \qquad (28)
\end{aligned}$$

which proves the stated expression for $\mathbf{C}_{\mathrm{dev}}^{(w)}(t,s;k,\ell)$. □

G Dynamic programming for transition quantities

We describe how to compute the sense-transition quantities without enumerating all possible intermediate sense paths. In particular, the sense-pair probabilities $q^{(w)}(t,s;k,\ell)$ and the cross-period covariance quantities $\Xi^{(w)}(t,s;k,\ell)$, and hence the deviation components $\mathbf{C}_{\mathrm{dev}}^{(w)}(t,s;k,\ell)$, can be computed by dynamic programming.

Computational cost. We analyze the computational cost of computing non-adjacent quantities induced by the Markov chain after GMM fitting and adjacent period OT solving have already been completed, and hence the corresponding transport maps $A_{t,k,\ell}^{(w)}$ and transition probabilities $P_t^{(w)}(k,\ell)$ are available. A direct pathwise computation requires enumerating latent component paths across periods, leading to an exponential increase in computational cost with the number of periods T. We therefore exploit dynamic programming to aggregate contributions from intermediate latent components without explicit path enumeration. For one word, let $K = \max_t K_t^{(w)}$, and let d denote the embedding dimension. The memory costs below refer only to additional memory required during computation and exclude storage of the precomputed mixture parameters, adjacent period transport maps and transition probabilities, as well as the final outputs.

Considering the computation of all quantities indexed by (t,s,k,ℓ), the joint component probabilities $q(t,s;k,\ell)$ can be obtained by either explicit enumeration of latent component paths or dynamic programming. A naive enumeration of all latent component paths yields the upper bound $O(T^3K^T)$, whereas the dynamic programming recursion requires only $O(T^2K^3)$. In terms of memory, naive path enumeration can be implemented with $O(T)$ memory by generating one path at a time, while the dynamic programming recursion requires $O(K^2)$ memory for the current transition-probability matrix.

For the computation of $\Xi^{(w)}(t,s;k,\ell)$, naive enumeration of all latent component paths yields the upper bound $O(T^3K^Td^3)$ in time and $O(T+d^2)$ in memory. In contrast, the dynamic programming recursion requires $O(T^2K^3d^3)$ time and $O(K^2d^2)$ memory. The conditional weights $r_{t,s}(j \mid k,\ell)$ can be computed in $O(1)$ time for each (t,s,k,ℓ,j) without additional memory, provided that the recursion for $R_{t,s}$ used to compute $q(t,s;k,\ell)$ is carried out alongside the Ξ recursion.

Once $\Xi^{(w)}(t,s;k,\ell)$ has been computed, the deviation component

$$\mathbf{C}_{\mathrm{dev}}^{(w)}(t,s;k,\ell) = \Sigma_{t,k}^{(w)} + \Sigma_{s,\ell}^{(w)} - \Xi^{(w)}(t,s;k,\ell) - \Xi^{(w)}(t,s;k,\ell)^\top$$

requires only $O(d^2)$ time and $O(d^2)$ memory for each (t,s,k,ℓ), yielding $O(T^2K^2d^2)$ time overall. This is lower-order than the $O(T^2K^3d^3)$ cost of computing Ξ, and therefore the additional computational cost of obtaining $\mathbf{C}_{\mathrm{dev}}$ is negligible in comparison.

Proposition G.1 (Dynamic programming recursions)**.** *For $t \le s$, let*

$$R_{t,s}^{(w)}(k,\ell) := \Pr\Big(Z_s^{(w)} = \ell \mid Z_t^{(w)} = k\Big).$$

Then

$$R_{t,t}^{(w)} = I_{K_t^{(w)}}, \qquad R_{t,r+1}^{(w)} = R_{t,r}^{(w)} P_r^{(w)}, \quad r = t, \dots, s-1, \tag{29}$$

and hence

$$q^{(w)}(t,s;k,\ell) = \pi_{t,k}^{(w)} R_{t,s}^{(w)}(k,\ell). \tag{30}$$

For $t < s$, define the backward smoothing probability

$$r_{t,s}^{(w)}(j \mid k,\ell) := \Pr\Big(Z_{s-1}^{(w)} = j \mid Z_t^{(w)} = k, Z_s^{(w)} = \ell\Big).$$

Then, whenever $R_{t,s}^{(w)}(k,\ell) > 0$,

$$r_{t,s}^{(w)}(j \mid k,\ell) = \frac{R_{t,s-1}^{(w)}(k,j) P_{s-1}^{(w)}(j,\ell)}{R_{t,s}^{(w)}(k,\ell)}. \tag{31}$$

The recursion for $\Xi^{(w)}$ is initialized by

$$\Xi^{(w)}(t,t;k,\ell) = \mathbf{1}_{\{k=\ell\}} \Sigma_{t,k}^{(w)}, \tag{32}$$

and, for $s > t$,

$$\Xi^{(w)}(t,s;k,\ell) = \sum_{j=1}^{K_{s-1}^{(w)}} r_{t,s}^{(w)}(j \mid k,\ell)\, \Xi^{(w)}(t,s-1;k,j) \left(A_{s-1,j,\ell}^{(w)}\right)^{\top}. \tag{33}$$

This recursion is evaluated for positive-mass endpoint pairs. Zero-weight terms in the sum are omitted, and we set $\Xi^{(w)}(t,s;k,\ell) = 0$ for zero-mass endpoint pairs.

Proof. The recursion for $R_{t,s}^{(w)}$ follows from the Markov property. For $r = t, \dots, s-1$,

$$\begin{aligned} R_{t,r+1}^{(w)}(k,\ell) &= \Pr\Big(Z_{r+1}^{(w)} = \ell \mid Z_t^{(w)} = k\Big) \\ &= \sum_{j=1}^{K_r^{(w)}} \Pr\Big(Z_r^{(w)} = j \mid Z_t^{(w)} = k\Big) \Pr\Big(Z_{r+1}^{(w)} = \ell \mid Z_r^{(w)} = j\Big) \\ &= \sum_{j=1}^{K_r^{(w)}} R_{t,r}^{(w)}(k,j) P_r^{(w)}(j,\ell), \end{aligned} \tag{34}$$

which gives equation 29. Moreover,

$$\begin{aligned} q^{(w)}(t,s;k,\ell) &= \Pr\Big(Z_t^{(w)} = k\Big) \Pr\Big(Z_s^{(w)} = \ell \mid Z_t^{(w)} = k\Big) \\ &= \pi_{t,k}^{(w)} R_{t,s}^{(w)}(k,\ell), \end{aligned} \tag{35}$$

which proves equation 30.

Next, by the definition of conditional probability,

$$\begin{aligned} r_{t,s}^{(w)}(j \mid k,\ell) &= \frac{\Pr\Big(Z_{s-1}^{(w)} = j \mid Z_t^{(w)} = k\Big) \Pr\Big(Z_s^{(w)} = \ell \mid Z_{s-1}^{(w)} = j, Z_t^{(w)} = k\Big)}{\Pr\Big(Z_s^{(w)} = \ell \mid Z_t^{(w)} = k\Big)} \\ &= \frac{R_{t,s-1}^{(w)}(k,j) P_{s-1}^{(w)}(j,\ell)}{R_{t,s}^{(w)}(k,\ell)}, \end{aligned} \tag{36}$$

where the final equality follows from the Markov property.

