# MANGO:Multi-Angle Neural Gated Operators for Chirp-Perturbed PDEs

**Yunlong Zhu**[1,2] **Zheng Wang**[1,2] **Eun-Hu Kim**[1] **Zunwei Fu**[1,2*]
[1]School of Mathematics and Statistics, Linyi University
[2]College of Information Communication Technology, The University of Suwon

## Abstract

We introduce MANGO (Multi-Angle Neural Gated Operator), a spectral neural operator for chirp-perturbed PDEs whose solutions exhibit a local spectral content that rotates continuously with position. MANGO maintains multiple fractional Fourier transform (FRFT) branches whose angles are learnable parameters, optimized jointly with the spectral weights and a position-dependent softmax gate that combines branches at every spatial location. The architecture is matched to the structure of the problem class: the FRFT at angle $\alpha_0$ diagonalizes the chirp-perturbed PDE whose intrinsic angle is $\alpha_0$, and MANGO's learnable angles let the architecture discover $\alpha_0$ from data when it is unknown or varies across the domain—neither of which classical FRFT methods accommodate. We further introduce a Mihlin–Hörmander regularizer on the spectral weights of each non-Fourier branch, encouraging each branch to act as a bounded operator on $L^p$ for $1 < p < \infty$. We construct four chirp-perturbed PDE benchmarks with closed-form ground truth, positioned as identifiability tests. On these benchmarks, MANGO recovers the intrinsic spectral structure of each PDE from supervised solution data alone and achieves state-of-the-art performance against six neural-operator baselines.

## 1 Introduction

A wide class of variable-coefficient PDEs share a distinctive structural feature: their solutions exhibit a local spectral content that rotates continuously with position. The paraxial wave equation in graded-index optical media, the quantum harmonic oscillator, convection–diffusion in linear shear flows, and the Fokker–Planck equation for the Ornstein–Uhlenbeck process are all instances where quadratic-phase propagation, harmonic-oscillator dynamics, or Fokker–Planck/Ornstein– Uhlenbeck structure arise naturally [20, 22, 24]. Each reduces, after a short calculation, to a chirp-perturbed form in which a constant-coefficient base operator is conjugated by a Gaussian chirp factor, producing a variable-coefficient operator whose linear-in-$x$ convection is paired with a quadratic potential, and whose solutions have a local oscillation rate that varies continuously across the domain [4, 22]. Numerical methods that perform well on this class need a representation whose spectral basis is not the same at every spatial point.

Spectral neural operators [14, 16] do not have this property. The integral kernel of a spectral neural operator is parameterized by a learned linear transformation of a fixed, truncated basis: at every spatial point the same basis is applied, and the learned weights act on a small number of modes. This is efficient when the solution admits a compact representation in the chosen basis— translation-invariant problems, periodic boundaries—but on chirp- perturbed PDEs the local frequency content at one point is not the local frequency content at another, and a single basis must compensate by retaining many high-order modes [16, 23]. Recent extensions of the family make the architecture adaptive in adjacent ways: geometry-deformed variants accommodate non-rectangular domains while preserving a single

*Corresponding author: zwfu@suwon.ac.kr

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

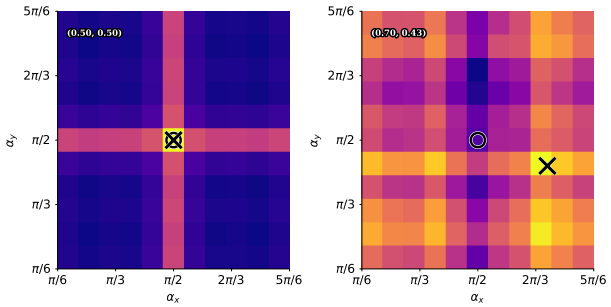


Figure 1: Optimal spectral basis for chirp-perturbed PDE solution
The black ∘ marks the standard Fourier basis $(\pi/2, \pi/2)$. Left (vanilla PDE): $\times$ and ∘ coincide at $(0.50\pi,\ 0.50\pi)$ — the Fourier basis is optimal, as expected. Right (chirp-perturbed PDE): the optimum shifts to $(0.70\pi,\ 0.43\pi)$, far from the Fourier basis, with substantial energy also concentrated at neighboring non-Fourier angles.

global basis on the deformed coordinates [17, 18]; modulation-based methods adapt within a fixed basis through pointwise or channel-wise coupling functions [6, 7, 12, 13]; attention-based operator learners bypass the fixed-basis constraint entirely, but at the cost of the explicit spectral structure that gives Fourier operators their discretization invariance and resolution generalization [3, 9, 30]. None of these changes the identity of the basis as a function of spatial position. Figure 1 makes this mismatch concrete in a representative PDE perturbed by a chirp. For a solution of the standard 2D Laplace equation, the energy concentration over fractional Fourier angles $(\alpha_x, \alpha_y) \in [\pi/6, 5\pi/6]^2$ is sharply peaked at $(\pi/2, \pi/2)$: the Fourier basis is optimal, as expected. For a solution of the chirp-perturbed 2D Laplace equation, the optimum shifts to $(0.70\pi,\ 0.43\pi)$, far from the Fourier basis, with substantial energy also concentrated at neighboring non-Fourier angles. No single Fourier basis is optimal for chirp-perturbed PDEs, and the optimum is not constant across the family.

