Uni-Cheb: A Transform-Compatible Learnable Chebyshev Filter for Multimodal Spectral Modulation

Shuoqiu Duan[1], Jiasen Gao[1], Xiaoliang Chen[1]*
[1]School of Computer and Software Engineering, Xihua University,
Chengdu 610039, P. R. China

Abstract

While spatiotemporal modeling has emerged as an effective approach in multimodal learning, it struggles to resolve the "spectral dilemma" where distinct tasks demand contradictory frequency views. For example, deception detection requires global macro-patterns, while intent recognition depends on local micro-transients. To address this, we propose **Uni-Cheb**, a unified spectral operator centered on the Learnable Chebyshev Filter (LCF). Designed as a transform-compatible, plug-and-play component, LCF maintains a consistent mathematical form to adaptively modulate frequency components across different spectral transforms. Motivated by Chebyshev approximation, LCF employs learnable response-coordinate warping to focus on task-relevant bands. Guided by task-specific signal priors, the LCF seamlessly integrates with the Discrete Fourier Transform (DFT) to capture global temporal patterns or the Discrete Wavelet Transform (DWT) to capture localized semantic shifts, all while maintaining a negligible computational footprint. Furthermore, it serves as an effective spectral preconditioner to mitigate distribution shifts in cross-domain transfer learning. Extensive experiments across multiple benchmarks demonstrate that Uni-Cheb consistently improves state-of-the-art baselines. By rendering frequency modulation both task-adaptive and highly efficient, Uni-Cheb provides an effective, transform-compatible spectral modulation operator for multimodal learning.

1 Introduction

Spectral representation learning offers a theoretically grounded paradigm for decomposing complex signals into interpretable frequency components. This approach reveals latent structures that are often entangled within the spatio-temporal domain. In the context of multimodal learning, such a perspective is particularly vital because signals from different modalities inherently exhibit vastly distinct intrinsic characteristics.

Despite the theoretical advantages of spectral analysis, dominant modeling paradigms in multimodal learning [Liu et al., 2023, Zhu et al., 2024] remain entrenched in the spatio-temporal domain, largely overlooking the discriminative power of the frequency domain. Consequently, these methods often struggle with fine-grained tasks where critical signals are obscured by noise or entangled with dominant modalities. For instance, detecting deceptive micro-expressions [Ji et al., 2025] or recognizing subtle user intents [Hu et al., 2025] requires isolating specific frequency bands that are often lost during standard spatio-temporal aggregation [Yang et al., 2025].

Recognizing this potential, the spectral perspective has recently gained traction. Wang et al. [2025] introduced frequency-domain forecasting to mitigate autocorrelation bias, while Pei et al. [2025] and Cui et al. [2025] utilized spectral filtering for adversarial purification and causal confounding

*Correspongding author: chenxl@mail.xhu.edu.cn.

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

mitigation, respectively. However, many existing spectral methods rely on fixed frequency partitions or predefined spectral responses. Specifically, approaches like FreDF [Wang et al., 2025] depend on the standard DFT, diffusion models in Pei et al. [2025] employ fixed high-low frequency splitting, and causal frameworks in Cui et al. [2025] rely on pre-defined scale bands. Even adaptive approaches, such as Yang et al. [2025], typically employ spectral masking rather than jointly learning the response shape and its evaluation coordinates. Spectral informativeness concerns whether a task contains useful frequency-domain structure, whereas the *Spectral Dilemma* concerns which spectral representation is preferable when such structure is present. For example, global deception detection requires capturing holistic behavioral anomalies, whereas local intent recognition relies on transient shifts. These differences motivate task-adaptive response learning within the selected spectral representation.

In this work, we propose Uni-Cheb, a spectral modulation framework centered on the Learnable Chebyshev Filter (LCF). The unification of Uni-Cheb lies at the modular level: the spectral transform is selected according to task-specific signal priors, while the same transform-compatible LCF formulation is applied afterward. Motivated by Chebyshev approximation theory, LCF employs a compact polynomial expansion to parameterize learnable spectral responses. Crucially, we introduce a learnable scaling parameter (α) to perform *response-coordinate warping*, allowing the filter to dynamically stretch or compress the coordinates at which its response is evaluated. This shared response-learning formulation can be applied to Fourier- or wavelet-domain representations without redesigning the filtering module.

To evaluate this shared formulation, we extend our evaluation across three distinct multimodal domains that impose diverse spectral constraints. First, guided by the prior that deception cues manifest globally, we apply the LCF with the Discrete Fourier Transform (DFT). Here, the LCF adaptively modulates global temporal patterns [Ji et al., 2025]. Second, for intent recognition where signals are transient and non-stationary [Hu et al., 2025], we couple the LCF with the Discrete Wavelet Transform (DWT), enabling modulation of localized time-frequency sub-bands. Finally, in the context of transfer learning, we posit that domain shifts often manifest as frequency simplicity bias [He et al., 2024]. We demonstrate that the LCF can serve as a spectral preconditioner, modulating source and target spectral representations for cross-domain alignment.

In summary, our main contributions are as follows:

- We propose the Learnable Chebyshev Filter (LCF), a lightweight, transform-compatible spectral module that combines learnable Chebyshev response coefficients with response-coordinate scaling. The same filter formulation can be reused across different spectral transforms with a small parameter overhead (e.g., merely $\sim$0.002M parameters).
- We introduce a modular spectral modulation design. Guided by task-specific signal priors, we integrate the same LCF formulation across three distinct settings: DFT for deception detection, DWT for intent recognition, and spectral preconditioning for transfer learning, demonstrating its versatile applicability.
- We conduct extensive experiments across multiple benchmarks, demonstrating consistent improvements over strong task-specific baselines. Our analyses further examine the contributions of learnable scaling and task-dependent spectral transform selection.

2 Related Work

2.1 Multimodal Representation Learning

Recent advances in multimodal learning have predominantly focused on scaling spatial-temporal representations, leveraging Large Language Models (LLMs) to unify heterogeneous signals. For instance, Zhu et al. [2024] extended contrastive learning to bind distinct modalities into a shared semantic space, while Lian et al. [2025] established benchmarks for explaining emotion nuances using Multimodal LLMs. To improve performance in fine-grained tasks, various feature alignment and modulation techniques have been explored: Ji et al. [2025] employed hierarchical optimal transport to transfer affective knowledge for deception detection, and Hu et al. [2025] introduced adaptive re-calibration to modulate modalities for intention recognition. However, Chaudhuri et al. [2025] theoretically identified that current fusion mechanisms remain prone to geometric representation collapse. Crucially, these paradigms remain heavily entrenched in the spatial-temporal domain, often lacking explicit mechanisms to resolve spectral structures—such as separating micro-tremors from

sensor noise in frequency space. Consequently, they may remain susceptible to high-frequency artifacts, which can limit their effectiveness under domain distribution shifts.

2.2 Deep Learning in the Frequency Domain

Recognizing the limitations of spatial-temporal modeling, recent studies have increasingly integrated spectral transformations to capture global dependencies. In sequence modeling, Wang et al. [2025] introduced FreDF for frequency-domain forecasting, while Fons et al. [2025] proposed LSCD conditioned on Lomb-Scargle periodograms to handle irregular sampling. To enhance generalization, Hua et al. [2025] developed Fourier Position Embeddings leveraging spectral properties. In multimodal reasoning, Yang et al. [2025] employed spectral band masking for adaptive recommendation, whereas Cui et al. [2025] utilized multi-scale frequency memory to mitigate confounders. Regarding robustness and efficiency, Pei et al. [2025] utilized low-frequency preservation against perturbations, while Xiao et al. [2024] and He et al. [2024] explored efficient spectral parameterization.

However, many existing paradigms rely on fixed spectral partitions and predefined filtering responses (e.g., high-low frequency splits), treating frequency information as a static prior. Even adaptive approaches [Yang et al., 2025] often utilize spectral masking rather than jointly learning the response shape and its evaluation coordinates. This lack of flexibility limits task-adaptive modulation across tasks with different spectral preferences. We address these limitations with the Learnable Chebyshev Filter (LCF), a compact spectral response module with learnable response-coordinate warping. As a transform-compatible, plug-and-play module, LCF can be coupled with different transforms (e.g., DFT or DWT) based on task-specific signal priors, without requiring transform-specific architectural redesign.

3 Methodology

In this section, we present Uni-Cheb, a learnable spectral modulation operator designed to handle the spectral heterogeneity of multimodal signals. A key advantage of our proposed **Learnable Chebyshev Filter (LCF)** module is its **transform-compatible formulation**. While the underlying transform (e.g., DFT or DWT) is selected according to task-specific signal characteristics, LCF applies the same response-learning formulation across the resulting spectral representations.

We first formulate the generalized spectral representation problem. Then, we derive the LCF, detailing its Chebyshev response parameterization and learnable response-coordinate warping. Finally, we demonstrate the broad applicability of the operator by instantiating it across three distinct spectral scenarios: global basis modulation (via DFT) for deception detection, local basis modulation (via DWT) for intent recognition, and spectral preconditioning for cross-domain transfer learning.

3.1 Problem Formulation and Spectral Transformation

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$ denote a multimodal dataset, where $X_i \in \mathbb{R}^{L \times C}$ represents an input representation with signal support L (e.g., temporal length T or spatial resolution $H \times W$) and channel dimension C. We define a modular spectral filtering pipeline comprising a forward transform $\mathcal{T}$, a modulation function $\mathcal{H}$, and an inverse transform $\mathcal{T}^{-1}$.

Formally, the filtering process is defined as:

$$\hat{X} = \mathcal{T}^{-1}(\mathcal{H}(\mathcal{T}(X))) \tag{1}$$

where $\mathcal{T}$ maps the input features to a generalized spectral domain $S \in \mathbb{F}^{K \times C}$. Here, $\mathbb{F}$ represents the value field (complex $\mathbb{C}$ for Fourier-based instantiations or real $\mathbb{R}$ for Wavelet-based ones), and K denotes the number of spectral coefficients. $\mathcal{H}$ applies the Learnable Chebyshev Filter (LCF), and $\mathcal{T}^{-1}$ reconstructs the refined feature $\hat{X}$ back to the original spatiotemporal domain for downstream processing. LCF is applied to each modality independently before the original downstream fusion, classification, or alignment module.

3.2 The Learnable Chebyshev Filter (LCF)

The core challenge is to design a filter that adaptively emphasizes task-relevant spectral components without prior knowledge of the noise distribution. We propose LCF, which combines a structured Chebyshev response parameterization with learnable response-coordinate warping.

Learnable Response-Coordinate Construction. Crucially, relying on unconstrained learnable weights can lead to excessive spectral flexibility and overfitting. To address this, we construct the response evaluation coordinates using the roots of Chebyshev polynomials. The resulting angular parameterization provides a structured basis for learning compact spectral responses (see Appendix Section E for the approximation-theoretic background).

Specifically, we define the canonical angular coordinates $\phi \in \mathbb{R}^K$ via $\phi_k = \frac{2k-1}{2K}\pi$ for $k = 1, \ldots, K$, where K denotes the number of valid frequency bins (e.g., $K = \lfloor T/2 \rfloor + 1$ for DFT). To enable task-adaptive frequency modulation, a learnable scaling parameter α is introduced to obtain the response evaluation coordinates θ:

$$\theta_k = \mathrm{Softplus}(\alpha) \cdot \phi_k = \mathrm{Softplus}(\alpha) \cdot \frac{2k-1}{2K}\pi \tag{2}$$

where $\mathrm{Softplus}(\alpha)$ is initialized to 1, giving an initial scaling factor of 1.0. The parameter α adjusts the coordinates at which the Chebyshev response is evaluated, enabling task-adaptive spectral modulation without resampling or rearranging the underlying transform coefficients.

