LEARNING PARTIAL SEQUENCE ALIGNMENTS THROUGH SHARED GROUP STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Related sequences often observe different portions of a shared underlying process. Aligning them requires identifying both overlapping regions and correspondences within those regions. Yet individually plausible pairwise matches can be mutually incompatible when considered across a group. We introduce a supervised groupwise partial alignment framework that makes collective correspondence compatibility an explicit learning objective. The framework processes sequence groups jointly to estimate overlap boundaries and monotone alignments, organizing their directed soft correspondences into a common relation. Relations expressible as inner products of shared representations have a positive-semidefinite Gram structure. We therefore penalize the directed group relation's distance to the symmetric positive-semidefinite cone as a soft compatibility constraint. Experiments on CMU ARCTIC speech and simulated stratigraphic sequences with relative-geological-time ground truth show lower mean alignment error and more accurate overlap estimation than the evaluated classical, differentiable, and learned baselines. Ablations show that the group objective further improves direct and composed correspondence accuracy in both domains. These results support learning partial correspondences according to both their pairwise accuracy and their compatibility with the other alignments in the group.

1 INTRODUCTION

Sequence alignment makes it possible to compare recordings and measurements through the content they share, rather than through their sampling indices. Corresponding events may occur at different positions, last for different durations, or produce different local responses. Records can also begin and end at different stages of a process, leaving only part of their content in common. The task is therefore to determine both where correspondence exists and how the shared portions should be matched. Dynamic programming provides a structured solution to temporal warping in speech and seismic imaging (Sakoe & Chiba, 1978; Hale, 2013), while differentiable formulations allow alignment to participate in learning (Cuturi & Blondel, 2017). Partial alignment methods further account for observations that have no counterpart (Dvornik et al., 2021). Yet repeated patterns can leave several plausible pairwise paths even after temporal order is taken into account, especially when overlap boundaries are uncertain. When related observations form a sequence group, independent pairwise alignment discards additional evidence: correspondences involving other observations can help distinguish locally plausible alternatives.

The sequences form a group as observations or realizations of a common progression of content or events, not merely because their samples look similar. This progression can leave corresponding changes even when local responses are shifted, altered, or absent. Identifying a change in A with one in B constrains how both relate to C wherever they share that content (Figure 1). What should be preserved is a common account of these relationships, not identical signals.

One way to use these relations is to infer a joint alignment over all sequences. A dense K-way dynamic program has L^K states for K sequences of length L. Existing groupwise methods such as GCTW and RF-DTAN avoid this construction through alternative formulations (Zhou & De la Torre, 2016; Shapira Weber & Freifeld, 2023). We represent each ordered pair by a distribution over monotone partial paths and couple their learning through a relation assembled across the sequence group. Computing paths separately need not mean learning them independently.

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.

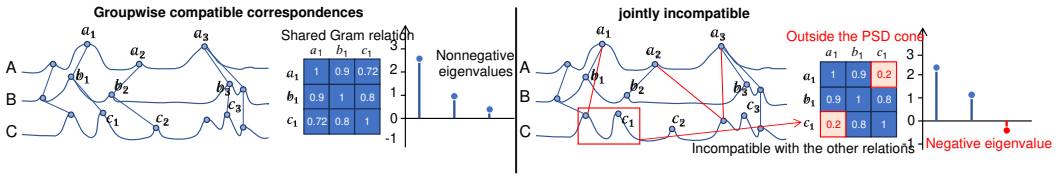


Figure 1: **Pairwise plausibility does not imply groupwise compatibility.** Both schematic matrices are symmetric, and every two-position principal submatrix is positive semidefinite (PSD). Only the left matrix admits a shared Gram representation across all three positions; the right has a negative eigenvalue. This motivates a soft Gram-compatibility prior, not a requirement that valid alignments be exact symmetric PSD relations.

Pairwise plausibility does not imply groupwise compatibility, even when opposite directions agree. Reusing a representation for each position across all of its relations yields a positive semidefinite (PSD) Gram matrix of inner products. Symmetry alone is insufficient: the symmetric relation in Figure 1 (right) has a negative eigenvalue and admits no such shared representation. Related Gram and spectral formulations have been used in joint correspondence estimation and learned matching (Chen et al., 2014; Wang et al., 2022). We use this structure as a soft prior on the group relation assembled from directed correspondence marginals.

We introduce a supervised groupwise partial alignment framework with two objectives. A group-conditioned network scores each ordered sequence pair. The *Partial Alignment Likelihood* (PAL) objective supervises complete monotone partial paths, jointly learning overlap boundaries and internal correspondences. The resulting path distributions yield directed correspondence marginals that retain matching and overlap uncertainty and form a group relation. The squared distance of this relation to the symmetric PSD cone defines the *Group Relational Compatibility* (GRC) objective, which couples predictions across pairs. The relation guides learning; inference decodes each directed alignment on its own lattice using group-conditioned scores.

Our contributions are:

- **Structured partial-path learning.** We introduce a supervised groupwise partial alignment framework that represents each ordered pair by a distribution over monotone partial paths. PAL jointly learns overlap boundaries and within-overlap correspondences. Group-conditioned scores incorporate evidence from other sequences, while decoding retains efficient pairwise dynamic programming.
- **Group relational compatibility.** GRC assembles directed path occupancy marginals into a common relation, preserving uncertainty in both overlap participation and matching. Its squared distance to the symmetric PSD cone defines a soft compatibility objective that couples alignment learning across pairs. The relation guides learning without imposing a shared timeline or replacing the pairwise outputs.
- **Evidence for relational learning.** Experiments on CMU ARCTIC speech and simulated stratigraphic sequences show gains in alignment accuracy and shared-region recovery over the evaluated pairwise and groupwise baselines. Controlled comparisons distinguish GRC's contribution from feature-level group interaction. GRC improves direct and composed correspondence accuracy and reduces disagreement between the two routes in both domains.

2 Related Work

2.1 Pairwise and Partial Sequence Alignment

Dynamic time warping (DTW) formulates pairwise temporal alignment as a minimum-cost monotone path (Sakoe & Chiba, 1978). Subsequent work has strengthened this formulation through richer local representations and differentiable optimization: ShapeDTW incorporates neighborhood shape information into matching costs (Zhao & Itti, 2018), while Soft-DTW replaces the hard path mini-

mum with a smooth relaxation suitable for gradient-based learning (Cuturi & Blondel, 2017). TimePoint learns sparse keypoints and descriptors for efficient DTW-based alignment (Shapira Weber et al., 2025). These methods retain the basic view of alignment as a structured relation defined for a sequence pair.

Local alignment identifies shared subsequences rather than forcing complete-sequence correspondence, as exemplified by Smith–Waterman (Smith & Waterman, 1981). Extensions have addressed unmatched content and learnable path inference in complementary ways. Drop-DTW augments dynamic programming with dropping operations so that shared content can be aligned while unmatched elements are excluded (Dvornik et al., 2021). DecDTW instead treats temporal warping as a differentiable optimization layer, allowing losses on the predicted warp to train upstream representations (Xu et al., 2023). D3TW learns action alignment from ordered transcripts (Chang et al., 2019), while DEDAL learns protein-sequence alignment from alignment supervision through differentiable programming (Llinares-López et al., 2023). Together, these developments broaden the admissible correspondences and the ways alignment can participate in learning, while their native alignment object remains a relation between two sequences. Our setting retains this structured pairwise prediction viewpoint, but learns each partial alignment in the context of a larger sequence group rather than as an isolated pair.

2.2 Joint and Groupwise Sequence Alignment

Groupwise alignment couples multiple related sequences through shared representations, transformations, or collection-level objectives. Canonical time warping couples temporal alignment with correlated representations (Zhou & De la Torre, 2009). GCTW jointly estimates temporal transformations and correlated projections for multiple sequences (Zhou & De la Torre, 2016), while DCTW extends canonical time warping with nonlinear learned representations (Trigeorgis et al., 2016). In this family, multiple observations are coordinated through a common representation in which temporally corresponding content becomes comparable.