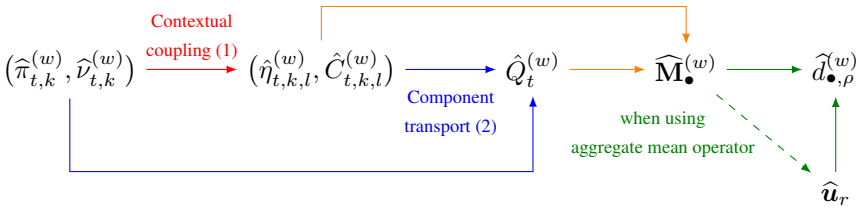

Figure 5: Construction pipeline of the estimated semantic geometry. The pipeline maps estimated sense distributions to geometric representations. $\bullet \in \{\mathrm{full}, \mathrm{mean}, \mathrm{dev}\}, \rho \in \{\mathrm{tr}, 1, \ldots, d\}$.

It remains to establish the recursion for $\Xi^{(w)}$. From the definition of $M_{t,s}^{(w)}$ in equation 17, for $s > t$,

$$M_{t,s}^{(w)} = M_{t,s-1}^{(w)} \left(A_{s-1,Z_{s-1}^{(w)},Z_s^{(w)}}^{(w)} \right)^\top. \tag{37}$$

For the initial case, we use the empty-product convention $M_{t,t}^{(w)} := I_d$, so that

$$\Xi^{(w)}(t,t;k,k) = \Sigma_{t,k}^{(w)}.$$

For $s > t$, conditioning on $Z_{s-1}^{(w)}$ gives

$$\begin{aligned}
&\mathbb{E}\Big[M_{t,s}^{(w)} \mid Z_t^{(w)} = k, Z_s^{(w)} = \ell\Big] \\
&= \sum_{j=1}^{K_{s-1}^{(w)}} r_{t,s}^{(w)}(j \mid k, \ell) \mathbb{E}\Big[M_{t,s-1}^{(w)} \Big| Z_t^{(w)} = k, Z_{s-1}^{(w)} = j, Z_s^{(w)} = \ell\Big] \left(A_{s-1,j,\ell}^{(w)} \right)^\top.
\end{aligned} \tag{38}$$

Since $M_{t,s-1}^{(w)}$ depends only on $Z_t^{(w)}, \ldots, Z_{s-1}^{(w)}$, the Markov property gives

$$\begin{aligned}
&\mathbb{E}\Big[M_{t,s-1}^{(w)} \mid Z_t^{(w)} = k, Z_{s-1}^{(w)} = j, Z_s^{(w)} = \ell\Big] \\
&\qquad = \mathbb{E}\Big[M_{t,s-1}^{(w)} \mid Z_t^{(w)} = k, Z_{s-1}^{(w)} = j\Big].
\end{aligned} \tag{39}$$

Multiplying equation 38 by $\Sigma_{t,k}^{(w)}$ yields

$$\Xi^{(w)}(t,s;k,\ell) = \sum_{j=1}^{K_{s-1}^{(w)}} r_{t,s}^{(w)}(j \mid k, \ell)\, \Xi^{(w)}(t, s-1; k, j) \left(A_{s-1,j,\ell}^{(w)} \right)^\top, \tag{40}$$

which proves equation 33. □

H STATISTICAL RECOVERY

This appendix studies how estimation errors propagate through the construction of the semantic geometry and establishes the error rates for the second-moment operators and semantic distances under the Gaussian specialization given in Theorem 3.7. Figure 5 illustrates the construction pipeline from estimated sense distributions and prevalences to the semantic distances. Figure 6 summarizes the error propagation from the estimated Gaussian mixture parameters to the estimated semantic geometry. We first establish the estimation rate for the Gaussian mixture parameters and then propagate these errors through each step of the construction to obtain the convergence rate of the second-moment operators and the induced distances.

Recovery of Gaussian mixture parameters. We assume throughout the theoretical analysis that the true number of mixture components is known. This assumption does not impose a substantive restriction. When a known finite upper bound $K_{\max}$ on the number of components is available, an

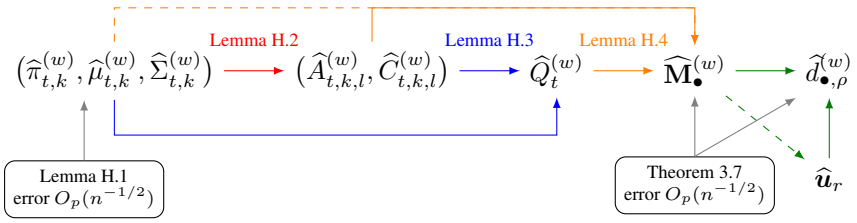

Figure 6: Error propagation and recovery guarantees for the estimated semantic geometry. $\bullet \in \{\text{full}, \text{mean}, \text{dev}\}, \rho \in \{\text{tr}, 1, \ldots, R^{(w)}\}$.

estimate $\widehat{K}_t^{(w)}$ can be obtained by a penalized likelihood model-selection procedure. Under the regularity conditions in Definition B.2, the model-selection consistency established by Huang et al. (2017, Theorem 3) implies that

$$\Pr\left(\widehat{K}_t^{(w)} = K_t^{(w)}\right) \to 1 \qquad (n \to \infty).$$

Thus, the true component number is recovered with probability tending to one as the sample size increases. We therefore condition the subsequent analysis on the event of correct component-number recovery and focus on the estimation error of the Gaussian mixture parameters and its propagation through the subsequent stages.

Gaussian mixture parameter estimation. Taking the true component numbers $K_t^{(w)}$ as known, we next consider the estimation of the Gaussian mixture parameters. For a Gaussian mixture distribution

$$\gamma = \sum_{k=1}^{K} \pi_k \nu_k, \qquad \nu_k = \mathcal{N}(m_k, \Sigma_k),$$

define

$$\theta_k = \left(m_k^\top, \sigma_k(i,j) : i = 1, \ldots, d,\ j = 1, \ldots, i\right)^\top,$$

where $\sigma_k(i,j)$ denotes the (i,j)-th entry of Σ_k. The full Gaussian mixture parameter vector is

$$\theta = (\pi_1, \ldots, \pi_K, \theta_1^\top, \ldots, \theta_K^\top)^\top.$$

The full vector θ is retained for the componentwise parameter-error bounds below.

For the theoretical guarantee with fixed component counts, we use the penalized maximum likelihood estimator of Chen & Tan (2009). To specify it, let

$$p\ell(\theta) = \ell_n(\theta) + p_n(\theta),$$

where $\ell_n(\theta)$ denotes the ordinary log-likelihood based on the observed sample, and

$$p_n(\theta) = -\frac{1}{n} \sum_{k=1}^{K} \left(\text{tr}(\Sigma_k^{-1}) + \log|\Sigma_k|\right).$$

Then, the estimator is

$$\widehat{\theta} \in \arg\max_{\theta} p\ell(\theta).$$

Under the regularity conditions in Definition B.2, Theorem 2 of Chen & Tan (2009) gives the asymptotic normality in a nonredundant local parameterization. In particular, after aligning component labels, the full constrained parameter vector satisfies

$$\left\|\widehat{\theta}_t^{(w)} - \theta_t^{(w)}\right\| = O_p(n^{-1/2}).$$

To quantify the parameter estimation errors, define

$$E_\mu^{(w)} := \max_{t,k} \left\|\widehat{\mu}_{t,k}^{(w)} - \mu_{t,k}^{(w)}\right\|_2,$$

$$E_\Sigma^{(w)} := \max_{t,k} \left\|\widehat{\Sigma}_{t,k}^{(w)} - \Sigma_{t,k}^{(w)}\right\|_2,$$

$$E_\pi^{(w)} := \max_{t,k} \left|\widehat{\pi}_{t,k}^{(w)} - \pi_{t,k}^{(w)}\right|.$$

