Resolving Time-Frequency Ridge Crossings via Frequency-Rate Lifting

Pingping Pan
Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen)
Shenzhen 518107, China

You Li*
Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen)
Shenzhen 518107, China

Yunjian Zhang*
Shenzhen Kaihong Digital Industry Development Company Ltd.
Shenzhen 518000, China
and Harbin Institute of Technology (Chongqing)
Chongqing 401151, China
yunjzhang@xmu.edu.cn

Mujiangshan Wang
Shenzhen Kaihong Digital Industry Development Company Ltd.
Shenzhen 518000, China
and Harbin Institute of Technology (Chongqing)
Chongqing 401151, China

Abstract

Ridge crossings, where two components share the same instantaneous frequency (IF) at a given time, are a fundamental bottleneck in multi-component time-frequency (TF) mode decomposition. We trace this difficulty to *insufficient representational dimensionality* and resolve it by lifting signals into the 3-D time-frequency-frequency-rate (TFFR) space $(t, f, \dot{f})$, where crossing ridges separate because they typically differ in the instantaneous frequency rate (IFR). We formalize this via a separability proposition and confirm it empirically through Monte Carlo simulation (unresolved-crossing rate: 69.00% in 2-D $\rightarrow$ 6.12% in 3-D). Based on this principle, we propose **TFFR-Net**, a query-based masked transformer that jointly predicts per-ridge masks and dense IFR maps. An IFR feedback loop injects the estimated frequency rate back into each query embedding, enabling the decoder to perform instance discrimination in TFFR space rather than the TF plane. On a synthetic benchmark with up to six chirp components under broadband noise (0–25 dB SNR), TFFR-Net consistently reduces IF MAE by at least 5.8% and Dice loss by at least 5.4% relative to competing methods. Evaluation on real-world signals (bat echolocation, power-system oscillation, earthquake vibration) demonstrates effective generalization beyond synthetic training data.

1 Introduction

Time-frequency (TF) analysis transforms a one-dimensional signal into a two-dimensional energy distribution that reveals how the signal spectra evolves over time. Multi-component signals arise in structural health monitoring [1], radar micro-Doppler analysis [2], and biomedical signal processing [3]. Each component traces a *ridge* in the TF plane whose trajectory equals the instantaneous frequency (IF) [4]. Extracting individual ridges from their superposition, known as multi-component TF mode decomposition, is essential for source separation, mode analysis, and parameter estimation.

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Classical ridge extraction relies on peak picking or reassignment [5], synchrosqueezing [6, 7], or variational mode decomposition [8]. These methods excel when ridges are well separated but degrade at crossings, low signal-to-noise ratios (SNR), or when the number of components is unknown. Recent deep-learning approaches learn a direct mapping from raw signals to component masks [9, 10], yet they treat the TF plane as a flat image and cannot distinguish ridges that share the same frequency at a given time.

We argue that the root cause is a *representation bottleneck*: the 2-D (t, f) plane has insufficient dimensionality to separate crossing components, because two ridges can share the same TF pixel while belonging to distinct physical modes. This information loss is inherent to the 2-D representation and difficult to be resolved by improving the downstream algorithm alone. We resolve this bottleneck by *lifting* the representation to the three-dimensional time-frequency-frequency-rate (TFFR) space $(t, f, \dot{f})$, where $\dot{f}(t) \triangleq d\mathrm{IF}/dt$ is the instantaneous frequency rate (IFR). At a crossing in (t, f), two components typically have different IFRs and therefore remain separated in $(t, f, \dot{f})$. We formalize this observation in Section 3.2: a simultaneous IF-and-IFR crossing requires two independent transcendental equations to hold at the same instant, which constrains the admissible parameters to a vanishingly small (measure-zero) set. Monte Carlo simulation confirms that the unresolved-crossing rate drops from 69.00% in 2-D to 6.12% in 3-D.

We cast TF mode decomposition as instance segmentation and adopt a Mask2Former-style [11] query-based architecture. Each learnable query produces a binary TF mask, an active/inactive score and a per-pixel IFR map. The IFR is integrated through a dedicated frequency-rate head, a per-instance IFR loss, and a feedback loop that projects each query's mean frequency rate back into its positional embedding, guiding the decoder to discriminate ridges along the additional IFR axis.

Our contributions are as follows:

1. We formalize why 2-D TF representations cannot separate crossing ridges and prove that adding the IFR as a third coordinate reduces the crossing probability to a theoretically negligible level (Proposition 1), confirmed empirically by a Monte Carlo unresolved-crossing-rate drop from 69.00% to 6.12%.
2. We design TFFR-Net, a query-based masked transformer with an IFR feedback loop that conditions each query on its estimated frequency rate, enabling the decoder to distinguish overlapping ridges via TFFR-space discrimination. A complex-valued multi-scale front-end that generalizes the short-time Fourier transform (STFT) provides learnable analysis bandwidth.
3. On a synthetic benchmark spanning 0–25 dB SNR, TFFR-Net outperforms five competing methods, including four classical approaches and a Mask2Former-based deep learning method. Quantitative evaluation on real-world signals further validates generalization.

2 Related work

Classical time-frequency ridge extraction The reassignment method [5] sharpens blurred TF distributions by moving energy toward the IF. Synchrosqueezing [6, 12] achieves a similar effect within a wavelet framework and admits an inverse transform for mode recovery. Ridge detection is typically performed by greedy peak tracking or penalized path optimization [13]. Variational nonlinear chirp mode decomposition (VNCMD) [14] formulates mode extraction as a variational optimization problem under a nonlinear chirp model. Ridge path regrouping with intrinsic chirp component decomposition (RPRG+ICCD) [15] extracts multiple ridge paths from the TF representation and refines each component by iterative chirp fitting. Adaptive chirp mode pursuit (ACMD) [16] iteratively demodulates each component under a local chirp assumption. Improved variational generalized nonlinear mode decomposition (IVGNMD) [17] extends the variational framework with a generalized nonlinear model for separating crossed and dispersive modes. These methods require careful parameter tuning and assume locally separated ridges.

Deep learning for TF analysis Izacard et al. [9] pioneer data-driven frequency estimation by training a neural network to predict a pseudo-spectrum from raw complex samples. TFA-Net [10] extends this with a complex-valued front-end but only outputs a global segmentation mask without distinguishing individual components. TFD-Net [18] adapts Mask2Former [11] with deformable

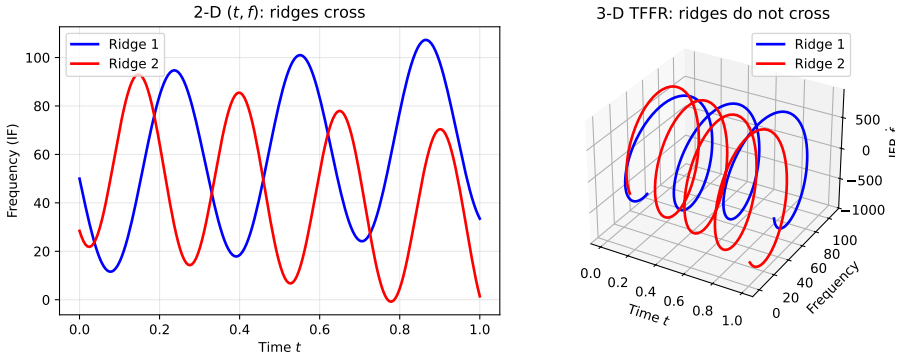


Figure 1: **Ridge crossings arise from insufficient representational dimensionality.** Two chirp ridges cross in the 2-D (t, f) plane (left), but after *lifting* to the 3-D TFFR space $(t, f, \dot{f})$ (right), the same ridges separate along the frequency-rate axis because their IFRs differ at the crossing instant.

attention [19] to the TF plane, formulating mode decomposition as instance segmentation. However, treating TF decomposition as purely 2-D segmentation ignores frequency-rate information and cannot well resolve crossing ambiguities. Our work addresses this by *lifting* the representation to a higher-dimensional space, a strategy that proves effective across machine learning, from kernel methods to Fourier feature mappings [20] and coordinate embeddings in neural radiance fields [21]. In our setting, the additional coordinate is the physically meaningful IFR, and we show analytically and empirically that the lifting eliminates the vast majority of crossing ambiguities (Section 3.2).

