RELIABILITY-CONDITIONED PRODUCT-TANGENT LEARNING FOR DEGRADED ACOUSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synchronized acoustic descriptors offer complementary views of one signal, but they do not fail uniformly: impulsive, co-channel, and additive interference perturb their coordinates, norms, and temporal structure in different ways. A corrupted high-energy descriptor can therefore dominate Euclidean relation learning even when another domain retains useful evidence. We study this failure as reliability-conditioned geometric representation learning. The proposed RMAL-CRNN first estimates coordinate reliability, then maps every descriptor domain to its own learned spherical tangent space. Head-specific tangent-metric attention learns cross-domain relations only after this correction, and a factorized nonnegative sparse affinity refines the fused sequence over time. Controlled simulations isolate the response of these stages to individual and compound degradations. Across a public ten-class UAV benchmark and an independent eight-class Glasgow/ACSAC 2022 acoustic-identity release, the complete model reaches 86.30% and 93.88% on the ten-class benchmark at -5 and 0 dB, and 81.18% at 0 dB on Glasgow versus 74.28% for CRNN. Multi-seed ablations and Holm-corrected paired tests consistently favor the complete path. These results support reliability-before-relation learning for the evaluated descriptors, data, and corruption regimes; they do not imply universal robustness to unseen domains or degradations.

1 INTRODUCTION

Acoustic observations are often represented by several synchronized descriptors rather than one homogeneous vector. Time-, frequency-, cepstral-, and time–frequency-domain transforms expose different signal structure, and modern audio networks can learn rich temporal or spectrotemporal representations (Cakir et al., 2017; Kong et al., 2020; Gong et al., 2021). Complementarity alone is not enough under corruption. An impulse may disrupt energy-localized coordinates, co-channel interference may inflate a domain norm, and diffuse noise may erase temporal structure. Since the domains differ in scale, interpretation, and sensitivity, the same perturbation does not reduce their reliability equally. The learning problem is thus not merely how to fuse several views, but how to keep a damaged view from determining the relations among all of them.

Recent acoustic models improve representation capacity through pretraining, attention, spatial reconstruction, relation-aware scene modeling, and disentanglement across cities or devices (Hou et al., 2024; Yang & Li, 2024; Tan et al., 2024). These advances address important variability, yet synchronized handcrafted descriptors are usually treated as Euclidean channels or tokens. Under unequal corruption, an expressive fusion operator can still compare domains before it knows which coordinates are trustworthy. Fixed concatenation has the same weakness in a simpler form: high magnitude is indistinguishable from high reliability. We ask a narrower question. Can reliability be estimated before relation learning while each descriptor retains a domain-specific geometry?

Several mature ideas inform the answer, but none is the claim of this paper by itself. Robust Student-t statistics attenuate large standardized residuals (Lange et al., 1989); manifold and subspace methods preserve non-Euclidean or self-expressive structure (Roweis & Saul, 2000; Tenenbaum et al., 2000; Vidal, 2011); geometric deep learning embeds such structure in trainable models (Bronstein et al., 2017; Chakraborty et al., 2022); and attention or differentiable affinity layers learn relations end to end (Vaswani et al., 2017; Veličković et al., 2018; Ji et al., 2017). What remains unresolved

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.

here is their ordering under compound acoustic degradation. Reliability must precede geometric comparison, the domains must not be collapsed into one tangent plane, and temporal refinement must not obscure whether the cross-domain relation itself was well formed.

We instantiate this ordering in a robust manifold-attention low-rank convolutional recurrent network (RMAL-CRNN). Figure 1 previews the path. Our contribution is threefold:

- We formulate compound-corruption learning as reliability-conditioned relation estimation across synchronized, noncommensurate descriptor domains. The formulation distinguishes descriptor identity from modality or dataset domain and makes unequal reliability an explicit learning variable.
- We construct a product of learned spherical tangent spaces, then evaluate cross-domain relations with head-specific tangent-metric attention. The projection follows the corresponding spherical log-map direction blockwise; the attention distribution has a finite-bandwidth concentration envelope; and the subsequent raw temporal affinity has an explicit rank budget with nonnegative proximal thresholding.
- We align evidence with the proposed order without conflating protocols. Controlled individual and compound degradations probe stage-specific behavior and representation geometry; a separate SNR-conditioned public-data study reports multi-seed comparisons, structural ablations, confidence intervals, effect sizes, and Holm-corrected paired tests. All conclusions are restricted to the evaluated data and corruption ranges.

2 Related Work

Acoustic representation learning. Convolutional-recurrent networks, large-scale pretrained audio networks, and spectrogram transformers learn temporal or time–frequency representations directly (Cakir et al., 2017; Kong et al., 2020; Gong et al., 2021). More recent work models cooperative scene–event relations, reconstructs spatial channels without labels, or disentangles city and device factors (Hou et al., 2024; Yang & Li, 2024; Tan et al., 2024). We instead study synchronized fixed descriptors whose physical roles and corruption sensitivities differ. The method is compatible with learned front ends in principle, but that transfer is not evaluated here.

Geometry and robust subspaces. Local manifold learning and subspace clustering established complementary ways to preserve neighborhood or self-expression structure (Roweis & Saul, 2000; Tenenbaum et al., 2000; Elhamifar & Vidal, 2013; Liu et al., 2013). Geometric deep networks extend learning beyond Euclidean inputs (Bronstein et al., 2017; Chakraborty et al., 2022). Our construction does not claim these primitives as new. It uses robust coordinate weights before mapping each descriptor to a separate learned spherical tangent block, so a common ambient dimension does not erase domain identity.

Structured attention and affinity. Standard attention learns token interactions from queries and keys (Vaswani et al., 2017); graph attention and deep self-expression add structural constraints (Veličković et al., 2018; Ji et al., 2017). Squared-distance attention also has known stability distinctions from unconstrained dot-product attention (Kim et al., 2021). Here the queries and keys are projected onto learned head-specific tangent planes and compared by squared Euclidean distance in those planes. A separate factorized, nonnegative sparse relation then operates over time. This separation lets the controlled experiments ask whether cross-domain and temporal relations respond to different failure modes.

3 Reliability-Conditioned Product-Tangent Learning

3.1 Problem Setup

Let $\mathbf{r}_t \in \mathbb{R}^L$ denote the t-th acoustic frame and let K fixed transforms produce synchronized descriptor blocks

$$\mathbf{x}_t^{(k)} = \mathcal{F}_k(\mathbf{r}_t) \in \mathbb{R}^{d_k}, \qquad \mathbf{x}_t = \operatorname{concat}_{k=1}^{K} \mathbf{x}_t^{(k)}. \tag{1}$$

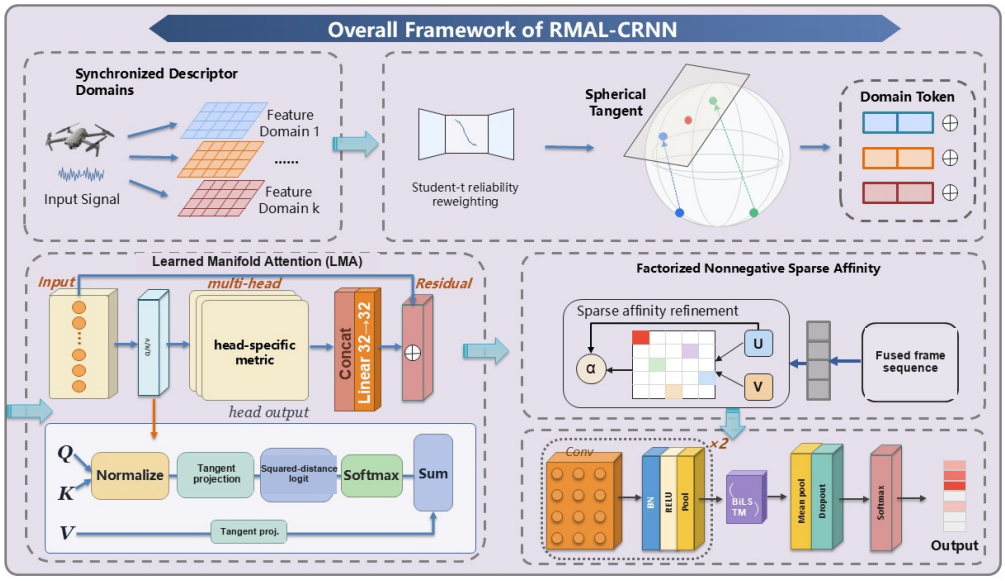


Figure 1: Ordered RMAL-CRNN path. Coordinate reliability precedes domain-specific tangent projection and cross-domain relation learning. The fused frame sequence then parameterizes learned factors for temporal affinity refinement.