Lemma H.1 (Gaussian mixture parameter recovery)**.** *For known K, under Definition B.2 and the penalty and regularity conditions of Chen & Tan (2009), the penalized estimator described above satisfies*

$$E_\mu^{(w)}, E_\Sigma^{(w)}, E_\pi^{(w)} = O_p(n^{-1/2}).$$

Proof. By the definition of $\theta_t^{(w)}$ and the symmetry of the covariance matrices,

$$\begin{aligned}\left\|\widehat{\theta}_t^{(w)} - \theta_t^{(w)}\right\|^2 \geq & \sum_{k=1}^{K_t^{(w)}} \left|\widehat{\pi}_{t,k}^{(w)} - \pi_{t,k}^{(w)}\right|^2 \\ & + \sum_{k=1}^{K_t^{(w)}} \left\|\widehat{\mu}_{t,k}^{(w)} - \mu_{t,k}^{(w)}\right\|_2^2 \\ & + \frac{1}{2}\sum_{k=1}^{K_t^{(w)}} \left\|\widehat{\Sigma}_{t,k}^{(w)} - \Sigma_{t,k}^{(w)}\right\|_2^2.\end{aligned}$$

Thus, for each t,

$$\max_k \left|\widehat{\pi}_{t,k}^{(w)} - \pi_{t,k}^{(w)}\right|, \max_k \left\|\widehat{\mu}_{t,k}^{(w)} - \mu_{t,k}^{(w)}\right\|_2, \max_k \left\|\widehat{\Sigma}_{t,k}^{(w)} - \Sigma_{t,k}^{(w)}\right\|_2 = O_p(n^{-1/2}).$$

Since the number of periods is fixed, taking the maximum over t gives

$$E_\pi^{(w)}, E_\mu^{(w)}, E_\Sigma^{(w)} = O_p(n^{-1/2}).$$

□

Perturbation of Gaussian coupling maps and costs. Define the perturbation errors of the Gaussian coupling maps and transport costs by

$$E_A^{(w)} := \max_{t,k,\ell} \left\|\widehat{A}_{t,k,\ell}^{(w)} - A_{t,k,\ell}^{(w)}\right\|_2,$$

and

$$E_C^{(w)} := \max_{t,k,\ell} \left|\widehat{C}_{t,k,\ell}^{(w)} - C_{t,k,\ell}^{(w)}\right|.$$

Lemma H.2 (Perturbation of Gaussian coupling maps and costs)**.** *Under the regularity conditions in Definition B.2, as $n \to \infty$ with $E_\mu^{(w)}, E_\Sigma^{(w)} \to 0$,*

$$E_A^{(w)} = O\left(E_\Sigma^{(w)}\right),$$

and

$$E_C^{(w)} = O\left(E_\mu^{(w)} + E_\Sigma^{(w)}\right).$$

Proof. We use the following perturbation bounds. For matrices X, Y and perturbations $\Delta X, \Delta Y$ satisfying

$$\|\Delta X\|_2, \|\Delta Y\|_2 \leq E,$$

we have

$$\|(X + \Delta X)(Y + \Delta Y) - XY\|_2 \leq 2\max\{\|X\|_2, \|Y\|_2\}E + o(E).$$

For $X \succ 0$ and $\|\Delta X\|_2 \leq E$,

$$\|(X + \Delta X)^{-1} - X^{-1}\|_2 \leq \frac{1}{\lambda_{\min}(X)^2}E + o(E),$$

$$\|(X + \Delta X)^{1/2} - X^{1/2}\|_2 \leq \frac{1}{2\lambda_{\min}(X)^{1/2}}E + o(E),$$

$$\|(X + \Delta X)^{-1/2} - X^{-1/2}\|_2 \leq \frac{1}{2\lambda_{\min}(X)^{3/2}}E + o(E).$$

Under Definition B.2,

$$\lambda_{\min}(\Sigma_{t,k}^{(w)}) > 0.$$

Applying the above perturbation bounds successively to

$$A_{t,k,\ell}^{(w)} = \left(\Sigma_{t,k}^{(w)}\right)^{-1/2} \left[\left(\Sigma_{t,k}^{(w)}\right)^{1/2} \Sigma_{t+1,\ell}^{(w)} \left(\Sigma_{t,k}^{(w)}\right)^{1/2}\right]^{1/2} \left(\Sigma_{t,k}^{(w)}\right)^{-1/2}.$$

gives

$$\left\|\widehat{A}_{t,k,\ell}^{(w)} - A_{t,k,\ell}^{(w)}\right\|_2 = O\Big(\left\|\widehat{\Sigma}_{t,k}^{(w)} - \Sigma_{t,k}^{(w)}\right\|_2 + \left\|\widehat{\Sigma}_{t+1,\ell}^{(w)} - \Sigma_{t+1,\ell}^{(w)}\right\|_2\Big).$$

Since T and $K_t^{(w)}$ are fixed, taking the maximum over t, k, ℓ preserves the order, and hence

$$E_A^{(w)} = O\Big(E_\Sigma^{(w)}\Big).$$

Likewise, consider the transport cost

$$\begin{aligned} C_{t,k,\ell}^{(w)} = \; & \left\|\mu_{t,k}^{(w)} - \mu_{t+1,\ell}^{(w)}\right\|^2 \\ & + \mathrm{Tr}\left(\Sigma_{t,k}^{(w)} + \Sigma_{t+1,\ell}^{(w)} - 2\left[(\Sigma_{t,k}^{(w)})^{1/2}\Sigma_{t+1,\ell}^{(w)}(\Sigma_{t,k}^{(w)})^{1/2}\right]^{1/2}\right). \end{aligned}$$

The squared mean term is Lipschitz in the mean parameters on the bounded parameter space, while the covariance terms are controlled by the above matrix perturbation bounds. In particular, since d is fixed, the trace terms can be bounded in terms of the corresponding spectral norm perturbations. Consequently,

$$\begin{aligned} \left|\widehat{C}_{t,k,\ell}^{(w)} - C_{t,k,\ell}^{(w)}\right| = O\Big(& \left\|\widehat{\mu}_{t,k}^{(w)} - \mu_{t,k}^{(w)}\right\|_2 + \left\|\widehat{\mu}_{t+1,\ell}^{(w)} - \mu_{t+1,\ell}^{(w)}\right\|_2 \\ & + \left\|\widehat{\Sigma}_{t,k}^{(w)} - \Sigma_{t,k}^{(w)}\right\|_2 + \left\|\widehat{\Sigma}_{t+1,\ell}^{(w)} - \Sigma_{t+1,\ell}^{(w)}\right\|_2\Big). \end{aligned}$$

Again, since T and $K_t^{(w)}$ are fixed, taking the maximum over t, k, ℓ preserves the order. Thus,

$$E_C^{(w)} = O\Big(E_\mu^{(w)} + E_\Sigma^{(w)}\Big).$$

□

Perturbation of the sense transport plans. We next consider the perturbation of the optimal sense transport plans induced by estimation errors in the sense prevalences and the transport costs. Define

$$E_Q^{(w)} := \max_t \left\|\widehat{Q}_t^{(w)} - Q_t^{(w)}\right\|_{\mathrm{F}}.$$

The perturbation $E_\pi^{(w)}$ corresponds to the perturbation of the right-hand side of the equality constraints in equation 2, while $E_C^{(w)}$ corresponds to the perturbation of its objective coefficients.