A specific tool from harmonic analysis is matched to this problem class. The fractional Fourier transform (FRFT) is a one-parameter family of unitary transforms indexed by an angle $\alpha$ in the time–frequency plane [20, 22], computable in continuous/factorized form via a chirp–FFT–chirp decomposition [21] or in exact unitary discrete form via a cached Hermite–Gaussian eigendecomposition. Originally introduced by Namias [20], the transform was put on a rigorous $L^p$ foundation by Chen et al. [4] and extended to higher dimensions by Fu et al. [5]. The FRFT at angle $\alpha_0$ *diagonalizes* the chirp-perturbed PDE whose intrinsic angle is $\alpha_0$: in the basis of FRFT-$\alpha_0$, the variable-coefficient operator acts as an ordinary Fourier multiplier, and the $L^p$ multiplier theory of Chen et al. [4] transfers directly to the operator's stability analysis. This is the architectural buy that other rotation-of-time–frequency parameterizations do not provide. The diagonalization fails at any other angle, and the failure becomes severe as the angle approaches 0 or $\pi$. The FRFT thus solves the chirp-perturbed problem in principle: if the angle is known, and if that angle is the same everywhere in the domain. Both assumptions fail in practice: the angle is a property of the PDE coefficients, which are themselves what one is trying to learn from observation; and physical chirp problems generically have intrinsic angles that vary across the spatial domain, including paraxial propagation in inhomogeneous media and variable-coefficient diffusion in graded materials [22, 24].

We embed this structure into a neural operator architecture in which the FRFT angles are themselves trained from data. The Multi-Angle Neural Gated Operator (MANGO) maintains $K$ spectral branches whose angles $\{\alpha_1, \ldots, \alpha_K\}$ are learnable parameters optimized jointly with the spectral weights and the position-dependent softmax gate that combines branch output at every spatial location. Our contributions are as follows:

**MANGO architecture** We introduce a multi-branch mixture of FRFT spectral convolutions with learnable angles, combined by a position-dependent softmax gate, with axis-separable tensor-product extension to two dimensions. The learnable angles are a deliberate departure from a fixed-angle-grid mixture: they let the architecture realize a position-dependent effective basis through both gate-weighted branch selection and gradient-driven angle adaptation, simultaneously (Section 3).

**Branch-wise stability regularizer.** We introduce a regularizer on the spectral weights of each branch, motivated by Chen et al.'s extension of the Mihlin–Hörmander multiplier theorem to the FRFT setting [4]. The regularizer encourages each FRFT branch to act as a bounded operator on $L^p$ for $1 < p < \infty$. It does not bound the composed nonlinear network or its generalization error, but provides a tractable, theory-informed constraint on the linear spectral layers. The regularizer is included in the training objective at all non-Fourier angles.

**Identifiability benchmarks.** We define four chirp-perturbed PDE benchmarks (Laplace, Heat, Poisson, Cauchy–Riemann) with closed-form ground truth, positioned explicitly as *identifiability tests* that make pointwise chirp-rate recovery checkable. To evaluate generalizability beyond analytical chirps, we further test MANGO on standard unperturbed Burgers (where MANGO provably and empirically reduces to FNO) and on five real-world physical systems from RealPDEBench [11] (four incompressible Navier–Stokes flows and swirl-stabilized combustion), where MANGO achieves the lowest relative $L^2$ error on all five systems.

**Spectral structure recovery.** The trained MANGO model recovers the intrinsic spectral structure of each chirp-perturbed PDE from supervised solution data alone. Recovery proceeds through gate concentration on the branch whose initial angle is closest to the PDE's intrinsic angle, through angle adaptation in which a branch's learnable $\alpha$ drifts toward the intrinsic angle, or through a combination of the two.

## 2 Related Work

### 2.1 Spectral Neural Operators and Their Limitations

The Fourier Neural Operator (FNO) [16] parameterizes the integral kernel of a neural operator via a learned linear transformation of a fixed truncated Fourier basis. This construction makes the layer grant discretization invariance and zero-shot super-resolution to the resulting operator. FNO has been extended along several directions: factorized Fourier representations that decompose the spectral convolution along spatial axes for higher dimensions [27]; wavelet- and multiwavelet-based parameterizations that trade global frequency localization for spatial–frequency localization [8, 28]; spectral bases adapted to non-rectangular geometry through Laplace-transform decompositions [2]; and hybrid graph–Fourier architectures for unstructured meshes and arbitrary geometries [18]. The general theory of integral neural operators that subsumes these instantiations has also been established [14].

### 2.2 Neural Operators for Perturbed and Variable-Coefficient PDEs

**Generalized geometries and learned deformations.** For PDEs on irregular domains or with deforming boundaries, GINO [18] combines a graph neural operator that encodes irregular point clouds with an FNO operating on a regularized latent grid. Geo-FNO [17] learns a smooth deformation that warps an arbitrary domain to a canonical rectangular grid where an FNO is applied, transferring the learned solution back through the deformation's inverse. These approaches preserve a single global Fourier basis on the latent representation; the variability they accommodate is geometric rather than spectral, and the spectral content of the solution remains uniformly represented across the deformed domain.