The full observation is $\mathbf{X} = [\mathbf{x}_1^\top; \ldots; \mathbf{x}_T^\top] \in \mathbb{R}^{T \times D_\Sigma}$, where T is the frame count, $D_\Sigma = \sum_k d_k$, and y is the class label. These blocks are transforms of the same frame, not independent modalities or dataset domains. In our experiments $K = 6$ and every block is 32-dimensional, but the formulation permits unequal d_k. We seek a classifier $p(y \mid \mathbf{X})$ that remains useful when a corruption changes block statistics and coordinate reliability unequally.

For a unit base point $\mathbf{p}_k \in \mathbb{S}^{d-1}$, define the spherical tangent space

$$\mathcal{T}_{\mathbf{p}_k}\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \mathbf{p}_k^\top \mathbf{u} = 0\}. \tag{2}$$

The target representation is the product

$$\mathcal{M} = \prod_{k=1}^{K} \mathcal{T}_{\mathbf{p}_k}\mathbb{S}^{d-1}, \qquad \dim(\mathcal{M}) = K(d-1). \tag{3}$$

Separate base points let all blocks share an ambient dimension without imposing one geometry on statistically different descriptors.

3.2 Ordered Representation Path

Figure 1 shows the complete path. Reliability weighting is applied before any cross-domain comparison. Each corrected block is aligned and projected to a learned spherical tangent space; tangent-metric attention then estimates relations among the K domains. Only after this cross-domain fusion do we construct a compact sparse affinity over the T frames. The ordering is not cosmetic: the controlled experiments separately probe coordinate outliers, input-scale sensitivity, temporal noise, and their compound action.

3.3 Geometric Student-t Tangent-Subspace Projection (GST)

Heavy-tailed deviations motivate a Student-t-shaped reliability score (Lange et al., 1989). For domain k, learn a center $\boldsymbol{\mu}_k \in \mathbb{R}^{d_k}$, positive coordinate scale $\boldsymbol{\sigma}_k$, and degrees of freedom $\nu_k > 0$. With stabilizer $\epsilon > 0$, all-ones vector $\mathbf{1}$, elementwise division $\oslash$, and Hadamard product $\odot$, the

standardized residual $\boldsymbol{\rho}_t^{(k)}$, reliability vector $\mathbf{g}_t^{(k)}$, and corrected block $\widehat{\mathbf{x}}_t^{(k)}$ are

$$\begin{aligned}\boldsymbol{\rho}_t^{(k)} &= (\mathbf{x}_t^{(k)} - \boldsymbol{\mu}_k) \oslash (\boldsymbol{\sigma}_k + \epsilon \mathbf{1}),\\ \mathbf{g}_t^{(k)} &= (\nu_k + 1)\mathbf{1} \oslash \left(\nu_k \mathbf{1} + \boldsymbol{\rho}_t^{(k)} \odot \boldsymbol{\rho}_t^{(k)}\right),\\ \widehat{\mathbf{x}}_t^{(k)} &= \mathbf{g}_t^{(k)} \odot \mathbf{x}_t^{(k)}. \end{aligned} \tag{4}$$

Each entry satisfies $0 < g_{t,j}^{(k)} \le 1 + 1/\nu_k$. Coordinates near the learned center may be mildly amplified, whereas large standardized residuals are downweighted. This score is relative and learned; it is not a calibrated corruption probability.

The corrected block is aligned to $\mathbb{R}^d$ by a learned map $\mathbf{A}_k$, then projected orthogonally by $\Pi_{\mathbf{p}_k}$:

$$\mathbf{h}_t^{(k)} = \mathbf{A}_k \widehat{\mathbf{x}}_t^{(k)}, \qquad \mathbf{A}_k \in \mathbb{R}^{d \times d_k}, \tag{5}$$

$$\mathbf{p}_k = \frac{\overline{\mathbf{p}}_k}{\|\overline{\mathbf{p}}_k\|_2}, \qquad \Pi_{\mathbf{p}_k} = \mathbf{I}_d - \mathbf{p}_k \mathbf{p}_k^\top, \tag{6}$$

$$\mathbf{u}_t^{(k)} = \Pi_{\mathbf{p}_k} \mathbf{h}_t^{(k)}. \tag{7}$$

The final domain token $\mathbf{z}_t^{(k)}$ is normalized and receives a learned bias $\mathbf{b}_k$ in the same tangent plane:

$$\mathbf{z}_t^{(k)} = \sqrt{d}\, \frac{\mathbf{u}_t^{(k)}}{\|\mathbf{u}_t^{(k)}\|_2 + \epsilon} + \Pi_{\mathbf{p}_k} \mathbf{b}_k \in \mathcal{T}_{\mathbf{p}_k} \mathbb{S}^{d-1}. \tag{8}$$

Stacking gives $\mathbf{Z}_t = [\mathbf{z}_t^{(1)}, \dots, \mathbf{z}_t^{(K)}]^\top \in \mathbb{R}^{K \times d}$. Appendix E proves the precise geometric interpretation: for nondegenerate angles, each projected block is collinear with the corresponding spherical logarithmic coordinate and differs only by a positive scale. The statement concerns direction; normalization and the learned bias subsequently operate within the tangent block.

3.4 Learned Manifold Attention (LMA) with Tangent Metrics

We add a learned domain embedding $\mathbf{e}_k$ to retain source identity, $\widetilde{\mathbf{z}}_t^{(k)} = \mathbf{z}_t^{(k)} + \mathbf{e}_k$. For head $h \in \{1, \dots, N_h\}$, let $d_h = d/N_h$, normalize a learned base point $\mathbf{p}_h \in \mathbb{S}^{d_h - 1}$, and set $\Pi_{\mathbf{p}_h} = \mathbf{I} - \mathbf{p}_h \mathbf{p}_h^\top$. The matrices $\mathbf{W}_q^h$, $\mathbf{W}_k^h$, and $\mathbf{W}_v^h$ are learned query, key, and value maps. Queries and keys are normalized before projection; values are projected into the same head tangent plane:

$$\begin{aligned}\overline{\mathbf{q}}_i^h &= \frac{\mathbf{W}_q^h \widetilde{\mathbf{z}}_t^{(i)}}{\max(\|\mathbf{W}_q^h \widetilde{\mathbf{z}}_t^{(i)}\|_2, \epsilon)}, \quad \mathbf{q}_i^h = \Pi_{\mathbf{p}_h} \overline{\mathbf{q}}_i^h,\\ \overline{\mathbf{k}}_j^h &= \frac{\mathbf{W}_k^h \widetilde{\mathbf{z}}_t^{(j)}}{\max(\|\mathbf{W}_k^h \widetilde{\mathbf{z}}_t^{(j)}\|_2, \epsilon)}, \quad \mathbf{k}_j^h = \Pi_{\mathbf{p}_h} \overline{\mathbf{k}}_j^h,\\ \mathbf{v}_j^h &= \Pi_{\mathbf{p}_h} \mathbf{W}_v^h \widetilde{\mathbf{z}}_t^{(j)}. \end{aligned} \tag{9}$$

Instead of a dot product, the logit is the negative squared distance induced by the flat metric of the learned tangent plane:

$$\ell_{ij}^h = -\gamma_h \|\mathbf{q}_i^h - \mathbf{k}_j^h\|_2^2, \qquad \gamma_h = \exp(\overline{\gamma}_h) > 0, \tag{10}$$

$$a_{ij}^h = \frac{\exp(\ell_{ij}^h)}{\sum_{m=1}^K \exp(\ell_{im}^h)}. \tag{11}$$

With output map $\mathbf{W}_o$ and LayerNorm operator LN, the attended tokens $\mathbf{y}_{t,i}$ are

$$\mathbf{y}_{t,i} = \mathrm{LN}\left(\widetilde{\mathbf{z}}_t^{(i)} + \mathbf{W}_o \,\mathrm{concat}_{h=1}^{N_h} \sum_{j=1}^K a_{ij}^h \mathbf{v}_j^h\right). \tag{12}$$

Normalization and orthogonal projection imply $\|\mathbf{q}_i^h - \mathbf{k}_j^h\|_2^2 \le 4$. Hence the logits are bounded in $[-4\gamma_h, 0]$, and Appendix F gives a finite-bandwidth upper bound on the largest attention weight and

a corresponding entropy lower bound. The head base point removes one learned direction from each preprojected difference:

$$\|\Pi_{\mathbf{p}_h}(\overline{\mathbf{q}} - \overline{\mathbf{k}})\|_2^2 = \|\overline{\mathbf{q}} - \overline{\mathbf{k}}\|_2^2 - \left((\overline{\mathbf{q}} - \overline{\mathbf{k}})^\top \mathbf{p}_h\right)^2. \tag{13}$$

Different heads therefore learn complementary rank-$(d_h - 1)$ metrics. The distinction from standard attention (Vaswani et al., 2017) is the learned tangent projector and its squared-distance kernel in that plane, not attention as a general operation.