A complementary line learns input-dependent transformations for aligning collections of signals. DTAN amortizes temporal warping with a neural network, enabling learned transformations to be applied to unseen observations (Shapira Weber et al., 2019). RF-DTAN relates the per-sequence transformations through an inverse-consistency averaging objective and extends this approach to variable-length signals (Shapira Weber & Freifeld, 2023). Related correspondence-learning methods have also used temporal cycle consistency to couple matches across observations (Dwibedi et al., 2019; Hadji et al., 2021). GTCC extends cycle-based learning to repeated actions (Donahue & Elhamifar, 2024). These approaches demonstrate several ways in which information shared across a collection can guide alignment or correspondence learning. Our work draws on both pairwise and groupwise formulations: it preserves directed, pair-specific partial alignments as the prediction objects while allowing the sequence group to influence how those predictions are learned jointly. Thus, the distinction is not whether multiple sequences are available, but how their shared information enters the alignment problem.

3 Problem Formulation

Let

$$\mathcal{X} = \{X^{(1)}, \dots, X^{(K)}\}, \qquad X^{(k)} = (x_1^{(k)}, \dots, x_{T_k}^{(k)}),$$

denote a group of K related sequences, where $x_p^{(k)} \in \mathbb{R}^d$ and the sequence lengths T_k may differ.

For an ordered pair $i \neq j$, we seek a partial monotone alignment between $X^{(i)}$ and $X^{(j)}$. We represent it by

$$y_{ij} = \big(a_{ij}, b_{ij}, q_{ij}\big),$$

where

$$I_{ij} = \{a_{ij}, \dots, b_{ij}\} \subseteq \{1, \dots, T_i\}$$

is the source interval participating in the alignment, and

$$q_{ij} : I_{ij} \to \{1, \dots, T_j\}, \qquad p \leq p' \Longrightarrow q_{ij}(p) \leq q_{ij}(p'),$$

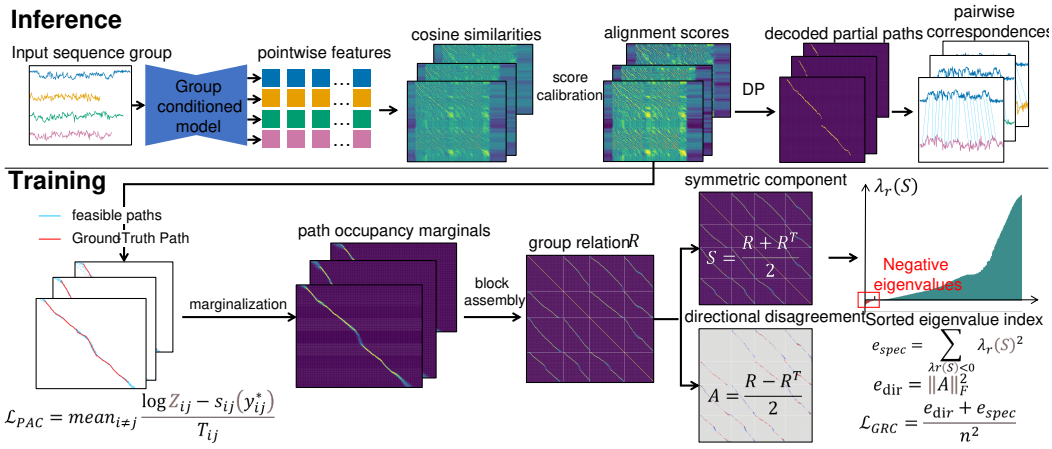


Figure 2: **Framework overview.** The same group-conditioned scoring model supports pairwise DP decoding at inference (top). Training (bottom) combines PAL path supervision with GRC on the group relation assembled from predicted occupancy marginals.

specifies the correspondence within that interval. The mapped range of q_{ij} determines the corresponding overlap in $X^{(j)}$. Both the overlap boundaries and the correspondence are unknown and must be estimated; positions outside the overlapping regions are not required to participate in the alignment.

Given the full sequence group $\mathcal{X}$, our task is to estimate the collection of partial alignments

$$\widehat{\mathcal{Y}} = \{\widehat{y}_{ij}\}_{i \neq j}.$$

Each alignment is estimated in the context of the entire group: when aligning a particular pair, the other related sequences may provide additional evidence about which regions are shared and how their contents correspond. We therefore learn the pairwise partial alignments from the sequence group jointly, while retaining the overlap and correspondence for each sequence pair as the quantities to be estimated.

We retain directed partial alignments between sequence pairs as the prediction objects. Our evaluation focuses on whether group information improves these correspondences within content that is alignable across the group.

4 Method

Figure 2 summarizes our framework. A group-conditioned scoring model defines distributions over partial alignment paths. The Partial Alignment Likelihood (PAL) supervises their overlap boundaries and internal correspondences, while Group Relational Compatibility (GRC) regularizes a group relation assembled from their posterior occupancies. Training first fits the path model with PAL and then introduces GRC. At inference, the same scoring model supplies a pairwise dynamic-programming decoder.

4.1 Group-Conditioned Pairwise Matching

A shared multiscale encoder–decoder with two multi-head group-set interaction (GSI) modules based on cross-sequence attention (Vaswani et al., 2017) produces group-conditioned, unit-normalized pointwise representations $F_i = f_{\theta,i}(\mathcal{X}) \in \mathbb{R}^{T_i \times D}$. For each ordered pair $i \neq j$, we compute

$$E_{ij}(p, q) = \alpha \langle F_i(p), F_j(q) \rangle + \beta - \log T_j, \qquad \alpha > 0. \tag{1}$$

Here, α and β are learned calibration parameters, and the target-length offset is applied at each matched source position.

4.2 Jointly Locating and Aligning Shared Regions

For an ordered pair $i \neq j$, the admissible monotone partial paths are

$$\mathcal{Y}_{ij} = \left\{ y = (a, b, q_a, \ldots, q_b) \;\middle|\; \begin{array}{l} 1 \le a < b \le T_i, \\ q_p \in \{1, \ldots, T_j\}, \quad a \le p \le b, \\ q_p \le q_{p+1}, \quad a \le p < b \end{array} \right\}. \tag{2}$$

Each path matches every position in its source interval, allowing repeated target positions and forward skips. An additive score defines the normalized path distribution:

$$s_{ij}(y) = \sum_{p=a}^{b} E_{ij}(p, q_p), \qquad Z_{ij} = \sum_{y \in \mathcal{Y}_{ij}} \exp s_{ij}(y), \qquad p_{ij}(y \mid \mathcal{X}) = \frac{\exp s_{ij}(y)}{Z_{ij}}. \tag{3}$$

Dynamic programming computes Z_{ij} without enumerating paths. For an annotated path $y^\star_{ij}$ with target coordinates rounded to the nearest lattice positions, PAL averages the source-length-normalized conditional negative log-likelihood over ordered pairs:

$$\mathcal{L}_{\mathrm{PAL}}(\mathcal{X}) = \frac{1}{K(K-1)} \sum_{i \neq j} \frac{\log Z_{ij} - s_{ij}(y^\star_{ij})}{T_i}. \tag{4}$$

Competing paths differ in their endpoints and internal matches, so PAL jointly supervises overlap and correspondence without directly assessing compatibility across pairs.

4.3 From Pairwise Uncertainty to a Group Relation

Group-level regularization uses occupancy marginals rather than a decoded path, obtained through the log-partition gradient identity (Wainwright & Jordan, 2008):

$$M_{ij}(p, q) = \sum_{y \in \mathcal{Y}_{ij}} p_{ij}(y \mid \mathcal{X}) \mathbf{1}[(p, q) \in y] = \frac{\partial \log Z_{ij}}{\partial E_{ij}(p, q)}. \tag{5}$$

The row sum $\sum_q M_{ij}(p, q) = \Pr_{p_{ij}}(a \le p \le b)$ gives the probability that source position p participates, preserving overlap uncertainty alongside matching uncertainty. Directional marginals are kept separate because M_{ji} need not equal M_{ij}^T. Select original positions $\Omega_i \subseteq \{1, \ldots, T_i\}$ and reuse the same Ω_i in every block involving sequence i. With $n = \sum_i |\Omega_i|$, define

$$R_\Omega \in \mathbb{R}^{n \times n}, \qquad [R_\Omega]_{ij} = \begin{cases} \mathrm{I}_{|\Omega_i|}, & i = j, \\ M_{ij}[\Omega_i, \Omega_j], & i \neq j. \end{cases} \tag{6}$$

Identity blocks encode fixed self-correspondences, not feature similarities; off-diagonal blocks retain the original occupancies without pair-level normalization or forward–reverse averaging. Selection controls relation size without pooling observations.