Filter Synthesis via Truncated Series. The LCF operator is parameterized by a learnable filter bank $\mathbf{B} \in \mathbb{C}^{M \times C}$, where M represents the order of the Chebyshev expansion. The frequency-dependent filter response $\mathbf{W} \in \mathbb{C}^{K \times C}$ is synthesized by evaluating the truncated Chebyshev series over the adaptive grid θ:

$$\mathbf{W}_{k,c} = \sum_{m=0}^{M-1} \mathbf{B}_{m,c} \cos(m \cdot \theta_k) \tag{3}$$

This formulation provides a compact parameterization of the spectral response. By optimizing the coefficients $\mathbf{B}$ alongside the scaling parameter α, the model jointly learns the response shape and its evaluation coordinates. Functionally, the modulation $\mathcal{H}(\mathbf{S})$ is realized as an element-wise product $\hat{\mathbf{S}} = \mathbf{S} \odot \mathbf{W}$. For Fourier-based instantiations, complex-valued response weights can adjust both the magnitude and phase of spectral coefficients, while wavelet-based instantiations use real-valued response weights.

3.3 Task-Specific Instantiations of the LCF Operator

We demonstrate the versatility of LCF by instantiating the transform pair $(\mathcal{T}, \mathcal{T}^{-1})$ for distinct signal characteristics. Detailed implementation procedures for these instantiations are provided in Appendix Section A.

Instantiation I: Global Modulation for Deception Detection. For video-based deception detection, discriminative cues (e.g., micro-expressions) manifest as quasi-periodic global rhythms. We employ the Discrete Fourier Transform (DFT).

Forward Transform: Given an input feature sequence $\mathbf{X} \in \mathbb{R}^{T \times C}$ (where T is the temporal length and C is the feature dimension), we transform it into the frequency domain via the real-to-complex FFT, preserving both amplitude and phase information:

$$\mathbf{S}_{\text{DFT}} = \mathrm{FFT}(\mathbf{X}) \in \mathbb{C}^{K \times C} \tag{4}$$

where $K = \lfloor T/2 \rfloor + 1$ represents the number of valid frequency bins.

Chebyshev Modulation: The complex-valued spectrum is directly modulated by our frequency-dependent filter response matrix $\mathbf{W}$ via element-wise complex multiplication:

$$\hat{\mathbf{S}}_{\text{DFT}} = \mathbf{S}_{\text{DFT}} \odot \mathbf{W} \in \mathbb{C}^{K \times C} \tag{5}$$

This complex multiplication adjusts the magnitude and phase of the Fourier coefficients through the learned filter response.

Inverse Transform: We mathematically reconstruct the refined spatiotemporal sequence by applying the Inverse FFT to the modulated complex spectrum:

$$\hat{\mathbf{X}} = \mathrm{IFFT}(\hat{\mathbf{S}}_{\mathrm{DFT}}) \in \mathbb{R}^{T \times C} \tag{6}$$

Instantiation II: Local Modulation for Intent Recognition. For intent recognition, discriminative cues are embedded in fine-grained spatial transients. We employ the 2D Discrete Wavelet Transform (DWT).

Forward Transform: Given a spatial feature map $\mathbf{X}_{spatial} \in \mathbb{R}^{H \times W \times C}$ (reshaped from the sequence $\mathbf{X}$, where $T = H \times W$), we utilize the Haar wavelet to decompose it into four sub-bands, obtaining the real-valued wavelet coefficients:

$$\mathbf{C}_{sub} = \mathrm{DWT}_{sub}(\mathbf{X}_{spatial}) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C} \tag{7}$$

where $sub \in \{LL, LH, HL, HH\}$ indexes the approximation and detail components.

Chebyshev Modulation: For each sub-band, LCF learns independent Chebyshev response coefficients and a scaling parameter α_{sub} to generate a sub-band-specific real-valued weight matrix $\mathbf{W}_{sub}$. We apply modulation directly to the coefficients of each component:

$$\hat{\mathbf{C}}_{sub} = \mathbf{C}_{sub} \odot \mathbf{W}_{sub} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C} \tag{8}$$

Here, $\mathbf{W}_{sub}$ performs spectral reweighting. For instance, the model can learn a larger magnitude for $\mathbf{W}_{HH}$ to highlight latent feature components sensitive to high-frequency diagonal edges associated with subtle intent gestures.

Inverse Transform: We reconstruct the refined spatial features using the Inverse 2D DWT (IDWT). This layer combines the modulated sub-bands to recover the original spatial structure before the backbone's original fusion module:

$$\hat{\mathbf{X}}_{spatial} = \mathrm{IDWT}(\hat{\mathbf{C}}_{LL}, \hat{\mathbf{C}}_{LH}, \hat{\mathbf{C}}_{HL}, \hat{\mathbf{C}}_{HH}) \in \mathbb{R}^{H \times W \times C} \tag{9}$$

Instantiation III: Spectral Alignment for Transfer Learning. Beyond resolving the intra-domain spectral dilemma, we extend Uni-Cheb to address *spectral distribution shifts* in cross-domain scenarios, such as transferring affective knowledge from a source domain (e.g., the **DFEW** dataset) to the target task of deception detection. While Ji et al. [2025] employs H-OTKT for spatial alignment, direct application on raw features faces *noise entanglement*, where the optimal transport (OT) objective may inadvertently force the alignment of task-irrelevant background artifacts or sensor noise.

To address this, we instantiate LCF as a *Spectral Preconditioner*. Instead of aligning noisy spatial features, we reformulate the transfer objective to operate explicitly in the *refined spectral domain*. Let $\mathbf{Z}_{\mathcal{S}}$ and $\mathbf{Z}_{\mathcal{T}}$ denote the latent feature representations from the source and target domains, respectively. We define the LCF-modulated spectral operator as $\Phi_{\mathrm{LCF}}(\mathbf{Z}) = \mathrm{FFT}(\mathbf{Z}) \odot \mathbf{W}$, where $\mathbf{W} \in \mathbb{C}^{K \times C}$ is the effective frequency-dependent filter response matrix synthesized via the Chebyshev basis expansion.

To satisfy the real-valued metric space requirements of standard optimal transport algorithms, the spectral-guided alignment loss is defined symmetrically on the spectral magnitudes:

$$\mathcal{L}_{align} = \mathcal{D}_{\text{H-OTKT}}\left(|\Phi_{\mathrm{LCF}}(\mathbf{Z}_{\mathcal{S}})|, |\Phi_{\mathrm{LCF}}(\mathbf{Z}_{\mathcal{T}})|\right) \tag{10}$$

where $|\cdot|$ denotes the element-wise magnitude of the complex spectra, and $\mathcal{D}_{\text{H-OTKT}}$ represents the hierarchical Sinkhorn distance. The transport cost is computed on the magnitudes of the LCF-modulated spectra, encouraging alignment of the source and target spectral magnitude distributions. Minimizing the alignment loss jointly with the downstream task objective allows the learned response to adapt to cross-domain differences while preserving task-relevant information.

4 Experiments

4.1 Datasets

Deception Detection Datasets. To validate the effectiveness, generalizability, and robustness of Uni-Cheb, evaluation is performed on two benchmark datasets: (1) Real Life Trial [Pérez Rosas et al.,

2015], (2) DOLOS [Guo et al., 2023]. For detailed information about the dataset, please refer to Appendix Section C.

Intent Recognition Datasets. To evaluate the model's performance on multimodal intent recognition, we employ two large-scale benchmark datasets: (1) MIntRec [Zhang et al., 2022] and (2) MIntRec2.0 [Zhang et al., 2024]. For detailed statistics and class definitions, please refer to Appendix Section C.

Transfer Learning Dataset. To validate the effectiveness of affective knowledge transfer, we employ the large-scale *in-the-wild* dataset DFEW [Jiang et al., 2020] as the source domain. For specific data selection (e.g., single-labeled subset) and detailed statistics, please refer to Appendix Section C.

4.2 Experimental Details

All experiments were implemented using PyTorch 2.7.0 on Ubuntu 22.04 with Python 3.9. Deception Detection and Transfer Learning experiments were conducted on an NVIDIA RTX 4090 (24GB) GPU, while Intent Recognition utilized a V100 (32GB) GPU. To ensure strict reproducibility, all detailed training hyperparameters, including the optimizer, exact learning rates, batch sizes, and epochs, are comprehensively documented in Appendix Section J. Furthermore, a detailed system-level analysis of computational complexity, including precise parameter counts and inference latency profiling, is provided in Appendix Section E.5. All other baseline-specific architectural settings remain identical to their original configurations.

4.3 Baseline

To validate Uni-Cheb, we benchmark against task-specific SOTA methods.

Deception Detection and Transfer Learning. We adopt AFFAKT [Ji et al., 2025], which employs Hierarchical Optimal Transport (H-OTKT) for spatial knowledge transfer. This baseline allows us to evaluate the contribution of LCF-based spectral modulation to cross-domain alignment.

Intent Recognition. We utilize the Dynamic-attention Allocation Fusion (DAF) framework by MVCL-DAF [Hu et al., 2025]. While DAF achieves SOTA by optimizing spatial attention weights, comparing against it verifies that our spectral reweighting (Instantiation II) captures fine-grained transient semantics more effectively than purely spatial mechanisms.

4.4 Experimental Results

Table 1: Comparison results on RLT and DOLOS datasets. By integrating LCF, our method consistently outperforms the baseline (AFFAKT). **Bold**: Optimal; Underline: Sub-optimal.

Modality	Method	RLT			DOLOS		
		F1	ACC	AUC	F1	ACC	AUC
Visual	AFFAKT	87.60	86.70	87.89	71.02	67.64	72.12
	+ LCF	**93.75**	**93.75**	**96.35**	**76.00**	**73.00**	**77.77**
	Improv.	*+6.15*	*+7.05*	*+8.46*	*+4.98*	*+5.36*	*+5.65*
Audio	AFFAKT	73.16	74.40	73.96	69.82	61.98	63.91
	+ LCF	**83.33**	**81.60**	**84.20**	**71.10**	**63.64**	**65.45**
	Improv.	*+10.17*	*+7.20*	*+10.24*	*+1.28*	*+1.66*	*+1.54*
Fused	AFFAKT	84.12	84.27	85.63	71.49	68.10	72.89
	+ LCF	**91.67**	**92.00**	**91.43**	**74.07**	**72.16**	**76.30**
	Improv.	*+7.55*	*+7.73*	*+5.80*	*+2.58*	*+4.06*	*+3.41*

Table 2: Evaluation of Spectral-Guided H-OTKT. Comparison on RLT and DOLOS datasets. **Bold**: Optimal; Underline: Sub-optimal.

Modality	Method	RLT			DOLOS		
		F1	ACC	AUC	F1	ACC	AUC
Visual	H-OTKT	87.60	86.70	87.89	71.02	67.64	72.12
	+ LCF	**90.50**	**89.77**	**90.33**	**75.60**	**71.92**	**76.76**
	Improv.	*+2.90*	*+3.07*	*+2.44*	*+4.58*	*+4.28*	*+4.64*
Audio	H-OTKT	73.16	74.40	73.96	69.82	61.98	63.91
	+ LCF	**76.92**	**77.63**	**78.57**	**70.87**	**65.06**	**65.88**
	Improv.	*+3.76*	*+3.23*	*+4.61*	*+1.05*	*+3.08*	*+1.97*
Fused	H-OTKT	84.12	84.27	85.63	71.49	68.10	72.89
	+ LCF	**86.62**	**86.38**	**87.28**	**76.92**	**73.08**	**77.43**
	Improv.	*+2.50*	*+2.11*	*+1.65*	*+5.43*	*+4.98*	*+4.54*

Deception Detection: Global Spectral Modulation. We first evaluate Uni-Cheb's capacity to resolve global spectral signatures on the RLT and DOLOS benchmarks (Table 1). On the RLT visual modality, Uni-Cheb achieves a **93.75%** F1 score, outperforming the AFFAKT baseline by **6.15%**. Consistent gains are observed across Audio (**+10.17%**) and Fused (**+7.55%**) streams. The learned Chebyshev response and coordinate scaling allow LCF to adapt its spectral emphasis to the downstream task. This task-adaptive modulation provides a flexible means of learning global spectral patterns, with extended comparisons against alternative spectral methods detailed in Section H.