3.5 Factorized Nonnegative Sparse Affinity Refinement (FNSAR)

After concatenating the K attended tokens, a projection yields one frame vector $\mathbf{f}_t \in \mathbb{R}^{d_f}$, and $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_T]^\top$. Learned factor maps $\mathbf{W}_u, \mathbf{W}_v \in \mathbb{R}^{r_\ell \times d_f}$ construct a directed temporal relation with rank budget r_ℓ:

$$\begin{aligned} \mathbf{U} &= \mathbf{F}\mathbf{W}_u^\top && \in \mathbb{R}^{T \times r_\ell}, \\ \mathbf{V} &= \mathbf{F}\mathbf{W}_v^\top && \in \mathbb{R}^{T \times r_\ell}, \\ \mathbf{C}_{\text{raw}} &= r_\ell^{-1}\mathbf{U}\mathbf{V}^\top && \in \mathbb{R}^{T \times T}, \\ \mathbf{C} &= \mathrm{ReLU}\left(\mathbf{C}_{\text{raw}} - \vartheta \mathbf{1}\mathbf{1}^\top\right), \quad \vartheta = \exp(\overline{\vartheta}) > 0. \end{aligned} \tag{14}$$

Here $\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{T \times T}$, ReLU acts entrywise, and the learned positive threshold is ϑ. The raw affinity satisfies $\mathrm{rank}(\mathbf{C}_{\text{raw}}) \le r_\ell$. Entrywise thresholding is the proximal map of a nonnegative ℓ_1 penalty, so increasing ϑ can only remove support; the rank claim is made only for the pre-threshold matrix. The refined output retains a direct residual path, with logistic sigmoid σ setting the learned mixing coefficient ζ:

$$\mathbf{f}_t^{\text{out}} = (1 - \zeta)\mathbf{f}_t + \zeta\ \mathrm{LN}\left(\sum_{s=1}^{T} C_{ts}\mathbf{f}_s\right), \qquad \zeta = \sigma(\overline{\zeta}) \in (0, 1). \tag{15}$$

Appendix G proves the rank, proximal-support, and LayerNorm-residual bounds. This construction is related to differentiable self-expression (Elhamifar & Vidal, 2013; Liu et al., 2013; Ji et al., 2017), but operates on the already fused temporal sequence and explicitly enforces nonnegative sparse support.

3.6 Training and Cost

The output sequence is classified by a CRNN. With cross-entropy $\mathcal{L}_{\text{cls}}$ and nonnegative factor and sparsity weights λ_{fac} and λ_{sps}, training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \frac{\lambda_{\text{fac}}}{2r_\ell}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) + \lambda_{\text{sps}}\|\mathbf{C}\|_1. \tag{16}$$

The cross-domain attention costs $O(TK^2d)$ after linear projections. Temporal affinity formation and aggregation cost $O(T^2 r_\ell)$ and $O(T^2 d_f)$; the method remains quadratic in sequence length but avoids iterative singular-value decomposition. Appendix D records the complete configuration and hyperparameter study.

4 Experiments

The experiments address three questions. **Q1:** Do the stages respond to the failure modes they were designed for? **Q2:** Does the complete path improve classification across the evaluated SNRs? **Q3:** Are its gains consistent across matched random seeds and structural removals? Full signal-generation controls, feature definitions, implementation settings, and secondary diagnostics appear in Appendix A.

4.1 Protocol

The controlled study combines a class-dependent rotating-source model with coherent paths, modulation, turbulence, sparse impulses, co-channel interference, and diffuse Gaussian noise. The latter

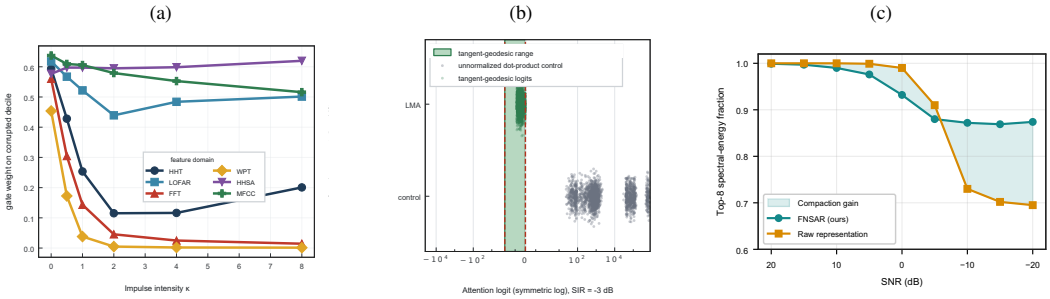


Figure 2: Controlled mechanism diagnostics assembled from separately extracted vector panels. (a) GST reliability across descriptor domains versus impulse intensity. (b) Tangent-geodesic LMA logits and an unnormalized dot-product control at SIR $= -3$ dB. (c) Top-8 spectral-energy fraction before and after FNSAR.

three are varied independently through impulse intensity κ, signal-to-interference ratio (SIR), and SNR, then superimposed for the compound condition. Six synchronized 32-dimensional descriptor domains are extracted: HHT, LOFAR, FFT, WPT, HHSA, and MFCC. These controls are normalized simulation parameters, not claimed as universal physical constants.

The public-data study combines two complementary benchmarks. The Kielce release contributes a balanced ten-class subset with equal numbers of quadrotor and hexarotor types. Glasgow/ACSAC 2022 (Diao et al., 2022) contributes eight acoustic-identity groups from DJI Mini 2 recordings (44.1-kHz, 16-bit, single-channel WAV) captured at 1 m and 5 m; these groups include propeller configurations and are not eight UAV models. Released SNR conditions are evaluated directly after the common six-domain feature extraction. Each benchmark uses a fixed split, shared optimization settings, and repeated random-seed runs; results are mean accuracy with sample standard deviation. Appendix D records the selected hyperparameters and model compositions.

The controlled study and the two public benchmarks support different claims. Q1 uses the full simulator to isolate stage responses to individual and compound degradations. Q2–Q3 use SNR-conditioned public-data evaluations; impulse and co-channel controls are not applied to those clips. Accordingly, Tables 1–2 provide low-SNR classification evidence, not end-to-end compound-corruption classification or transfer evidence for the simulator.

4.2 Q1: Controlled Mechanism Tests

Figure 2 places direct diagnostics from all three stages on a common visual baseline. Under impulses, GST changes descriptor reliability selectively rather than scaling every domain together. Under co-channel interference, the tangent-geodesic LMA logits remain inside their finite admissible interval, whereas the unnormalized dot-product control spans a much wider range. Under additive noise, FNSAR preserves a larger top-8 spectral-energy fraction than the raw representation at low SNR. Full descriptor-sensitivity, fusion-fidelity, peak-attention, affinity, and representation-drift plots appear in Appendix B. These diagnostics test stage behavior; they are not end-to-end classification results.

The compound condition is the harder mechanism test because all three failures act simultaneously. Figure 3 shows two complementary UMAP diagnostics from the controlled study. At the sample level, compound degradation disperses and overlaps class embeddings; after the complete model, the plotted clusters become more compact and separated. At the descriptor level, corruption disturbs the relative organization of the six domains, while reliability weighting and tangent projection tighten the domain clusters. The UMAP fitting configuration is unavailable, so these panels are descriptive visualizations rather than reproducible quantitative evidence or proofs of separability.