4.4 Relational Compatibility through a Shared-Representation Prior

Although matching scores use shared feature inner products, the resulting directional path marginals need not form a Gram matrix. GRC acts on their assembled relation R_Ω, using an idealized relation that reuses a representation $h_{i,p}$ for each selected position across all pairs:

$$C_{(i,p),(j,q)} = \langle h_{i,p}, h_{j,q} \rangle, \qquad C = HH^\mathsf{T} \succeq 0, \tag{7}$$

where H stacks these reference representations, not additional model variables. We use this structure as a soft prior by penalizing the squared distance from R_Ω to the symmetric PSD cone:

$$\mathcal{L}_{\mathrm{GRC}}(\mathcal{X}; \Omega) = \frac{1}{n^2} \min_{C \in \mathbb{S}^n_+} \|R_\Omega - C\|_F^2. \tag{8}$$

The reference matrix C is constrained only to be symmetric PSD; it need not be an alignment matrix or preserve the identity self-blocks. To evaluate this distance, write

$$S_\Omega = \frac{R_\Omega + R_\Omega^\mathsf{T}}{2}, \qquad A_\Omega = \frac{R_\Omega - R_\Omega^\mathsf{T}}{2},$$

and let $S_\Omega = U \operatorname{diag}(\gamma_1, \ldots, \gamma_n) U^\mathsf{T}$. Orthogonality of the symmetric and skew-symmetric components, together with the standard Frobenius projection onto the PSD cone (Higham, 1988), gives

$$C^\star = U \operatorname{diag}\big((\gamma_1)_+, \ldots, (\gamma_n)_+\big) U^\mathsf{T},$$
$$\mathcal{L}_{\mathrm{GRC}} = \frac{1}{n^2} \left[\underbrace{\|A_\Omega\|_F^2}_{e_{\mathrm{dir}}} + \underbrace{\sum_{r:\gamma_r<0} \gamma_r^2}_{e_{\mathrm{spec}}} \right], \tag{9}$$

where $(u)_+ = \max(u, 0)$. Equation equation 9 evaluates the single structural distance in Equation equation 8; its two contributions are fixed by the orthogonal decomposition, not independently chosen regularizers. Symmetry alone does not ensure joint Gram-compatibility: an indefinite symmetric relation still lies outside the cone. Because the projection acts on the assembled group relation, the correction to one pair can depend on predictions involving other sequences. GRC encourages compatibility under this shared-representation prior, rather than imposing exact reciprocity or transitivity on the directional alignments. Penalizing only the negative eigenvalues of S_Ω leaves the skew-symmetric part of R_Ω unconstrained. GRC instead measures the joint correction required to reach the symmetric PSD cone. Because the spectral term is evaluated on the assembled group relation, an update to one pair can depend on predictions involving other sequences. GRC encourages compatibility under this shared-representation model, rather than imposing exact reciprocity or transitivity on the directional alignments.

4.5 Joint Learning and Pairwise Inference

Staged training. A PAL-only warm-start grounds paths and overlap boundaries in correspondence supervision before the group-level prior is introduced. The subsequent joint stage retains PAL while GRC refines compatibility among these relations. For a batch $\mathcal{B}$, let the nonempty subset $\mathcal{B}_G \subseteq \mathcal{B}$ contain the groups selected for GRC. The joint objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{\mathcal{X} \in \mathcal{B}} \mathcal{L}_{\mathrm{PAL}}(\mathcal{X}) + \lambda_G \frac{1}{|\mathcal{B}_G|} \sum_{\mathcal{X} \in \mathcal{B}_G} \mathcal{L}_{\mathrm{GRC}}(\mathcal{X}; \Omega), \tag{10}$$

where $\lambda_G > 0$ controls regularization strength. PAL remains active because compatibility alone does not ensure correct correspondences. GRC requires no additional group-level annotations. PAL updates both representations and score calibration; GRC updates the encoder and cross-sequence interactions through occupancy marginals and matching scores, with gradients to α and β stopped only on the GRC branch.

Inference. Scores computed from the full input group are used to decode each ordered pair by dynamic programming:

$$\widehat{y}_{ij} = \arg\max_{y \in \mathcal{Y}_{ij}} s_{ij}(y). \tag{11}$$

Decoding uses the same path scores as training, without occupancy marginalization, group-relation construction, or PSD projection.

5 Experiments

We evaluate direct and composed correspondence accuracy and shared-region recovery on real speech and synthetic stratigraphic sequences.

5.1 Experimental Setup

Datasets. *CMU ARCTIC (CMU)* (Kominek & Black, 2004) provides speech recordings of shared text prompts. We use fixed groups of $K = 4$ recordings, represented by 39-dimensional MFCC-based features (Davis & Mermelstein, 1980) at their native sequence lengths. Continuous reference correspondences are derived from the automatic LAB phone alignments, while complete recordings

remain available as input. Prompt-disjoint training, validation, and test splits contain 2,229, 474, and 469 groups, respectively.

Synthetic Stratigraphic Impedance Alignment (SSIA) contains groups of $K = 8$ impedance traces of length 512, sampled from synthetic structural volumes generated following Wu et al. (2020). Reference correspondences match equal relative geological time (RGT) values (Wu & Hale, 2015) through monotonic interpolation; only impedance is supplied as input. Training uses 10,000 groups from 87 volumes. Validation and test contain 500 groups each; they are group-disjoint but share the same 10 additional volumes, which are excluded from training. Dataset construction and preprocessing are detailed in Appendix A.1.

Evaluation. Evaluation uses a reference-defined group-common domain: source positions with valid reference correspondences to every other group member. This domain is not supplied at inference. Errors and tolerances use native target coordinates—MFCC frame indices for CMU and trace indices for SSIA. PathMAE measures mean absolute target-index error against continuous references on protocol-valid positions. $\mathrm{F}_1@\tau$ summarizes precision and recall of valid correspondences within τ index units, with $\tau \in \{0, \ldots, 20\}$. IoU measures agreement between predicted and reference mapped target intervals under the common-domain protocol.

For distinct sequences i, j, k, TransitiveMAE and TransitiveP95 summarize the mean and 95th-percentile errors of predicted $i \to j \to k$ compositions against the reference $i \to k$ mapping. They measure reference accuracy, not prediction-only cycle disagreement. Results are aggregated hierarchically, with equal parent-prompt weighting for CMU and volume-level aggregation for SSIA. Full metric definitions and aggregation rules are given in Appendix A.2.

Baselines. We compare classical pairwise methods, Raw-DTW and ShapeDTW (Sakoe & Chiba, 1978; Zhao & Itti, 2018); partial-alignment, differentiable, and path-learning methods, Drop-DTW, Soft-DTW, DecDTW-Core, and the path-supervised DecDTW-Sup variant (Dvornik et al., 2021; Cuturi & Blondel, 2017; Xu et al., 2023); and joint/groupwise methods, GCTW, DCTW, and RF-DTAN (Zhou & De la Torre, 2016; Trigeorgis et al., 2016; Shapira Weber & Freifeld, 2023). All methods use the same splits, reference correspondences, and evaluation protocol, with validation-based configuration or checkpoint selection and no test-set tuning. Final baseline configurations are listed in Appendix A.3.

Training and reporting. Controlled PAL-only and PAL+GRC continuations start from the same PAL warm-start checkpoint for each seed. Both use validation alignment loss for checkpoint selection and early-stop independently. We report means over three random seeds for our models. Further details of the shared PAL and PAL+GRC network are provided in Appendix A.