Instance segmentation with transformers DETR [22] introduces set prediction with learnable queries and bipartite matching. Deformable DETR [19] improves convergence with multi-scale deformable attention. Mask2Former [11] unifies segmentation tasks via masked cross-attention; DI-MaskDINO [23] further addresses detection-segmentation imbalance. The sparse-set formulation is also applied beyond images, e.g., to 3-D occupancy prediction [24]. Our decoder follows the Mask2Former design and adds per-query IFR prediction and IFR feedback.

3 Method

3.1 Preliminaries

We consider a multi-component analytic signal $s(t) = \sum_{k=1}^{K} A_k \exp(2\pi j\, \phi_k(t))$, where K is the number of components, $A_k > 0$ is the k-th amplitude, $j = \sqrt{-1}$, and each phase follows

$$\phi_k(t) = r_k \cos(f_{1,k}\, t + \theta_k) + f_{3,k}\, t + f_{2,k}\, t^2. \tag{1}$$

Here r_k is the frequency-modulation amplitude, $f_{1,k}$ is the modulation rate, θ_k is the initial phase, $f_{3,k}$ is the center frequency, and $f_{2,k}$ controls the linear frequency drift. The IF and IFR are obtained analytically as

$$\mathrm{IF}_k(t) = \phi_k'(t) = -r_k f_{1,k} \sin(f_{1,k}\, t + \theta_k) + f_{3,k} + 2f_{2,k}\, t, \tag{2}$$

$$\mathrm{IFR}_k(t) = \mathrm{IF}_k'(t) = -r_k f_{1,k}^2 \cos(f_{1,k}\, t + \theta_k) + 2f_{2,k}. \tag{3}$$

The continuous IF trajectory is rasterized onto a discrete TF grid of F frequency bins and T time steps using Bresenham's line algorithm [25], producing a binary mask $\mathbf{M}_k \in \{0, 1\}^{F \times T}$. Each ridge pixel (f, t) is annotated with the analytical frequency rate $\mathrm{IFR}_k(t)$. We formulate multi-component mode decomposition as instance segmentation: each ridge is a distinct instance, assigned a binary mask and an IFR map.

3.2 TFFR separability analysis

The core difficulty of TF mode decomposition is *mode separability*: two components *cross* in the 2-D TF plane when $\mathrm{IF}_k(t_0) = \mathrm{IF}_l(t_0)$ for some t_0. This is a single equation in one unknown, so real

solutions generically exist and the ridges share the same TF pixel. We resolve this by *lifting* to the 3-D TFFR space $(t, f, \dot{f})$, where a crossing additionally requires the IFRs to coincide, imposing a second independent constraint on t_0.

Proposition 1 (3-D separability for generic chirps)**.** *Let two chirp components share the same modulation rate ($f_{1,k} = f_{1,l} \triangleq f_1$) but differ in at least one of $\{r, \theta, f_2, f_3\}$. At a TF crossing time t_0 where $\mathrm{IF}_k(t_0) = \mathrm{IF}_l(t_0)$, their IFR residual is*

$$\Delta\dot{f}(t_0) \triangleq \mathrm{IFR}_k(t_0) - \mathrm{IFR}_l(t_0) = f_1^2\big[r_l\cos(f_1 t_0{+}\theta_l) - r_k\cos(f_1 t_0{+}\theta_k)\big] + 2(f_{2,k} - f_{2,l}). \quad (4)$$

Requiring $\Delta\dot{f}(t_0){=}0$ simultaneously forces t_0 to satisfy two independent transcendental equations. By the implicit function theorem, this imposes a codimension-1 constraint on the parameter space $(r_k, r_l, \theta_k, \theta_l, f_{2,k}, f_{2,l})$: the admissible set has Lebesgue measure zero, i.e., simultaneous IF-and-IFR crossings can occur for special parameter choices but have zero probability under any continuous parameter distribution. See Appendix A for the full proof.

Geometric interpretation and generality In the 2-D TF plane, each ridge traces a planar curve $\gamma_k = \{(t, \mathrm{IF}_k(t))\}$. Two such curves generically meet at isolated crossing points: this is a codimension-1 event in the parameter space, hence common. After lifting, each ridge becomes a space curve $\Gamma_k = \{(t, \mathrm{IF}_k(t), \mathrm{IFR}_k(t))\}$ in $\mathbb{R}^3$. Because two 1-D curves in a 3-D ambient space require a codimension-2 coincidence to intersect, their meeting is a far more restrictive event that occurs with zero probability under any continuous parameter distribution. Figure 1 illustrates the key intuition: at a TF crossing, the two ridges diverge along the frequency-rate axis.

Although the algebraic analysis above is stated for the chirp model of Eq. (1), the codimension argument is purely topological: for *any* twice-differentiable IF law, two space curves in $\mathbb{R}^3$ have codimension 2, so simultaneous IF-and-IFR crossings require a codimension-2 coincidence regardless of the signal model. The chirp model therefore serves as a concrete, analytically tractable instance of this broadly applicable geometric principle.

Empirical verification via Monte Carlo simulation To quantify this improvement beyond the analytic argument, we conduct a Monte Carlo study on a discrete grid. We generate $N_{\text{mc}}{=}5{,}000$ signals with $N{=}256$ samples under the synthetic protocol, drawing K uniformly from the integers 2 through 6 (Appendix B). With duration $T = N/f_s$, a pair is unresolved in 2-D when $|\Delta f| < \varepsilon_f = f_s/N$ at some time, and remains unresolved in 3-D when it also satisfies $|\Delta\dot{f}| < \varepsilon_r = \varepsilon_f/T = f_s^2/N^2$. In 2-D, **69.00%** of trials are unresolved, confirming that TF crossings are extremely common. After lifting to 3-D TFFR, the unresolved-crossing rate drops to **6.12%** (see Appendix C and Fig. 5). The residual 6.12% is consistent with the theoretical prediction: continuous-parameter theory guarantees zero probability of *exact* IF-and-IFR coincidence, but on a discrete grid two distinct ridges can still fall within the same frequency and rate bins when their IFR difference is small relative to the grid resolution.

Motivated by this separability analysis, we design an architecture that jointly predicts IF masks and IFR maps, enabling the decoder to exploit the lifted TFFR representation for ridge separation.