4.3 Q2: Robust Classification

Table 1 compares sequence, convolutional, attention, forecasting, and manifold-oriented models on both benchmarks. RMAL-CRNN has the highest listed mean throughout. On Kielce it reaches

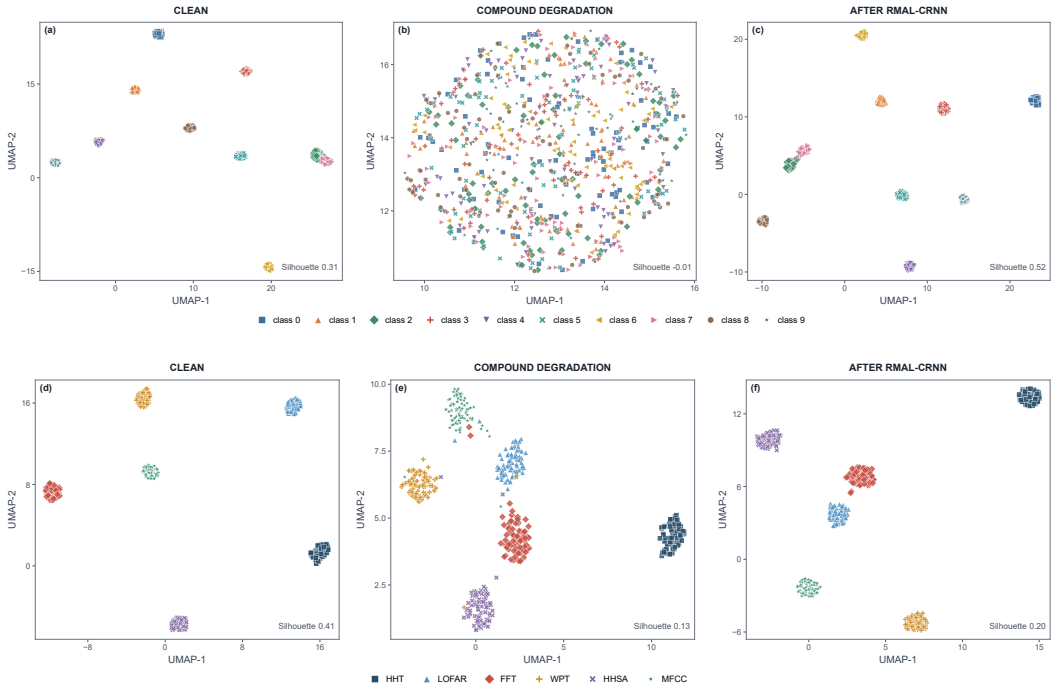


Figure 3: UMAP diagnostics under compound degradation. Top: clean, corrupted, and RMAL-CRNN sample embeddings. Bottom: clean, corrupted, and reliability-conditioned product-tangent descriptor embeddings.

86.30% and 93.88% at -5 and 0 dB, exceeding the strongest baseline by 2.27 and 3.03 points. On Glasgow it reaches 81.18% at 0 dB versus 74.28% for CRNN, and remains ahead at lower SNRs. The Glasgow results show that the reliability-conditioned path retains an advantage when the task is acoustic identity recognition under released noise conditions; they do not establish transfer across platforms or recording protocols.

Figure 4 complements the aggregate results with class-wise behavior at 0 dB. In the reported run, HHT and HHSA exhibit broader off-diagonal errors, and direct concatenation does not recover the strongest individual-domain structure. RMAL-CRNN produces the most concentrated diagonal among the displayed matrices. This panel is descriptive and is not used in the paired inference below.

4.4 Q3: Component Necessity

We remove one operation at a time with fixed data and training settings. Parameter counts span only 333.3k–339.4k and are not monotonic with accuracy. Table 2 shows losses of 1.41, 1.50, and 1.80 points without FNSAR, LMA, and GST, and 6.96 points against CRNN. Changes use unrounded paired run differences, explaining the apparent 0.01 discrepancy from subtracting the rounded RMA-CRNN means. Comparison and ablation summarize separate repeated-run sets; paired tests use the multi-seed ablation set, and the summaries are not pooled.

Matched paired tests provide a second check across the paired multi-seed runs. All reported 95% confidence intervals for the complete-minus-control difference have positive lower bounds. After Holm–Bonferroni correction, the adjusted p-values remain significant, with positive paired effect sizes. The full statistics are in Table 4. This evidence supports the dependency-aware complete path for the evaluated runs. It does not, by itself, prove that the same ordering is optimal for every descriptor family or corruption distribution.

Table 1: Accuracy across SNRs (%, mean ± sample standard deviation). Panel (a) uses the ten-class Kielce benchmark; panel (b) uses the eight-class Glasgow/ACSAC 2022 acoustic-identity benchmark.

(a) Kielce, ten classes

Model (literature index)	−5 dB	0 dB	5 dB	10 dB
MAtt ((Chakraborty et al., 2022; Vaswani et al., 2017))	69.65±1.07	80.33±1.43	84.50±1.41	88.65±1.17
InceptionTime ((Ismail Fawaz et al., 2020))	78.90±0.79	86.30±0.42	90.75±0.70	94.65±0.50
CNN-Transformer ((Vaswani et al., 2017))	79.55±0.97	86.83±0.48	91.23±0.70	93.85±0.45
CRNN ((Cakir et al., 2017))	80.80±0.82	87.15±0.33	91.80±0.34	94.68±0.64
SEResNet1D ((Hu et al., 2020))	81.98±0.49	89.28±0.13	92.73±0.52	95.78±0.46
TCN ((Bai et al., 2018))	83.70±1.22	90.85±0.49	93.15±0.19	96.28±0.39
TimeMixer++ ((Wang et al., 2025))	84.03±1.21	89.48±0.78	93.68±0.60	96.68±0.49
MambaSL ((Gu & Dao, 2023))	82.53±1.15	89.28±1.07	93.20±0.18	95.98±0.36
RMAL-CRNN (ours)	86.30±0.29	93.88±0.36	96.08±0.35	97.48±0.51

(b) Glasgow/ACSAC 2022 (Diao et al., 2022), eight classes

Model (literature index)	−15 dB	−10 dB	−5 dB	0 dB
MAtt ((Chakraborty et al., 2022; Vaswani et al., 2017))	30.35±1.38	41.68±1.78	54.43±1.23	65.12±0.41
InceptionTime ((Ismail Fawaz et al., 2020))	38.90±1.27	54.25±1.51	66.85±0.72	75.02±1.31
CNN-Transformer ((Vaswani et al., 2017))	38.32±0.34	55.10±0.57	69.12±0.66	78.53±0.40
CRNN ((Cakir et al., 2017))	35.35±1.15	51.38±1.02	66.50±1.75	74.28±0.62
SEResNet1D ((Hu et al., 2020))	40.75±0.19	56.12±0.88	69.38±1.04	78.50±1.13
SPDNet ((Huang & Van Gool, 2017))	34.62±0.21	46.03±1.45	61.85±1.45	70.88±2.43
TCN ((Bai et al., 2018))	36.35±1.31	50.30±0.52	65.05±1.50	74.95±0.58
BiLSTM ((Graves & Schmidhuber, 2005))	36.42±0.57	50.68±1.84	65.83±1.14	73.95±0.44
RMAL-CRNN (ours)	45.28±1.21	62.70±0.91	74.10±0.88	81.18±1.64

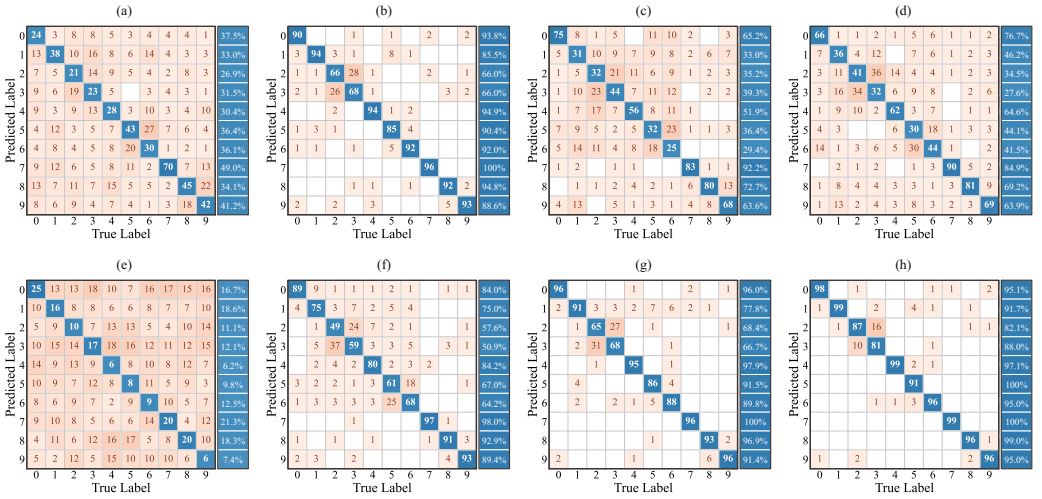


Figure 4: Confusion matrices at 0 dB for individual descriptors, concatenation, and RMAL-CRNN. Percentages in parentheses are the reported overall accuracies.

5 Limitations

The evidence uses six fixed descriptors, two public UAV benchmarks, and matched synthetic degradations; it does not establish transfer to learned front ends, other platforms, unseen corruptions, broader reverberation, or multimodal fusion. UMAP and mechanism diagnostics are descriptive, while the theory characterizes representation structure rather than classification accuracy. Cross-dataset transfer and corruption mismatch remain the main empirical gaps.