Intent Recognition: Modeling Transient Micro-Semantics. Intent recognition poses a distinct spectral challenge, as signals manifest as non-stationary, transient semantic abruptions (e.g., micro-

Table 3: Performance comparison on MIntRec and MIntRec 2.0 datasets. All metrics are in %. **Bold** indicates best results.

Method	MIntRec				MIntRec 2.0			
	ACC	WF1	WP	R	ACC	WF1	WP	R
MVCL-DAF	74.72	74.61	75.07	71.94	57.80	55.05	55.82	42.03
+ LCF (Ours)	**77.08**	**76.88**	**77.80**	**75.63**	**60.40**	**59.35**	**60.63**	**54.52**
Improv.	*+2.36*	*+2.27*	*+2.73*	*+3.69*	*+2.60*	*+4.30*	*+4.81*	*+12.49*

expressions). Uni-Cheb demonstrates superior adaptability to these transients (Table 3), yielding Accuracy gains of **+2.36%** on MIntRec and **+2.60%** on MIntRec 2.0. Notably, Recall on MIntRec 2.0 surges by **+12.49%**. The DWT-based instantiation applies independently learned Chebyshev responses to localized wavelet sub-bands, allowing LCF to modulate local feature patterns before downstream fusion. Analyses of α convergence and M sensitivity are provided in Appendix Sections F and G.

Transfer Learning: Spectral Modulation and Alignment. Finally, we evaluate Uni-Cheb for cross-domain affective transfer (DFEW to DOLOS). As detailed in Table 2, the standard H-OTKT baseline achieves 71.02% F1 on the DOLOS visual modality. By incorporating LCF as a spectral preconditioner, Spectral-Guided H-OTKT improves performance to **75.60%** (**+4.58%**). In this setting, LCF modulates the source and target spectral representations before magnitude-based alignment, enabling task-adaptive spectral modulation within the transfer framework.

Additional paired evaluations across five random seeds confirm consistent improvements over the original task-specific baselines on all four benchmarks, with two-sided paired tests yielding $p < 0.001$ across all reported metrics (Appendix I.2).

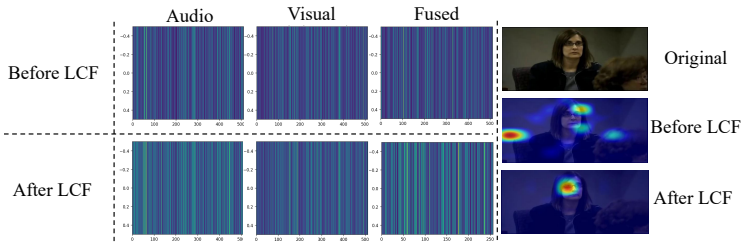


Figure 1: Multimodal Signal Evolution. Left: Feature response curves (Audio, Visual, Fused) before and after LCF. Right: Corresponding spectral heatmaps. The modulated features exhibit more concentrated spectral responses and changes in spatial attention patterns following LCF.

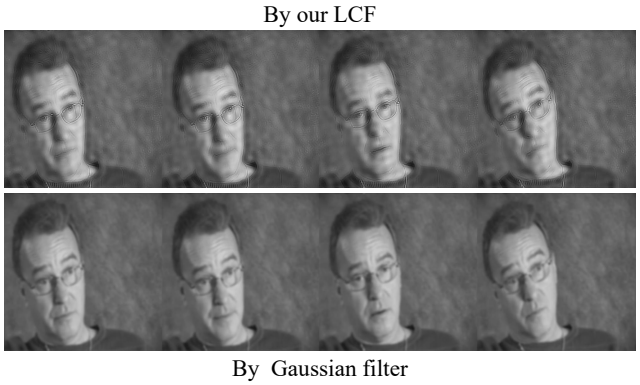


Figure 2: Visualization diagrams for detailed analysis after LCF module filtering. The upper section shows the images processed by our LCF module, while the lower section displays the results obtained using Gaussian filtering.

4.5 Case Study: Interpretable Visualization of LCF

To interpret the internal mechanism of the LCF module within the specific context of deception detection, we provide a comprehensive visualization analysis from two complementary perspectives: multimodal signal evolution (Figure 1) and spatial spectral filtering (Figure 2).

Multimodal Signal Evolution (Figure 1). We examine feature evolution to illustrate how LCF reshapes spectral representations. Compared with the original features, the LCF-processed features exhibit more concentrated spectral energy in the displayed examples. The attention heatmaps also show increased focus on facial regions, particularly around the eyes, illustrating changes in the learned feature representations.

Preservation of Discriminative Micro-Cues (Figure 2). Complementing the spectral analysis, Figure 2 compares the visual outputs of LCF and Gaussian filtering. In the displayed examples, LCF retains more visible facial detail, including regions around the eyes and mouth, while Gaussian filtering produces smoother outputs. These qualitative observations complement the quantitative evaluation of LCF.

5 Ablation Study

Table 4: Ablation study of LCF mechanisms on the Visual modality. We denote the components as ①: Learnable Chebyshev Response (fixed $s = 1$) and ②: Learnable Scaling Factor α.

Settings		RLT (Visual)			DOLOS (Visual)		
①	②	F1	ACC	AUC	F1	ACC	AUC
✗	✗	87.60	86.70	87.89	71.02	67.64	72.12
✓	✗	91.85	91.10	93.50	74.50	70.85	75.80
✓	✓	**93.75**	**93.75**	**96.35**	**76.00**	**73.00**	**77.77**

Table 5: Comparison of spectral bases across different benchmarks. Baseline refers to AFFAKT for Deception Detection and MVCL-DAF for Intent Recognition.

Method	RLT			DOLOS			MIntRec				MIntRec 2.0			
	F1	ACC	AUC	F1	ACC	AUC	ACC	WF1	WP	R	ACC	WF1	WP	R
Baseline*	87.60	86.70	87.89	71.02	67.64	72.12	74.72	74.61	75.07	71.94	57.80	55.05	55.82	42.03
Uni-Cheb (DFT)	**93.75**	**93.75**	**96.35**	**76.00**	**73.00**	**77.77**	76.40	76.33	76.67	74.35	60.26	59.19	60.54	54.12
Uni-Cheb (DWT)	91.67	92.00	95.20	74.15	71.20	75.90	**77.08**	**76.88**	**77.80**	**75.63**	**60.40**	**59.35**	**60.63**	**54.52**
Gap (DWT-DFT)	*-2.08*	*-1.75*	*-1.15*	*-1.85*	*-1.80*	*-1.87*	*+0.68*	*+0.55*	*+1.13*	*+1.28*	*+0.14*	*+0.16*	*+0.09*	*+0.40*

In this section, we conduct a systematic ablation analysis to evaluate the effectiveness of the proposed Uni-Cheb framework. We specifically examine the contributions of the Learnable Chebyshev Filter (LCF) components and task-dependent spectral transform preferences, demonstrating how the transform-compatible LCF formulation can be integrated with different spectral transforms.

5.1 Impact of Each Component

We decouple the contributions of the Chebyshev response parameterization and the adaptive scaling parameter α to evaluate their individual efficacy (Table 4). Without learnable scaling, the Chebyshev filter improves the RLT F1 score from 87.60% to 91.85%, demonstrating the effectiveness of the structured response parameterization.

Incorporating α further raises performance to 93.75%. By adjusting the response evaluation coordinates, the scaling parameter provides additional flexibility for task-adaptive spectral modulation. Detailed comparisons and sensitivity analyses are provided in Appendices H and G.

Alternative Response Parameterizations. We further compare LCF with nine alternative learnable spectral response parameterizations across four benchmarks, including parameter-matched Spline and RBF filters, alternative polynomial and cosine responses, learnable smooth filters, and an α-only control. Uni-Cheb achieves the highest mean performance across all 14 reported metrics.

Compared with the α-only control, Uni-Cheb improves F1 by 2.17 and 1.77 percentage points on RLT and DOLOS, respectively, and Accuracy by 1.47 and 1.21 points on MIntRec and MIntRec2.0. The parameter-matched comparisons further support the effectiveness of the structured Chebyshev response beyond additional learnable capacity. Complete results and experimental settings are provided in Appendix H.3.

Cross-Architecture Evaluation. We additionally evaluate LCF with PECL, MulT, FreeLunch, and a FLatten-based focused linear-attention encoder. These experiments extend the evaluation to different feature encoders, fusion architectures, and transfer mechanisms. Notably, incorporating LCF into the FLatten-based encoder improves mean Accuracy by 4.77, 4.47, 4.72, and 2.60 percentage points on RLT, DOLOS, MIntRec, and MIntRec2.0, respectively. Improvements are also observed with PECL, MulT, and FreeLunch. The complete cross-architecture results are reported in Appendix H.4.

5.2 Task-Dependent Spectral Transform Preferences

A core hypothesis of this work is that distinct multimodal tasks favor different spectral representations. To examine this task-dependent preference, we compare LCF instantiated with a Global Basis (Fourier/DFT) and a Local Basis (Wavelet/DWT) across deception detection and intent recognition. This comparison also evaluates whether the same LCF formulation can operate with different spectral transforms.

Experimental results on deception benchmarks primarily highlight the flexibility of our method. Uni-Cheb outperforms the corresponding baseline with either transform, indicating that the LCF response-learning formulation can be used in both spectral representations. Beyond this overall improvement, we observe a clear preference for the Global Basis. As shown in Table 5, the Fourier basis (DFT) surpasses the Wavelet basis (DWT) by 2.08 percentage points in F1 on the RLT dataset. This observation is consistent with the global support of Fourier representations, which can capture temporal variations over the full feature sequence. Such representations provide a suitable way to modulate longer-range patterns potentially relevant to deception detection.

In contrast, intent recognition exhibits a preference for the Local Basis. While both spectral variants outperform the baseline, Table 5 shows that DWT yields higher Recall (+1.28 percentage points) than DFT on MIntRec. This observation is consistent with the localized support of wavelets, which can represent variations at different positions and scales. In Uni-Cheb, the DWT decomposes encoded features into sub-bands that are independently modulated by the learned Chebyshev responses and then reconstructed for downstream fusion. This design makes local frequency information available to the original intent-recognition architecture without modifying its fusion mechanism.

The observed results yield two main insights. First, improvements from both variants support the applicability of LCF across the evaluated spectral transforms. Second, the relative preference for DFT in deception detection and DWT in intent recognition highlights task-dependent spectral characteristics, consistent with the motivation behind the spectral dilemma. Statistical comparisons across five independent random seeds are provided in Appendix Section I. Together, these findings illustrate the value of a transform-compatible design: the same response-learning module can be combined with a task-appropriate spectral transform while preserving the downstream architecture.

6 Conclusion

In this paper, we address the spectral dilemma in multimodal learning by proposing Uni-Cheb, a learnable spectral modulation operator built around the Learnable Chebyshev Filter (LCF). Motivated by Chebyshev approximation theory, LCF combines a compact spectral response parameterization with learnable response-coordinate scaling. Quantitative evaluations across four benchmarks demonstrate that Uni-Cheb consistently improves performance over state-of-the-art baselines. Furthermore, our experiments reveal task-dependent preferences for Fourier and Wavelet transforms. These findings highlight the effectiveness of our modular, transform-compatible design for spectral modulation across heterogeneous multimodal tasks.