3.3 Architecture overview

Figure 2 illustrates the TFFR-Net pipeline: a learnable multi-scale front-end that generalizes STFT produces an adaptive TF representation; a U-Net backbone with temporal gradient fusion and a deformable pixel decoder produce mask features; a masked transformer decoder with IFR feedback outputs per-query masks and IFR maps.

3.4 Complex-valued multi-scale front-end

The raw complex signal $\mathbf{x} \in \mathbb{R}^{2\times T}$ (real and imaginary channels) is fed to three parallel complex convolutions [26] with kernel sizes $1{\times}31$, $1{\times}63$, and $1{\times}95$:

$$\mathbf{X}_m = \big|\mathrm{ComplexConv}_m(\mathbf{x})\big|, \quad m \in \{1, 2, 3\}. \quad (5)$$

The modulus converts outputs to real-valued TF-like maps; concatenation along the filter axis yields a $3n_f$-channel tensor ($n_f{=}8$), refined by a convolutional block attention module (CBAM) [27].

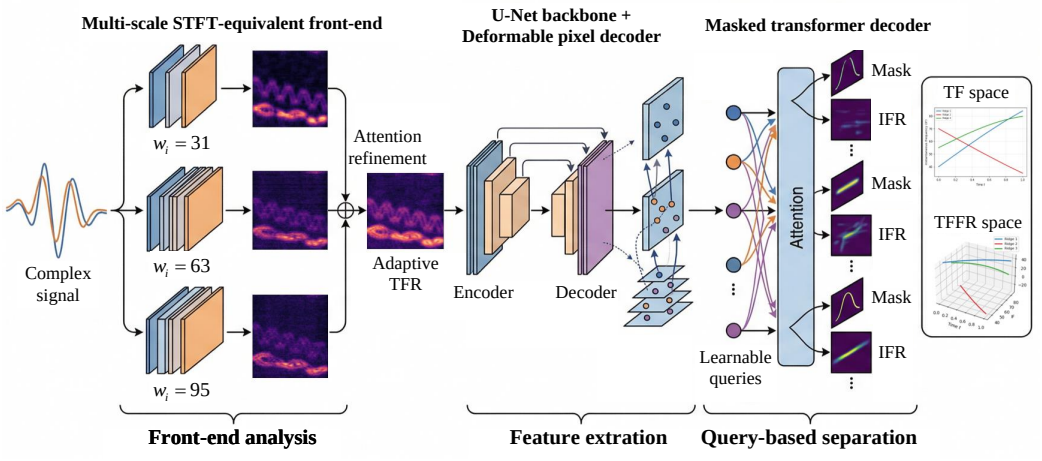


Figure 2: **TFFR-Net architecture.** The complex-valued multi-scale front-end (Sec. 3.4) generalizes multi-scale STFT. A U-Net backbone fuses temporal gradients. The deformable pixel decoder produces mask features for L=4 decoder layers. Each of Q=10 queries predicts a mask and an IFR map. An IFR feedback loop projects per-query mean IFR back into query embeddings.

Each branch generalizes a short-time Fourier transform (STFT) with analysis window length equal to the convolution kernel size (31, 63, or 95 samples, respectively). A complex convolution with stride 1 and n_f output filters computes the inner product between the signal and n_f complex templates at each time step. When the kernels are constrained to be orthogonal Fourier bases, the modulus output exactly recovers the STFT magnitude spectrogram; the classical STFT is therefore a special case of this formulation. In practice, we use standard Xavier uniform initialization and let the network learn kernels that go beyond fixed Fourier bases, adapting the effective window shape and bandwidth to the data in a fully data-driven manner. The three branches thus implement a learnable multi-scale front-end with different time-frequency trade-offs: narrower kernels (length 31) favor high temporal resolution for rapidly varying chirps, while wider kernels (length 95) favor high frequency resolution for slowly varying components. This enables TFFR-Net to handle signals whose components span a wide range of chirp rates without manual bandwidth selection.

3.5 Backbone and temporal gradient fusion

The attention-enhanced features are processed by a U-Net-style encoder-decoder backbone [28] with 15 convolutional layers and additive skip connections. The encoder progressively doubles the channel width ($24 \to 48 \to 96 \to 192$) while halving spatial resolution; the decoder mirrors this structure, producing multi-scale feature maps at four resolutions that capture both fine ridge detail and coarse spectral context.

Temporal gradient fusion The IFR is, by definition, the time derivative of the IF. To inject this temporal-variation information, we compute the finite-difference temporal gradient of the U-Net's full-resolution output $\mathbf{B} \in \mathbb{R}^{3n_f \times F \times T}$, where $3n_f$=24 is the front-end channel count: $\nabla_t \mathbf{B}_{:,:,t} = \mathbf{B}_{:,:,t+1} - \mathbf{B}_{:,:,t}$. We then concatenate $[\mathbf{B}; \nabla_t \mathbf{B}]$ channel-wise and fuse via a 1×1 convolution, replacing the highest-resolution feature in the multi-scale feature pyramid. This provides the pixel decoder with features that explicitly encode how spectral content changes across adjacent time steps, serving as a direct proxy for the frequency rate.

The four-scale feature maps are fused by a multi-scale deformable attention pixel decoder [19], which applies deformable cross-attention across scales to produce mask features $\mathbf{F}_{\text{mask}} \in \mathbb{R}^{C \times F \times T}$ (where C denotes the hidden dimension) and multi-scale memory tensors for the transformer decoder.

3.6 Masked transformer decoder with IFR feedback

Inspired by the query-based decoding paradigm of Mask2Former [11], we employ Q=10 learnable queries (sufficient for up to six components with margin), each represented by a content embedding

$\mathbf{q}_i^{\text{con}} \in \mathbb{R}^C$ and a positional embedding $\mathbf{q}_i^{\text{pos}} \in \mathbb{R}^C$. The decoder consists of L=4 layers. Each layer performs: (1) masked cross-attention, where each query attends only to TF locations within its predicted mask from the previous layer (restricting attention to the spatial region of interest); (2) self-attention among queries (enabling competition and coordination); and (3) a feed-forward network. At every layer, three prediction heads project the updated query embeddings:

Mask head A multi-layer perceptron (MLP) maps each query embedding to a mask embedding $\mathbf{e}_i^{\text{mask}} \in \mathbb{R}^C$, producing raw logits via dot product $\hat{\mathbf{M}}_i^{\text{raw}} = \mathbf{e}_i^{\text{mask}} \cdot \mathbf{F}_{\text{mask}}$. A lightweight convolutional refinement module with a residual connection yields the final mask: $\hat{\mathbf{M}}_i = \hat{\mathbf{M}}_i^{\text{raw}} + \mathcal{R}(\hat{\mathbf{M}}_i^{\text{raw}})$.

Frequency-rate head A separate MLP produces a frequency-rate embedding $\mathbf{e}_i^{\text{ifr}} \in \mathbb{R}^C$, and the per-pixel IFR prediction is $\hat{\mathbf{R}}_i = \mathbf{e}_i^{\text{ifr}} \cdot \mathbf{F}_{\text{mask}}$.

Classification head A linear layer predicts an active/inactive score $\hat{c}_i \in \mathbb{R}$ per query.