Table 2: Structural ablations at 0 dB (%, mean ± standard deviation).

Variant	Composition	Accuracy	Change
RMAL-CRNN	GST + LMA + FNSAR	93.89 ± 0.33	–
RMA-CRNN	GST + LMA	92.47 ± 0.49	-1.41
RL-CRNN	GST + FNSAR	92.39 ± 0.52	-1.50
MAL-CRNN	LMA + FNSAR	92.09 ± 0.38	-1.80
CRNN	baseline	86.93 ± 0.35	-6.96

6 Conclusion

Compound degradation makes relation learning difficult because synchronized descriptors lose reliability at different rates. RMAL-CRNN addresses that asymmetry before fusion: it reweights coordinates, preserves separate product-tangent blocks, learns cross-domain relations with head-specific tangent metrics, and only then refines temporal affinity. Controlled mechanism tests support the intended stage responses, while two separate public-data studies support low-SNR classification and component necessity. The broader implication is modest but useful: when input families are noncommensurate and fail unequally, reliability should be treated as a prerequisite for relation estimation rather than as a correction applied after fusion.

The immediate next step is to move beyond synchronized descriptors from a single acoustic modality toward genuinely heterogeneous multimodal inputs. Audio, vision, radio-frequency, and other sensing streams require modality-specific reliability estimation, geometry-aware cross-modal alignment, and mechanisms that remain stable when one modality is missing or corrupted. Cross-dataset transfer, unseen degradation mixtures, learned front ends, and more efficient long-sequence affinity models are complementary directions for testing how far the reliability-before-relation principle generalizes.

Statements

AI use. Generative AI assisted with language, literature checks, figure labels, and LaTeX quality control; it did not generate data, select runs, or alter results, and the authors verified all assisted content. **Ethics.** UAV acoustic recognition has safety and dual-use risks; deployment requires authorization, data governance, false-alarm assessment, and human oversight. **Reproducibility.** The appendices provide feature construction, degradation equations, selected settings, statistics, and proofs. Exact dataset counts, a persistent locator, the complete training schedule, and UMAP settings are unavailable and are not claimed; no public-code claim is made.

References

Jérôme Antoni. Cyclostationarity by examples. *Mechanical Systems and Signal Processing*, 23(4): 987–1036, 2009. doi: 10.1016/j.ymssp.2008.10.010.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

A. P. C. Bresciani, Jean Maillard, and Leandro D. de Santana. Physics-based scintillations for outdoor sound auralization. *Journal of the Acoustical Society of America*, 154(2):1179–1190, 2023. doi: 10.1121/10.0020666.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017.

Emre Cakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.

Rudrasis Chakraborty, Javier Bouza, Jonathan Manton, and Baba C. Vemuri. Manifoldnet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2022. doi: 10.1109/TPAMI.2020.3003846.

Yufeng Diao, Yichi Zhang, Guodong Zhao, and Mohamed Khamis. Drone authentication via acoustic fingerprint. In *Proceedings of the 38th Annual Computer Security Applications Conference*, pp. 658–668. ACM, 2022. doi: 10.1145/3564625.3564653. URL https://doi.org/10.1145/3564625.3564653.

Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

Jun Fang, Anthony Finn, Rick Wyber, et al. Acoustic detection of unmanned aerial vehicles using biologically inspired vision processing. *Journal of the Acoustical Society of America*, 151(2): 968–981, 2022.

Z. F. T. Gan, V. Tumelero Valente, Kenneth S. Brentner, and Eric Greenwood. Broadband noise modulation of multirotor aircraft. *Journal of the Acoustical Society of America*, 157(2):924–939, 2025. doi: 10.1121/10.0035646.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Interspeech*, pp. 571–575, 2021.

Alex Graves and Juergen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. In *2005 IEEE International Joint Conference on Neural Networks*, volume 4, pp. 2048–2053, 2005. doi: 10.1109/IJCNN.2005.1556215.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. doi: 10.48550/arXiv.2312.00752.

Yuchen Hou, Boseong Kang, Audrius Mitchell, Wenwu Wang, Jian Kang, and Dick Botteldooren. Cooperative scene-event modelling for acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:68–82, 2024. doi: 10.1109/TASLP.2023.3323135.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020.

Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2036–2042, 2017.

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*, 34:1936–1962, 2020. doi: 10.1007/s10618-020-00710-y.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, 2017.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571, 2021.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Yizhen Tan, Haoyu Ai, Shanshan Li, and Mark D. Plumbley. Acoustic scene classification across cities and devices via feature disentanglement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1286–1297, 2024. doi: 10.1109/TASLP.2024.3353578.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Rene Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025.

Bo Yang and Xuelong Li. Self-supervised learning of spatial acoustic representation with cross-channel signal reconstruction and multi-channel conformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4211–4225, 2024. doi: 10.1109/TASLP.2024.3458811.

A Experimental and Implementation Details

A.1 Frame and Descriptor Construction

Let $r[n]$ be the real-valued sampled signal at discrete sample index n. The t-th length-L frame with hop Δ is

$$\mathbf{r}_t = \big[r[(t-1)\Delta], \ldots, r[(t-1)\Delta + L - 1]\big]^\top. \tag{17}$$

Each frame is transformed into HHT, LOFAR, FFT, WPT, HHSA, and MFCC blocks. The implementation standardizes each block to 32 coordinates and retains fixed block boundaries throughout training. The six blocks are synchronized descriptions of the same frame. They should not be interpreted as independent sensors or missing-modality inputs.

A.2 Controlled Source and Degradation Model

At continuous time τ, the class-c source $s_c(\tau)$ combines harmonic orders, modulated broadband fluctuations, and periodic ringing, consistent with measured multirotor modulation (Fang et al., 2022; Gan et al., 2025) and rotating-machine cyclostationarity (Antoni, 2009):

$$\begin{aligned} s_c(\tau) = {} & \varpi_{\mathrm{B}} \sin(2\pi f_{\mathrm{B}}^{(c)} \tau) + \varpi_{2\mathrm{B}} \sin(4\pi f_{\mathrm{B}}^{(c)} \tau) \\ & + \varpi_{\mathrm{w}} w(\tau) \big[1 + \varpi_{\mathrm{M}} \sin(2\pi f_{\mathrm{M}}^{(c)} \tau)\big] \\ & + \varpi_{\mathrm{r}} \sin(2\pi f_{\mathrm{b}} \tau) \exp[-\beta_0 (\tau \bmod 1/f_{\mathrm{R}}^{(c)})]. \end{aligned} \tag{18}$$

Here $f_{\mathrm{B}}^{(c)}$, $f_{\mathrm{M}}^{(c)}$, and $f_{\mathrm{R}}^{(c)}$ are the class-dependent blade-passing, modulation, and repetition rates; f_{b} is the ringing frequency; $w(\tau)$ is 1-kHz low-pass noise; and the ϖ coefficients and β_0 control relative component strengths and ringing decay. The normalized controls are $(\varpi_{\mathrm{B}}, \varpi_{2\mathrm{B}}, \varpi_{\mathrm{w}}, \varpi_{\mathrm{M}}, \varpi_{\mathrm{r}}, \beta_0) = (1, 0.5, 0.4, 0.8, 0.8, 30)$; classes differ through the blade-passing, modulation, and transient rates.

After coherent delayed paths, a turbulence envelope, and Doppler phase, the three controlled additive degradations are applied to form $r_{\mathrm{sim}}(\tau)$:

$$\begin{aligned} r_{\mathrm{sim}}(\tau) = {} & \Re\left\{ \mathrm{HT}\left[\sum_{p=0}^{P} \chi_p \widetilde{s}_c(\tau - \delta_p) g_{\mathrm{turb}}(\tau) e^{\mathrm{i}\Phi_{\mathrm{dop}}(\tau)} \right] \right\} \\ & + \varepsilon_{\mathrm{imp}}(\tau; \kappa) + J(\tau; \mathrm{SIR}) + v(\tau; \mathrm{SNR}). \end{aligned} \tag{19}$$

Here P is the maximum path index, χ_p and δ_p are path gain and delay, $\widetilde{s}_c$ is the propagated class source, g_{turb} is the turbulence envelope, Φ_{dop} is Doppler phase, HT forms the analytic representation, and $\Re\{\cdot\}$ takes its real part. The terms $\varepsilon_{\mathrm{imp}}$, J, and v denote sparse impulses, a competing source, and diffuse noise; κ, SIR, and SNR independently control their strengths. Their simultaneous application defines the compound condition. The random envelope is motivated by physics-based acoustic scintillation models (Bresciani et al., 2023); exact simulator sampling distributions and parameter ranges were not preserved with the experimental record and are not reconstructed here.