5.2 Alignment Performance

Table 1 compares direct correspondence accuracy, shared-region recovery, and composed accuracy across both benchmarks. PAL already achieves lower PathMAE and TransitiveMAE and higher IoU than all evaluated external baselines. Adding GRC further improves these measures, reaching PathMAE values of 1.350 on CMU and 6.540 on SSIA, compared with 1.447 for DecDTW-Sup and 7.991 for Soft-DTW, respectively. The higher IoU values show that the improvement extends beyond pointwise localization to recovery of the mapped shared range.

PAL+GRC also reduces mean and tail composed errors relative to PAL and obtains the lowest TransitiveP95 on SSIA; DecDTW-Sup retains the lowest TransitiveP95 on CMU. Figure 3 complements these results with the full tolerance-dependent recovery curves. Both PAL variants lead across the positive-tolerance range, with GRC providing additional improvement over PAL. Together, these evaluations cover correspondence localization, shared-region recovery, and the accuracy of mappings composed across sequences.

5.3 Analysis of Group Relational Learning

Separating feature interaction from relational supervision. To distinguish GRC from the backbone's access to other sequences, we enable or remove GSI, the cross-sequence attention modules,

Table 1: **Main alignment comparison.** P-MAE, T-MAE, and T-P95 denote PathMAE, TransitiveMAE, and TransitiveP95 in native target indices. Composed errors are reference-based; IoU evaluates the mapped shared interval. Our rows are three-seed means. Best values are bold.

	CMU				SSIA			
Method	P-MAE ↓	IoU ↑	T-MAE ↓	T-P95 ↓	P-MAE ↓	IoU ↑	T-MAE ↓	T-P95 ↓
Raw-DTW	2.228	0.890	2.801	8.458	15.624	0.932	18.350	65.797
ShapeDTW	2.037	0.893	2.569	7.284	8.473	0.951	9.822	42.501
Soft-DTW	2.256	0.889	2.829	8.655	7.991	0.944	9.296	40.000
Drop-DTW	2.451	0.900	3.225	10.427	9.184	0.949	10.684	41.750
DecDTW-Core	6.826	0.886	10.492	43.307	9.588	0.942	11.611	48.477
DecDTW-Sup	1.447	0.886	1.732	**4.741**	14.738	0.921	14.524	58.923
GCTW	11.253	0.886	11.915	50.361	17.396	0.923	16.700	64.957
DCTW	2.684	0.893	2.919	9.344	17.514	0.928	16.149	58.181
RF-DTAN	11.951	0.886	12.410	48.706	22.431	0.915	23.475	72.624
PAL	1.373	0.965	1.703	5.870	7.110	0.966	8.618	34.666
PAL+GRC	**1.350**	**0.968**	**1.668**	5.712	**6.540**	**0.969**	**8.054**	**32.174**

and train each architecture with PAL alone or PAL+GRC. Within each GSI setting, the two objective branches share a PAL warm-start and model-selection rules. Table 4 shows that GRC lowers direct and composed errors in both datasets even without GSI. With GRC enabled, retaining GSI additionally improves PathMAE, $F_1@2$, and IoU in both domains. These results distinguish two contributions: GSI exchanges information during feature construction, whereas GRC regularizes the relations induced by the predicted correspondences. The benefit of GRC is therefore not explained solely by the cross-attention architecture.

Accuracy and agreement of correspondence routes. We examine whether GRC improves relations across pairs by comparing direct predictions with ground truth, composed predictions with ground truth, and direct predictions with their composed counterparts. All three comparisons use support jointly valid for PAL and PAL+GRC. GRC reduces each error in both datasets (Table 5); direct–composed MAE decreases from 12.774 to 11.720 on SSIA and from 0.928 to 0.908 on CMU. The reduction in disagreement accompanies improved accuracy of both routes, connecting the alignment gains to more coherent correspondence predictions within the sequence group.

Scope of relational regularization. We study how the number of jointly constrained sequences affects GRC while retaining eight sequences for input, PAL supervision, and evaluation on SSIA. Table 2 reports $K_{\mathrm{GRC}} = 2, \ldots, 7$. Broader scopes yield lower direct and composed errors overall: The largest PathMAE decrease is from 8.924 at $K_{\mathrm{GRC}} = 2$ to 6.978 at 3. from $K_{\mathrm{GRC}} = 3$ to 7, PathMAE decreases from 6.978 to 6.533 and TransitiveP95 from 33.678 to 31.750. The main error measures settle into a similar performance range at $K_{\mathrm{GRC}} = 5$–7, with small fluctuations across metrics. This pattern supports using a broader collection of predicted relations for joint regularization.

Table 2: **Joint regularization scope on SSIA.** Input, PAL supervision, and evaluation each use $K = 8$. K_{GRC} specifies the number of sequences included in relation construction. Entries are means. Additional settings are given in Appendix B.3.

K_{GRC}	PathMAE ↓	$F_1@2$ ↑	IoU ↑	TransitiveMAE ↓	TransitiveP95 ↓
2	8.924	0.584	0.967	9.496	40.584
3	6.978	0.592	0.969	8.192	33.678
4	6.757	0.591	0.969	8.151	33.286
5	6.570	0.593	0.969	8.009	32.363
6	6.577	0.590	0.969	8.014	32.250
7	6.533	0.588	0.969	7.952	31.750

5.4 Robustness and Qualitative Analysis

Common overlap and warm-start uncertainty. We ask whether GRC helps when overlap is limited or initial predictions are uncertain. We stratify test groups separately by common overlap and frozen PAL warm-start uncertainty, using validation-defined strata shared by both models. Higher uncertainty corresponds to larger PAL errors. PAL+GRC improves PathMAE, $F_1@2$, IoU, and composed accuracy in every stratum of both datasets (Appendix B.4).

Qualitative correspondence behavior. Figure 4 illustrates two characteristic alignment challenges. In CMU, several baseline alignments extend into leading or trailing regions outside the shared labelled utterance. PAL identifies the shared speech region while leaving the surrounding nonshared content unmatched; PAL+GRC further refines correspondence within the utterance. In SSIA, locally similar impedance patterns can occur at different RGT values, and selected baselines incorrectly align such repeated structures. PAL recovers a more coherent alignment over the shared RGT range, while PAL+GRC corrects residual local mismatches toward the reference. Together, these examples illustrate PAL's joint learning of overlap and internal alignment, followed by additional group-informed refinement through GRC.

6 Discussion and Limitations

The controlled comparisons suggest that group information is useful beyond feature exchange: relations among predictions can themselves provide learning signals. GRC lets pairwise predictions constrain and complement one another alongside path supervision. Learning can therefore draw on group structure without requiring recovery of a single joint alignment.

Learning-based alignment also changes how task adaptation is performed. Classical methods such as DTW and ShapeDTW rely on choices of distances, descriptors, and path constraints (Sakoe & Chiba, 1978; Zhao & Itti, 2018), with hyperparameters that can be selected using validation data. The distinction is therefore not simply between methods that use data and methods that do not. Our framework shifts more of this adaptation from manual specification to learned representations and matching scores, guided by correspondence supervision and relational information. This flexibility comes at the cost of annotated training data and additional computation.

Cross-domain transfer remains an important limitation. Our models are trained separately on speech and stratigraphic sequences: the experiments demonstrate applicability of the same formulation across domains, not zero-shot generalization of a single trained model. One-dimensional observations can differ substantially in sampling, signal statistics, channel semantics, sequence length, and correspondence structure. Self-supervised training on synthetic warps has enabled transfer to real-world time-series alignment (Shapira Weber et al., 2025). Larger collections of sequence groups spanning multiple domains, sampling regimes, and multichannel observations could support pretraining of more transferable correspondence representations. This offers a direction from dataset-specific alignment models toward broadly pretrained models that can be adapted across tasks.

7 Conclusion

We studied partial sequence alignment from groups of related observations, where isolated pairwise objectives leave useful relational information unexploited. PAL unifies shared-region identification and monotone correspondence estimation within a structured path objective. GRC further couples their learning through the group relational structure induced by soft pairwise predictions. Experiments on speech and simulated stratigraphic sequences demonstrate gains in direct correspondence accuracy, shared-region recovery, and reference-based composed accuracy, while controlled analyses support the contribution of relational learning beyond feature-level group interaction. The central message is not simply to use more sequences: when observations naturally form a group, their indirect relationships provide valuable information for learning correspondences that isolated pairwise alignment leaves unused.