Acknowledgment

This work is supported by the Guizhou Provincial Key Technology Support Program (Grant no. Qian-Ke-He-Zhicheng (2026) General 225) and the Chengdu Key Research and Development Support Program(Grant no. 2026-XT00-00005-GX).

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, pages 12449–12460, 2020.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. *arXiv preprint arXiv:2505.22483*, 2025.

Xiaoyu Cui, Weixing Chen, and Jiandong Su. A multiscale frequency domain causal framework for enhanced pathological analysis. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Elizabeth Fons, Alejandro Sztrajman, Yousef El-Laham, Luciana Ferrer, Svitlana Vyetrenko, and Manuela Veloso. Lscd: Lomb–scargle conditioned diffusion for time series imputation. In *Forty-second International Conference on Machine Learning*, 2025.

Felix A Gers and Jürgen Schmidhuber. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194. IEEE, 2000.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai Kin Kong, Bingquan Shen, and Alex Kot. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22135–22145, 2023.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5961–5971, October 2023.

Xilin He, Jingyu Hu, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Muhammad Haris Khan, and Linlin Shen. Towards combating frequency simplicity-biased learning for domain generalization. *Advances in Neural Information Processing Systems*, 37:31078–31102, 2024.

Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. Adaptive multimodal fusion: Dynamic attention allocation for intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17267–17275, 2025.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Forty-second International Conference on Machine Learning*, 2025.

Zihan Ji, Xuetao Tian, and Ye Liu. Affakt: A hierarchical optimal transport based method for affective facial knowledge transfer in video deception detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1336–1344, 2025.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 4296–4313, 2022.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Badri N Patro, Vinay P Namboodiri, and Vijay S Agneeswaran. Spectformer: Frequency and attention is what you need in a vision transformer. In *2025 IEEE/CVF winter conference on applications of computer vision (WACV)*, pages 9543–9554. IEEE, 2025.

Gaozheng Pei, Ke Ma, Yingfei Sun, Qianqian Xu, and Qingming Huang. Diffusion-based adversarial purification from the perspective of the frequency domain. In *Forty-second International Conference on Machine Learning*, 2025.

Verónica Pérez Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66, 2015.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, pages 10078–10093, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6558–6569, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zipeng Xiao, Siqi Kou, Zhongkai Hao, Bokai Lin, and Zhijie Deng. Amortized fourier neural operators. *Advances in Neural Information Processing Systems*, 37:115001–115020, 2024.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations*, 2021.

Wei Yang, Rui Zhong, Yiqun Chen, Chi Lu, and Peng Jiang. Structured spectral reasoning for frequency-adaptive multimodal recommendation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697, 2022.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. Mintrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations*, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024.

A Detailed Experimental Implementation of Spectral Instantiations

In this section, we provide the step-by-step experimental procedures for integrating Uni-Cheb into the baseline architectures for the three instantiations.

A.1 Instantiation I: Global Modulation for Deception Detection

Following the baseline architecture in Ji et al. [2025], the Uni-Cheb module is inserted immediately after the feature extraction backbone and before the classification head.

1. **Feature Extraction:** The visual features $\mathbf{F}_v$ are extracted using the pre-trained VideoMAE [Tong et al., 2022], and audio features $\mathbf{F}_a$ are extracted using W2V2 [Baevski et al., 2020].
2. **Spectral Transformation (FFT):** A 1D Fast Fourier Transform (FFT) is applied along the temporal dimension T to convert the sequential features into the frequency domain: $\mathbf{S}_v = \mathrm{FFT}(\mathbf{F}_v)$ and $\mathbf{S}_a = \mathrm{FFT}(\mathbf{F}_a)$.
3. **LCF Modulation:** To accommodate modality-specific characteristics, each modality maintains its own independent learnable filter bank. Specifically, LCF synthesizes distinct complex filter response matrices $\mathbf{W}_v$ and $\mathbf{W}_a$ for the visual and audio modalities, respectively. The spectral modulation is performed independently via element-wise complex multiplication: $\hat{\mathbf{S}}_v = \mathbf{S}_v \odot \mathbf{W}_v$ and $\hat{\mathbf{S}}_a = \mathbf{S}_a \odot \mathbf{W}_a$.
4. **Reconstruction (IFFT):** An Inverse FFT (IFFT) is applied to transform the modulated spectra back to the temporal domain, yielding refined features $\hat{\mathbf{F}}_v = \mathrm{IFFT}(\hat{\mathbf{S}}_v)$ and $\hat{\mathbf{F}}_a = \mathrm{IFFT}(\hat{\mathbf{S}}_a)$.
5. **Classification:** The processed features are concatenated and fed into the downstream MLP classifier.

A.2 Instantiation II: Local Modulation for Intent Recognition

We adopt the multimodal architecture proposed by Hu et al. [2025], where Uni-Cheb is inserted between the unimodal encoders and the Dynamic Attention Fusion (DAF) module.

1. **Encoding:** Inputs from visual, audio, and textual modalities are processed by the Transformer Encoder [Vaswani et al., 2017], BiPeephole LSTM [Gers and Schmidhuber, 2000], and BERT Encoder [Devlin et al., 2019] respectively, generating latent representations $\mathbf{H}_v, \mathbf{H}_a, \mathbf{H}_t$.
2. **Wavelet Decomposition (DWT):** A Discrete Wavelet Transform (DWT) is applied to the feature maps of each modality, decomposing them into multiple frequency sub-bands.
3. **Sub-band LCF Modulation:** The LCF is applied independently to each wavelet sub-band. The filter weights modulate the energy of each sub-band via element-wise multiplication.
4. **Reconstruction (IDWT):** An Inverse DWT (IDWT) is performed on the modulated sub-bands to reconstruct the refined feature representations $\hat{\mathbf{H}}_v, \hat{\mathbf{H}}_a, \hat{\mathbf{H}}_t$.
5. **Fusion (DAF):** The reconstructed features are directly input into the DAF module for cross-modal fusion and classification.

A.3 Instantiation III: Spectral Alignment for Transfer Learning

Uni-Cheb is integrated into the preprocessing stage of the H-OTKT framework [Ji et al., 2025].

1. **Input Preparation:** The source domain features $\mathbf{Z}_{\mathcal{S}}$ and target domain features $\mathbf{Z}_{\mathcal{T}}$ are taken as inputs.
2. **Spectral Projection (FFT):** Both source and target features are transformed into the frequency domain via FFT: $\mathbf{S}_{\mathcal{S}} = \mathrm{FFT}(\mathbf{Z}_{\mathcal{S}})$ and $\mathbf{S}_{\mathcal{T}} = \mathrm{FFT}(\mathbf{Z}_{\mathcal{T}})$.
3. **Spectral Filtering (LCF):** LCF is applied to the spectra of both domains. The spectra are modulated by the learnable complex filter response to adjust the magnitude and phase of the frequency components.

4. **Hierarchical Optimal Transport:** The hierarchical Sinkhorn distance is computed between the magnitudes of the LCF-modulated source and target spectra to obtain the alignment loss.
5. **Inverse Projection (IFFT):** While the optimal transport explicitly supervises the spectral alignment via $\mathcal{L}_{align}$, the refined representations are concurrently transformed back to the spatiotemporal domain via IFFT for downstream processing.

B Notations

Table 6: Notations and their corresponding descriptions used in Uni-Cheb.

Notations	Description	Notations	Description
$\mathbb{R}, \mathbb{C}$	Real / Complex number space	y	Task-specific class label
$\mathcal{D}$	Multimodal dataset $\{(\mathbf{X}_i, y_i)\}$	$\mathbf{X}$	Input feature sequence ($T \times C$)
$\mathcal{T}, \mathcal{T}^{-1}$	Forward / Inverse spectral transform	$\mathcal{H}(\cdot)$	LCF modulation function
$\mathbf{S}$	Generalized spectral representation	$\hat{\mathbf{X}}$	Reconstructed refined feature
α	Learnable response-coordinate scaling parameter	ϕ_k	Canonical angular nodes
θ_k	Scaled response evaluation coordinates	$\mathbf{B}$	Learnable Chebyshev response coefficients
$\mathbf{W}$	Learned spectral response	$\lvert\mathbf{W}\rvert$	Magnitude of filter response
$\mathbf{S}_{\text{DFT}}$	Complex Spectrum in DFT	$\hat{\mathbf{S}}_{\text{DFT}}$	Modulated spectrum (DFT)
$\mathbf{X}_{\text{spatial}}$	Reshaped spatial feature map	$\mathbf{C}_{sub}$	Wavelet coefficients (DWT)
sub	Wavelet sub-band index	$\hat{\mathbf{C}}_{sub}$	Reweighted wavelet coefficients
$\mathbf{Z}_{\mathcal{S}}, \mathbf{Z}_{\mathcal{T}}$	Source / Target latent features	$\mathbf{W}_{\text{align}}$	Spectral alignment response
$\Phi_{\text{LCF}}(\cdot)$	LCF-modulated spectral operator	$\mathcal{L}_{\text{align}}$	Spectral alignment loss
$\mathcal{D}_{\text{H-OTKT}}$	Hierarchical Sinkhorn distance	$\odot$	Element-wise multiplication

To facilitate readability and ensure clarity across the diverse spectral instantiations (i.e., global DFT, local DWT, and cross-domain transfer), we summarize the frequently used mathematical notations in Table 6.

C Datasets

Deception Detection Datasets. We conduct the experiments on two most widely used datasets in deception detection task, Real Life Trial (RLT) dataset and DOLOS dataset:

- **Real Life Trial (RLT)** dataset is a popular real-world dataset collected from public court trials, which consists of 121 videos including 61 deceptive and 60 truthful video clips. As it is a real-world dataset, the Real Life Trial dataset has more noise on both the video and audio.
- **DOLOS** is the largest game-show deception detection dataset recently proposed in the field, containing rich deceptive dialogues. The dataset consists of 1,675 video clips featuring 213 subjects (141 male and 72 female participants), with each clip lasting 2-19 seconds.

Intent Recognition Datasets. We conduct the experiments on two most widely used datasets in Intent Recognition task:

- **MIntRec** [Zhang et al., 2022] dataset contains 2224 high-quality samples and 20 intent labels. The samples cover text, visual, and acoustic modalities.
- **MIntRec2.0** [Zhang et al., 2024] represents a large-scale benchmark dataset designed for multimodal intent recognition. This dataset encompasses 1245 high-quality dialogues, aggregating a total of 15,040 samples that incorporate text, video, and audio modalities. A distinguishing characteristic of the MIntRec 2.0 dataset is its fine-grained intent classification, featuring a robust categorization system with 30 distinct intent categories. Furthermore, the dataset is specifically engineered to include “out-of-range" samples that naturally arise in conversational exchanges, thereby providing a more rigorous testing scenario for evaluating model performance.

Facial Expression Recognition Datasets. DFEW is a large-scale real-world dataset collected from over 1,500 movies, consisting of 16,372 video clips annotated with seven basic emotions (anger,

disgust, fear, happy, sad, surprise, neutral). As a real-world dataset, DFEW contains significant variations in illumination, pose, and occlusion, making it highly challenging for facial expression recognition tasks.

D Further Visualization Study

D.1 Cross-Domain Visualization Analysis in Deception Detection

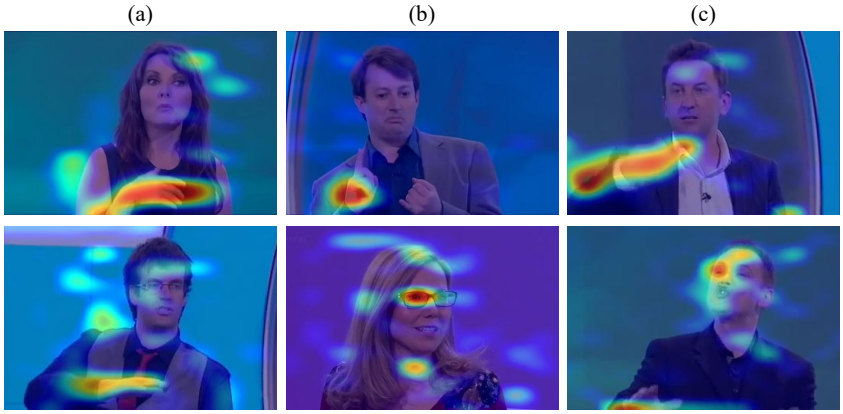


Figure 3: More visualization cases. We visualize the cases trained on the RLT dataset and tested on the DOLOS dataset.