A.3 Public-Data Classes

Table 3: Balanced public-data subset used in the experiments.

Structure	UAV models
Quadrotor (X4)	MATRICE 300; Mavic 3; Mavic Air 2S; Mavic Air 2; Mavic 2 Zoom
Hexarotor (X6)	Yuneec H520; Yuneec H520E RTK; DJI S900; DJI F550; Tarot 690S

The subset, preprocessing, and split are held fixed across the reported comparisons. A formal bibliographic record, persistent access URL, exact split counts, and redistribution terms were not preserved with the experimental record. We therefore identify the collection by its releasing team and institution and do not claim that the present draft alone is sufficient to reconstruct the split.

B Additional Mechanism Diagnostics

These panels expand the three diagnostics in Figure 2. Figure 5 shows unequal descriptor sensitivity and selective GST correction, Figure 6 isolates the scale response of LMA against an unnormalized control, and Figure 7 reports the descriptor and temporal-compaction quantities used for FNSAR.

C Complete Paired Statistics

Table 4: Paired Student t-tests between RMAL-CRNN and control variants at 0 dB over the paired multi-seed runs (Holm–Bonferroni corrected). Differences use unrounded paired runs.

Control	Difference	95% CI	t	d_z	Holm p
RMA-CRNN	$+1.41$	$[+0.77, +2.05]$	$+5.41$	$+2.05$	1.6×10^{-3}
RL-CRNN	$+1.50$	$[+1.03, +1.97]$	$+7.78$	$+2.94$	7.1×10^{-4}
MAL-CRNN	$+1.80$	$[+1.20, +2.40]$	$+7.35$	$+2.78$	7.1×10^{-4}
CRNN	$+6.96$	$[+6.69, +7.22]$	$+63.95$	$+24.17$	3.9×10^{-9}

D Implementation and Sensitivity

The selected attention head count is $N_h = 4$, factor rank $r_\ell = 8$, and fusion dimension $d_f = 64$. The learning rate is 10^{-3} and batch size is 32. These values come from the repeated scans summarized in Figure 8; the scan is a sensitivity analysis, not evidence that the settings are globally optimal.

The comparison architectures cover sequence, convolutional, attention, and manifold-oriented alternatives. InceptionTime (Ismail Fawaz et al., 2020) stacks six Inception modules; CNN-Transformer uses two one-dimensional convolution layers followed by a four-layer, four-head Transformer encoder; SEResNet1D is an 18-layer residual network with squeeze-and-excitation attention (Hu et al., 2020); TCN (Bai et al., 2018) uses dilated causal convolutions with dilation factors $(1, 2, 4, 8)$; and TimeMixer++ follows the multiscale pattern architecture of Wang et al. (2025). All models share the same split and training protocol. Primary citations or complete external configurations for the implementation labels MAtt, CNN-Transformer, and MambaSL were not preserved with the experimental record.

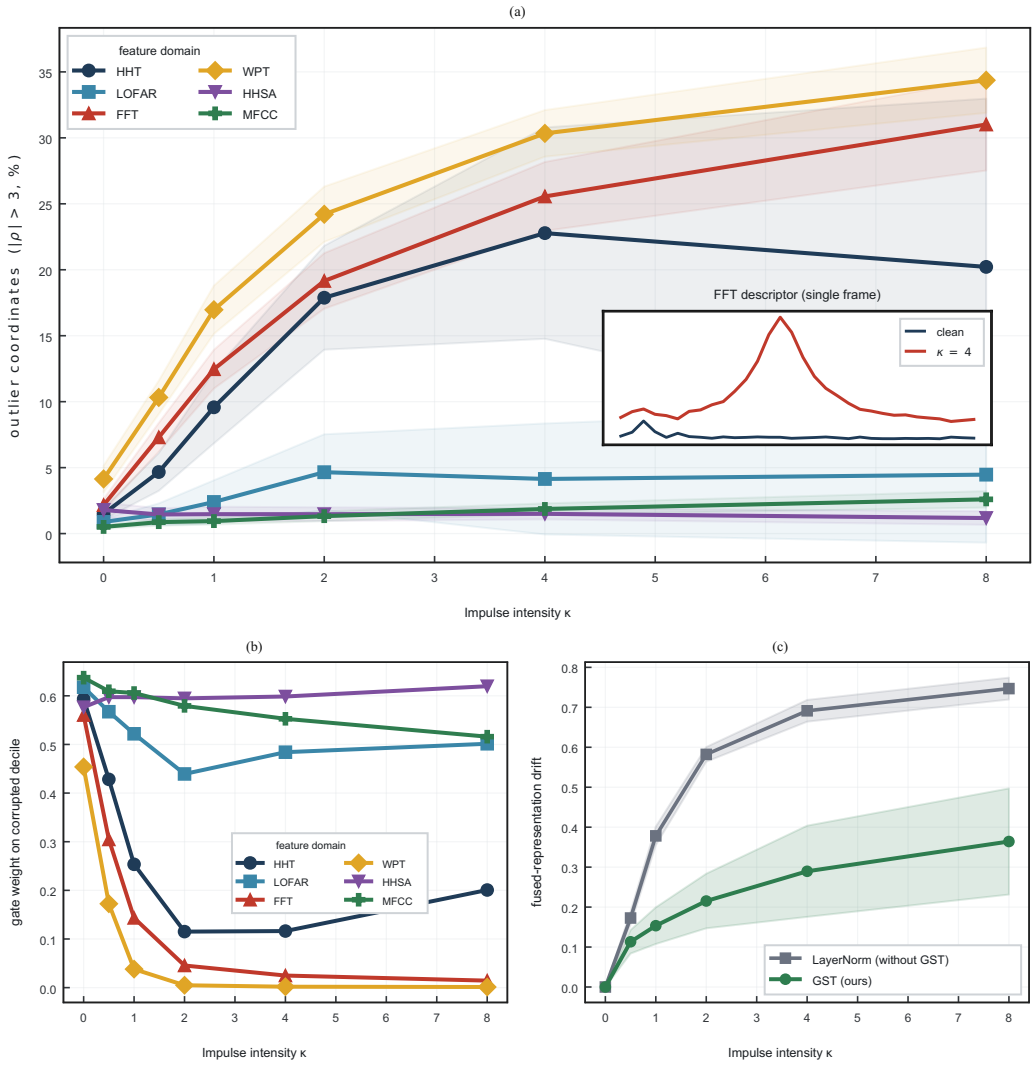


Figure 5: Impulsive corruption affects the six descriptor domains unequally. Learned reliability weights change selectively, and the corrected fused representation drifts more smoothly from its clean reference than the LayerNorm control in the evaluated range.

TimeMixer++ Macro-F1 is 84.00 ± 1.16, 89.44 ± 0.77, 93.68 ± 0.60, and 96.67 ± 0.49 at -5, 0, 5, and 10 dB. MambaSL Macro-F1 is 82.58 ± 1.19, 89.30 ± 1.05, 93.20 ± 0.17, and 95.97 ± 0.36. Means and sample standard deviations are recomputed from four logged seeds.

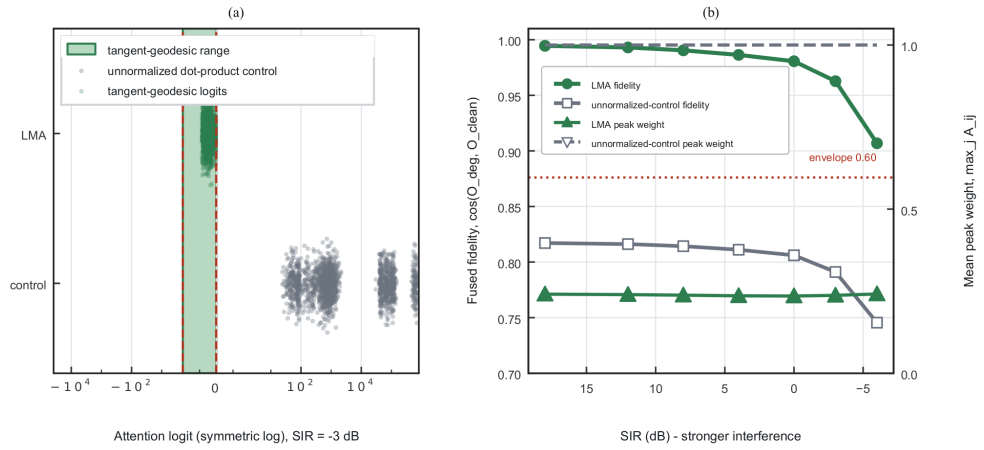


Figure 6: Tangent-metric attention under co-channel interference. The diagnostic fixes $\gamma = 0.50$; normalized projected logits stay within their admissible interval, whereas the unnormalized dot-product control is scale-sensitive.