AI USE STATEMENT

We used large language models to assist with searching for and locating relevant files and literature, and with implementing and tuning baseline methods on our datasets. All AI-assisted content was manually checked by the authors. The authors remain responsible for the final paper, code, experiments, and conclusions.

REFERENCES

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3546–3555, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Chang_D3TW_Discriminative_Differentiable_Dynamic_Time_Warping_for_Weakly_Supervised_Action_CVPR_2019_paper.html.

Yuxin Chen, Leonidas Guibas, and Qixing Huang. Near-optimal joint object matching via convex relaxation. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 100–108. PMLR, 2014. URL https://proceedings.mlr.press/v32/chend14.html.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903. PMLR, 2017.

S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980. doi: 10.1109/TASSP.1980.1163420.

Gerard Donahue and Ehsan Elhamifar. Learning to predict activity progress by self-supervised video alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18667–18677, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Donahue_Learning_to_Predict_Activity_Progress_by_Self-Supervised_Video_Alignment_CVPR_2024_paper.html.

Mikita Dvornik, Isma Hadji, Konstantinos G. Derpanis, Animesh Garg, and Allan D. Jepson. Drop-DTW: Aligning common signal between sequences while dropping outliers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13782–13793, 2021.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1801–1810, 2019.

Isma Hadji, Konstantinos G. Derpanis, and Allan D. Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11068–11077, 2021.

Dave Hale. Dynamic warping of seismic images. *Geophysics*, 78(2):S105–S115, 2013. doi: 10.1190/geo2012-0327.1.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988. doi: 10.1016/0024-3795(88)90223-6.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. doi: 10.1007/s10618-020-00710-y.

John Kominek and Alan W. Black. The CMU arctic speech databases. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 223–224, Pittsburgh, PA, USA, 2004.

Felipe Llinares-López, Quentin Berthet, Mathieu Blondel, Olivier Teboul, and Jean-Philippe Vert. Deep embedding and alignment of protein sequences. *Nature Methods*, 20:104–111, 2023. doi: 10.1038/s41592-022-01700-2.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Ron Shapira Weber and Oren Freifeld. Regularization-free diffeomorphic temporal alignment nets. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30794–30826. PMLR, 2023.

Ron Shapira Weber, Shahar Benishay, Andrey Lavrinenko, Shahaf E. Finder, and Oren Freifeld. TimePoint: Accelerated time series alignment via self-supervised keypoint and descriptor learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54275–54299. PMLR, 2025. URL https://proceedings.mlr.press/v267/shapira-weber25a.html.

Ron A. Shapira Weber, Matan Eyal, Nicki Skafte Detlefsen, Oren Shriki, and Oren Freifeld. Diffeomorphic temporal alignment nets. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6570–6581, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/db98dc0dbafde48e8f74c0de001d35e4-Abstract.html.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981. doi: 10.1016/0022-2836(81)90087-5.

George Trigeorgis, Mihalis A. Nicolaou, Stefanos Zafeiriou, and Björn W. Schuller. Deep canonical time warping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5110–5118, 2016. doi: 10.1109/CVPR.2016.552.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Runzhong Wang, Junchi Yan, and Xiaokang Yang. Neural graph matching network: Learning Lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5261–5279, 2022. doi: 10.1109/TPAMI.2021.3078053.

Xinming Wu and Dave Hale. Horizon volumes with interpreted constraints. *Geophysics*, 80(2): IM21–IM33, 2015. doi: 10.1190/geo2014-0212.1.

Xinming Wu, Zhicheng Geng, Yunzhi Shi, Nam Pham, Sergey Fomel, and Guillaume Caumon. Building realistic structure models to train convolutional neural networks for seismic structural interpretation. *Geophysics*, 85(4):WA27–WA39, 2020. doi: 10.1190/geo2019-0375.1.

Ming Xu, Sourav Garg, Michael Milford, and Stephen Gould. Deep declarative dynamic time warping for end-to-end learning of alignment paths. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UClBPxIZqnY.

Jiaping Zhao and Laurent Itti. shapeDTW: Shape dynamic time warping. *Pattern Recognition*, 74: 171–184, 2018. doi: 10.1016/j.patcog.2017.09.020.

Feng Zhou and Fernando De la Torre. Canonical time warping for alignment of human behavior. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper_files/paper/2009/hash/2ca65f58e35d9ad45bf7f3ae5cfd08f1-Abstract.html.

Feng Zhou and Fernando De la Torre. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):279–294, 2016. doi: 10.1109/TPAMI.2015.2414429.

APPENDIX

A EXPERIMENTAL DETAILS

A.1 DATASETS AND PREPROCESSING

CMU ARCTIC. The CMU benchmark uses seven speakers from CMU ARCTIC (Kominek & Black, 2004), with each group containing $K = 4$ recordings of the same parent prompt. The training, validation, and test splits contain 511 parent prompts and 2,229 groups, 110 parent prompts and 474 groups, and 109 parent prompts and 469 groups, respectively. All groups associated with a parent prompt belong to the same split. Speakers may appear in multiple splits; the evaluation is therefore prompt-disjoint rather than speaker-disjoint.

Audio is processed at 16 kHz using a 25 ms window, a 10 ms frame shift, 23 Mel filters, and 13 static MFCC coefficients. Static coefficients are mean-normalized using speaker-specific statistics estimated exclusively from the training split, without variance normalization. First- and second-order temporal derivatives are then included to obtain 39-dimensional features. Each utterance retains its native length, giving an input of shape $T_i \times 39$. Batch construction introduces padding and masking only; padded positions are excluded from alignment and evaluation. The benchmark does not resample all recordings to a fixed sequence length.

SSIA. SSIA contains groups of $K = 8$ impedance traces extracted from synthetic structural volumes generated following Wu et al. (2020). Each trace retains its original 512 positions. Training contains 10,000 groups from 87 volumes. Validation and test contain 500 groups each, drawn from the same additional set of 10 held-out volumes. Training volumes are disjoint from the held-out volumes; validation and test are group-disjoint but not volume-disjoint. RGT is used only to construct supervision and evaluation references, not as a model input.

Each impedance trace is independently z-score normalized at loading time. Clipping and scaling are then applied using the 99th percentile of the absolute normalized values. This preprocessing changes amplitudes only and preserves the original 512 sequence indices.

A.2 EVALUATION PROTOCOL

Reference correspondences and evaluation domain. For an ordered source–target pair $i \to j$, let $q_{ij}^{\mathrm{GT}}(p)$ and $\widehat{q}_{ij}(p)$ denote the reference and predicted target coordinates for source position p. CMU references are obtained by interpolating matched automatic LAB phone-alignment intervals into continuous native MFCC frame coordinates. SSIA references match equal RGT values through monotone interpolation. Positions without valid shared labelled content or RGT coverage are excluded from the reference domain.

Both benchmarks use the group-common source domain

$$\mathcal{D}_i = \left\{ p \in \{1, \ldots, T_i\} : q_{ij}^{\mathrm{GT}}(p) \text{ is valid for every } j \neq i \right\}. \tag{12}$$

For each ordered pair $i \to j$, let $\mathcal{D}_{ij}$ denote the source positions with valid reference correspondences. The group-common evaluation domain is $\mathcal{D}_i = \bigcap_{j \neq i} \mathcal{D}_{ij}$. A position outside $\mathcal{D}_i$ may have a reference correspondence for an individual pair, but it does not have references to every other group member and is therefore outside the scope of this evaluation. This reference-defined domain is used for evaluation and is not supplied at inference. All errors are measured in the target sequence's native index space: one native MFCC frame step for CMU and one position on the original 512-point trace for SSIA. These units measure correspondence displacement, not signal-amplitude differences, and do not represent a common physical-time unit across datasets.