To rigorously evaluate the generalization capability and robustness of the proposed framework, we extend our visual analysis beyond the within-domain settings discussed in the main text. Specifically, we visualize the attention maps of a model trained on the **RLT** dataset (high-stakes courtroom scenarios) and tested on the **DOLOS** dataset (low-stakes gaming environments).

As illustrated in Figure 3, despite the distribution shift between domains, the model attends to localized facial and hand regions in the displayed examples. **Hand Gestures.** Subfigures (a)–(d) and (f) show attention concentrated around hand movements, suggesting that these regions contribute to the learned representations. **Ocular Cues.** Subfigures (e) and (f) show attention near the eyes, indicating that ocular regions are also emphasized in the displayed examples.



Figure 4: Visualization Examples of the Model's Attentional Regions.

D.2 Visualization Analysis in Deception Detection

To provide interpretability into the learned representations, we visualize the attention maps in Figure 4. As observed in the top row, Uni-Cheb places more concentrated attention on ocular and facial regions in the displayed examples. These attention patterns suggest that the learned representations emphasize localized facial information relevant to the downstream task. In comparison, the baseline AFFAKT (bottom row) exhibits more diffuse responses, including attention to background regions and clothing.

These qualitative differences complement the cross-domain performance comparisons reported in the main text.

D.3 Visualization Analysis in Intent Recognition



Figure 5: Visualization Examples of the Model's Attentional Regions.

To provide interpretability into the learned representations, we visualize the attention maps in Figure 5, comparing our Uni-Cheb against the baseline MVCL-DAF.

As observed in the first row (By Uni-Cheb), the model shows concentrated attention around the ocular region and other facial details. Such regions contain fine spatial structures, including edges and local texture variations, which can be represented by high-frequency wavelet sub-bands. Compared with MVCL-DAF, the displayed examples show more spatially concentrated attention under Uni-Cheb. This qualitative comparison complements the intent-recognition results reported in the main text.

E Theoretical Analysis of Spectral Approximation

In this section, we present the approximation-theoretic motivation for the Learnable Chebyshev Filter (LCF). We first review the classical minimax property of Chebyshev nodes and then discuss the use of a truncated Chebyshev expansion for parameterizing spectral responses. These properties motivate the structured and compact formulation adopted in LCF.

E.1 Preliminaries and Problem Formulation

For the approximation-theoretic analysis, consider an ideal spectral response function $H^*(\lambda)$ defined on the normalized domain $\lambda \in [-1, 1]$, with $H^* \in C[-1, 1]$.

Instead of learning unconstrained independent weights for all K frequency bins (which is prone to overfitting and high-frequency noise), LCF evaluates a continuous polynomial approximator $P_{M-1}(\lambda)$ of degree $M-1$ over a set of discrete sampling nodes $\mathcal{X} = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$.

The approximation quality over the domain is measured by the uniform norm (minimax error):

$$E_{M-1}(H^*) = \inf_{P \in \mathbb{P}_{M-1}} \|H^* - P\|_\infty = \inf_{P \in \mathbb{P}_{M-1}} \sup_{\lambda \in [-1,1]} |H^*(\lambda) - P(\lambda)|, \tag{11}$$

where $\mathbb{P}_{M-1}$ denotes the space of polynomials of degree at most $M-1$, and $E_{M-1}(H^*)$ is the absolute best uniform approximation error achievable.

E.2 Classical Minimax Property of Chebyshev Nodes

The canonical response coordinates are motivated by classical polynomial interpolation theory. In particular, Chebyshev nodes minimize the maximum absolute value of the monic nodal polynomial, a property underlying standard interpolation error bounds.

Assume $H^* \in C^K([-1, 1])$, and let P_{K-1} interpolate H^* at K distinct nodes. The classical interpolation error is bounded by:

$$\|H^* - P_{K-1}\|_\infty \leq \frac{\|H^{*(K)}\|_\infty}{K!} \|\omega\|_\infty, \tag{12}$$

where $\omega(\lambda) = \prod_{i=1}^{K}(\lambda - \lambda_i)$ is the **nodal polynomial**. Since $\|H^{*(K)}\|_\infty$ depends purely on the target signal's intrinsic complexity, minimizing the worst-case error bound requires minimizing the infinity norm of the monic polynomial $\omega(\lambda)$:

$$\min_{\{\lambda_i\}} \max_{\lambda\in[-1,1]} |\omega(\lambda)| = \min_{\{\lambda_i\}} \max_{\lambda\in[-1,1]} \left|\prod_{i=1}^{K}(\lambda - \lambda_i)\right|. \tag{13}$$

[**Minimax Property of Scaled Chebyshev Polynomials**] The unique monic polynomial of degree K that minimizes the infinity norm on $[-1, 1]$ is the scaled Chebyshev polynomial of the first kind $\tilde{T}_K(\lambda) = 2^{-(K-1)}T_K(\lambda)$. Consequently, the optimal evaluation nodes minimizing the worst-case nodal bound are the roots of $T_K(\lambda)$.

Proof. The proof proceeds by contradiction.

1. **Properties**: $\tilde{T}_K(\lambda)$ attains its maximum absolute value $\|\tilde{T}_K\|_\infty = 1/2^{K-1}$ with alternating signs at $K + 1$ extrema points $\eta_j = \cos(\frac{j\pi}{K})$ for $j = 0, \ldots, K$. Specifically, $\tilde{T}_K(\eta_j) = (-1)^j/2^{K-1}$.

2. **Contradiction Hypothesis**: Suppose there exists another monic polynomial $Q(\lambda)$ of degree K such that $\|Q\|_\infty \le \|\tilde{T}_K\|_\infty$ and $Q \not\equiv \tilde{T}_K$. Let $R(\lambda) = \tilde{T}_K(\lambda) - Q(\lambda)$. Since both are monic, the λ^K terms cancel, leaving $R(\lambda)$ as a polynomial of degree at most $K - 1$.

3. **Sign Alternation and Root Counting**: At the extrema η_j, we evaluate $R(\eta_j) = \tilde{T}_K(\eta_j) - Q(\eta_j)$. Since $|Q(\eta_j)| \le \|\tilde{T}_K\|_\infty = |\tilde{T}_K(\eta_j)|$, it follows that $R(\eta_j)$ weakly alternates signs across the $K + 1$ points ($\eta_0 > \eta_1 > \cdots > \eta_K$). Consequently, $R(\lambda)$ must have at least one root in each of the K intervals $[\eta_j, \eta_{j-1}]$. If $R(\eta_j) \neq 0$ for all j, $R(\lambda)$ has strictly K distinct roots. If $R(\eta_j) = 0$ for some internal extremum $j \in \{1, \ldots, K - 1\}$, then since $\tilde{T}_K(\eta_j)$ is a global extremum and $\|Q\|_\infty \le \|\tilde{T}_K\|_\infty$, the polynomial $Q(\lambda)$ must be tangent to $\tilde{T}_K(\lambda)$ at η_j. This implies η_j is also a local extremum of $R(\lambda)$ (i.e., $R'(\eta_j) = 0$), making it a root of multiplicity at least 2. In all cases, counting multiplicities, $R(\lambda)$ has at least K roots. Since the degree of $R(\lambda)$ is at most $K - 1$, the Fundamental Theorem of Algebra dictates that $R(\lambda) \equiv 0$. Therefore, $Q(\lambda) = \tilde{T}_K(\lambda)$, which contradicts the assumption that $Q \not\equiv \tilde{T}_K$. Hence, uniqueness is rigorously proven. □

Solving $T_K(\lambda) = 0$ yields the Chebyshev roots $\lambda_i = \cos\left(\frac{2i-1}{2K}\pi\right)$, corresponding to the canonical angular coordinates $\phi_k = \frac{2k-1}{2K}\pi$ used in LCF.

E.3 Filter Synthesis via Truncated Chebyshev Expansion

LCF synthesizes its spectral response using a low-order Chebyshev expansion of degree $M - 1$ ($M < K$), with learnable coefficients $B_{m,c}$. For channel c, the response is expressed as:

$$S_{M-1}(\lambda) = \sum_{m=0}^{M-1} B_{m,c}T_m(\lambda) \tag{14}$$

For classical polynomial interpolation, let $I_{M-1}H^*$ denote the degree-$(M - 1)$ interpolant of H^* at M Chebyshev nodes. Lebesgue's lemma gives:

$$\|H^* - I_{M-1}H^*\|_\infty \le (1 + \Lambda_{M-1})E_{M-1}(H^*), \tag{15}$$

where Λ_{M-1} denotes the Lebesgue constant of the interpolation operator, which grows as $\mathcal{O}(\log M)$ for Chebyshev nodes.

In LCF, the coefficients $B_{m,c}$ are optimized using downstream task objectives. The classical interpolation bound therefore motivates the choice of a structured Chebyshev parameterization, while the effectiveness of the learned response is established empirically.

E.4 Learnable Response-Coordinate Warping

The learnable scaling parameter α defines $s = \mathrm{Softplus}(\alpha)$ and adjusts the response evaluation coordinates as $\theta_k = s\phi_k$. This operation changes where the Chebyshev response is evaluated while leaving the underlying spectral coefficients unchanged.

In the polynomial coordinate, the mapping corresponds to $x_k = \cos(s\phi_k)$. For $s \neq 1$, these coordinates do not generally coincide with the original Chebyshev roots, so the classical interpolation optimality result does not directly extend to the learned response. Instead, the scaling parameter and Chebyshev coefficients are jointly optimized through the downstream task objective, enabling task-adaptive spectral modulation.

E.5 Computational Complexity and Efficiency Analysis

To systematically evaluate the computational overhead of spectral transformations (e.g., FFT/DWT and complex-valued operations), we conducted an end-to-end hardware profiling. We deployed our models on a single NVIDIA RTX 4090 GPU to measure the exact parameter counts, inference latency, and VRAM footprints. To ensure a unified and fair comparison, all inference latencies are normalized to **milliseconds per sample (ms/sample)**. Furthermore, to isolate the efficiency of our Fourier-based design, we implemented a "Pure Wavelet" variant of Uni-Cheb that strips away all cross-subband convolution networks, relying purely on element-wise Chebyshev filtering.

Table 7: System-Level Computational Overhead Analysis on Deception Detection and Intent Recognition. All latencies are normalized to milliseconds per sample (ms/sample). "Added Params" refers to the extra parameters introduced compared to the respective baseline.

Task	Model Configuration	Total Params (M)	Added Params (M)	Latency (ms/sample)	VRAM (GB)
Deception	Baseline (No Filter)	85.608 M	-	35.31 ms	16.41 GB
	Ours: Wavelet Uni-Cheb	85.612 M	+ 0.004 M	35.31 ms	17.26 GB
	Ours: Fourier Uni-Cheb	**85.610 M**	**+ 0.002 M**	**35.63 ms**	**16.90 GB**
Intent	Baseline (No Filter)	740.025 M	-	53.94 ms	17.84 GB
	Ours: Wavelet Uni-Cheb	740.050 M	+ 0.025 M	60.97 ms	17.89 GB
	Ours: Fourier Uni-Cheb	**740.036 M**	**+ 0.011 M**	**57.37 ms**	**17.85 GB**

As detailed in Table 7, the proposed Fourier-based Uni-Cheb is exceptionally lightweight. In Deception Detection, it introduces merely **0.002 M** additional parameters (a $\sim 0.002\%$ increase over the 85.6M baseline) and incurs an imperceptible latency overhead of $+0.32$ **ms per sample**. In Intent Recognition, which requires filtering across three independent modalities, the Wavelet-based variant introduces **+0.025 M** parameters and **+7.03 ms per sample** in inference latency.