IFR feedback loop At each decoder layer, we compute a per-query IFR summary by mask-weighted averaging of the predicted IFR map, project it via a small MLP, and add it as a positional offset:

$$\bar{r}_i = \frac{\sum_{f,t} \sigma(\hat{M}_i)_{f,t} \cdot \hat{R}_{i,f,t}}{\sum_{f,t} \sigma(\hat{M}_i)_{f,t} + \epsilon}, \qquad \mathbf{q}_i^{\text{pos}} := \mathbf{q}_i^{(0)} + \text{MLP}(\bar{r}_i), \tag{6}$$

where $\mathbf{q}_i^{(0)}$ is the initial positional embedding, $\sigma(\cdot)$ denotes the sigmoid function, and $\epsilon = 10^{-6}$ prevents division by zero. This conditions the next cross-attention layer on the estimated frequency-rate range of each query, effectively lifting the query's search space from the 2-D TF plane to the 3-D TFFR space. As a result, two queries claiming ridges that overlap in (t, f) but have different IFRs receive distinct positional offsets, enabling the decoder to route each query toward its target ridge even at crossing regions. In Section D, we empirically verify that this mechanism leads to complementary IFR specialization across queries.

3.7 Training objectives

Hungarian matching Since ridges have no canonical ordering, we assign predicted queries to ground-truth instances via bipartite matching. We construct a $Q \times K$ cost matrix from Dice cost and binary cross-entropy cost between all query–ground-truth (GT) pairs. Because the continuous IF trajectory may fall between discrete frequency bins, energy can leak to adjacent grid cells; we therefore additionally evaluate ± 1-pixel frequency shifts and select the alignment minimizing cost for each pair. The optimal one-to-one assignment $\hat{\pi}$ is found via the Hungarian algorithm [29]; unmatched queries are assigned the "inactive" label.

Loss functions For matched pairs, the total loss aggregated over all L=4 decoder layers is

$$\mathcal{L} = \lambda_{\text{ins}} \mathcal{L}_{\text{ins}} + \lambda_{\text{ifr}} \mathcal{L}_{\text{ifr}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}, \tag{7}$$

where $\mathcal{L}_{\text{ins}}$ combines Dice and binary cross-entropy (BCE) losses on instance masks with weights α_{dice}=0.2 and α_{bce}=0.8, penalizing both overlap errors and per-pixel classification errors; $\mathcal{L}_{\text{ifr}}$ is a smooth-ℓ_1 regression loss computed only on GT ridge pixels, ensuring that the IFR prediction is supervised where ground-truth frequency-rate values are available; $\mathcal{L}_{\text{cls}}$ is binary cross-entropy with adaptive positive-class weighting for active/inactive query classification, balancing the variable number of active components ($K \in \{1, \ldots, 6\}$) against Q=10 total queries. We set λ_{ins}=2.0, λ_{ifr}=1.0, λ_{cls}=0.1.

4 Experiments

4.1 Setup

Comparison methods We compare against five methods spanning classical iterative decomposition and deep learning, including **VNCMD** [14], which formulates mode extraction as variational optimization under a nonlinear chirp model; **RPRG+ICCD** [15], which extracts ridge paths from

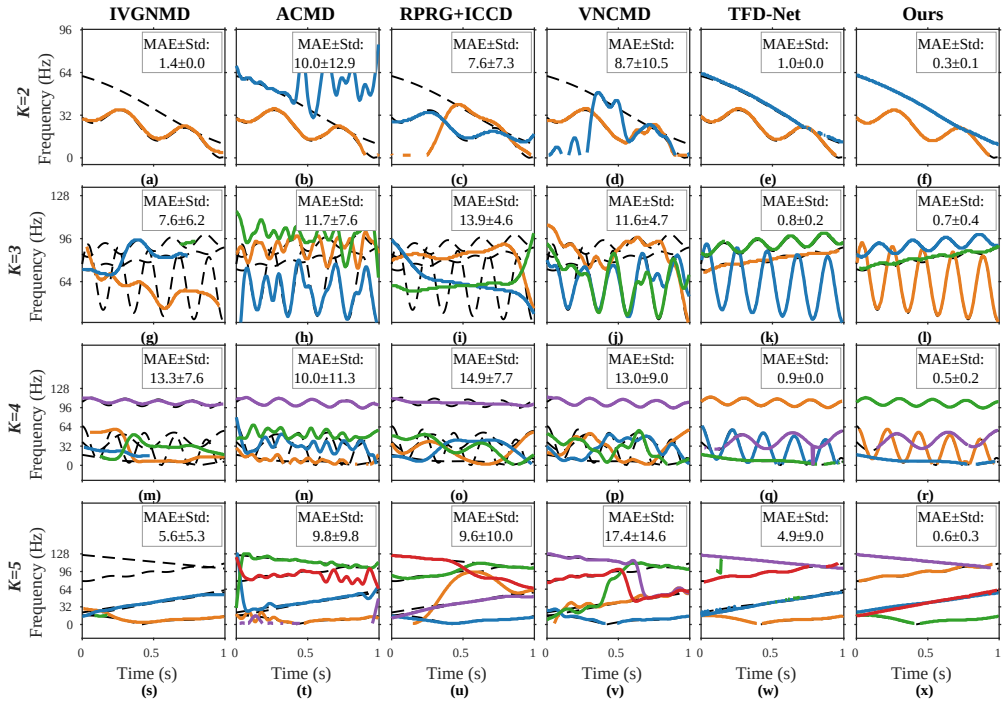


Figure 3: **Qualitative mode decomposition on synthetic signals (20 dB SNR).** Rows correspond to increasing component counts ($K{=}2, 3, 4, 5$); columns compare six methods (IVGNMD, ACMD, RPRG+ICCD, VNCMD, TFD-Net, TFFR-Net). Colored solid lines denote predicted IF ridges; black dashed lines indicate ground-truth trajectories. Each panel reports IF MAE$\pm$Std (Hz). TFFR-Net (rightmost column) yields the most accurate ridge estimates, particularly at crossing regions where classical methods exhibit ridge swapping or deviation and TFD-Net shows occasional mode switching.

the TF plane and refines each component via intrinsic chirp fitting; **ACMD** [16], which iteratively demodulates each component under a local chirp assumption; **IVGNMD** [17], which extends the variational framework with a generalized nonlinear model; and **TFD-Net** [18], a Mask2Former-style instance segmentation network operating on the 2-D TF plane without IFR prediction.

Metrics *Synthetic (supervised).* We evaluate at six fixed SNR levels (0, 5, 10, 15, 20, 25 dB) using three metrics after Hungarian matching: **Dice Loss** ($\downarrow$): $1 -$ Dice coefficient between predicted and GT instance masks; **IF MAE** ($\downarrow$): mean absolute error of IF on GT ridge pixels; **IF Std** ($\downarrow$): standard deviation of IF errors.

4.2 Qualitative results

Synthetic signals Figure 3 visualizes decomposition results on representative synthetic test signals at 20 dB SNR with varying component counts. The four classical methods (IVGNMD, ACMD, RPRG+ICCD, VNCMD) produce swapped or deviated ridges near crossing regions, where two IF trajectories share the same frequency at a given time instant. IVGNMD and VNCMD often introduce spurious oscillations that deviate from the true trajectories, while ACMD and RPRG+ICCD exhibit discontinuities when ridges approach each other. TFD-Net, operating purely in the 2-D TF plane, yields substantially better results than the classical approaches but still shows occasional mode switching at crossings, especially for higher component counts. TFFR-Net consistently resolves these crossing ambiguities: predicted ridges closely track the ground-truth trajectories (black dashed lines), yielding the lowest IF MAE$\pm$Std across all panels. This confirms the analysis in Section 3.2: the additional IFR dimension disambiguates ridges that overlap in (t, f) but differ in frequency rate $\dot{f}$.