PathMAE. The pointwise correspondence error is

$$e_{ij}(p) = \left|\widehat{q}_{ij}(p) - q_{ij}^{\mathrm{GT}}(p)\right|. \tag{13}$$

Let $\mathcal{V}_{ij} \subseteq \mathcal{D}_i$ contain positions with valid predicted and reference correspondences. PathMAE averages the target-index deviation over this support:

$$\mathrm{PathMAE}_{ij} = \frac{1}{|\mathcal{V}_{ij}|} \sum_{p \in \mathcal{V}_{ij}} e_{ij}(p). \tag{14}$$

The reference coordinates remain continuous during evaluation.

Tolerance-dependent correspondence recovery. For tolerance τ, a predicted correspondence is correct only when it is valid and its target-index error is at most τ. Let $\mathcal{P}_{ij}$ denote the evaluated predicted correspondence positions. The correctly recovered positions are

$$\mathcal{C}_{ij}^{\tau} = \{p \in \mathcal{P}_{ij} \cap \mathcal{D}_i : \widehat{q}_{ij}(p) \text{ is valid},\ e_{ij}(p) \le \tau\}. \tag{15}$$

Precision measures the fraction of evaluated predictions that are correct, while recall measures the fraction of reference correspondences recovered within tolerance:

$$P_{ij}^{\tau} = \frac{|\mathcal{C}_{ij}^{\tau}|}{|\mathcal{P}_{ij}|}, \qquad R_{ij}^{\tau} = \frac{|\mathcal{C}_{ij}^{\tau}|}{|\mathcal{D}_i|}, \qquad \mathrm{F}_{1,ij}@\tau = \frac{2P_{ij}^{\tau}R_{ij}^{\tau}}{P_{ij}^{\tau} + R_{ij}^{\tau}}. \tag{16}$$

We report $\tau = 0, \ldots, 20$. Correctness is assessed against the reference target coordinate for the same source position.

Shared mapped-interval IoU. Let $\widehat{J}_{ij}^{\mathrm{com}}$ and $J_{ij}^{\mathrm{com,GT}}$ denote the predicted and reference mapped shared intervals under the common-domain protocol. Their intersection-over-union is

$$\mathrm{IoU}_{ij} = \frac{\left|\widehat{J}_{ij}^{\mathrm{com}} \cap J_{ij}^{\mathrm{com,GT}}\right|}{\left|\widehat{J}_{ij}^{\mathrm{com}} \cup J_{ij}^{\mathrm{com,GT}}\right|}. \tag{17}$$

Here, $|\cdot|$ measures interval extent in native target coordinates. IoU evaluates recovery of the shared mapped region, not overlap between complete sequence extents.

Composed correspondence accuracy. For three distinct sequences i, j, k, the predicted $i \to j \to k$ correspondence is

$$\widehat{q}_{ijk}^{\mathrm{comp}}(p) = \overline{q}_{jk}\big(\widehat{q}_{ij}(p)\big), \tag{18}$$

where $\overline{q}_{jk}$ continuously interpolates the predicted $j \to k$ mapping over its valid support. The composed error is measured directly against the $i \to k$ reference:

$$e_{ijk}^{\mathrm{tr}}(p) = \left|\widehat{q}_{ijk}^{\mathrm{comp}}(p) - q_{ik}^{\mathrm{GT}}(p)\right|. \tag{19}$$

Let $\mathcal{V}_{ijk}^{\mathrm{tr}} \subseteq \mathcal{D}_i$ contain positions for which the first prediction, the interpolated second prediction, and the direct reference are valid. The mean composed error is

$$\mathrm{TransitiveMAE}_{ijk} = \frac{1}{|\mathcal{V}_{ijk}^{\mathrm{tr}}|} \sum_{p \in \mathcal{V}_{ijk}^{\mathrm{tr}}} e_{ijk}^{\mathrm{tr}}(p). \tag{20}$$

TransitiveP95 reports the 95th percentile of valid pointwise composed errors, characterizing the error-distribution tail. Both metrics evaluate ground-truth-referenced composed accuracy. They are neither sums of two pairwise MAEs nor prediction-only cycle-disagreement measures.

Aggregation. Results are summarized hierarchically through directed pairs or ordered triples and sequence groups. CMU uses equal parent-prompt weighting, while SSIA uses volume-level aggregation. A K-sequence group contains $K(K-1)$ directed pairs and $K(K-1)(K-2)$ ordered triples before validity filtering: 12 pairs and 24 triples for CMU, and 56 pairs and 336 triples for SSIA. Lower PathMAE, TransitiveMAE, and TransitiveP95 are better; higher F-scores and IoU are better.

A.3 Baseline and Model Configurations

Non-learned and optimization-based baselines. Table 3 summarizes the final configurations of Raw-DTW (Sakoe & Chiba, 1978), ShapeDTW (Zhao & Itti, 2018), Soft-DTW (Cuturi & Blondel, 2017), Drop-DTW (Dvornik et al., 2021), DecDTW-Core (Xu et al., 2023), and GCTW (Zhou & De la Torre, 2016). These configurations perform alignment without training a neural representation. Internal analysis grids are algorithm-specific computational representations; they do not redefine the benchmark's native sequence coordinates.

Table 3: **Final configurations of non-learned and optimization-based baselines.** Here, r is the descriptor radius, γ is the Soft-DTW smoothing parameter, and M and λ are the DecDTW warp discretization and regularization parameters. Evaluation uses native sequence coordinates.

Method	CMU ARCTIC	SSIA
Raw-DTW	Native 39-dimensional MFCCs; closed-end DTW	Normalized impedance; closed-end DTW
ShapeDTW	Descriptor radius $r = 4$	Descriptor radius $r = 16$
Soft-DTW	$\gamma = 1.5$; 256-point internal analysis grid	$\gamma = 0.3$; 512-point grid
Drop-DTW	Drop cost $= 3.2$	Drop cost $= 0.03$
DecDTW-Core	$M = 256$; $\lambda = 0.08$; subsequence mode off	$M = 256$; $\lambda = 0.05$
GCTW	128-point grid; canonical dimension 4; 10 outer and 2 inner iterations; MCCA regularization 0.1	128-point grid; canonical dimension 1; 10 outer and 2 inner iterations; MCCA regularization 0.1; temporal regularization 0.01

SSIA baselines operate on per-curve-normalized impedance. GCTW and DecDTW-Sup additionally use a deterministic five-channel local descriptor derived from each normalized trace $z(t)$:

$$\mathbf{d}(t) = \left[z(t),\ \mathrm{MA}_9(z)(t),\ \nabla z(t),\ \nabla^2 z(t),\ z(t) - \mathrm{MA}_9(z)(t)\right]^{\mathsf{T}}. \tag{21}$$

Here, MA_9 is a nine-point moving average, and the first and second derivatives are obtained by successive numerical-gradient operations. The descriptor expands each position of a single trace from one channel to five; it does not combine five traces or introduce a learned representation. Descriptor construction preserves the original 512 positions before any method-specific analysis-grid conversion.

DecDTW-Sup. Both dataset implementations use a shared residual one-dimensional convolutional encoder followed by the DecDTW alignment solver (Xu et al., 2023). The encoder has 192 hidden channels and produces 96-dimensional embeddings. CMU uses 39-channel MFCC input, whereas SSIA uses the five-channel impedance descriptor in Equation equation 21. The solver uses 128 analysis positions and 256 warp discretization points.

DCTW. A shared frame-wise multilayer perceptron contains two hidden layers of width 128 (Trigeorgis et al., 2016). The latent dimension is 8 for CMU and 4 for SSIA. Alignment is estimated on a 128-position internal grid using the group consensus procedure.

RF-DTAN. The CMU implementation uses a three-block temporal convolutional backbone operating on 39-channel MFCC features. The SSIA implementation uses an InceptionTime backbone (Ismail Fawaz et al., 2020) on the normalized single-channel impedance trace. Both variants use four recurrent temporal transformations with 16 tessellation segments (Shapira Weber & Freifeld, 2023).

Our alignment network. PAL and PAL+GRC share the same group-conditioned network. A five-stage one-dimensional residual encoder uses channel widths $(64, 128, 192, 256, 320)$, with

$(1, 1, 2, 2, 2)$ residual blocks at the corresponding stages. Two eight-head group-set interaction modules are inserted at the coarsest encoder resolution and the first decoder resolution. Each module allows a sequence representation to attend to the other members of its input group.