This extreme parameter efficiency stems directly from the mathematical formulation of the modules. Because the Fourier transform operates on a unified global spectrum, it requires only a single complex-valued filter bank. In contrast, the spatial-frequency localisation of the Wavelet transform requires independent filters for each spatial subband, which slightly increases the parameter footprint. These measurements demonstrate that LCF introduces a modest computational overhead relative to the corresponding baseline architectures.

F Empirical Convergence of the Spectral Scaling Factor

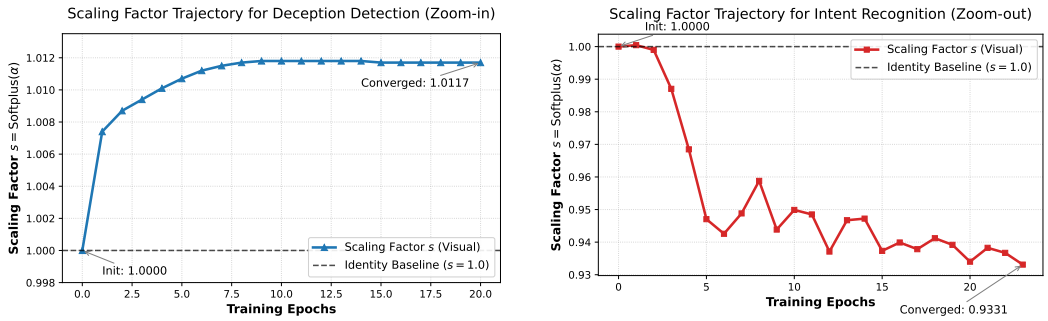
A key component of Uni-Cheb is the learnable response-coordinate scaling parameter α. To examine its optimization behavior, we tracked the spectral scaling factor $s = \mathrm{Softplus}(\alpha)$ during training across different tasks.

As illustrated in Figure 6, the learned scaling factor s converges to different values across the evaluated tasks.

Specifically, under the Chebyshev response-coordinate mapping $\theta_k = s\phi_k$, the scaling factor s determines the evaluation coordinates of the learned filter response:

- **Deception Detection ($s > 1$):** As shown in Figure 6a, the scaling factor converges to 1.0117, expanding the angular coordinates at which the Chebyshev response is evaluated.
- **Intent Recognition ($s < 1$):** As shown in Figure 6b, the scaling factor converges to 0.9331, compressing the response evaluation coordinates.

These trajectories illustrate how the response-coordinate scaling parameter adapts to different downstream tasks during training.



(a) Deception Detection (b) Intent Recognition

Figure 6: Training trajectories of the spectral scaling factor $s = \mathrm{Softplus}(\alpha)$, initialized at 1.0, for the visual modality. The learned scaling factor converges to 1.0117 for deception detection and 0.9331 for intent recognition, reflecting task-dependent response-coordinate scaling.

G Hyperparameter Sensitivity Analysis of M

This section investigates the sensitivity of Uni-Cheb's performance to the hyperparameter M, which denotes the number of terms in the Chebyshev expansion (with maximum polynomial degree $M-1$). The selection of M is a pivotal factor in modulating the balance between spectral resolution and model capacity. To assess the generalisability of the proposed framework, systematic evaluations were conducted across $M \in \{2, 3, 5, 10, 20\}$ on both the small-scale Real-Life Trial (RLT) dataset and the larger-scale MIntRec benchmark. The expansion size M was selected using validation data only; $M = 2$ was then fixed for the reported test evaluations.

Empirical Analysis on the RLT Dataset. Given the restricted sample volume of the RLT dataset (121 video clips), maintaining appropriate structural regularisation is essential to prevent overfitting. The experimental results for the visual modality are summarised in Table 8.

Table 8: Hyperparameter sensitivity analysis of the number of filter banks (M) on the RLT dataset (Visual Modality). Results indicate that the configuration with $M = 2$ achieves a favorable balance between spectral representation and regularisation.

Metric (%)	M=2	M=3	M=5	M=10	M=20
Accuracy	**93.75**	92.00	88.00	88.00	88.00
F1-Score	**93.75**	92.31	88.89	88.89	88.89
AUC	**96.35**	92.80	92.64	94.24	94.24

The performance trajectories suggest a correlation between spectral capacity and model robustness:

- **Regularisation Effectiveness at $M = 2$:** Under the current experimental setting, the highest Accuracy (93.75%) and AUC (96.35%) are yielded by the most constrained configuration ($M = 2$). In low-data regimes, the LCF module with a minimal expansion degree appears to function as an effective structural prior. By limiting the degrees of freedom, the model may reduce sensitivity to sample-specific spectral variations.
- **Impact of Increased Capacity ($M \geq 3$):** A marginal decline in performance is observed as M increases. Specifically, an Accuracy reduction of approximately 6% occurs when M scales from 2 to 5. This phenomenon suggests that higher-order polynomials may introduce excessive spectral capacity, potentially leading the network to incorporate sample-specific spectral artefacts rather than extracting generalisable behavioural cues.

Empirical Analysis on the MIntRec Dataset. To further examine whether these observations extend to larger-scale scenarios, evaluations were performed on the fine-grained MIntRec benchmark. The results are detailed in Table 9.

The empirical findings on MIntRec are consistent with the trends observed in the RLT dataset. The configuration with $M = 2$ achieves the most competitive results across all metrics, including an

Table 9: Hyperparameter sensitivity analysis of the number of filter banks (M) on the MIntRec dataset. The data suggests that $M = 2$ remains a robust choice across multiple metrics under the defined conditions.

Metric (%)	M=2	M=3	M=5	M=10	M=20
ACC	**77.08**	75.16	74.61	75.06	75.06
WF1	**76.88**	75.88	74.33	75.17	74.28
WP	**77.80**	75.30	74.78	75.76	74.77
R	**75.63**	74.50	74.14	74.61	73.97

Accuracy of 77.08% and a Recall of 75.63%. Notably, as M increases beyond 2, a plateau and subsequent fluctuation in performance are recorded, with Recall experiencing a slight decline to 73.97% at $M = 20$. This further indicates that an ultra-lightweight LCF might provide sufficient flexibility for capturing essential spectral dynamics while maintaining the necessary regularisation.

In conclusion, these results indicate that the hyperparameter M modulates the trade-off between spectral resolution and model capacity. Within the scope of these experiments, the $M = 2$ configuration demonstrates stable performance across different dataset scales, suggesting its effectiveness in regularising the spectral response against idiosyncratic noise.

H Extended Comparisons with Alternative Spectral Methods

To comprehensively evaluate the Learnable Chebyshev Filter (LCF) within the broader landscape of spectral learning, we compare our framework against both classical signal processing filters and recent deep frequency-domain architectures. This evaluation spans two distinct spectral scenarios: global modulation for deception detection and local modulation for intent recognition.

To this end, we replace the LCF module with the following representative deep frequency paradigms while maintaining the original spatial-temporal backbones (e.g., AFFAKT and MVCL-DAF) identical:

- **FNet Lee-Thorp et al. [2022]:** A lightweight architecture that replaces self-attention with standard, unparameterised discrete Fourier transforms for token mixing.
- **AFNO Guibas et al. [2021]:** An Adaptive Fourier Neural Operator that frames token mixing as continuous global convolution using block-diagonal weight matrices in the frequency domain.
- **SpectFormer Patro et al. [2025]:** A spectral-spatial transformer that combines frequency-domain representations with attention mechanisms.
- **FreDF Wang et al. [2025]:** A framework specifically designed to forecast and learn representations entirely within the frequency domain.

As shown in Tables 10 and 11, LCF achieves favorable results compared with the evaluated spectral baselines. These comparisons support the effectiveness of the proposed structured, learnable spectral response parameterization.

H.1 Deception Detection

We first evaluate alternative spectral mechanisms on the visual modality of the deception detection datasets. We integrated multiple baselines: classical filters (**Butterworth**, **Gaussian**) and advanced deep frequency architectures (**FNet** Lee-Thorp et al. [2022], **SpectFormer** Patro et al. [2025], **AFNO** Guibas et al. [2021], **FreDF** Wang et al. [2025], and a **Simple Learnable Mask**).

As shown in Table 10, LCF outperforms the evaluated spectral alternatives on RLT and DOLOS. These comparisons highlight three aspects of the learned spectral response design:

- **Classical Filters (Butterworth/Gaussian):** These filters use predefined spectral response forms, whereas LCF learns its response coefficients and evaluation coordinates jointly from the downstream task objective. On RLT, the Gaussian filter achieves 90.00% Accuracy, compared with 93.75% for LCF.

Table 10: Comparison results of different spectral filters and architectures on the RLT and DOLOS datasets (Visual Modality). Results are reported as mean±std across 5-fold cross-validation. **Bold**: Optimal; Underline: Sub-optimal.

Method / Architecture	RLT (Visual)			DOLOS (Visual)		
	F1	ACC	AUC	F1	ACC	AUC
AFFAKT (Base)	87.60±5.16	86.70±5.58	87.89±5.16	71.02±2.33	67.64±1.99	72.12±2.92
+ FNet Lee-Thorp et al. [2022]	88.15±4.20	87.20±4.15	92.30±4.10	72.40±4.25	69.10±4.10	73.80±4.35
+ SpectFormer Patro et al. [2025]	88.45±4.55	87.60±4.30	92.80±4.25	71.80±4.55	68.40±4.30	74.10±4.50
+ AFNO Guibas et al. [2021]	89.25±3.90	88.80±3.85	93.50±3.70	72.80±3.95	69.50±3.85	74.20±3.90
+ FreDF Wang et al. [2025]	88.89±4.47	88.00±4.39	93.92±4.39	72.15±4.60	68.60±4.45	73.40±4.70
+ Learnable Mask	87.90±4.80	86.90±4.60	91.00±4.50	71.20±4.80	67.80±4.65	72.30±4.85
+ Gaussian Filter	90.31±4.76	90.00±3.23	94.36±3.23	73.50±5.15	70.85±4.95	74.80±5.25
+ Butterworth Filter	88.89±7.30	88.00±4.27	94.40±4.27	72.30±6.30	68.80±5.80	73.50±6.50
+ LCF (Ours)	**93.75**±5.34	**93.75**±4.06	**96.35**±4.06	**76.00**±3.48	**73.00**±2.99	**77.77**±2.99

- **Deep Spectral Architectures (e.g., AFNO, FreDF):** These architectures offer alternative frequency-domain modeling strategies. Under the evaluated deception-detection settings, LCF achieves stronger results with a compact, structured response parameterization.
- **Effectiveness of LCF:** LCF achieves 93.75% Accuracy on RLT and 73.00% on DOLOS. Its learnable response coefficients and coordinate scaling parameter jointly adapt the spectral response to the downstream task. The results support the effectiveness of this lightweight parameterization compared with the evaluated alternatives.

H.2 Intent Recognition

To further situate LCF, we compared our method against advanced unconstrained frequency-domain architectures—including FNet, SpectFormer, and AFNO—as well as classical filtering baselines on the fine-grained MIntRec benchmarks.

Table 11: Comprehensive performance comparison with extended frequency-domain baselines on Intent Recognition. LCF consistently outperforms advanced spectral architectures (e.g., FreDF, SpectFormer) and static filters, particularly in preserving Macro Recall (R).