**PDEs with explicit perturbation terms.** Methods for PDEs whose coefficients deviate from a tractable base case by a structured term have been classical in numerical analysis for decades. Recent learning-based work has considered PDEs whose coefficient fields are themselves drawn from a distribution: multipole graph kernel networks [15] capture long-range interactions in such settings at linear complexity, and DeepONet [19] generalizes across coefficient distributions for a fixed PDE family. Foundation-model approaches such as PDEformer-1 [31] and Unisolver [33] extend this further by encoding the PDE's symbolic form as input to the network, accommodating coefficient variability at the level of equation structure rather than parameter distribution. These approaches address generality across PDE families rather than efficiency on a specific structured family; they are complementary to the architectural specialization MANGO targets.

# 3 Methodology

## 3.1 Background and Notation

**Neural operator framework.** A neural operator approximates a mapping $\mathcal{G}^\dagger : \mathcal{A} \to \mathcal{U}$ between two function spaces of input and output fields, $\mathcal{A} = \{a : \Omega \to \mathbb{R}^{d_a}\}$ and $\mathcal{U} = \{u : \Omega \to \mathbb{R}^{d_u}\}$, defined on a domain $\Omega \subset \mathbb{R}^d$ [14]. We realize the surrogate $\mathcal{G}_\theta^\dagger = V \circ M_L \circ \cdots \circ M_1 \circ P$ as a composition of a pointwise lifting layer $P$, $L$ latent-space blocks, and a pointwise projection layer $V$. Each block is a pre-norm residual unit, the standard architectural choice in modern spectral neural operators following residual-network design [10]:

$$\mathbf{v}_l = \mathrm{FFN}\big(\mathrm{LN}(\mathbf{v}_{l-1}^{\mathrm{mid}})\big) + \mathbf{v}_{l-1}^{\mathrm{mid}}, \quad \mathbf{v}_{l-1}^{\mathrm{mid}} = F^{\mathrm{mixer}}\big(\mathrm{LN}(\mathbf{v}_{l-1})\big) + \mathbf{v}_{l-1}, \tag{1}$$

where LN denotes LayerNorm [1] and the spatial mixer $F^{\mathrm{mixer}}$ identifies the operator family. Spectral neural operators instantiate the mixer as a transform–multiply–inverse-transform sandwich on a chosen basis.

**The fractional Fourier transform.** For $\alpha \in (0, \pi)$, the fractional Fourier transform (FRFT) of order $\alpha$ is the unitary operator $\mathcal{F}_\alpha$ acting on $L^2(\mathbb{R})$ by

$$\mathcal{F}_\alpha[f](u) = A_\alpha \int_{-\infty}^{\infty} f(x)\, e^{i\pi(x^2+u^2)\cot\alpha - 2\pi i\, xu/\sin\alpha}\, dx, \qquad A_\alpha = \sqrt{1 - i\cot\alpha}, \tag{2}$$

with the principal branch of the square root chosen so that $A_{\pi/2} = 1$ (recovering the standard Fourier transform) and $A_\alpha \to 1$ as $\alpha \to \pi/2$ from either side. The endpoints $\alpha \in \{0, \pi\}$ are handled by extension as limiting cases of Eq. (2); $\mathcal{F}_\alpha$ interpolates between the identity at $\alpha \to 0$ and the standard Fourier transform $\mathcal{F}$ at $\alpha = \pi/2$. The transform admits the *chirp–FFT–chirp factorization*

$$\mathcal{F}_\alpha \;=\; A_\alpha \left(M_\alpha \circ S_{1/\sin\alpha} \circ \mathcal{F} \circ M_\alpha\right), \qquad M_\alpha[f](x) = e^{i\pi x^2 \cot\alpha}\, f(x), \tag{3}$$

where $S_\lambda f(u) = f(\lambda u)$ denotes rescaling and the scalar $A_\alpha$ multiplies the operator output to ensure unitarity [4].[2] The inverse takes the symmetric form $\mathcal{F}_\alpha^{-1} = \overline{A_\alpha}\left(M_\alpha^{-1} \circ \mathcal{F}^{-1} \circ S_{\sin\alpha} \circ M_\alpha^{-1}\right)$. While Eq. (3) motivates the continuous diagonalization theory below, in our discrete implementation we use an eigendecomposition-based discrete FRFT (Candan–Kutay–Ozaktas family; see Appendix D.2) rather than a chirp–FFT factorization: we construct the $\alpha$-independent orthonormal discrete Hermite–Gaussian eigenbasis of the unitary DFT (cached once per resolution in `complex128`), so $\alpha$ enters *only* through the eigenvalue phases $e^{-i\alpha n}$. This guarantees exact unitarity at every $\alpha$ without resampling and yields analytic, bounded angle gradients $\partial/\partial\alpha$ with no clamping or straight-through estimators, at an $O(N^2)$ cached basis-projection cost that accounts for only $12\%$ of epoch runtime.