The following lemma is based on a standard basis-stability argument for linear programs. The cost perturbation $E_C^{(w)}$ does not appear in the leading term of the bound, but it is essential for ensuring that the support of the optimal transport plan $Q_t^{(w)}$ remains unchanged under small perturbations. Once the support is fixed, the perturbation of $Q_t^{(w)}$ is determined solely by the perturbation of the marginal prevalences.

Lemma H.3 (Perturbation of the sense transport plans)**.** *Suppose that the optimal sense transport plan $Q_t^{(w)}$ is unique and satisfies*

$$\left|\mathrm{supp}\left(Q_t^{(w)}\right)\right| = K_t^{(w)} + K_{t+1}^{(w)} - 1.$$

Then, as $n \to \infty$ with $E_C^{(w)}, E_\pi^{(w)} \to 0$,

$$E_Q^{(w)} = O\Big(E_\pi^{(w)}\Big). \tag{41}$$

Proof. For each consecutive period pair $(t, t+1)$, the equality constraints in equation 2 have rank

$$K_t^{(w)} + K_{t+1}^{(w)} - 1,$$

since one of the marginal constraints is redundant. Let $b_t^{(w)}$ denote the right-hand side obtained by removing one redundant marginal constraint and stacking the remaining marginal probabilities.

Let $A_t^{(w)}$ denote the corresponding constraint matrix, and let $q_t^{(w)}$ denote the vector obtained by stacking all entries of $Q_t^{(w)}$. Then the equality constraints can be written as

$$A_t^{(w)} q_t^{(w)} = b_t^{(w)}.$$

By the support condition,

$$\left|\operatorname{supp}\left(Q_t^{(w)}\right)\right| = K_t^{(w)} + K_{t+1}^{(w)} - 1,$$

the columns of $A_t^{(w)}$ corresponding to the nonzero entries of $Q_t^{(w)}$ form a nonsingular square submatrix

$$B_t^{(w)} \in \mathbb{R}^{(K_t^{(w)}+K_{t+1}^{(w)}-1)\times(K_t^{(w)}+K_{t+1}^{(w)}-1)}.$$

Thus, the nonzero entries of $q_t^{(w)}$ are given by

$$q_{t,B}^{(w)} = \left(B_t^{(w)}\right)^{-1} b_t^{(w)}.$$

The uniqueness of $Q_t^{(w)}$ together with the support condition implies that the corresponding basic feasible solution is nondegenerate and that all nonbasic reduced costs are strictly positive. Therefore, for sufficiently small $E_C^{(w)}$ and $E_\pi^{(w)}$, the same basis $B_t^{(w)}$ remains optimal. Consequently, the basic entries of $\widehat{q}_t^{(w)}$ are given by

$$\widehat{q}_{t,B}^{(w)} = \left(B_t^{(w)}\right)^{-1} \widehat{b}_t^{(w)},$$

and hence

$$\widehat{q}_{t,B}^{(w)} - q_{t,B}^{(w)} = \left(B_t^{(w)}\right)^{-1} \left(\widehat{b}_t^{(w)} - b_t^{(w)}\right).$$

Since the nonbasic variables remain zero,

$$\begin{aligned}
\left\|\widehat{Q}_t^{(w)} - Q_t^{(w)}\right\|_{\mathrm{F}} &= \left\|\widehat{q}_{t,B}^{(w)} - q_{t,B}^{(w)}\right\|_2 \\
&= \left\|\left(B_t^{(w)}\right)^{-1} \left(\widehat{b}_t^{(w)} - b_t^{(w)}\right)\right\|_2 \\
&\le \left\|\left(B_t^{(w)}\right)^{-1}\right\|_2 \left\|\widehat{b}_t^{(w)} - b_t^{(w)}\right\|_2 \\
&= O\left(E_\pi^{(w)}\right).
\end{aligned}$$

Taking the maximum over t yields

$$E_Q^{(w)} = O\left(E_\pi^{(w)}\right).$$

□

Perturbation of the induced second-moment operators. Finally, we propagate the preceding estimation errors to the induced second-moment operators. For $\bullet \in \{\mathrm{full}, \mathrm{mean}, \mathrm{dev}\}$, define

$$E_{M,\bullet}^{(w)} := \max_{t,s} \left\|\widehat{\mathbf{M}}_\bullet^{(w)}(t,s) - \mathbf{M}_\bullet^{(w)}(t,s)\right\|_2.$$

Lemma H.4 (Gaussian second-moment operator perturbation)**.** *As $n \to \infty$ with*

$$E_Q^{(w)},\ E_\mu^{(w)},\ E_\Sigma^{(w)},\ E_A^{(w)} \to 0,$$

we have

$$E^{(w)}_{M,\mathrm{mean}} = O\Big(E^{(w)}_{Q} + E^{(w)}_{\mu}\Big),$$
$$E^{(w)}_{M,\mathrm{dev}} = O\Big(E^{(w)}_{Q} + E^{(w)}_{\Sigma} + E^{(w)}_{A}\Big),$$

and

$$E^{(w)}_{M,\mathrm{full}} = O\Big(E^{(w)}_{Q} + E^{(w)}_{\mu} + E^{(w)}_{\Sigma} + E^{(w)}_{A}\Big).$$

Proof. By equation 8, the mean and deviation operators are finite sums of terms weighted by the sense-transition probabilities. For the mean component, $\mathbf{C}^{(w)}_{\mathrm{mean}}(t,s;k,\ell)$ is determined by the corresponding sense means. Therefore, we obtain

$$E^{(w)}_{M,\mathrm{mean}} = O\Big(E^{(w)}_{Q} + E^{(w)}_{\mu}\Big).$$

For the deviation component, the covariance terms contribute $O(E^{(w)}_{\Sigma})$. Moreover, the cross-covariance terms involve $\Sigma^{(w)}_{t,k}$ multiplied by conditional expectations of finite products of the coupling maps $A^{(w)}$. Since the number of periods is fixed, the perturbation of these products is $O(E^{(w)}_{A})$. Combining these perturbations with the perturbation of the transition probabilities gives

$$E^{(w)}_{M,\mathrm{dev}} = O\Big(E^{(w)}_{Q} + E^{(w)}_{\Sigma} + E^{(w)}_{A}\Big).$$

Finally, by equation 6,

$$E^{(w)}_{M,\mathrm{full}} = O\Big(E^{(w)}_{Q} + E^{(w)}_{\mu} + E^{(w)}_{\Sigma} + E^{(w)}_{A}\Big).$$

□

Proof of distance statistical recovery. The preceding results establish the estimation rate for the Gaussian mixture parameters and propagate it through the coupling construction, the sense transport plans, and the induced second-moment operators. Combining these results with the second-moment operator-to-distance perturbation bound of Ezoe & Hisano (2026) yields the desired distance recovery rates.

Proof of Theorem 3.7. Under Assumptions 2.1 and 3.6, combining Lemma H.1, Lemma H.2, Lemma H.3, and Lemma H.4 yields

$$E^{(w)}_{M,\bullet} = O_p\Big((n^{(w)})^{-1/2}\Big).$$

We next relate the second-moment operator perturbation to the induced distance perturbation. The same argument as in Ezoe & Hisano (2026) applies, without the need for an orthogonal rotation between the population and estimated second-moment operators.