A skip-connected decoder restores the original temporal resolution and projects each position to a 192-dimensional ℓ_2-normalized embedding. Learned affine calibration of cosine similarity, together with the target-length offset defined in the Method section, produces the local matching scores used by structured partial-path inference.

The architecture is shared across datasets, with 39 input channels for CMU and one normalized impedance channel for SSIA. CMU uses $K = 4$ variable-length sequences, whereas SSIA uses $K = 8$ sequences of length 512. PAL+GRC retains the same inference architecture as PAL: GRC is an additional group-level training objective and introduces no separate network module at inference.

The implementation, training and evaluation scripts, and configuration files used in the experiments are available at the anonymous repository: `https://anonymous.4open.science/r/PAL_GRC-D7CE/`.

B Supplementary Experimental Results

B.1 Additional Alignment Results

Figure 3 reports correspondence recovery over the full tolerance range $\tau = 0, \ldots, 20$. Unlike a single-threshold score, these curves show how recovery changes as the allowed target-index error increases. Both PAL variants lead across the positive-tolerance range on CMU and SSIA, with PAL+GRC providing additional improvement over PAL.

The CMU curves approach their plateaus at smaller tolerances, whereas SSIA exhibits more gradual increases across the evaluated range. The curves complement the main comparison by showing that the alignment gains are not confined to one localization threshold. Tolerances are expressed in each benchmark's native target-index units, rather than a shared physical-time scale.

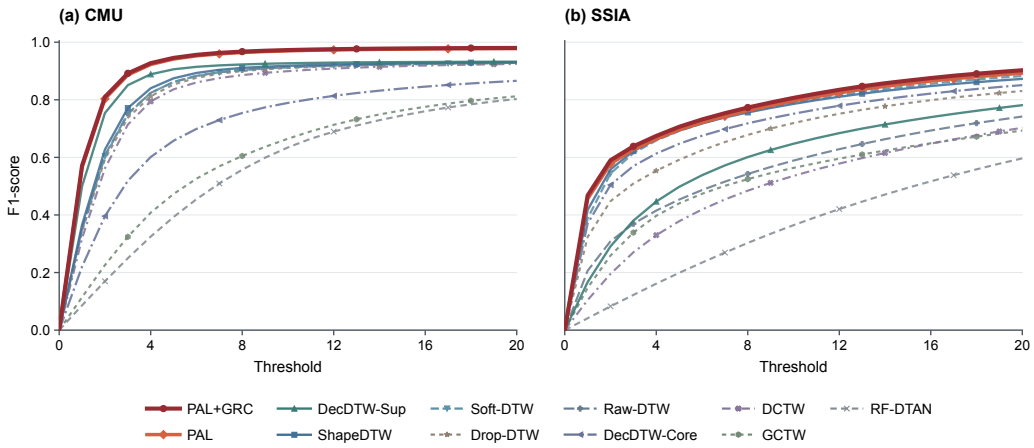


Figure 3: **Tolerance-dependent correspondence recovery.** $F_1@\tau$ for $\tau = 0, \ldots, 20$ on (a) CMU and (b) SSIA. A valid predicted correspondence is correct when its absolute target-index error is at most τ. Tolerances are measured in native MFCC frame indices for CMU and trace indices for SSIA.

B.2 Controlled Ablations

We separate architecture-level group interaction from objective-level relational regularization using a 2×2 comparison. GSI controls the cross-sequence attention modules, while GRC controls the relational objective. All four configurations retain PAL supervision. Within each GSI setting, the

Table 4: **GSI × GRC controlled ablation.** All configurations use PAL. GSI controls cross-sequence attention; GRC controls relational regularization. Within each GSI setting, the two objective branches share their PAL warm-start and selection rules. P-MAE, T-MAE, and T-P95 denote PathMAE, TransitiveMAE, and TransitiveP95. Entries are means.

Dataset	GSI	GRC	P-MAE ↓	F_1@2 ↑	IoU ↑	T-MAE ↓	T-P95 ↓
CMU	Off	Off	1.390	0.799	0.963	1.689	5.662
	Off	On	1.360	0.805	0.965	1.661	5.560
	On	Off	1.373	0.804	0.965	1.703	5.870
	On	On	1.350	0.809	0.968	1.668	5.712
SSIA	Off	Off	7.005	0.582	0.961	8.653	35.237
	Off	On	6.770	0.586	0.961	8.592	34.483
	On	Off	7.110	0.578	0.966	8.618	34.666
	On	On	6.540	0.589	0.969	8.054	32.174

PAL-only and PAL+GRC branches continue from the same PAL warm-start under the same model-selection rules.

Table 4 gives the complete metric set. Without GSI, adding GRC reduces CMU PathMAE from 1.390 to 1.360 and TransitiveMAE from 1.689 to 1.661. On SSIA, the corresponding errors decrease from 7.005 to 6.770 and from 8.653 to 8.592. GRC also lowers TransitiveP95 in both datasets. These improvements show that relational regularization remains effective without cross-sequence attention.

With GSI enabled, GRC further improves direct localization, shared-region recovery, and composed accuracy. On SSIA, PathMAE decreases from 7.110 to 6.540, TransitiveMAE from 8.618 to 8.054, and TransitiveP95 from 34.666 to 32.174. Conversely, adding GSI to a model already trained with GRC improves PathMAE, F_1@2, and IoU in both datasets. The two components therefore provide distinct means of using group information: feature exchange before path prediction and relational regularization of the resulting correspondences.

B.3 Relational Diagnostics

Direct and composed correspondence routes. For distinct sequences i, j, k, we compare the direct prediction from i to k with the prediction obtained by composition through j:

$$\begin{aligned} d_{ik}(p) &= \widehat{q}_{ik}(p), \\ c_{ijk}(p) &= \overline{q}_{jk}\big(\widehat{q}_{ij}(p)\big), \\ r_{ik}(p) &= q_{ik}^{\mathrm{GT}}(p), \end{aligned} \tag{22}$$

where $\overline{q}_{jk}$ denotes continuous interpolation of the predicted mapping over its valid support. We evaluate three pointwise errors:

$$\begin{aligned} e_{\text{D-GT}}(p) &= |d_{ik}(p) - r_{ik}(p)|, \\ e_{\text{C-GT}}(p) &= |c_{ijk}(p) - r_{ik}(p)|, \\ e_{\text{D-C}}(p) &= |d_{ik}(p) - c_{ijk}(p)|. \end{aligned} \tag{23}$$

The first two evaluate reference accuracy; the third evaluates agreement between the direct and composed predictions. The diagnostic covers all test groups and uses source positions jointly valid for PAL and PAL+GRC.

Table 5 shows that GRC reduces all three MAEs on both benchmarks. On SSIA, the composed-to-reference error decreases from 11.040 to 10.089 alongside a reduction in direct–composed disagreement from 12.774 to 11.720. The same direction of improvement holds on CMU. Thus, the direct and indirect routes become more accurate and more closely aligned with one another.

Joint regularization scope. The scope experiment retains eight input sequences and the full eight-sequence PAL and evaluation tasks. K_{GRC} determines how many sequences participate jointly in construction of the regularized relation. The complete results appear in Table 2.

Table 5: **Accuracy and agreement of correspondence routes.** Direct predictions (D), compositions through a third sequence (C), and references (GT) are compared on support jointly valid for PAL and PAL+GRC across all test groups. All entries are MAEs in native target indices (↓). D–C measures prediction agreement; D–GT and C–GT measure reference accuracy.

	CMU		SSIA	
Comparison ↓	PAL	PAL+GRC	PAL	PAL+GRC
Direct–GT	1.352	1.341	7.117	6.501
Composed–GT	1.567	1.548	11.040	10.089
Direct–Composed	0.928	0.908	12.774	11.720

Table 6: **Performance in the low-common-overlap stratum.** Stratum thresholds are determined from validation references and applied unchanged to the test set. P-MAE denotes PathMAE; T-MAE denotes the stratum-level composed-error summary. Entries are means.