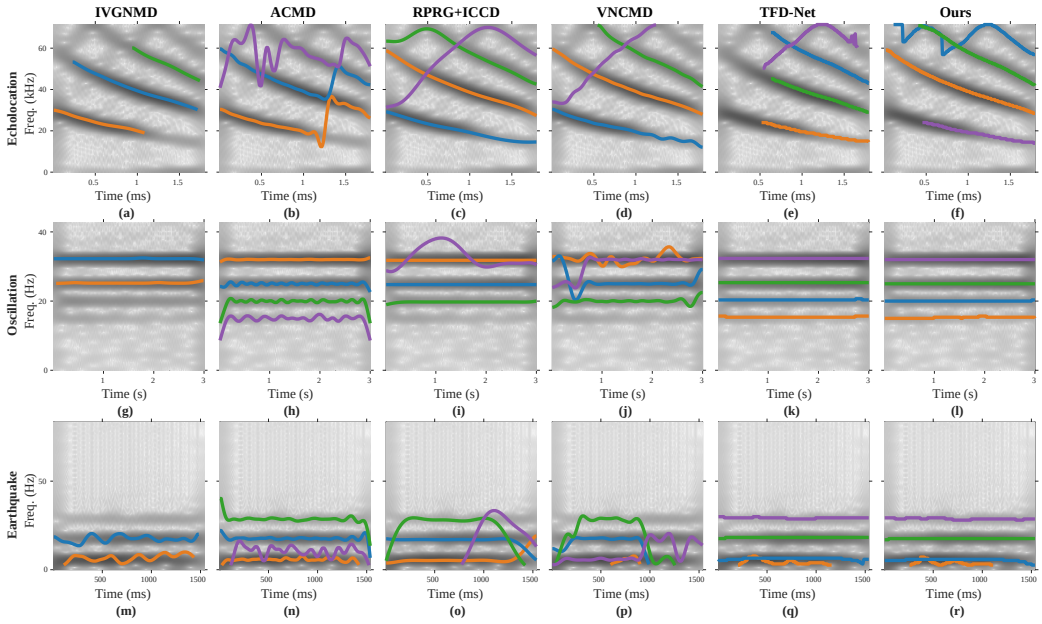


Figure 4: **Mode decomposition on real-world signals.** Rows: bat echolocation (143 kHz sampling rate, nonlinear frequency sweeps), subsynchronous oscillation (closely spaced spectral modes), and earthquake floor vibration (broadband transient, low SNR). Columns compare six methods. TFFR-Net produces smooth, well-separated ridges that closely follow spectral energy concentrations of STFT, while classical methods exhibit spurious oscillations or fail to resolve closely spaced modes.

Real-world signals To evaluate generalization beyond synthetic chirps, we test on three real-world signals: (i) a bat echolocation call [30] sampled at 143 kHz, containing multiple rapidly sweeping harmonics; (ii) a subsynchronous oscillation recording from a power system [31] with closely spaced spectral modes; and (iii) a floor vibration signal from a structural earthquake response [32] exhibiting broadband transient energy. Although no ground-truth IF annotations are available, we overlay the extracted ridges on the STFT spectrogram (grayscale background in Figure 4), which provides a visual reference for evaluating how well each method tracks the spectral energy concentrations.

IVGNMD and VNCMD produce highly oscillatory ridge estimates and fail to separate closely spaced modes in the oscillation and earthquake signals. ACMD and RPRG+ICCD yield smoother estimates but introduce discontinuities when ridges approach each other. TFD-Net resolves the dominant modes but occasionally merges weaker components in densely populated TF regions. TFFR-Net produces the smoothest and most complete ridge trajectories across all three signals, faithfully tracking spectral energy concentrations even in regions with closely spaced modes.

4.3 Quantitative comparison

Table 1 reports quantitative metrics across six SNR levels for all six methods.

Classical vs. learning-based methods The four classical methods (ACMD, VNCMD, IVGNMD, RPRG+ICCD) all exhibit high Dice loss (0.75–0.94) across all SNR levels. Among them, RPRG+ICCD achieves the lowest Dice at low SNR (0.901 at 0 dB vs. 0.942 for ACMD) owing to its explicit ridge-path regrouping step, but its Dice plateaus at high SNR (0.779 at 25 dB) while ACMD continues to improve (0.750). IF errors remain significantly larger for all classical methods: even at 25 dB, ACMD, RPRG+ICCD, and VNCMD reach IF MAE of at least 10, and IVGNMD exceeds 23. This gap reflects the fundamental limitation of iterative demodulation, since at crossings the local chirp assumption breaks down, causing ridge swapping or divergence. In contrast, the learning-based methods (TFD-Net and TFFR-Net) achieve substantially lower errors by directly predicting instance-level masks without sequential demodulation.

Table 1: Quantitative comparison across six SNR levels. Best in **bold**, second best underlined. All metrics ↓

		ACMD	VNCMD	IVGNMD	RPRG	TFD-Net	**Ours**
Dice↓	0 dB	0.942	0.917	0.929	0.901	0.648	**0.613**
	5 dB	0.908	0.859	0.887	0.840	0.510	**0.469**
	10 dB	0.862	0.812	0.786	0.803	0.404	**0.358**
	15 dB	0.812	0.778	0.758	0.785	0.329	**0.281**
	20 dB	0.775	0.761	0.753	0.781	0.276	**0.230**
	25 dB	0.750	0.750	0.748	0.779	0.248	**0.198**
IF MAE↓	0 dB	20.22	19.55	31.54	19.33	9.45	**8.87**
	5 dB	18.12	15.01	34.07	14.86	6.05	**5.70**
	10 dB	15.06	12.91	25.75	12.77	4.03	**3.62**
	15 dB	12.52	11.55	24.74	11.76	3.14	**2.56**
	20 dB	10.93	11.13	23.44	11.32	2.40	**2.02**
	25 dB	10.00	10.98	23.14	11.17	1.96	**1.66**
IF Std↓	0 dB	19.79	16.45	18.94	16.58	10.50	**9.88**
	5 dB	18.51	14.02	22.62	13.83	8.14	**7.68**
	10 dB	15.70	12.65	18.29	12.26	5.92	**5.18**
	15 dB	13.05	11.65	17.94	11.56	4.66	**3.57**
	20 dB	11.04	11.07	17.24	10.97	3.47	**2.70**
	25 dB	10.00	11.24	16.93	11.03	2.69	**1.97**

Table 2: Ablation study. Mask: no IFR; +Pred: add IFR pred.; +FB: add IFR feedback (full model); Bkb.I/S: backbone head. All metrics ↓