Method / Architecture	MIntRec				MIntRec 2.0			
	ACC	WF1	WP	R	ACC	WF1	WP	R
MVCL-DAF (Base)	74.72	74.61	75.07	71.94	57.80	55.05	55.82	42.03
+ FNet Lee-Thorp et al. [2022]	74.38	73.82	74.34	69.02	57.55	57.43	58.65	52.83
+ SpectFormer Patro et al. [2025]	74.38	74.37	74.73	71.69	58.44	57.99	58.34	53.19
+ AFNO Guibas et al. [2021]	75.28	75.03	75.60	73.92	58.88	58.14	58.46	53.70
+ FreDF Wang et al. [2025]	75.06	74.89	75.61	70.73	59.12	58.65	59.14	52.97
+ Simple Learnable Mask	74.83	74.59	75.59	73.38	58.39	57.86	58.16	53.02
+ Gaussian Filter	76.18	75.26	76.87	74.46	58.19	57.03	58.31	52.73
+ Butterworth Filter	76.18	75.96	76.92	72.98	58.68	58.03	58.24	53.68
+ LCF (Ours)	**77.08**	**76.88**	**77.80**	**75.63**	**60.40**	**59.35**	**60.63**	**54.52**

As shown in Table 11, LCF consistently outperforms the evaluated spectral architectures across multiple metrics on MIntRec and MIntRec2.0. For example, AFNO achieves Recall(R) scores of 73.92% and 53.70%, respectively, compared with 75.63% and 54.52% for LCF. These results further support the effectiveness of the compact Chebyshev response parameterization for fine-grained intent recognition.

Table 12: Controlled comparisons of learnable spectral response parameterizations on RLT and DOLOS (Visual Modality). All metrics are reported in % as mean±std across five independent seeds.

Parameterization	RLT (Visual)			DOLOS (Visual)		
	F1	ACC	AUC	F1	ACC	AUC
Spline	84.52±0.21	84.17±0.26	85.63±0.27	69.25±0.29	65.75±0.34	70.38±0.31
RBF	87.90±0.18	87.14±0.24	89.13±0.32	72.50±0.25	69.08±0.28	73.74±0.29
Ordinary Cosine	91.12±0.17	90.84±0.21	92.83±0.27	73.66±0.28	70.44±0.30	74.86±0.35
Learnable Gaussian	89.50±0.15	89.20±0.21	93.20±0.29	72.10±0.24	69.50±0.29	73.20±0.37
Learnable Butterworth	87.80±0.34	87.40±0.36	91.50±0.39	70.50±0.37	67.20±0.47	71.40±0.48
Smoothness-Regularized Mask	88.40±0.19	87.45±0.19	91.55±0.23	71.65±0.18	68.25±0.17	72.70±0.18
Legendre	89.87±0.24	89.47±0.27	91.51±0.24	73.00±0.25	69.99±0.25	74.13±0.31
Monomial	89.42±0.28	89.07±0.31	91.10±0.32	72.61±0.33	69.51±0.31	73.61±0.41
α-only	91.51±0.20	91.23±0.21	93.72±0.22	74.11±0.22	70.81±0.22	75.58±0.21
Uni-Cheb (Ours)	**93.68±0.20**	**93.81±0.25**	**96.28±0.18**	**75.88±0.27**	**73.12±0.23**	**77.79±0.23**

Table 13: Controlled comparisons of learnable spectral response parameterizations on MIntRec and MIntRec2.0. All metrics are reported in % as mean±std across five independent seeds.

Parameterization	MIntRec				MIntRec 2.0			
	ACC	WF1	WP	R	ACC	WF1	WP	R
Spline	73.26±0.13	73.13±0.14	73.56±0.12	69.94±0.27	59.27±0.09	58.41±0.18	59.02±0.25	53.84±0.29
RBF	75.28±0.13	75.15±0.14	75.43±0.14	72.47±0.26	58.44±0.09	57.58±0.18	58.18±0.19	52.91±0.27
Ordinary Cosine	74.99±0.23	74.63±0.24	74.74±0.19	72.09±0.30	58.94±0.17	58.29±0.19	58.69±0.21	52.66±0.27
Learnable Gaussian	74.96±0.21	74.70±0.21	75.06±0.20	71.63±0.29	57.20±0.25	56.40±0.25	57.10±0.24	51.80±0.44
Learnable Butterworth	74.06±0.31	73.76±0.34	73.98±0.32	71.82±0.44	57.60±0.40	56.80±0.33	57.30±0.39	52.10±0.49
Smoothness-Regularized Mask	74.57±0.21	74.24±0.21	74.66±0.22	72.49±0.21	58.75±0.17	58.15±0.18	58.55±0.22	53.30±0.22
Legendre	74.31±0.26	73.66±0.26	74.77±0.27	71.31±0.26	58.70±0.22	56.48±0.19	56.95±0.28	52.16±0.28
Monomial	74.73±0.33	73.88±0.33	74.45±0.31	71.57±0.33	58.35±0.31	56.07±0.30	56.47±0.36	51.76±0.31
α-only	75.46±0.38	75.10±0.37	75.12±0.38	72.60±0.37	59.11±0.19	57.86±0.16	58.52±0.20	53.14±0.21
Uni-Cheb (Ours)	**76.93±0.12**	**76.73±0.11**	**77.64±0.12**	**75.48±0.11**	**60.32±0.15**	**59.41±0.17**	**60.56±0.15**	**54.40±0.22**

H.3 Controlled Comparisons with Learnable Response Parameterizations

We compare LCF with nine alternative learnable spectral response parameterizations across four benchmarks. Within each task, all variants use the same backbone, spectral transform, insertion point, data split, optimizer, training schedule, model-selection protocol, and coordinate-scaling configuration. Experiments use five independent seeds ($\{42, 10, 100, 2024, 3407\}$), with results reported as mean $\pm$ sample standard deviation. These five-seed experiments were conducted independently of the DFT/DWT comparisons in Appendix I, although both used the same random seed identifiers.

Spline and RBF are parameter-matched to LCF, while Ordinary Cosine, Legendre, and Monomial provide alternative response parameterizations. We further consider learnable Gaussian and Butterworth responses, a Smoothness-Regularized Mask, and an α-only control. The latter uses a fixed response $W_0(x) = 1 + 0.1x$ with learnable coordinate scaling.

All control-specific hyperparameters are selected using validation data only. The RBF bandwidth is selected from $\{0.1, 0.25, 0.5, 1.0, 2.0\}$. The Smoothness-Regularized Mask uses low-resolution control points with a second-order difference penalty, whose regularization coefficient is selected from $\{0.01, 0.1, 1, 10\}$, with $\lambda = 0.01$ selected on all four datasets. Ordinary Cosine uses a cosine-series response evaluated on uniformly normalized warped coordinates, retaining the same task-specific transform and scaling mechanism. The Learnable Gaussian and Learnable Butterworth controls optimize their response parameters and differ from the fixed Gaussian and Butterworth filters evaluated in Section H.

As shown in Tables 12 and 13, Uni-Cheb achieves the highest mean performance across all reported metrics on the four benchmarks. The parameter-matched Spline and RBF comparisons indicate that the improvements are not explained solely by additional learnable response parameters. Comparisons with alternative polynomial, cosine, and smooth response families further support the effectiveness of the proposed response parameterization. The α-only control additionally highlights the contribution

of jointly learning response coefficients and coordinate scaling. For the four primary metrics (F1 on RLT and DOLOS, and Accuracy on MIntRec and MIntRec2.0), Uni-Cheb outperforms the stronger of the Legendre and Monomial controls by 3.81, 2.88, 2.20, and 1.62 percentage points, respectively. Two-sided paired t-tests on seed-matched runs yield $p < 0.001$ for all four comparisons. Relative to the α-only control, the corresponding improvements are 2.17, 1.77, 1.47, and 1.21 percentage points, with $p < 0.01$ for all four comparisons.

H.4 Cross-Backbone and Transfer-Mechanism Evaluations

We evaluate the compatibility of LCF with additional deception-detection backbones and transfer mechanisms. For PECL [Guo et al., 2023], LCF modulates the projected audio and visual representations independently before the original adapters and cross-modal fusion. For FreeLunch [Yang et al., 2021], LCF is applied before distribution calibration, while the original calibration mechanism, classifier, and evaluation settings are retained.

Table 14: Cross-backbone and transfer-mechanism evaluations on RLT and DOLOS. All metrics are in %. Results with Uni-Cheb are reported as mean±std over five seeds; baseline values are reported as provided in the corresponding experiments.

Dataset	Method	ACC	F1	AUC
	PECL Backbone			
RLT	PECL [Guo et al., 2023]	69.39	71.02	74.24
	PECL + Uni-Cheb	**80.95±5.06**	**81.48±4.36**	**80.73±5.15**
DOLOS	PECL	65.97	70.84	63.53
	PECL + Uni-Cheb	**70.18±4.61**	**72.73±6.80**	**67.83±2.44**
	FreeLunch Transfer Mechanism			
RLT	FreeLunch [Yang et al., 2021]	80.10	74.73	84.97
	FreeLunch + Uni-Cheb	**86.90±4.32**	**81.48±5.15**	**86.54±5.81**
DOLOS	FreeLunch	62.58	66.86	66.28
	FreeLunch + Uni-Cheb	**68.42±4.10**	**72.00±6.16**	**72.47±6.02**

For intent recognition, we integrate LCF into MulT [Tsai et al., 2019] before its original cross-modal attention blocks, retaining the backbone's fusion architecture. The evaluation covers MIntRec and MIntRec2.0, including all four metrics.

Table 15: Evaluation of Uni-Cheb with MulT on intent recognition benchmarks. All metrics are in %. Results with Uni-Cheb are mean±std over five seeds; MulT baseline values are reported as provided in the original comparison.

Dataset	Method	ACC	WF1	WP	R
MIntRec	MulT [Tsai et al., 2019]	72.52	71.80	72.60	67.44
	MulT + Uni-Cheb	**74.11±0.19**	**74.27±0.27**	**75.76±0.23**	**72.62±0.28**
MIntRec2.0	MulT	56.95	54.26	54.49	40.65
	MulT + Uni-Cheb	**59.02±0.22**	**58.47±0.17**	**59.15±0.25**	**53.38±0.25**

We further evaluate Uni-Cheb with a FLatten-based focused linear-attention encoder [Han et al., 2023]. For deception detection, the encoder processes temporal VideoMAE features, with DFT-based LCF inserted before pooling and classification. For intent recognition, the FLatten-based encoder replaces the visual temporal encoder in MVCL-DAF, while the text and audio encoders, alignment, fusion module, and DWT-based LCF remain unchanged. Both the baseline and LCF variants use the same data splits, optimization settings, model-selection rules, and five random seeds.

Table 16: Evaluation of Uni-Cheb with a FLatten-based focused linear-attention encoder. All metrics are in % and reported as mean±std across five independent seeds.

Dataset	Metric	**FLatten-based** [Han et al., 2023]	**+ Uni-Cheb**
RLT	ACC	85.71±6.89	**90.48±4.71**
	F1	83.33±8.93	**90.00±6.54**
	AUC	84.13±6.46	**92.29±6.71**
DOLOS	ACC	67.46±4.48	**71.93±4.03**
	F1	69.39±5.55	**73.68±5.11**
	AUC	68.36±5.48	**73.93±4.41**
MIntRec	ACC	71.46±0.45	**76.18±0.23**
	WF1	70.52±0.36	**75.94±0.16**
	WP	70.59±0.54	**76.16±0.35**
	R	66.67±0.34	**73.78±0.28**
MIntRec2.0	ACC	57.85±0.31	**60.45±0.37**
	WF1	57.91±0.42	**58.95±0.32**
	WP	58.38±0.46	**59.04±0.29**
	R	52.82±0.38	**54.29±0.26**

Across the evaluated backbones and transfer mechanisms, incorporating LCF improves the reported performance metrics. The PECL and MulT results extend the evaluation beyond the original AFFAKT and MVCL-DAF architectures, while FreeLunch provides an alternative to optimal-transport-based transfer. The FLatten-based experiments further demonstrate the compatibility of LCF with focused linear-attention encoders. These findings support the applicability of the shared spectral response module across different feature encoders, fusion architectures, and transfer mechanisms.