For the word-local mean-path basis, under Assumption 3.3, the same argument as in Proposition C.4 of Ezoe & Hisano (2026) yields

$$E^{(w)}_{u,r} := \min_{\epsilon\in\{-1,1\}} \Big\| \widehat{\boldsymbol{u}}^{(w)}_{r} - \epsilon \boldsymbol{u}^{(w)}_{r} \Big\|_2 \le \frac{2^{3/2}T}{\delta} E^{(w)}_{M,\mathrm{mean}}$$

for every $r \in [R^{(w)}]$. Hence,

$$E^{(w)}_{u,r} = O_p\Big((n^{(w)})^{-1/2}\Big),$$

since $T/\delta = O(1)$.

Moreover, the same argument as in Proposition C.7 of Ezoe & Hisano (2026) yields

$$\max_{t,s\in[T]} \Big| \widehat{d}^{(w)}_{\bullet,\mathrm{tr}}(t,s)^2 - d^{(w)}_{\bullet,\mathrm{tr}}(t,s)^2 \Big| \le d E^{(w)}_{M,\bullet},$$

and, for every $r \in [R^{(w)}]$,

$$\max_{t,s\in[T]} \left| \widehat{d}^{(w)}_{\bullet,r}(t,s)^2 - d^{(w)}_{\bullet,r}(t,s)^2 \right| \le E^{(w)}_{M,\bullet} + 2E^{(w)}_{u,r} \max_{t,s\in[T]} \|\mathbf{M}^{(w)}_{\bullet}(t,s)\|.$$

Since the population quantities are fixed, combining these bounds with

$$E^{(w)}_{M,\bullet} = O_p\Big((n^{(w)})^{-1/2}\Big)$$

gives

$$\max_{t,s\in[T]} \left| \widehat{d}^{(w)}_{\bullet,\rho}(t,s)^2 - d^{(w)}_{\bullet,\rho}(t,s)^2 \right| = O_p\Big((n^{(w)})^{-1/2}\Big)$$

for $\rho \in \{\mathrm{tr}, 1, \ldots, R^{(w)}\}$, which proves the theorem. □

I SUPPLEMENTARY EXPERIMENTS

The supplementary experiments follow the progression of Section 4. The synthetic study tests finite-sample recovery of the Gaussian estimator. DWUG provides standard LSC validation and examines the mean–deviation decomposition. Janus isolates multi-period profile recovery and compositional coherence. The Court analysis documents the natural text corpus and shows how numerical transition and mode readouts lead to inspectable passages. Each subsection supplies construction details, robustness results, or audit evidence for a claim made in the main text.

I.1 SYNTHETIC MIXTURE PROCESS AND SUPPLEMENTARY RECOVERY

This experiment tests finite-sample recovery under the Gaussian conditions of Theorem 3.7. It is a controlled check of the estimator rather than an additional application domain.

Population and estimation. We fix a three-period process in $d = 2$ with two components per period. Prevalences and centers are

$$\pi_1 = (.55, .45), \quad \mu_1 = \begin{pmatrix} -2.4 & 0 \\ 2.2 & 0 \end{pmatrix},$$
$$\pi_2 = (.40, .60), \quad \mu_2 = \begin{pmatrix} -1.9 & .35 \\ 2.7 & -.25 \end{pmatrix},$$
$$\pi_3 = (.28, .72), \quad \mu_3 = \begin{pmatrix} -1.3 & .70 \\ 3.1 & .45 \end{pmatrix}.$$

The full covariance matrices are

t	$\Sigma_{t,1}$	$\Sigma_{t,2}$
1	$\begin{pmatrix} .30 & 0 \\ 0 & .30 \end{pmatrix}$	$\begin{pmatrix} .35 & 0 \\ 0 & .25 \end{pmatrix}$
2	$\begin{pmatrix} 1.60 & .20 \\ .20 & .40 \end{pmatrix}$	$\begin{pmatrix} .45 & -.15 \\ -.15 & 1.80 \end{pmatrix}$
3	$\begin{pmatrix} 3.20 & .35 \\ .35 & .50 \end{pmatrix}$	$\begin{pmatrix} .55 & .20 \\ .20 & 3.50 \end{pmatrix}$.

Thus the population contains prevalence shifts, center displacement, and anisotropic within component change. Mean and deviation account for 0.873 and 0.127 of adjacent full path energy.

For each $n \in \{100, 200, 400, 800, 1600\}$, we draw n independent usages per period and repeat the experiment 50 times. We fit a full-covariance GMM independently in every period with $K = 2$ fixed, covariance regularization 10^{-6}, five initializations, and no access to sampled component labels. Fixing K matches Assumption 3.6. Because mixture labels are arbitrary, fitted components are aligned to the population components only when calculating errors.

Errors and supplementary results. We use

$$E_\pi = \max_{t,k} |\widehat{\pi}_{t,k} - \pi_{t,k}|, \quad E_\mu = \max_{t,k} \|\widehat{\mu}_{t,k} - \mu_{t,k}\|_2, \quad E_\Sigma = \max_{t,k} \|\widehat{\Sigma}_{t,k} - \Sigma_{t,k}\|_2.$$

For $\bullet \in \{\mathrm{full}, \mathrm{mean}, \mathrm{dev}\}$, operator error is

$$E_{M,\bullet} = \max_{t,s} \|\widehat{\mathbf{M}}_{\bullet}(t,s) - \mathbf{M}_{\bullet}(t,s)\|_2.$$

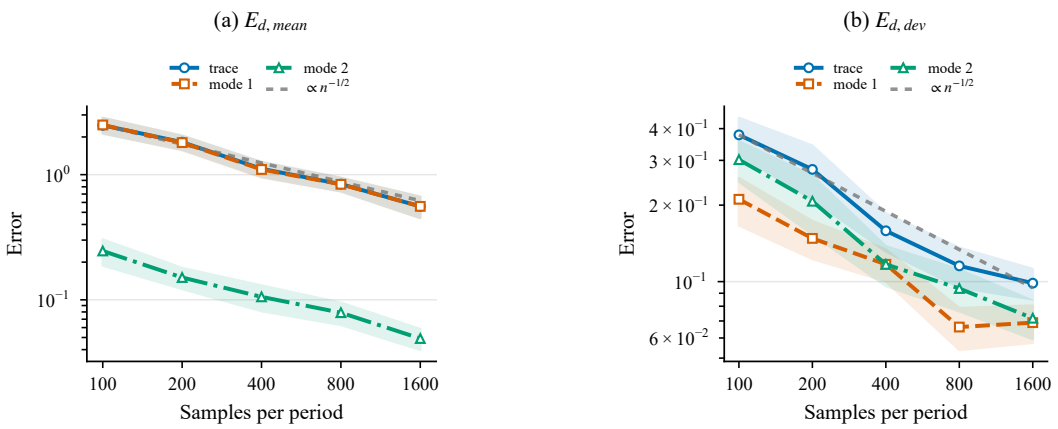

Figure 7: Squared-distance recovery for the (a) mean and (b) deviation processes. Trace and mode distances use the corresponding operator with the estimated word-local basis. Bands are 95% confidence intervals over 50 repetitions.

Squared-distance error is the maximum absolute error over period pairs for either the trace or one of the two word-local modes:

$$E_{d,\bullet} = \max_{t,s} \left| \widehat{d}_{\bullet,\rho}(t,s)^2 - d_{\bullet,\rho}(t,s)^2 \right|.$$

The fitted mode distances use eigenvectors estimated from the fitted mean-path operator. Curves report means and normal approximation 95% confidence intervals across repetitions. Table 4 estimates slopes by ordinary least squares on the five log mean errors.