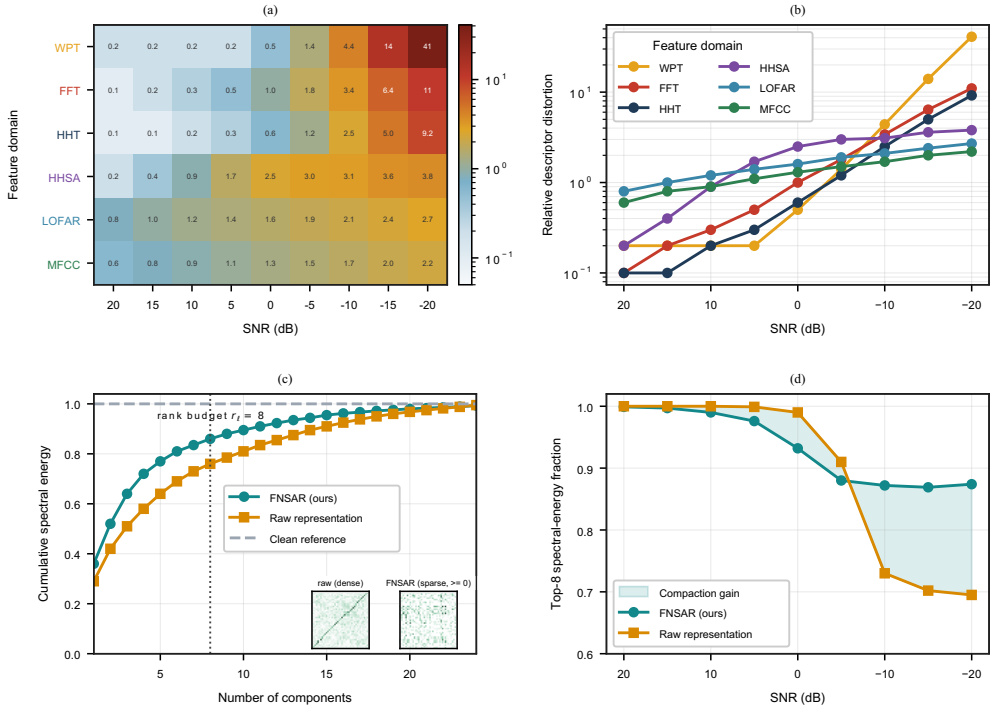


Figure 7: Descriptor sensitivity and factorized sparse temporal-affinity behavior under additive Gaussian noise.

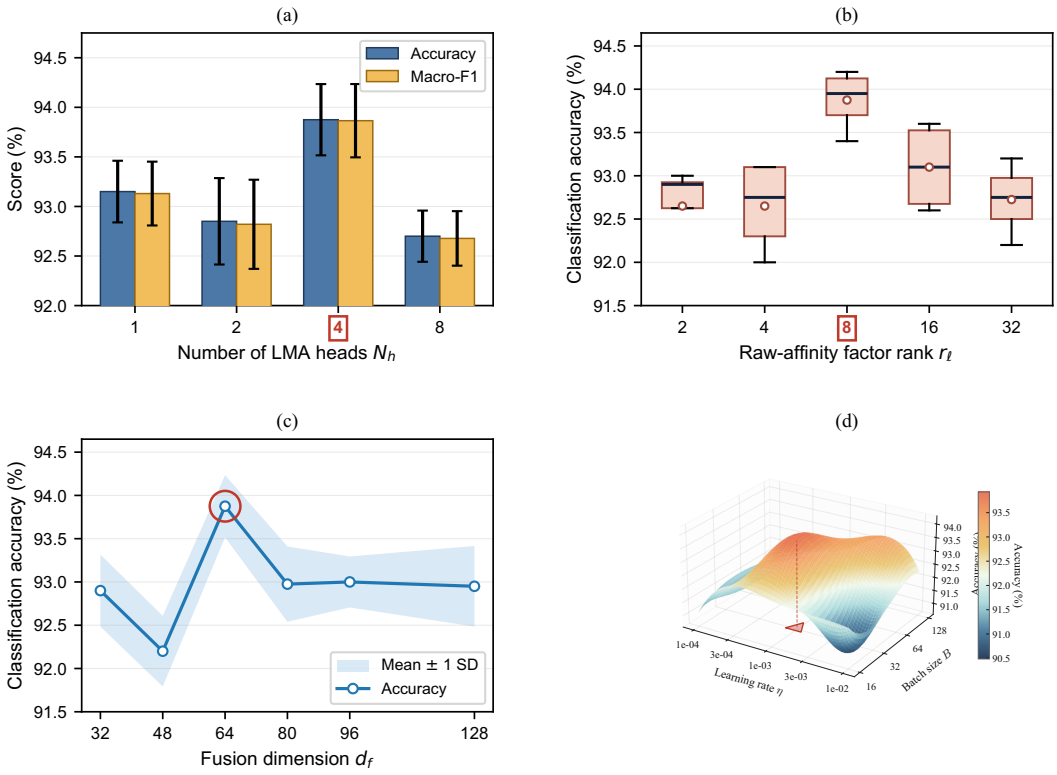


Figure 8: Sensitivity to attention heads, factor rank, fusion dimension, and the learning-rate/batch-size pair.

This appendix gives the complete proofs supporting the geometric interpretation and structural analysis in the main paper. It establishes the spherical log-map directional relation of the tangent projection, the concentration envelope of tangent-metric attention, and the factorized proximal structure of the temporal affinity.

E Blockwise Directional Equivalence on a Product of Spheres

Proposition S.1: Let $K \geq 1$ and $d \geq 2$. For every domain k, let the aligned gated vector in equation 5 satisfy $\mathbf{h}_t^{(k)} \neq \mathbf{0}$, and define

$$\boldsymbol{\upsilon}_t^{(k)} = \frac{\mathbf{h}_t^{(k)}}{\|\mathbf{h}_t^{(k)}\|_2}, \qquad \theta_{t,k} = \arccos\Big((\boldsymbol{\upsilon}_t^{(k)})^\top \mathbf{p}_k\Big),$$
$$\mathbf{u}_t^{(k)} = \Pi_{\mathbf{p}_k} \mathbf{h}_t^{(k)}, \qquad \omega_{t,k} = \frac{\theta_{t,k}}{\|\mathbf{h}_t^{(k)}\|_2 \sin\theta_{t,k}}. \tag{20}$$

Assume $\theta_{t,k} \in (0, \pi)$ for all k. On the product manifold $\mathcal{P} = (\mathbb{S}^{d-1})^K$, stack the base points, normalized aligned vectors, and tangent components as $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_K)$, $\boldsymbol{\upsilon}_t = (\boldsymbol{\upsilon}_t^{(1)}, \ldots, \boldsymbol{\upsilon}_t^{(K)})$, and $\mathbf{u}_t = [(\mathbf{u}_t^{(1)})^\top, \ldots, (\mathbf{u}_t^{(K)})^\top]^\top \in \mathbb{R}^{Kd}$. Then

$$\mathrm{Log}_{\mathbf{p}}(\boldsymbol{\upsilon}_t) = \boldsymbol{\Omega}_t \mathbf{u}_t, \qquad \boldsymbol{\Omega}_t = \mathrm{blkdiag}_{k=1}^{K}(\omega_{t,k} \mathbf{I}_d) \succ \mathbf{0}. \tag{21}$$

Hence each domain block of the orthogonal projection is collinear with, and points in the same direction as, the corresponding spherical logarithmic map. This blockwise relation is the product-manifold statement used by GST; normalization and the learned bias subsequently act within their corresponding tangent spaces.