		Mask	+Pred	**+FB**	Bkb.I	Bkb.S
Dice↓	0 dB	0.840	0.619	**0.613**	0.615	0.842
	5 dB	0.840	0.478	**0.469**	0.474	0.843
	10 dB	0.862	0.368	**0.358**	0.364	0.861
	15 dB	0.886	0.291	**0.281**	0.288	0.883
	20 dB	0.906	0.237	**0.230**	0.234	0.899
	25 dB	0.917	0.205	**0.198**	0.201	0.910
IF MAE↓	0 dB	10.52	9.27	**8.87**	9.31	11.62
	5 dB	8.96	6.12	**5.70**	6.19	9.94
	10 dB	11.62	3.97	**3.62**	4.00	12.25
	15 dB	15.47	2.91	**2.56**	2.84	16.08
	20 dB	18.29	2.10	**2.02**	2.19	18.83
	25 dB	18.57	1.87	**1.66**	1.82	18.56
IF Std↓	0 dB	10.96	10.41	**9.88**	10.47	11.74
	5 dB	10.41	8.31	**7.68**	8.49	11.38
	10 dB	13.31	5.72	**5.18**	5.79	14.29
	15 dB	16.73	4.17	**3.57**	4.03	16.67
	20 dB	18.33	2.78	**2.70**	2.93	18.93
	25 dB	18.98	2.36	**1.97**	2.25	18.98

TFFR-Net vs. TFD-Net TFFR-Net consistently outperforms TFD-Net at all SNR levels. The improvement is most pronounced for IF Std, which measures ridge estimation consistency: at 25 dB, IF Std drops from 2.69 to 1.97 (-26.8%), indicating that TFFR-Net produces more stable ridge trajectories with fewer outlier errors. IF MAE improves from 1.96 to 1.66 (-15.3%), and Dice loss decreases from 0.248 to 0.198 (-20.2%). At lower SNR, TFFR-Net maintains its advantage despite the noisier input: at 0 dB, Dice drops from 0.648 to 0.613, demonstrating robustness. These gains arise from the IFR prediction and feedback mechanism, which provides the decoder with frequency-rate information that is inaccessible to the purely 2-D TFD-Net.

4.4 Ablation study

Table 2 reports five model variants. *Mask only* removes all IFR supervision ($\lambda_{\text{ifr}}=0$); *+IFR pred.* adds per-instance IFR prediction; *+IFR fb.* further adds the IFR feedback loop (our full model); *Bkb. IFR* moves the IFR head to the backbone; *Bkb. seg* adds a dense segmentation head to the backbone, supervised by the union of all per-component ridge masks (i.e., the ideal TF representation containing all ridges), following the design of TFD-Net.

Without IFR loss (Mask only), the decoder degenerates: Dice *increases* with SNR (0.84 at 0 dB to 0.92 at 25 dB), indicating mode collapse from insufficient per-instance contrast. Adding IFR prediction resolves this collapse. The IFR feedback loop yields further consistent improvements: at 25 dB, IF MAE drops from 1.87 to 1.66 (-11.2%) and IF Std from 2.36 to 1.97 (-16.5%). Backbone-level IFR (Bkb. IFR) performs competitively but is consistently worse than decoder-level feedback (e.g., IF MAE 2.19 vs. 2.02 at 20 dB), confirming that coupling IFR estimation with query-level instance discrimination is more effective than providing it as a global backbone feature. The backbone segmentation head (Bkb. seg) also degenerates (Dice > 0.84), indicating that the 2-D TF segmentation objective at the backbone level interferes with the per-instance query-based decoding; this further motivates our choice of placing IFR supervision at the decoder level rather than in the backbone. In summary, IFR supervision prevents instance collapse, and the IFR feedback loop provides additional gains by creating a closed loop between IFR estimation and query disambiguation, consistent with the TFFR separability analysis in Section 3.2.

5 Conclusion

This paper introduced *frequency-rate lifting* as a principled solution to the ridge-crossing problem in 2-D time-frequency mode decomposition. By augmenting the TF plane with the IFR coordinate, crossing ridges are separated into distinct space curves whose simultaneous intersection is confined to a lower-dimensional parameter manifold (Proposition 1; Monte Carlo unresolved-crossing rate: 69.00%→6.12%). TFFR-Net operationalizes this insight through an IFR feedback loop that steers

each query toward its target ridge in the 3-D TFFR space. Across six SNR levels (0–25 dB), TFFR-Net consistently reduces IF MAE by at least 5.8% and Dice loss by at least 5.4% relative to TFD-Net, and generalizes effectively to three real-world signal types.

Limitations and future work TFFR-Net currently extracts TF ridges without reconstructing the underlying time-domain waveforms. Integrating a learnable inverse transform (e.g., a differentiable Griffin–Lim iteration or an unrolled VNCMD layer) into the current framework enables joint ridge extraction and component reconstruction, while reconstruction-based supervision further promotes accurate ridge separation. Additionally, more extensive validation on real-world signals with closely spaced and crossing ridges, together with domain-specific model adaptation (e.g., fine-tuning on target signal distributions for mechanical fault diagnosis or seismic monitoring), remain important directions.

Acknowledgments and Disclosure of Funding

This work was supported by the Natural Science Foundation of Guangdong Province, China [grant number 2026A1515010467]; and the Research Task Assignment Project from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No. GML-26420009.

References

[1] Yu-Zu Li and Sheng-En Fang. A vibration signal decomposition method for time-varying structures using empirical multi-synchroextracting decomposition. *Mechanical Systems and Signal Processing*, 224:112107, 2025.

[2] Honglei Zhang, Wenpeng Zhang, Yongxiang Liu, Wei Yang, and Shaowei Yong. Scatterer-level time-frequency-frequency rate representation for micro-motion identification. *Remote Sensing*, 15(20):4917, 2023.

[3] Amaya Rose Abraham, Anurag Nishad, Mihir Jaipuria, Abhay Upadhyay, and Sudeep Baudha. Dynamic Q value wavelet transform for IF estimation and seizure detection via QT plane-CNN framework. *EURASIP Journal on Advances in Signal Processing*, 2025(1):23, 2025.

[4] Patrick Flandrin. *Time-frequency/time-scale analysis*, volume 10. Academic press, 1998.

[5] François Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on signal processing*, 43 (5):1068–1089, 2002.

[6] Ingrid Daubechies, Jianfeng Lu, and Hau-Tieng Wu. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Applied and computational harmonic analysis*, 30(2): 243–261, 2011.

[7] Gaurav Thakur and Hau-Tieng Wu. Synchrosqueezing-based recovery of instantaneous frequency from nonuniform samples. *SIAM Journal on Mathematical Analysis*, 43(5):2078–2095, 2011.

[8] Konstantin Dragomiretskiy and Dominique Zosso. Variational mode decomposition. *IEEE transactions on signal processing*, 62(3):531–544, 2013.

[9] Gautier Izacard, Sreyas Mohan, and Carlos Fernandez-Granda. Data-driven estimation of sinusoid frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Pingping Pan, Yunjian Zhang, Zhenmiao Deng, Shaocan Fan, and Xiaohong Huang. TFA-Net: A deep learning-based time-frequency analysis tool. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9274–9286, 2023.

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[12] Sylvain Meignen, Thomas Oberlin, and Stephen McLaughlin. A new algorithm for multicomponent signals analysis based on synchrosqueezing: With an application to signal sampling and denoising. *IEEE transactions on Signal Processing*, 60(11):5787–5798, 2012.

[13] Nils Laurent and Sylvain Meignen. A novel ridge detector for nonstationary multicomponent signals: Development and application to robust mode retrieval. *IEEE Transactions on Signal Processing*, 69:3325–3336, 2021.