Table 4: Synthetic finite-sample recovery. The last column reports mean error at $n = 1600$.

Quantity	Log–log slope [95% CI]	Error at $n = 1600$
Prevalence E_π	-0.529 $[-0.599, -0.458]$	0.017
Center E_μ	-0.437 $[-0.565, -0.310]$	0.135
Covariance E_Σ	-0.445 $[-0.542, -0.348]$	0.337
Full operator $E_{M,\mathrm{full}}$	-0.532 $[-0.595, -0.469]$	0.570
Mean operator $E_{M,\mathrm{mean}}$	-0.534 $[-0.606, -0.463]$	0.581
Deviation operator $E_{M,\mathrm{dev}}$	-0.514 $[-0.654, -0.375]$	0.101
Full trace distance	-0.532 $[-0.588, -0.476]$	0.580
Full mode-1 distance	-0.540 $[-0.606, -0.473]$	0.542
Full mode-2 distance	-0.561 $[-0.716, -0.406]$	0.089
Mean trace distance	-0.542 $[-0.615, -0.468]$	0.558
Deviation trace distance	-0.514 $[-0.687, -0.341]$	0.099

The operator, trace-distance, and mode-distance intervals all contain the $n^{-1/2}$ reference rate. Parameter errors also decrease with n, with the center and covariance curves somewhat shallower over this finite range. Figure 7 separates mean- and deviation-process distances.

Pure mechanism checks. We test the decomposition with three populations in which only centers, covariances, or prevalences change. Center movement produces mean energy. Changing prevalences between fixed components also transports mass between their centers and therefore produces mean energy. The one-component covariance process produces deviation energy. With $n = 1600$ usages per period and 20 repetitions, Table 5 shows that CUSP assigns 99.89%, 98.95%, and 99.95% of path energy to the respective planted mechanisms. The remaining 1.05% mean allocation in the covariance-only process is finite-sample estimation error.

I.2 DWUG IMPLEMENTATION AND DECOMPOSITION

The English and German DWUG evaluations (Schlechtweg et al., 2021) use 9,107 and 9,125 XL-LEXEME usage embeddings (Cassotti et al., 2023), respectively. English contains 46 targets with 61–

Table 5: CUSP allocation of path energy in pure-mechanism populations, in percent. Each row changes only the stated population quantity. Entries are means $\pm$ 95% confidence-interval half-widths over 20 repetitions. The theoretically correct term is bold.

Population change	Mean term	Deviation term
Centers move	**99.89 $\pm$ 0.02**	0.11 $\pm$ 0.02
Covariances change	1.05 $\pm$ 0.40	**98.95 $\pm$ 0.40**
Prevalences change	**99.95 $\pm$ 0.01**	0.05 $\pm$ 0.01

100 usages per target–period. German contains 50 targets with 39–100. For every target and period, BIC selects $K \in \{1, \ldots, 5\}$ and diagonal or full covariance, jointly with covariance regularization in $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. Components with fitted weight below 0.01 are excluded, and all fits use seed 0. The DWUG vectors retain 1,024 coordinates after removal of the leading direction; no additional low-dimensional PCA projection is applied. Since the embedding dimension exceeds the number of usages per target–period, unregularized empirical full covariances are rank deficient. Positive covariance regularization makes the fitted Gaussian covariances nonsingular, but does not remove the statistical uncertainty of estimating them in this regime. The Gaussian recovery theorem assumes a correctly specified, fixed-dimensional model and the stated estimation conditions. In the lexical experiments, GMMs are regularized approximations: the decomposition remains exact for the fitted process, but the theorem does not guarantee recovery for normalized lexical data.

For each language, we pool all target–period embeddings, fit one centering and principal-direction transform, and hold it fixed across targets. The primary $k = 1$ representation removes the leading pooled panel direction before ℓ_2 normalization. Across the lexical experiments, PCA fitting uses random seed 0, and preprocessing fitting and transformation use single-threaded BLAS execution to control numerical variability. All methods in Table 1 use these same preprocessed embeddings. CUSP selects its mixture configuration by unsupervised BIC; baseline parameters follow the fixed configurations supplied with the code. These are different parameter-selection rules, not matched hyperparameter-search budgets. The $k = 0$ and $k = 3$ representations are sensitivity checks. Confidence intervals in Table 1 use 3,000 paired bootstrap samples of target words, and Table 6 reports all three preprocessing settings.

Table 6: Spearman correlation with DWUG graded-change judgments for the full, mean, and deviation traces. Each transform is fitted once to the corresponding language panel.

	$k = 0$			$k = 1$			$k = 3$		
Language	Full	Mean	Dev.	Full	Mean	Dev.	Full	Mean	Dev.
English	.772	.762	.673	.746	.742	.681	.762	.757	.706
German	.857	.857	.722	.815	.809	.674	.804	.798	.660

The mean trace remains close to the full trace under every preprocessing choice. The deviation trace is weaker but remains associated with the human score, and is comparatively large for heterogeneous words such as English *bar* and German *Rezeption*. The decomposition therefore identifies within component reorganization that the scalar full trace would otherwise conceal.

I.3 Janus construction and supplementary results

Cassotti & Tahmasebi (2025) generated the released Janus records with Llama 3 conditioned on a lemma, an Oxford English Dictionary sense definition, and a year. The release contains 131,811 sentences, 13,762 sense–year records, and 2,768 lemmas. We select 48 lemmas with four released sense pools and ten sentences per pool, yielding 1,920 fixed sentences. These populate six periods of 40 usages per lemma by sampling with replacement. The resulting 11,520 usage slots are not independent new sentences. Repeated contexts limit source diversity and can affect covariance estimates despite regularization. Sampling with replacement does not itself violate independence, but the enlarged set of usage slots should not be interpreted as an equally large collection of distinct historical observations. Resampling preserves any biases or irregularities in the released pools.

The experiment tests recovery of planted distributional changes and compositional coherence, not historical fidelity or recovery of the released sense identities.

Stable schedules keep the four pool prevalences at $(.25, .25, .25, .25)$. In gradual schedules, the old and new pool prevalences follow $(.90, .72, .54, .36, .18, 0)$ and its reverse. Abrupt schedules use $(.90, .90, .90, 0, 0, 0)$ and its reverse. The two background pools have nominal prevalence $.05$ in changed schedules, with finite counts allocated by largest remainder. Sixteen lemmas are assigned to each schedule. Released sense identifiers determine only these planted prevalences and are discarded before fitting. CUSP receives neither sense identifiers nor cross-period usage links.

We fit one pooled panel $k = 1$ transform to the original $1{,}920$ contexts and freeze it before constructing schedules. A separate diagonal GMM is then fitted to every lemma–period panel, with BIC choosing $K \in \{1, \ldots, 5\}$ and covariance regularization in $\{10^{-5}, 10^{-4}\}$ subject to minimum component weight $.01$ and at least three samples per component.

To test temporal coherence, pairwise OT independently optimizes every period-pair plan (Montariol & Allauzen, 2021; Kishino et al., 2025). For each consecutive triple, we compare its direct plan $q_{t,t+2}$ with the Markov composition of $q_{t,t+1}$ and $q_{t+1,t+2}$ and average the resulting ℓ_1 disagreement. CUSP defines the non-adjacent plan by this composition, so its error is numerical zero by construction. Table 7 separates schedule recovery from the composition check.