I Statistical Stability and Significance Analysis

I.1 Task-Dependent Transform Preferences

To evaluate the stability of the observed "Spectral Dilemma" (i.e., the task-specific preference for DFT or DWT bases), we conducted experiments across five independent random seeds ($\{42, 10, 100, 2024, 3407\}$). Table 17 reports the detailed accuracy and the results of a two-sided paired t-test.

The paired tests indicate consistent task-dependent transform preferences across the evaluated seeds, with DFT favored on RLT and DWT favored on MIntRec and MIntRec2.0.

Table 17: Detailed Accuracy (%) and standard deviations (mean±std across 5 independent random seeds), alongside Paired t-test results. Seed 1 corresponds to our default initialization. The results confirm that the performance advantages of the task-aligned bases (DFT for Deception, DWT for Intent) are consistent, stable, and statistically significant.

Task (Dataset)	Basis Configuration	Seed 1 (42)	Seed 2 (10)	Seed 3 (100)	Seed 4 (2024)	Seed 5 (3407)	Mean ± Std	Two-sided paired t-test (p-value)
Deception (RLT)	**Ours: Uni-Cheb (DFT)**	**93.75**	**93.10**	**93.50**	**93.60**	**93.80**	**93.55 ± 0.28**	**0.0057** ($p < 0.01$)
	Ours: Uni-Cheb (DWT)	91.67	92.20	91.80	92.90	92.10	92.13 ± 0.48	
Intent (MIntRec)	**Ours: Uni-Cheb (DWT)**	**77.08**	**77.02**	**76.78**	**76.90**	**76.86**	**76.93 ± 0.12**	**0.0009** ($p < 0.001$)
	Ours: Uni-Cheb (DFT)	75.85	75.40	75.90	75.96	75.69	75.76 ± 0.22	
Intent (MIntRec 2.0)	**Ours: Uni-Cheb (DWT)**	**60.40**	**60.35**	**60.15**	**60.28**	**60.32**	**60.30 ± 0.09**	**0.0141** ($p < 0.05$)
	Ours: Uni-Cheb (DFT)	60.26	59.63	59.85	59.72	59.58	59.81 ± 0.27	

I.2 Paired Baseline Comparisons Across Five Seeds

To assess the consistency of the improvements over the original task-specific baselines, we rerun the baselines and Uni-Cheb using identical data splits and five random seeds ($\{42, 10, 100, 2024, 3407\}$). We report the mean and standard deviation across runs for all evaluation metrics. The deception-detection comparisons use the visual modality, while the intent-recognition comparisons follow the original multimodal setting.

Table 18: Paired five-seed comparisons between the original deception-detection baseline and Uni-Cheb on the visual modality. Results are reported as mean±std (%). Gains are differences in mean performance, in percentage points.

Dataset	Method	ACC	F1	AUC
RLT	AFFAKT (Baseline)	86.61±0.27	87.52±0.26	87.89±0.23
	Uni-Cheb	**93.81±0.25**	**93.68±0.20**	**96.28±0.18**
	Gain	+7.20	+6.16	+8.39
DOLOS	AFFAKT (Baseline)	67.49±0.28	70.83±0.24	71.99±0.28
	Uni-Cheb	**73.12±0.23**	**75.88±0.27**	**77.79±0.23**
	Gain	+5.63	+5.05	+5.80

Table 19: Paired five-seed comparisons between the original intent-recognition baseline and Uni-Cheb. Results are reported as mean±std (%). Gains are differences in mean performance, in percentage points.

Dataset	Method	ACC	WF1	WP	R
MIntRec	MVCL-DAF (Baseline)	74.72±0.13	74.61±0.14	75.07±0.17	71.94±0.12
	Uni-Cheb	**76.93±0.12**	**76.73±0.11**	**77.64±0.12**	**75.48±0.11**
	Gain	+2.21	+2.12	+2.57	+3.54
MIntRec2.0	MVCL-DAF (Baseline)	57.71±0.19	55.06±0.20	55.78±0.22	42.09±0.28
	Uni-Cheb	**60.32±0.15**	**59.41±0.17**	**60.56±0.15**	**54.40±0.22**
	Gain	+2.61	+4.35	+4.78	+12.31

As shown in Tables 18 and 19, Uni-Cheb consistently improves upon the original task-specific baselines across all 14 dataset–metric comparisons. Two-sided paired t-tests using matched seeds and data splits yield $p < 0.001$ for all comparisons. These results demonstrate consistent performance improvements across independent runs.

J Detailed Experimental Setup

J.1 Deception Detection And Transfer Learning

For the original AFFAKT-based deception detection experiments on RLT and DOLOS, we utilize the Adam optimizer with a learning rate of 1×10^{-5}. The batch size is set to 4 for the training phase and 2 for the testing phase. Our model is trained for 20 epochs using a 5-fold cross-validation protocol. Following the baseline optimization strategy, the loss weighting hyperparameters λ (replacing the original baseline notation α to avoid confusion with our learnable response-coordinate scaling parameter) and η are configured to 0.95 and 0.01, respectively.

For the RLT dataset, the baseline spatial parameter ξ is set to 0.5 for the audio modality and 0.2 for the visual and fused modalities. For the DOLOS dataset, ξ is maintained at 0.2 across all modalities. The parameter ν is set to 0.05 for the visual and fused modalities in RLT, while for the audio modality in RLT and all modalities in DOLOS, ν is 0.1. Accuracy (ACC), F1 score, and area under the curve (AUC) are adopted as the evaluation metrics.

J.2 Intent Recognition

For the original MVCL-DAF-based intent recognition experiments on MIntRec and MIntRec2.0, the models are trained for a maximum of 100 epochs. The training batch size is 16, while the evaluation and testing batch sizes are 8. We adopt the AdamW optimizer with a learning rate decay strategy, utilizing a warmup proportion of 0.1 and a weight decay factor of 0.2. Depending on the specific dataset (MIntRec or MIntRec 2.0), the initial learning rate is set to either 2×10^{-5} or 1×10^{-5}. To prevent overfitting, an early stopping mechanism is implemented with a patience of 10 epochs (monitoring F1 score) or 8 epochs (monitoring Accuracy).

The architectural configuration includes a maximum depth of 5 for the multimodal dynamic fusion layer and a dropout probability of 0.5. We employ the Connectionist Temporal Classification (CTC) method for cross-modal alignment. For contrastive learning, the InfoNCE loss is used with a temperature scaling parameter set to 0.5.

J.3 Additional Experimental Protocols

The additional response-parameterization controls and cross-backbone evaluations follow the same task-specific training hyperparameters as the corresponding experiments described above. Specifically, experiments on RLT and DOLOS use the deception-detection settings in Section J.1, while experiments on MIntRec and MIntRec2.0 follow the intent-recognition settings in Section J.2. The optimizer, learning rate, batch sizes, and training schedule are kept consistent within each task. For cross-backbone evaluations, the original architecture-specific components and loss functions are retained where applicable. Implementation details and experimental results are provided in Sections H.3 and H.4.

K Scope of Applicability and Limitations

Uni-Cheb employs a transform-compatible spectral modulation framework with the following considerations for applicability and future extensions:

Dependence on Spectral Informativeness. The effectiveness of the Learnable Chebyshev Filter (LCF) depends on whether the selected spectral representation contains task-relevant information. When discriminative cues are predominantly static and are not well represented by the selected transform, spectral modulation may provide limited benefits. Uni-Cheb is therefore particularly suited to settings in which frequency-domain patterns contribute to prediction.

Dataset-Specific Effects. Improvements on RLT and DOLOS span two deception-detection settings, but the attention visualizations do not establish a unique physiological interpretation or exclude dataset-specific correlations.

Prior-Guided Transform Selection. Currently, Uni-Cheb selects spectral transforms according to task-specific signal characteristics (e.g., DFT for global temporal patterns and DWT for localized time-frequency patterns). This selection is external to LCF, whose response-learning formulation remains shared across the evaluated transforms. Exploring data-driven mechanisms to jointly select or combine multiple spectral transforms presents a direction for future work, particularly for heterogeneous signals containing both global and local temporal structure.

Future Work: Operator-Level Extensions. Beyond its current application to feature representations, a promising direction is to integrate Chebyshev-parameterized response learning into the internal kernel mappings of linear attention and the basis functions of KAN, extending LCF from feature preprocessing to operator-level adaptation.

L Broader Impacts

As a lightweight spectral modulation module, Uni-Cheb can be integrated into existing multimodal architectures with a small parameter overhead, although it introduces additional computation and memory usage. However, as a general-purpose spectral operator, its downstream applications, particularly in sensitive areas like deception detection or emotion recognition, must be deployed responsibly. We urge practitioners to be mindful of potential algorithmic biases and privacy concerns when applying such models to real-world human-centric systems (negative impact/mitigation).

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 main claims are supported by the experimental results in Section 4, the ablation studies in Section 5, and the approximation-theoretic analysis in Section E. The additional controlled comparisons and cross-backbone evaluations further support the effectiveness and applicability of LCF.

   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: We discuss the dependence on task-relevant spectral information and the current use of prior-guided transform selection in Section K, together with directions for extending the framework to more heterogeneous spectral representations.

   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: The assumptions, classical Chebyshev interpolation results, and associated error bounds are presented in Section E. The analysis provides the approximation-theoretic motivation for the structured response parameterization used by LCF.

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: The model formulations and implementation details are described in Sections A and 3. Dataset information and training configurations are provided in Sections C and J, including the additional experimental protocols in Section J.3.

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: We provide an anonymized GitHub repository link containing our code in Appendix M. All datasets used are publicly available academic benchmarks, with details provided in Appendix C.

   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: The experimental settings, including optimizers, learning rates, batch sizes, training schedules, and model-selection procedures, are described in Section J. Additional controlled comparisons follow the task-specific training protocols in Section J.3.

   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: We report mean and standard deviation across five independent runs for the controlled response-parameterization and cross-backbone evaluations in Sections H.3 and H.4. Two-sided paired t-tests for task-dependent spectral transform comparisons are reported in Table 17.

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: We explicitly specify the hardware used (NVIDIA RTX 4090 and V100 GPUs) in Section 4.2. A detailed system-level analysis covering total parameters, added parameters, inference latency, and VRAM footprints is provided in Appendix E.5 (Table 7).

   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: Our research strictly utilizes standard, publicly available academic datasets and adheres to the NeurIPS Code of Ethics.

   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: [Yes]

    Justification: We discuss the applicability of lightweight spectral modulation and potential risks related to privacy and algorithmic bias in human-centric applications in Section L.

    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: Our work proposes a lightweight spectral operator (Uni-Cheb) and does not release high-risk generative models (e.g., LLMs or image generators) or scraped datasets that require specific safeguards.

    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: [Yes]

Justification: We evaluate our method on standard, publicly available academic benchmarks (e.g., RLT, DOLOS, MIntRec, DFEW). We properly cite their original papers and creators in Appendix C.

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: [N/A]

Justification: This paper does not introduce or release any new datasets or novel collected assets; all evaluations are performed on existing public benchmarks.

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: The research does not involve any crowdsourcing experiments or new human subject data collection.

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: The study only utilizes pre-existing, publicly available datasets, and no new human subjects research was conducted that would require IRB approval.

    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: The core methodological development of this research (the Uni-Cheb operator) does not involve Large Language Models as an important or original component.

    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.