[14] Shiqian Chen, Xingjian Dong, Zhike Peng, Wenming Zhang, and Guang Meng. Nonlinear chirp mode decomposition: A variational method. *IEEE Transactions on Signal Processing*, 65(22): 6024–6037, 2017.

[15] Shiqian Chen, Xingjian Dong, Guanpei Xing, Zhike Peng, Wenming Zhang, and Guang Meng. Separation of overlapped non-stationary signals by ridge path regrouping and intrinsic chirp component decomposition. *IEEE Sensors Journal*, 17(18):5994–6005, 2017.

[16] Shiqian Chen, Yang Yang, Zhike Peng, Xingjian Dong, Wenming Zhang, and Guang Meng. Adaptive chirp mode pursuit: Algorithm and applications. *Mechanical Systems and Signal Processing*, 116:566–584, 2019.

[17] Hongbing Wang, Shiqian Chen, and Wanming Zhai. Improved variational generalized nonlinear mode decomposition for separating crossed chirp modes and dispersive modes of non-stationary signals in mechanical systems. *Mechanical Systems and Signal Processing*, 227:112407, 2025.

[18] Pingping Pan, Yunjian Zhang, and Jinyi Liu. TFD-Net: Towards intelligent time-frequency mode decomposition with practical applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19329–19336, 2026.

[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[20] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[22] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[23] Zhixiong Nan, Xianghong Li, Tao Xiang, and Jifeng Dai. Di-MaskDINO: A joint object detection and instance segmentation model. *Advances in Neural Information Processing Systems*, 37:60359–60381, 2024.

[24] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Ming-Ming Cheng. Opus: occupancy prediction using a sparse set. *Advances in Neural Information Processing Systems*, 37:119861–119885, 2024.

[25] Jack E Bresenham. Algorithm for computer control of a digital plotter. In *Seminal graphics: pioneering efforts that shaped the field*, pages 1–6. 1998.

[26] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. *arXiv preprint arXiv:1705.09792*, 2017.

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[28] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29, 2016.

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[30] Prestor A Saillant, James A Simmons, Steven P Dear, and Teresa A McMullen. A computational model of echo processing and acoustic imaging in frequency-modulated echolocating bats: the spectrogram correlation and transformation receiver. *The Journal of the Acoustical Society of America*, 94(5):2691–2712, 1993.

[31] Lin Jiang, Q Henry Wu, and JY Wen. Decentralized nonlinear adaptive control for multimachine power systems via high-gain perturbation observer. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51(10):2052–2059, 2004.

[32] SJ Dyke, Billie F Spencer Jr, P Quast, and MK Sain. Role of control-structure interaction in protective system design. *Journal of engineering mechanics*, 121(2):322–338, 1995.

Appendix

A Detailed analysis of Proposition 1

When $f_{1,k} = f_{1,l} = f_1$, the IF crossing condition $\mathrm{IF}_k(t_0) = \mathrm{IF}_l(t_0)$ expands to

$$r_k f_1 \sin(f_1 t_0 + \theta_k) - r_l f_1 \sin(f_1 t_0 + \theta_l) = (f_{3,k} - f_{3,l}) + 2(f_{2,k} - f_{2,l})\, t_0. \tag{8}$$

This is one transcendental equation in t_0; for generic parameters it admits isolated real solutions (TF crossings exist).

Subtracting the IFR equations gives

$$\Delta\dot{f}(t_0) = f_1^2 \big[r_l \cos(f_1 t_0 + \theta_l) - r_k \cos(f_1 t_0 + \theta_k)\big] + 2(f_{2,k} - f_{2,l}). \tag{9}$$

Setting $\Delta\dot{f}(t_0) = 0$ produces a second transcendental equation in t_0:

$$r_k f_1^2 \cos(f_1 t_0 + \theta_k) - r_l f_1^2 \cos(f_1 t_0 + \theta_l) = 2(f_{2,k} - f_{2,l}). \tag{10}$$

Equations (8) and (10) form a system of two independent transcendental equations in the single unknown t_0. Writing $\alpha = f_1 t_0$ and defining $\mathbf{u} = (r_k \sin(\alpha + \theta_k) - r_l \sin(\alpha + \theta_l),\ r_k \cos(\alpha + \theta_k) - r_l \cos(\alpha + \theta_l))$, the system requires this two-dimensional vector to equal a specific constant vector determined by f_2, f_3. As α varies, $\mathbf{u}(\alpha)$ traces a closed curve in $\mathbb{R}^2$; a simultaneous solution exists only when the target point lies on this curve, which is a codimension-1 condition on the four-dimensional parameter space $(r_k, r_l, \theta_k, \theta_l)$ (or equivalently on (f_2, f_3) for fixed modulation parameters). Because this curve is one-dimensional while the target point lives in $\mathbb{R}^2$, the set of parameters that place the target exactly on the curve has Lebesgue measure zero. Simultaneous IF-IFR crossings are therefore theoretically possible but confined to an infinitely thin slice of the parameter space.

B Implementation details

Network configuration n_f=8 filters per complex convolution branch (24 total), F=T=256 TF resolution, C=256 hidden dimension, Q=10 queries, L=4 decoder layers, 2 attention heads. Prediction heads use 3-layer MLPs ($256 \to 256 \to 256$). The IFR feedback MLP has architecture $1 \to 64 \to 256$. The MaskRefineHead uses three 3×5 convolutions with 4 channels and a residual skip.

Signal generation Parameters are drawn as follows: $f_{1,k} \sim \mathcal{U}(-30, 30)$, $f_{2,k} \sim \mathcal{U}(-50, 50)$, $r_k \sim \mathcal{U}(-0.5, 0.5)$, $f_{3,k} \sim \mathcal{U}(-128, 128)$, $\theta_k \sim \mathcal{U}(0, 2\pi)$. Amplitudes follow $A_k = 5|\mathcal{N}(0, 1)| + 0.1$. Each component has a random onset position and a minimum duration of 256 samples. The number of components $K \sim \mathcal{U}\{1, 6\}$.

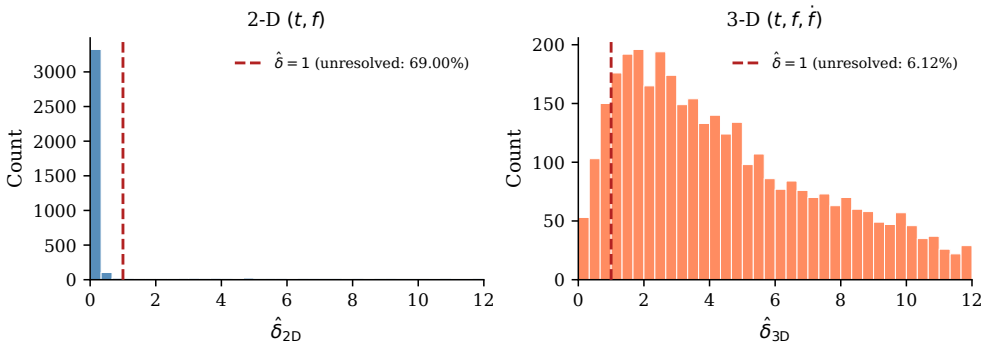


Figure 5: Distribution of the worst-case pairwise resolution distance across 5,000 Monte Carlo trials. The red dashed line marks the unresolved-crossing threshold ($\hat{\delta} = 1$). In the 2-D TF plane (left), 69.00% of trials fall below the threshold. In the 3-D TFFR space (right), the mass shifts above the threshold and the unresolved rate drops to 6.12%.