**Chirp-perturbed PDEs.** We define a linear differential operator $L$ on $\mathbb{R}$ to be *chirp-perturbed at angle* $\alpha_0$ if it can be written

$$L_{\alpha_0} \;=\; M_{\alpha_0} \circ L_0 \circ M_{\alpha_0}^{-1}, \tag{4}$$

where $L_0$ is a constant-coefficient operator and $M_{\alpha_0}$ is the chirp multiplier of Eq. (3). Expanding Eq. (4) produces a variable-coefficient operator whose coefficient structure pairs linear-in-$x$ convection with a quadratic-in-$x$ potential.

**Proposition 3.1** (FRFT diagonalizes the matched chirp operator)**.** *Let $L_{\alpha_0}$ be chirp-perturbed at angle $\alpha_0 \in (0, \pi)$ as in Eq.* (4)*, with $L_0$ a constant-coefficient differential operator with Fourier symbol $\widehat{L_0}(\xi)$. Let $\mathcal{M}_{\widehat{L_0}}$ denote the operator of multiplication by $\widehat{L_0}(\xi)$. Then for $f \in \mathcal{S}(\mathbb{R})$,*

$$\mathcal{F}_{\alpha_0}\, L_{\alpha_0}\, \mathcal{F}_{\alpha_0}^{-1} \;=\; \mathcal{F} \circ \mathcal{M}_{\widehat{L_0}} \circ \mathcal{F}^{-1}, \tag{5}$$

*i.e. in the FRFT-$\alpha_0$ basis, $L_{\alpha_0}$ acts as an ordinary Fourier multiplier with symbol $\widehat{L_0}(\xi)$.*

The proof is a direct calculation using Eq. (3) and (4); the full derivation is given in Appendix A.

Equation (5) fails at any angle $\alpha \neq \alpha_0$, with a residual that grows as the mismatch $|\alpha - \alpha_0|$ increases and that becomes severe near $\alpha = 0$ (where $\cot\alpha$ diverges) or $\alpha = \pi$ (where the cosecant rescaling diverges).

[2] The scalar $A_\alpha$ absorbs the Jacobian of the rescaling $S_{1/\sin\alpha}$; the product $|A_\alpha|^2 = 1/|\sin\alpha|$ matches the rescaling Jacobian and is what makes $\mathcal{F}_\alpha$ $L^2$-unitary for all $\alpha \in (0, \pi)$.

The qualitative diagonalization of Proposition 3.1 has a quantitative companion that controls the architectural budget: at the matched angle the truncated FRFT basis achieves exponential coefficient decay, while the standard Fourier basis is reduced to algebraic decay whenever $\alpha_0 \neq \pi/2$.

**Theorem 3.1** (Spectral decay separation)**.** *Let $\mathcal{L}_{\alpha_0}$ be chirp-perturbed at intrinsic angle $\alpha_0 \in (0, \pi)$ as in Eq.* (4)*, and let $u_{\alpha_0}(\cdot, t)$ denote the solution at depth $t > 0$ with initial data $f \in H^s(\mathbb{R})$, $s > 1/2$. Let $\hat{u}_k^{(\alpha)}$ denote the $k$-th coefficient of $u_{\alpha_0}(\cdot, t)$ in the truncated FRFT-$\alpha$ basis. There exist constants $C, c > 0$, depending on $s$, $\alpha_0$, $t$ but not on $k$, such that for all sufficiently large $|k|$:*

*(i)* Matched basis *($\alpha = \alpha_0$):* $\left|\hat{u}_k^{(\alpha_0)}\right| \leq C\,|k|^{-s}\, e^{-2\pi t\,|\csc\alpha_0|\,|k|}$.

*(ii)* Off-target FRFT basis *($\alpha \neq \alpha_0$):* $\left|\hat{u}_k^{(\alpha)}\right| \leq C\,|k|^{-s+\delta(\Delta\alpha)}\, e^{-c(\Delta\alpha)\,t\,|k|}$*, where $\Delta\alpha := \alpha - \alpha_0$, the rate-degradation function $\delta$ is continuous with $\delta(0) = 0$, and $c(\Delta\alpha) \to 2\pi|\csc\alpha_0|$ as $\Delta\alpha \to 0$.*

*(iii)* Standard Fourier basis *($\alpha = \pi/2$):* $\left|\hat{u}_k^{(F)}\right| \leq C\,|k|^{-s+\delta(\alpha_0-\pi/2)}$*, with $\delta(\alpha_0 - \pi/2) > 0$ strictly whenever $\alpha_0 \neq \pi/2$.*

**Corollary 3.1** (Mode-budget separation)**.** *For target relative error $\varepsilon > 0$, the matched FRFT basis requires $k_{\mathrm{FRFT}} = O(\log(1/\varepsilon))$ modes, whereas the standard Fourier basis requires $k_F = O(\varepsilon^{-1/(s-\delta)})$ modes: exponential versus algebraic in the target accuracy.*

Proofs of Theorem 3.1 and Corollary 3.1 are given in Appendix A.4. The continuity of $\delta$ at 0 in part (ii) is what allows the gradient on $\alpha_k$ to be informative even before exact alignment: a branch whose angle is close to $\alpha_0$ already enjoys near-exponential decay, and the loss landscape rewards angle drift toward $\alpha_0$ continuously rather than only at the matched point. Theorem 3.1 establishes the *architectural asset* MANGO's branches are designed to access; whether training actually reaches a configuration that accesses it is a separate empirical question addressed in Section 4, with a heuristic critical-point account given in Appendix A.4.7.