Proof: The product-manifold logarithmic map is the concatenation of its component maps. For domain k, the logarithmic map on $\mathbb{S}^{d-1}$ is

$$\log_{\mathbf{p}_k}(\boldsymbol{\upsilon}_t^{(k)}) = \frac{\theta_{t,k}}{\sin\theta_{t,k}} \Big(\boldsymbol{\upsilon}_t^{(k)} - \cos\theta_{t,k}\, \mathbf{p}_k\Big). \tag{22}$$

Because $\cos\theta_{t,k} = (\mathbf{h}_t^{(k)})^\top \mathbf{p}_k / \|\mathbf{h}_t^{(k)}\|_2$,

$$\boldsymbol{\upsilon}_t^{(k)} - \cos\theta_{t,k}\, \mathbf{p}_k = \frac{\Pi_{\mathbf{p}_k} \mathbf{h}_t^{(k)}}{\|\mathbf{h}_t^{(k)}\|_2} = \frac{\mathbf{u}_t^{(k)}}{\|\mathbf{h}_t^{(k)}\|_2}. \tag{23}$$

Substitution into equation 22 gives $\log_{\mathbf{p}_k}(\boldsymbol{\upsilon}_t^{(k)}) = \omega_{t,k} \mathbf{u}_t^{(k)}$. Stacking the K identities yields equation 21. Since every $\omega_{t,k} > 0$, the direction of each nonzero block is preserved. At $\theta_{t,k} = 0$, both component vectors vanish and no direction is defined; at $\theta_{t,k} = \pi$, the component logarithmic map is not unique.

The directional agreement has a simple essence: both constructions use the same high-dimensional orthogonal component $\Pi_{\mathbf{p}_k} \mathbf{h}_t^{(k)}$, and spherical curvature only multiplies that component by a positive blockwise scale. This result supports the projection term in equation 8, after which normalization and the learned bias remain within the same tangent block.

F Concentration Analysis of Tangent-Metric Attention

Proposition S.2: For head h, let

$$D_{\max}^h = \max_{1 \leq i,j \leq K} \|\mathbf{q}_i^h - \mathbf{k}_j^h\|_2^2 \leq 4. \tag{24}$$

Then the logits and weights in equation 10–equation 11 satisfy

$$-\gamma_h D_{\max}^h \leq \ell_{ij}^h \leq 0, \tag{25}$$

$$\max_j a_{ij}^h \leq \frac{1}{1 + (K-1) e^{-\gamma_h D_{\max}^h}}, \tag{26}$$

$$\mathsf{H}(\mathbf{a}_{i,\cdot}^h) \geq \log\Big(1 + (K-1) e^{-\gamma_h D_{\max}^h}\Big). \tag{27}$$

Proof: By equation 9, $\|\bar{\mathbf{q}}_i^h\|_2 \le 1$ and $\|\bar{\mathbf{k}}_j^h\|_2 \le 1$. Orthogonal projection is non-expansive, so $\|\mathbf{q}_i^h\|_2 \le 1$ and $\|\mathbf{k}_j^h\|_2 \le 1$. The triangle inequality gives $\|\mathbf{q}_i^h - \mathbf{k}_j^h\|_2 \le 2$, proving equation 24; equation 25 then follows from $\gamma_h > 0$.

Let $j^\star = \arg\max_j \ell_{ij}^h$. The softmax can be written as

$$\max_j a_{ij}^h = \frac{1}{1 + \sum_{m \neq j^\star} e^{\ell_{im}^h - \ell_{ij^\star}^h}}. \tag{28}$$

The logit interval implies $\ell_{im}^h - \ell_{ij^\star}^h \ge -\gamma_h D_{\max}^h$. Lower-bounding every term in the denominator proves equation 26. Finally, for any discrete distribution $\mathbf{a}$, $\mathsf{H}(\mathbf{a}) \ge -\log \max_j a_j$, which gives equation 27.

For finite γ_h and $K > 1$, the peak-weight bound is strictly below one and the entropy bound is positive. The proposition therefore gives an explicit concentration envelope for tangent-metric attention at a finite inverse bandwidth. Its distinctive geometric ingredient is the head-specific rank-$(d_h - 1)$ projector, which selects complementary tangent directions for cross-domain relations.

G Raw-Affinity Rank Budget, Proximal Sparsity, and Residual-Output Bound

Let $\mathbf{F} \in \mathbb{R}^{T \times d_f}$, $\mathbf{W}_u, \mathbf{W}_v \in \mathbb{R}^{r_\ell \times d_f}$, and define

$$\mathbf{U} = \mathbf{F}\mathbf{W}_u^\top \qquad \in \mathbb{R}^{T \times r_\ell}, \tag{29}$$

$$\mathbf{V} = \mathbf{F}\mathbf{W}_v^\top \qquad \in \mathbb{R}^{T \times r_\ell}, \tag{30}$$

$$\mathbf{C}_{\text{raw}} = \frac{1}{r_\ell}\mathbf{U}\mathbf{V}^\top \qquad \in \mathbb{R}^{T \times T}, \tag{31}$$

$$\mathbf{C} = \mathrm{ReLU}\big(\mathbf{C}_{\text{raw}} - \vartheta \mathbf{1}_T \mathbf{1}_T^\top\big) \qquad \in \mathbb{R}^{T \times T}. \tag{32}$$

Proposition S.3.1: The pre-threshold affinity satisfies

$$\mathrm{rank}(\mathbf{C}_{\text{raw}}) \le r_\ell, \tag{33}$$

independent of T.

Proof: Since $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{T \times r_\ell}$,

$$\mathrm{rank}(\mathbf{U}\mathbf{V}^\top) \le \min\{\mathrm{rank}(\mathbf{U}), \mathrm{rank}(\mathbf{V})\} \le r_\ell.$$

Multiplication by $1/r_\ell$ preserves rank, and the bound is attained when both factors have full column rank. Proposition S.3.1 characterizes the compact pre-threshold relation, while Propositions S.3.2–S.3.3 characterize the nonnegative sparse support and residual aggregation used by the network. Asymmetric entries encode directional temporal dependence, and diagonal entries retain frame self-information.

Proposition S.3.2: Let $g_\vartheta(\mathbf{G}) = \vartheta\|\mathbf{G}\|_1 + \iota_{\ge 0}(\mathbf{G})$, where $\iota_{\ge 0}$ is the indicator of the nonnegative orthant. Then

$$\mathbf{C} = \arg\min_{\mathbf{G} \ge 0} \frac{1}{2}\|\mathbf{G} - \mathbf{C}_{\text{raw}}\|_F^2 + \vartheta\|\mathbf{G}\|_1, \tag{34}$$

$$\mathbf{C} = \mathrm{prox}_{g_\vartheta}(\mathbf{C}_{\text{raw}}), \tag{35}$$

and

$$\vartheta_2 > \vartheta_1 \implies \mathrm{supp}(\mathbf{C}_{\vartheta_2}) \subseteq \mathrm{supp}(\mathbf{C}_{\vartheta_1}). \tag{36}$$

Proof: The objective is separable. For one entry $x = C_{\text{raw},ts}$, solve

$$\min_{g \ge 0} \ \frac{1}{2}(g - x)^2 + \vartheta g.$$

Strict convexity gives the unique solution $g^\star = \max\{x - \vartheta, 0\}$. Applying this result entry-wise proves equation 34–equation 35. Since $C_{ts} > 0$ exactly when $C_{\text{raw},ts} > \vartheta$, increasing ϑ can only remove support.

Proposition S.3.3: Let the aggregated column vector be $\mathbf{m}_t = (\mathbf{CF})^\top_{t,:} = \sum_{s=1}^{T} C_{ts}\mathbf{f}_s$ and

$$\mathrm{LN}(\mathbf{m}_t) = \boldsymbol{\psi}_{\mathrm{LN}} \odot \hat{\mathbf{m}}_t + \mathbf{b}_{\mathrm{LN}}, \qquad \|\hat{\mathbf{m}}_t\|_2 \le \sqrt{d_f}.$$

For finite LayerNorm parameters, define

$$G_{\mathrm{LN}} = \|\boldsymbol{\psi}_{\mathrm{LN}}\|_\infty \sqrt{d_f} + \|\mathbf{b}_{\mathrm{LN}}\|_2.$$

Then the output in equation 15 satisfies

$$\|\mathbf{f}_t^{\mathrm{out}}\|_2 \le (1-\zeta)\|\mathbf{f}_t\|_2 + \zeta G_{\mathrm{LN}}, \tag{37}$$

independent of the scale of $\mathbf{C}$.

Proof: The LayerNorm bound follows from

$$\|\,\mathrm{LN}(\mathbf{m}_t)\|_2 \le \|\boldsymbol{\psi}_{\mathrm{LN}}\|_\infty \|\hat{\mathbf{m}}_t\|_2 + \|\mathbf{b}_{\mathrm{LN}}\|_2 \le G_{\mathrm{LN}}.$$

Applying the triangle inequality to equation 15 gives

$$\|\mathbf{f}_t^{\mathrm{out}}\|_2 \le (1-\zeta)\|\mathbf{f}_t\|_2 + \zeta\|\,\mathrm{LN}(\mathbf{m}_t)\|_2,$$

and equation 37 follows. This establishes the LayerNorm-residual bound used by the temporal refinement.