Table 7: Janus recovery with unsupervised period local GMMs and pooled panel $k = 1$ preprocessing. Composition error is mean ℓ_1 disagreement.

Schedule	n	Schedule recovery	Pairwise OT	CUSP
Stable	16	Path length 0	9.8×10^{-18}	1.0×10^{-17}
Gradual	16	Profile TV .044	.253	1.8×10^{-17}
Abrupt	16	Peak accuracy 1.000	6.4×10^{-18}	1.1×10^{-17}
Overall	48	$\rho = .954$, AUROC 1.000	.084	1.3×10^{-17}

The pairwise inconsistency is concentrated in gradual replacement, where mass is redistributed over several adjacent steps. Stable schedules and the single-boundary abrupt schedules already compose to numerical precision. Janus therefore does not establish uniform superiority over pairwise OT. It shows accurate recovery of the planted temporal profiles and verifies the coherence added by CUSP when change unfolds over multiple periods.

I.4 Court corpus construction and fitting

The Court panel is built from the May 6, 2024 snapshot of CourtListener's bulk opinion data, maintained by Free Law Project (Free Law Project, 2024), and includes opinions filed from 1950 through 2024. Excerpt identifiers are the `id` values of individual CourtListener opinion records in that snapshot. The target vocabulary contains 60 unigram legal lemmas selected from the tables of contents of West Academic law textbooks.[2] The vocabulary was fixed before model fitting and ranking, and the complete list is included with the released code. The analysis uses eight decade bins from 1950 through 2020. The final bin contains opinions from 2020 through the May 2024 snapshot, while the preceding bins cover ten years each. Equalizing usage counts does not equalize calendar coverage. The reported distances compare pooled period distributions rather than annualized rates of semantic change, so the shorter final bin should be interpreted with this limitation.

Construction proceeds in five stages. First, we locate target-lemma matches and retain a window of approximately 200 words on each side. Up to $5{,}000$ candidate opinions are considered for each lemma–decade. At most 20 occurrences are scanned for an opinion–lemma pair, and at most six passages from that pair can be retained. Second, a fixed LLM prompt receives the marked occurrence and its surrounding window. It retains the occurrence only when the court is defining, applying, limiting, or otherwise construing the lemma as an object of legal reasoning. Captions, repeated procedural text, proper nouns, ordinary uses, and incidental mentions are discarded. The same prompt and decision rule are applied to every lemma and decade. Consequently, the analysis concerns the selected legal-reasoning usages, not changes in a word's prevalence between ordinary discourse and

[2] `https://www.westacademic.com/`

legal reasoning or its unrestricted general-language history. The exact prompts are included with the released code. The collection run used `gpt-5.6-luna` for both this retention decision and the subsequent verbatim extraction. It targeted up to 240 distinct retained opinions per lemma–decade, with a minimum target of 120. Third, a separate prompt copies one contiguous passage verbatim that contains the marked occurrence and completes the relevant reasoning. It does not summarize or rewrite the opinion. Fourth, XL-LEXEME embeds the marked target token in that extracted passage. Finally, if $N_{w,t}$ is the retained count for lemma w in decade t, we sample $n_w = \min_t N_{w,t}$ passages from every decade with seed 0. Each lemma therefore has equal counts across its eight periods, while n_w may differ between lemmas.

The procedure yields 322,222 passages before equalization, and the balanced panel contains 258,480. The full balanced panel is used once to fit the Court centering and leading principal direction. The same fitted $k = 1$ transform is then applied to every lemma and decade. For each lemma–decade, BIC selects a diagonal Gaussian mixture with $K \in \{1, \ldots, 5\}$ and minimum fitted weight .02. Of the 480 fits, 471 select $K = 5$, six select $K = 4$, two select $K = 3$, and one selects $K = 2$. Because BIC often selects the upper bound, the component cap determines the resolution of the fitted representation. The resulting components describe recurring usage structure and are not claimed to recover a unique inventory of linguistic senses. Alternative caps may change component assignments and the allocation between mean and deviation terms, while the decomposition remains exact for each fitted process.

Across panel fitted ABTT settings $k = 0, \ldots, 4$, a sensitivity analysis using a common $z \geq 1.1$ screen places *privacy* among the top five lemmas by enrichment and the top two by secondary-mode temporal separation. Its largest change peaks in either 1950→1960 or 1960→1970. Thus its prominence and multi-period character persist across these preprocessing choices, although the decade of its largest peak varies.

The released materials include the exact retention and extraction prompts, fixed vocabulary, configuration files, frozen preprocessing and model outputs, and scripts for reproducing the reported geometry, rankings, passage retrieval, and figures. The 100 MB submission limit prevents us from bundling the complete Court panel. We therefore provide the analysis code and configurations, the frozen transform fitted on the full panel, processed inputs for the two main-text lemmas, and numerical outputs for all 60 lemmas. The released scripts reproduce the reported Court figures and rankings and rerun the main text analyses. This subsection documents the corpus construction and fitting procedure for rebuilding the full panel from the cited CourtListener source. We will publicly release the complete balanced Court text and embedding panels upon acceptance. Table 8 summarizes the resulting corpus and balanced panel.

Table 8: Court corpus and balanced analysis-panel sizes. Per-decade counts are calculated after equalization within each lemma.

Quantity	Count
Distinct court opinions represented	100,473
Target lemmas	60
Retained passages before equalization	322,222
Passages in the balanced analysis panel	258,480
Median passages per lemma–decade	531
Range of equalized counts across lemmas	294–911
Passages per decade for *privacy*	651
Lemmas meeting the standardized-change threshold	53

I.5 Court rankings and passage analysis

This subsection follows the Court analysis in the order in which every result is selected. The within-lemma standardized profile first identifies an adjacent decade that is unusually large relative to the lemma's own history. At that peak, $E_{\max}$ identifies the component movement whose share of change is largest relative to its transported mass. The word-local analysis then asks whether the leading modes peak at different transitions. Representative source and target passages make these numerically selected component movements inspectable.

All numerical selections precede passage inspection. After the analysis fixes a component movement, each usage is assigned to the fitted component with the highest posterior probability. Representative passages within the selected source and target components are then ranked independently using a fixed score based on proximity to the component center and a surface form check favoring exact matches and forms beginning with the target lemma. The retrieved passages provide representative source and target usages for the attributed movement, and CourtListener identifiers link every excerpt to its underlying opinion. Named decisions situate the retrieved language historically.

Peak transition attribution by maximum enrichment. At a lemma's selected peak, combine the mean and deviation contributions as

$$a_{k\ell} = q_{k\ell}\,\mathrm{tr}(\mathbf{C}_{\mathrm{mean}}(k,\ell) + \mathbf{C}_{\mathrm{dev}}(k,\ell))\,, \qquad \alpha_{k\ell} = a_{k\ell}/d_t^2.$$

We define

$$E_{\max}^{(w)} = \max_{q_{k\ell}>0} \frac{\alpha_{k\ell}}{q_{k\ell}},$$

which identifies the component movement whose contribution share is most disproportionate to its transported-mass share. Since $a_{k\ell} = q_{k\ell}\,\mathrm{tr}(\mathbf{C}_{\mathrm{mean}} + \mathbf{C}_{\mathrm{dev}})$, the factor $q_{k\ell}$ cancels in the ratio. Thus $E_{\max}$ measures displacement contribution per unit of transported mass, normalized by the total squared change; there is no uncancelled inverse-mass factor. Numerically, the implementation excludes pairs with $q_{k\ell} \le 10^{-15}$ and floors the total squared change at 10^{-15}. This is a per-unit-mass attribution statistic, not a significance measure, and rare component pairs can still have uncertain estimated displacements. Because mixtures are fitted separately by period, k and ℓ are period local labels. Equal or unequal indices have no intrinsic semantic interpretation. Table 9 reports the ten highest values. The complete ranking of all 53 lemmas passing this threshold and the corresponding figure are included in the supplementary material.