### 3.2 The MANGO Architecture

Figure 2 illustrates the architecture. Each of the $L$ blocks applies the mixer $F_\ell^{\mathrm{mixer}}$ defined below alongside a $1\times1$ residual convolution within the pre-norm residual structure of Eq. (1). The 2D extension applies the 1D mixer sequentially along the two spatial axes, with independent learnable angles and gates per axis, totalling $2K$ angle parameters per layer.

**$K$-branch FRFT mixer.** The MANGO mixer replaces the single-basis spectral convolution of FNO with a $K$-branch mixture over the fractional Fourier family. For an input feature $\mathbf{v} \in \mathbb{R}^{N\times d_v}$ discretized at $N$ spatial points, the mixer in layer $\ell$ computes

$$F_\ell^{\mathrm{mixer}}(\mathbf{v})(x) \;=\; \mathrm{Re}\left\{\sum_{k=1}^{K} g_k^{(\ell)}(x)\left[\mathcal{F}^{-1}_{\alpha_k^{(\ell)}}\, W_k^{(\ell)}\, \mathcal{F}_{\alpha_k^{(\ell)}}\,\mathbf{v}\right](x)\right\}, \tag{6}$$

where $\alpha_1^{(\ell)}, \ldots, \alpha_K^{(\ell)} \in (0, \pi)$ are *learnable* branch angles, $W_k^{(\ell)} \in \mathbb{C}^{M\times d_v\times d_v}$ is the spectral weight tensor of branch $k$ acting on the top-$M$ FRFT modes, and $g_k^{(\ell)}(x)$ is the position-dependent gate weight defined below. The real part is taken before the result enters the FFN of Eq. (1), so latent fields remain real-valued. With the cached Hermite–Gaussian eigenbasis, each branch reduces to two dense unitary projections ($O(N^2 d_v)$) and a truncated mode-wise multiplication ($O(M d_v^2)$), so the full mixer scales as $O(K(N^2 d_v + M d_v^2))$ per layer (or $O(K(N \log N \cdot d_v + M d_v^2))$ under a chirp–FFT backend). The branch angles are initialized on a uniform grid in $[\pi/6, 5\pi/6]$ and optimized jointly with the spectral weights and the gate parameters.

**Discrete branches versus a continuous angle field.** A natural alternative is to parameterize a single FRFT angle $\alpha(\mathbf{x})$ as a continuous function of position via an MLP head. We do not adopt this design because the FRFT diagonalization of Proposition 3.1 and the $L^p$ multiplier theory of Theorem A.4 both require $\alpha$ to be constant across the domain on which the FRFT acts: a position-varying $\alpha(\mathbf{x})$ has no corresponding diagonalization theorem, and the multiplier-norm guarantees that motivate the regularizer of Section 3.3 would not apply branch-wise. The $K$-branch construction makes the angular discretization explicit and trainable, recovering position-dependence through the gate's spatial allocation rather than through a per-pixel angle field.