Dataset	Objective	P-MAE ↓	F_1@2 ↑	IoU ↑	T-MAE ↓
CMU	PAL	1.371	0.792	0.957	1.588
	PAL+GRC	1.343	0.799	0.960	1.552
SSIA	PAL	6.628	0.571	0.961	10.009
	PAL+GRC	6.009	0.583	0.964	9.102

From $K_{\mathrm{GRC}} = 3$ to 5, PathMAE decreases from 6.978 to 6.570 and TransitiveMAE from 8.192 to 8.009. At scopes of five to seven sequences, the errors occupy a narrower range, while TransitiveP95 continues to decrease from 32.363 to 31.750. The best F_1@2 occurs at $K_{\mathrm{GRC}} = 5$, and IoU remains 0.969 at the reported precision for $K_{\mathrm{GRC}} = 3, \ldots, 7$. The sweep therefore shows an overall benefit from broader relational scope, with different metrics exhibiting different degrees of saturation.

B.4 Robustness across Overlap and Uncertainty

Common-overlap strata. Partial alignment must identify both the shared region and the correspondence within it. We therefore evaluate whether GRC's gains depend on how much content is shared by the group. Low, medium, and high common-overlap strata are defined using validation-reference thresholds and then applied unchanged to the test groups. PAL and PAL+GRC are evaluated within the same strata.

PAL+GRC improves PathMAE, F_1@2, IoU, and the stratum-level TransitiveMAE summary in all three overlap strata of both benchmarks. Table 6 illustrates this behavior in the low-overlap stratum. On SSIA, PathMAE decreases from 6.628 to 6.009 and the composed error from 10.009 to 9.102, accompanied by improved F_1@2 and IoU. CMU exhibits improvements in all four measures as well. The benefits are therefore not confined to groups with extensive common coverage.

Warm-start predictive uncertainty. We examine initial alignment difficulty using uncertainty computed from frozen PAL warm-start predictions. Validation-derived tertiles define low, medium, and high uncertainty strata. The same frozen reference model and grouping rule are used for the PAL and PAL+GRC evaluations, so the strata do not change with the evaluated objective.

Higher warm-start uncertainty corresponds to more difficult alignment cases. Between the low- and high-uncertainty strata, PAL PathMAE rises from 1.200 to 1.527 on CMU and from 4.473 to 9.186 on SSIA. PAL+GRC improves localization and composed accuracy at each uncertainty level, together with higher F_1@2 and IoU. These results show that relational regularization benefits both initially reliable predictions and cases with greater initial ambiguity.

C Qualitative Visualization

C.1 Qualitative Correspondence Examples

Figure 4 compares the original observations, selected baselines, PAL, and PAL+GRC with the reference alignments. The examples highlight two distinct sources of alignment error: including regions without shared content and selecting incorrect matches within the shared region.

CMU: nonshared boundary regions. The recordings contain leading and trailing regions outside the shared labelled utterance. Several baseline alignments extend into these regions, introducing matches where correspondence should not be established. PAL localizes the shared speech content and leaves the surrounding nonshared regions unmatched. Within the selected utterance, PAL+GRC further refines the correspondences toward the reference. These examples illustrate the importance of estimating overlap boundaries together with the internal alignment, rather than forcing complete-recording correspondence.

SSIA: repeated-pattern ambiguity. Similar impedance patterns can occur at positions associated with different RGT values. In the displayed examples, some baseline correspondences connect locally similar structures that do not represent the same geological time. PAL recovers a more coherent correspondence over the shared RGT range, while PAL+GRC further corrects local mismatches. The examples illustrate how group-informed refinement improves correspondence even after the shared region has been identified.

Together, the two domains expose complementary challenges: CMU emphasizes where alignment should begin and end, whereas SSIA emphasizes ambiguous matching within shared content. PAL addresses overlap selection and monotone correspondence jointly, and GRC provides an additional relational learning signal for refining these predictions.

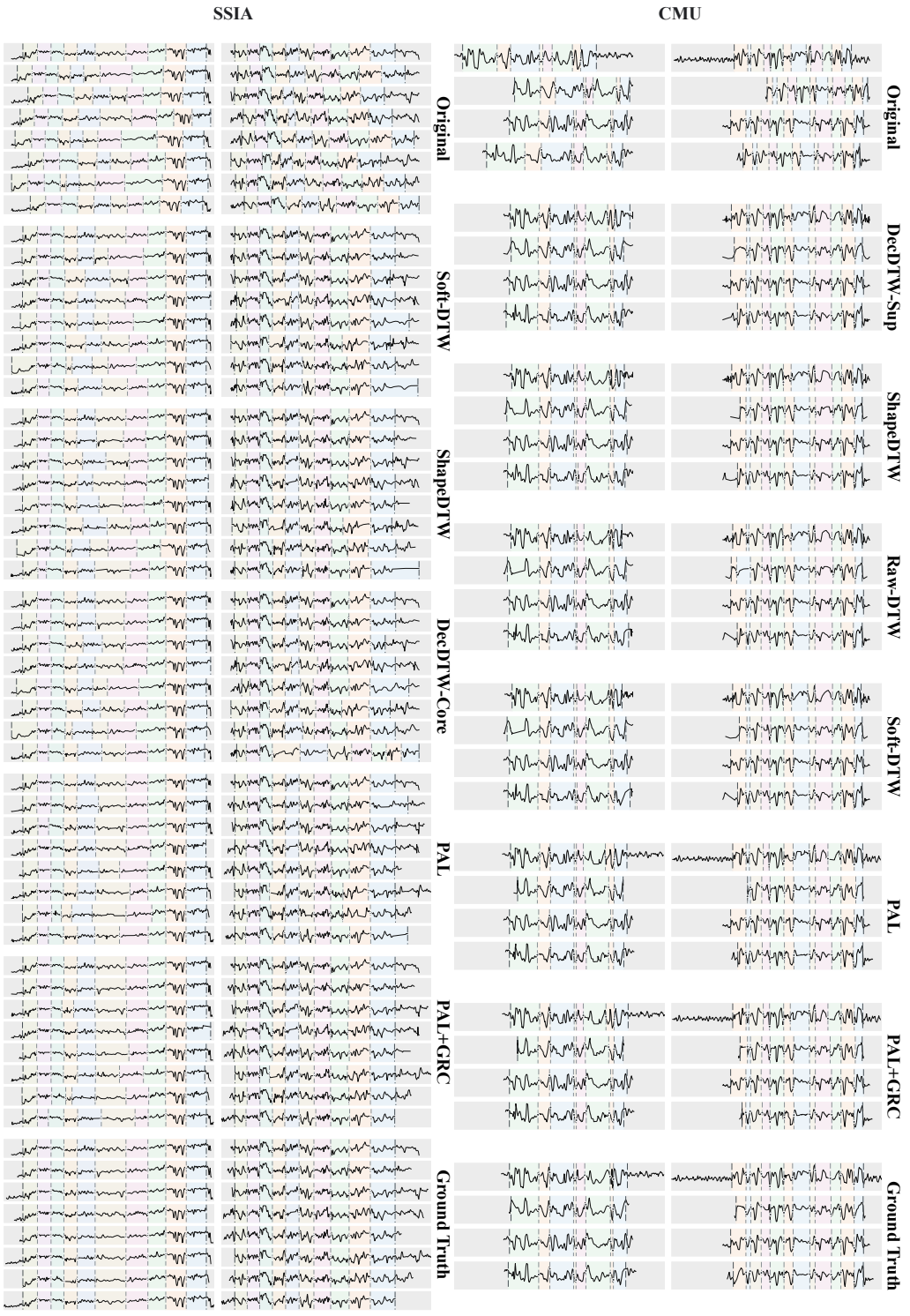


Figure 4: **Qualitative partial-correspondence comparisons.** Original observations, selected baselines, PAL, and PAL+GRC are shown alongside the references. CMU examples illustrate handling of nonshared leading and trailing regions; SSIA examples illustrate ambiguity between locally similar impedance patterns at different RGT values. PAL identifies the shared region and its correspondence, while PAL+GRC further refines local matches.