Training RMSprop optimizer with base LR 10^{-4} ($2\times$ for decoder heads), batch size 8, 5-epoch linear warmup, and ReduceLROnPlateau scheduler (factor 0.5, patience 7). Gradient norms are clipped at 1.0. Mixed-precision training is used. Early stopping patience is 15 epochs.

C Quantitative separability via Monte Carlo simulation

To quantify the improvement in the general multi-component case, we conduct a Monte Carlo study on a discrete grid. We use the synthetic-data protocol of Appendix B with $N = 256$ samples and sampling rate f_s, and draw the number of components K uniformly from the integers 2 through 6. The signal duration is

$$T = \frac{N}{f_s}. \tag{11}$$

One frequency bin gives the frequency threshold $\varepsilon_f = f_s/N$. Because IFR measures frequency change per unit time, its matching threshold is one frequency bin divided by the signal duration, $\varepsilon_r = \varepsilon_f/T$. Thus

$$\varepsilon_f = \frac{f_s}{N}, \qquad \varepsilon_r = \frac{\varepsilon_f}{T} = \frac{f_s^2}{N^2}. \tag{12}$$

An unresolved crossing in 2-D occurs when some component pair satisfies $|\Delta f| < \varepsilon_f$ at some time. The same pair remains unresolved in 3-D only if it also satisfies $|\Delta \dot{f}| < \varepsilon_r$:

$$\text{2-D:} \quad |\Delta f| < \varepsilon_f, \tag{13}$$

$$\text{3-D:} \quad |\Delta f| < \varepsilon_f, \qquad |\Delta \dot{f}| < \varepsilon_r. \tag{14}$$

For $N = f_s = 256$, these thresholds are 1 Hz and 1 Hz/s. Equivalently, a trial is unresolved when the worst-case pairwise resolution distance falls below one: $\hat{\delta}_{\text{2D}} = \min_{i<j} \min_t |\Delta f|/\varepsilon_f$ in 2-D and $\hat{\delta}_{\text{3D}} = \min_{i<j} \min_t \max\big(|\Delta f|/\varepsilon_f, |\Delta \dot{f}|/\varepsilon_r\big)$ in 3-D. Over $N_{\text{mc}} = 5{,}000$ independent trials, the unresolved-crossing rate decreases from 69.00% (3449/5000) in 2-D to 6.12% (306/5000) in 3-D (Table 3). Figure 5 shows the corresponding distributions. The residual 6.12% is consistent with the theoretical prediction: continuous-parameter theory guarantees zero probability of *exact* IF-and-IFR coincidence, but on a discrete grid two distinct ridges can still fall within the same frequency and rate bins when their IFR difference is small relative to the grid resolution.

D Inter-query specialization analysis

A natural question is whether the learnable queries actually exploit the IFR dimension for component separation. To investigate, we record the mean IF $\bar{f}_k$ and mean IFR $\bar{r}_k$ of each matched GT component across 500 synthetic test signals (1,750 matched query-component pairs), grouped by query index.

Table 3: Finite-resolution Monte Carlo analysis of ground-truth IF/IFR geometry ($N_{\text{mc}} = 5{,}000$; $K \sim \text{Unif}\{2, \ldots, 6\}$). An unresolved crossing requires $|\Delta f| < \varepsilon_f$ in 2-D, and also $|\Delta \dot{f}| < \varepsilon_r$ in 3-D.

Space	**Unresolved-crossing rate** $\downarrow$
2-D (t, f)	69.00% (3449/5000)
3-D $(t, f, \dot{f})$ (ours)	**6.12% (306/5000)**

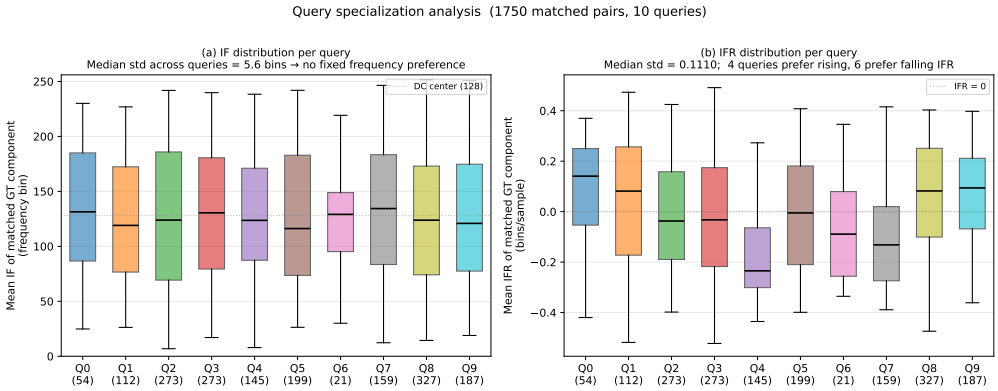


Figure 6: Query specialization in IF vs. IFR. (a) Per-query distribution of matched GT component mean IF: queries show no fixed frequency preference (overlapping boxes, median std $= 5.57$ bins). (b) Per-query distribution of matched GT component mean IFR: queries develop complementary positive/negative IFR preferences (4 positive, 6 negative median), confirming that the IFR feedback loop enables queries to specialize along the frequency-rate axis.

Figure 6 shows the per-query distributions. In the IF dimension (panel a), all 10 queries exhibit broadly overlapping distributions with similar medians (std of medians $= 5.57$ frequency bins), confirming that queries do not develop fixed frequency-band preferences. This is expected: the Hungarian matching assigns queries to components on a per-sample basis, and the IF range of randomly generated chirps is uniformly distributed.

In contrast, the IFR dimension (panel b) reveals a clear directional specialization: queries Q0, Q1, Q8, Q9 consistently match components with positive (rising) IFR, while Q2, Q3, Q4, Q5, Q6, Q7 prefer negative (falling) IFR (std of medians $= 0.11$ bins/sample). This asymmetry demonstrates that the IFR feedback loop (Eq. (6)) enables queries to develop complementary preferences along the frequency-rate axis, directly supporting our central claim that extending from 2-D TF to 3-D TFFR provides an additional axis of discrimination that queries learn to exploit.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state four contributions; Sections 3.1–4 support them with theory and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion identifies the need for more real-world data validation and the development of a learnable reconstruction module as current limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proposition 1 in Section 3.1 states assumptions; the full proof is in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full architecture details, hyperparameters, and data generation procedures are provided in the paper and Appendix B. The source code and datasets for this work are available at: `https://anonymous.4open.science/r/TFFR-92D2/`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The source code and datasets for this work are available at: `https://anonymous.4open.science/r/TFFR-92D2/`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All training details, hyperparameters, optimizer settings, and evaluation metrics are specified in Section 4 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The mean estimation error and the mean squared estimation error of the method are both reported in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments were conducted on a single NVIDIA GPU (RTX 6000). Training takes approximately 12 hours for 400 epochs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. No human subjects or sensitive data are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents foundational signal processing research with no direct negative societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No external datasets or pre-trained models are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The source code and datasets for this work are available at: `https://anonymous.4open.science/r/TFFR-92D2/`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.