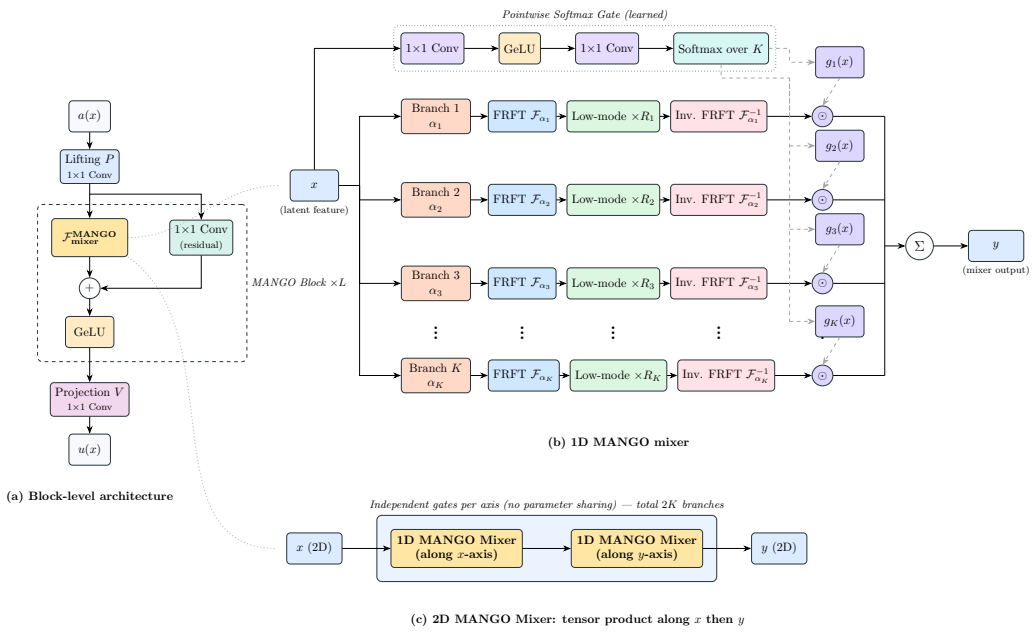


Figure 2: **MANGO architecture. (a)** Overall model: a lifting layer $P$ ($1\times1$ convolution), $L$ stacked MANGO blocks, and a projection layer $V$. **(b)** 1D MANGO mixer: $K$ FRFT branches with learnable angles $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, each applying a forward FRFT, low-mode spectral multiplication by learnable weights $R_k$, and an inverse FRFT. **(c)** 2D MANGO mixer: tensor-product application along $x$ then $y$, with independent gates per axis.

**Position-dependent softmax gate.** The gate $g^{(\ell)} : \Omega \to \Delta^{K-1}$ produces, at each spatial location, a probability distribution over the $K$ branches. We parameterize it as a small pointwise MLP $h^{(\ell)}$ with parameters $\phi^{(\ell)}$, followed by softmax normalization:

$$g_k^{(\ell)}(x) \;=\; \frac{\exp\big(h_k^{(\ell)}(\mathbf{v}(x);\phi^{(\ell)})\big)}{\sum_{j=1}^{K}\exp\big(h_j^{(\ell)}(\mathbf{v}(x);\phi^{(\ell)})\big)}, \qquad h^{(\ell)} : \mathbb{R}^{d_v} \to \mathbb{R}^K, \tag{7}$$

where $h^{(\ell)}$ is a two-layer MLP shared across spatial locations. The gate produces position-dependent mixture weights because $\mathbf{v}(x)$ varies with $x$. Together with the learnable branch angles, the architecture realizes a position-dependent effective basis in two simultaneous senses: which branch dominates varies across the domain, and which angle each branch occupies is itself optimized. While this spatial Mixture-of-Experts gating scaffolding is standard, our contribution lies in what the experts are: learnable unitary FRFT operators matched to chirp-perturbed PDEs via Proposition 3.1 and controlled by the multiplier regularizer below, turning expert selection into physical chirp-rate identification.

### 3.3 Mihlin–Hörmander Regularization

**Multiplier theorem for the FRFT.** The classical Mihlin–Hörmander multiplier theorem has been extended to the FRFT setting [4]: if a multiplier $m_\alpha \in C^1(\mathbb{R}\setminus\{0\})$ associated to the FRFT-$\alpha$ transform satisfies the supremum-norm bounds

$$\sup_{\xi\in\mathbb{R}} |m_\alpha(\xi)| \;\le\; B, \qquad \sup_{\xi\neq0} |\xi\, m'_\alpha(\xi)| \;\le\; B, \tag{8}$$

for some constant $B > 0$, then the operator $T_\alpha f = \mathcal{F}_\alpha^{-1}(m_\alpha\, \mathcal{F}_\alpha f)$ is bounded on $L^p(\mathbb{R})$ for $1 < p < \infty$, with operator norm controlled by $B$ and a $p$-dependent constant. The full statement, including the precise hypotheses on $m_\alpha$ near the origin, is recorded in Appendix A. The transferability of the classical Mihlin condition to the FRFT setting is what makes the regularizer below a meaningful constraint on the learned spectral weights.

Table 1: Comparison of neural operators on the four chirp-perturbed PDE benchmarks.

| Method | Laplace | Heat | Poisson | Cauchy-Riemann |
|---|---|---|---|---|
| FNO [16] (ICLR 2021) | $6.9\text{e}{-2}_{\pm 4.2\text{e}-3}$ | $6.3\text{e}{-2}_{\pm 8.5\text{e}-4}$ | $2.2\text{e}{-2}_{\pm 4.1\text{e}-4}$ | $6.3\text{e}{-2}_{\pm 9.2\text{e}-4}$ |
| LSM [29] (ICML 2023) | $6.6\text{e}{-2}_{\pm 1.2\text{e}-3}$ | $1.7\text{e}{-2}_{\pm 3.8\text{e}-4}$ | $1.9\text{e}{-2}_{\pm 3.5\text{e}-4}$ | $4.7\text{e}{-2}_{\pm 7.1\text{e}-4}$ |
| LaMO [26] (ICML 2025) | $4.1\text{e}{-2}_{\pm 8.6\text{e}-4}$ | $7.5\text{e}{-3}_{\pm 1.9\text{e}-4}$ | $1.7\text{e}{-2}_{\pm 3.1\text{e}-4}$ | $3.0\text{e}{-2}_{\pm 5.0\text{e}-4}$ |
| MINO [25] (TMLR 2025) | $6.8\text{e}{-2}_{\pm 9.4\text{e}-4}$ | $2.5\text{e}{-2}_{\pm 5.0\text{e}-4}$ | $2.0\text{e}{-2}_{\pm 3.9\text{e}-4}$ | $5.3\text{e}{-2}_{\pm 8.2\text{e}-4}$ |
| TFNO [13] (TMLR 2024) | $5.2\text{e}{-2}_{\pm 8.1\text{e}-4}$ | $1.2\text{e}{-2}_{\pm 2.8\text{e}-4}$ | $1.8\text{e}{-2}_{\pm 3.4\text{e}-4}$ | $3.8\text{e}{-2}_{\pm 6.0\text{e}-4}$ |
| HPM [32] (ICML 2025) | $3.2\text{e}{-2}_{\pm 6.5\text{e}-4}$ | $5.6\text{e}{-3}_{\pm 1.5\text{e}-4}$ | $1.6\text{e}{-2}_{\pm 2.8\text{e}-4}$ | $2.5\text{e}{-2}_{\pm 4.2\text{e}-4}$ |
| **MANGO (ours)** | $\mathbf{1.8\text{e}{-2}}_{\pm 1.7\text{e}-4}$ | $\mathbf{2.4\text{e}{-3}}_{\pm 8.0\text{e}-5}$ | $\mathbf{1.5\text{e}{-2}}_{\pm 1.4\text{e}-4}$ | $\mathbf{2.0\text{e}{-2}}_{\pm 1.9\text{e}-4}$ |