Table 9: Ten highest maximum-enrichment transitions under panel-fitted $k = 1$. Contribution α is the selected pair's share of variation at the selected peak.

Rank	Lemma	Peak transition (z)	Pair	α	$E_{\max}$
1	privacy	1960→1970 (1.43)	2→3	.371	15.10
2	search	1980→1990 (1.55)	3→4	.690	13.67
3	standing	1960→1970 (1.63)	3→4	.366	13.23
4	warrant	1950→1960 (2.26)	1→0	.335	11.11
5	trespass	2000→2010 (1.74)	1→4	.390	10.51
6	exercise	1960→1970 (1.36)	0→1	.587	10.33
7	effect	1970→1980 (1.53)	0→2	.239	10.17
8	due	1960→1970 (1.22)	3→0	.158	9.25
9	official	1950→1960 (2.06)	2→3	.119	7.60
10	liberty	1970→1980 (1.50)	2→0	.169	7.16

The following paragraphs examine the five highest ranked lemmas in Table 9. This selection is determined by the numerical ranking rather than by the readability of the retrieved passages. The complete ranking and component assignments are included in the supplementary material.

Privacy. The source says that "Professor Prosser classified these types of cases into four categories" (opinion `1179312`), while the target says that warrantless inspection poses "only a minimal threat to justifiable expectations of privacy" (opinion `1602980`). The pair moves from Prosser-style tort classification (Prosser, 1960) toward post-*Katz* expectation-of-privacy language (Supreme Court of the United States, 1967a). The preceding decade's leading pair moves from privacy as "a direct wrong of a personal character" (opinion `2604478`) toward the same Prosser component, making tort articulation, systematic classification, and expectation-of-privacy language successive rather than one retrospectively chosen contrast.

Search. The source applies a workers' compensation rule requiring a "conscientious work search" and asks whether an unsuccessful search reflects inability to work (opinion `1139619`). Targets describe a "pat-down search" and a "search of a jacket" (opinion `197788`). The transition is from *search* as a legally required effort to obtain work to physical inspection. The source is retained because work search is itself part of the legal disability test, not incidental job-seeking narration.

Standing. A source states that "Standing trees are a part of the realty" (opinion `1331324`). Targets ask whether a plaintiff has the "requisite standing to maintain the action" and describe standing as

a federal-court question under Article III (opinions 1513512, 357908). The pair separates the adjectival property sense from the jurisdictional doctrine.

Warrant. Sources state that facts "may warrant an inference of negligence" or "warrant an allowance of punitive damages" (opinions 2083592, 1437656). Targets discuss "probable cause for the issuance of a search warrant" and a "search warrant" (opinions 2216781, 2167941). Both pairs move from the verb licensing a conclusion to the noun denoting a judicial instrument.

Trespass. Sources plead a common-law trespass claim or state that "a cognizable claim for trespass occurs when personal property ... is injured or taken" (opinions 2421595, 1557045). Targets describe a "trespass-to-try-title action" or an "action in trespass to try title" (opinions 4293854, 1889942). The retrieved language supports redistribution between distinct legal constructions of *trespass*, without assigning it to a landmark decision.

Temporal separation of word-local modes. Let e_r be mode r's share of word-local mean-path energy. Among the five modes with the largest path-energy shares, order the two leading shares as $e_{(1)} \geq e_{(2)}$ and denote their peak decades by $\tau_{(1)}$ and $\tau_{(2)}$. For lemmas passing the standardized-change threshold, define

$$S_2 = e_{(2)} \log(1 + E_{\max})$$

when $\tau_{(1)} \neq \tau_{(2)}$. If the two leading modes peak together, the profile is classified as synchronized and receives no S_2 rank. Multiplying the second mode's energy share by a monotone transform of $E_{\max}$ favors histories with both a substantial second direction and a clearly attributable movement. The score is fixed only to prioritize passage analyses of temporally separated histories. It does not estimate the model, test significance, or determine whether change occurred. Table 10 reports the ten highest values.

Table 10: Ten highest secondary-mode temporal-separation scores. $e_{(1)}$ and $e_{(2)}$ are the two leading shares among the five highest-energy mean-path modes.

Rank	Lemma	S_2	$e_{(1)}$	$e_{(2)}$	Mode peaks
1	privacy	.837	.391	.301	1960→1970 / 1950→1960
2	standing	.788	.406	.297	1960→1970 / 1970→1980
3	effect	.721	.452	.299	1990→2000 / 1970→1980
4	trespass	.707	.359	.289	1970→1980 / 1990→2000
5	taking	.690	.387	.360	2010→2020 / 2000→2010
6	due	.683	.371	.294	1980→1990 / 1990→2000
7	search	.655	.481	.244	1980→1990 / 1990→2000
8	cause	.610	.387	.318	1950→1960 / 1970→1980
9	content	.581	.372	.308	1960→1970 / 2000→2010
10	burden	.567	.328	.308	1970→1980 / 1990→2000

Privacy, standing, trespass, and search also appear in the enrichment ranking, where their selected transitions are grounded by the passage analyses above. The high positions of *effect* and *taking* are consistent with broad, polysemous histories. We retain their quantitative positions without assigning them a single landmark-case narrative. Complete values for all 41 separated profiles are included in the supplementary material.

Privacy modes. Mode 2 (30%) peaks at 1950→1960. Its leading pair moves from a right of privacy "to be subordinated" to the public interest (opinion 1424064) toward an "actionable invasion" of privacy, and the action "sounds in tort" (opinion 1741636). Another target locates eavesdropping privacy in the Fourth Amendment (opinion 255989). Modes 3 and 4 (15% and 11%) peak at 1960→1970 and are led by the Prosser-to-expectations transition. Mode 4 also connects *Warden v. Hayden*'s statement (Supreme Court of the United States, 1967b) that the Fourth Amendment protects the right of privacy, "rather than any interest in the property seized" (opinion 1982603), to "no legitimate expectation of privacy" in a cargo hold (opinion 1378708). Mode 1 (39%) peaks at 1960→1970 and remains strong at 1970→1980. At the later step, one pair moves from the "rights of privacy of practitioners of the healing arts and of their patients" (opinion 2597129) toward statutory balancing of "personal privacy" against the "public interest" (opinion 1677943). Another connects family privacy under *Griswold* and *Roe* (opinion 2142114) to "no legitimate expectation of privacy" in plain view (opinion 2070690) (Supreme Court of the United States, 1965; 1973).

The modes therefore recover staggered and partly concurrent tort, constitutional, informational, and search-privacy usages rather than one undifferentiated event.

Expectation modes. The three leading modes carry 48%, 26%, and 18% of mean-path energy and all peak at $1960 \to 1970$. Their deterministically selected passages move from "an expectation of pay or compensation" and contributions made "without expectation of services or direct benefits" (opinions `2226138`, `1636296`) toward "a reasonable expectation of privacy infringed by the search and seizure" and an area "protected by an expectation of privacy" (opinions `1365600`, `1311660`). The contrast with *privacy* therefore comes from a fixed temporal criterion rather than retrospective case selection.