**Regularizer formulation.** We do not enforce the supremum-norm condition Eq. (8) directly. Supremum-norm constraints yield non-smooth penalties that are difficult to optimize under standard SGD; we therefore adopt an $L^2$ relaxation that tracks the same quantities (multiplier magnitude and weighted spectral derivative) but integrates rather than maximizes them across the truncated FRFT mode grid $\{\xi_m\}_{|m|\le M}$. For each non-Fourier branch, we add the resulting penalty to the training objective:

$$\mathcal{R}_{\mathrm{MH}}\big(\{W_k^{(\ell)}\}\big) \;=\; \sum_{\ell=1}^{L}\sum_{k=1}^{K} \mathbf{1}\big[\alpha_k^{(\ell)} \neq \pi/2\big] \cdot \frac{1}{M} \sum_{|m|\le M} \Big( \big\|W_k^{(\ell)}(\xi_m)\big\|_{\mathrm{F}}^2 + \big\|\xi_m \cdot \Delta_\xi W_k^{(\ell)}(\xi_m)\big\|_{\mathrm{F}}^2 \Big), \tag{9}$$

where $\Delta_\xi$ denotes the discrete spectral derivative across adjacent modes and $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm. Equation (9) is an $L^2$ relaxation of Eq. (8): the penalty is finite whenever the discrete sequences $\{W_k^{(\ell)}(\xi_m)\}$ and $\{\xi_m \Delta_\xi W_k^{(\ell)}(\xi_m)\}$ are bounded in $\ell^2$, which is implied by but does not imply the supremum-norm condition. The relaxation trades the exact $L^p$-boundedness guarantee of the original theorem for a smooth penalty compatible with stochastic gradient training. The total training loss is $\mathcal{L} = \mathcal{L}_{\text{rel-}L^2} + \beta\,\mathcal{R}_{\mathrm{MH}}$, with the regularizer weight $\beta$ tuned per benchmark on a logarithmic grid in $[10^{-5}, 10^{-2}]$ via held-out validation (Section 4.1). A six-decade sensitivity sweep (Appendix D.4) shows a two-decade plateau $\beta \in [10^{-5}, 10^{-3}]$ within $\sim 8\%$ of the optimum where every nonzero $\beta$ beats $\beta = 0$, making $\beta = 10^{-4}$ a reliable default without fine tuning.

# 4 Experiments

## 4.1 Experiment Setup

**Baselines.** We compare MANGO against six neural PDE solvers: the Fourier Neural Operator (FNO) [16], the Tensorized FNO (TFNO) [13], the Latent Spectral Model (LSM) [29], the Holistic Physics Solver (HPM) [32], the Mesh-Informed Neural Operator (MINO) [25], and the Latent Mamba Operator (LaMO) [26]. Every baseline is reproduced from its authors' official repository at a pinned commit (full provenance in Appendix D.3), reusing model code verbatim and tuning hyperparameters (including width/layers and method-specific spectral settings, with the authors' recommended configurations included as grid points) per benchmark on held-out validation splits. To isolate FRFT rotation from multi-branch capacity, we also evaluate a parameter-identical *Mixture-of-FNOs* control ($K = 8$, angles frozen at $\pi/2$).

**Benchmarks and Metric.** We evaluate on three suites: (i) four synthetic chirp-perturbed PDE benchmarks with closed-form ground truth—Laplace, Heat, Poisson, and Cauchy–Riemann (Appendix B)—designed as identifiability tests; (ii) standard unperturbed 1D Burgers to verify non-chirp reduction; and (iii) the five real-world physical systems of RealPDEBench [11] under its official real-world-finetuning protocol (3 seeds). Chirp-benchmark models are trained on NVIDIA A6000 GPUs and RealPDEBench models on A100 GPUs. We report relative $L^2$ error as the primary metric; further settings are in Appendix C.1.

## 4.2 Main Results

Table 1 reports the relative $L^2$ errors of MANGO and the six baselines across the four chirp-perturbed PDE benchmarks. Several baseline-level observations are worth noting. FNO is uniformly the

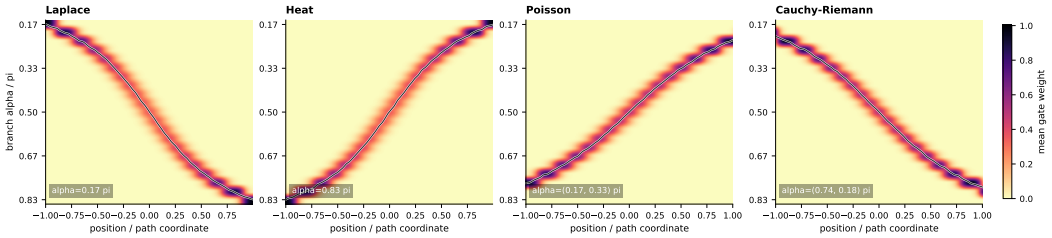


Figure 3: Gate concentration across the four chirp benchmarks. For each benchmark, the heatmap shows the test-set mean gate weight as a function of spatial position (x-axis) and branch angle $\alpha/\pi$ (y-axis).

weakest method, by an order of magnitude on Chirp-Laplace—direct empirical confirmation that a single fixed Fourier basis is poorly matched to chirp-perturbed problems, as the Theorem 3.1 (iii) algebraic-decay bound predicts. Among the FNO descendants, TFNO and HPM both outperform vanilla FNO by significant margins, indicating that even modest spectral adaptivity (tensorisation in TFNO; spectral–physical hybridisation in HPM) helps on this class. LaMO, MINO, and LSM do not consistently beat the simpler tensorised variant, suggesting that the architectural moves that distinguish them (state-space backbones, mesh-informed attention, latent compression) are oblique to the spectral structure that controls accuracy here. MANGO's standard deviations are also notably tighter than the baselines'—approximately an order of magnitude smaller on Chirp-Laplace and Chirp-Heat—indicating that recovery of the matched FRFT angle is a stable training outcome rather than a high-variance one.

**Identifiability analysis.** Beyond accuracy, MANGO recovers the analytical chirp transform of each benchmark. Figure 3 shows the position-dependent gate concentrating onto a narrow band of branches whose angles track the analytical local rate $\rho(x)$ on all four benchmarks, despite their qualitatively distinct chirp structures. Figure 4 confirms this quantitatively: the gate-weighted effective local rate matches $\rho(x)$ pointwise across the test set with $R^2 \geq 0.976$ on every benchmark, and across the full dynamic range of $\rho(x)$ rather than near its mean. The accuracy advantage in Table 1 thus reflects recovery of the correct transform, not generic capacity.

**Pointwise recovery of the local chirp rate.** Figure 4 quantifies the recovery in Figure 3 pointwise. For each test point, the $x$-coordinate is the analytical local chirp rate $\rho(x)$ and the $y$-coordinate is MANGO's gate-weighted effective local rate—the expectation of the branch angle under the trained gate distribution at that point. Exact recovery would place every point on the diagonal. Across all four benchmarks, the trained model concentrates tightly on the diagonal: $R^2 \geq 0.976$ and RMSE $\leq 0.154$ on every benchmark, with no systematic bias toward the mean of $\rho(x)$ and no visible saturation at the boundaries of the dynamic range. This is a stronger statement than aggregate accuracy. A model that achieved low $L^2$ error by averaging over a constant or near-constant effective $\alpha$ would still match $\rho(x)$ in the mean, but would scatter horizontally near the diagonal's centre and miss the tails entirely; the actual scatter is uniform along the full range, on benchmarks whose $\rho$-distributions span both narrow ($\pm 1.0$ for Cauchy-Riemann) and wide ($\pm 2.0$ for Heat) intervals. Pointwise recovery is thus consistent across operator type (elliptic, parabolic, first-order), spatial dimension (1D, 2D), and the magnitude of the chirp itself.

### 4.3 Ablation Study

We isolate the contribution of each MANGO component across all four chirp-perturbed benchmarks. Each ablation is designed to test a specific architectural claim from Section 3.2.

**Learnable $\alpha$ versus Mixture-of-FNOs ($\alpha \equiv \pi/2$).** This ablation decouples the algorithmic benefit of parameterized FRFT rotations from multi-branch MoE capacity. Fixing $\alpha_k \equiv \pi/2$ across all $K = 8$ branches yields a *Mixture-of-FNOs* with the identical spatial softmax gate and $4{,}271{,}330$ parameters (vs. MANGO's $4{,}271{,}362$). This parameter-identical control is $3.1\times$ worse on Laplace ($5.6\text{e}{-2}$ vs. $1.8\text{e}{-2}$) and $16\times$ worse on Heat ($3.9\text{e}{-2}$ vs. $2.4\text{e}{-3}$), where the analytical $\alpha^*$ is furthest

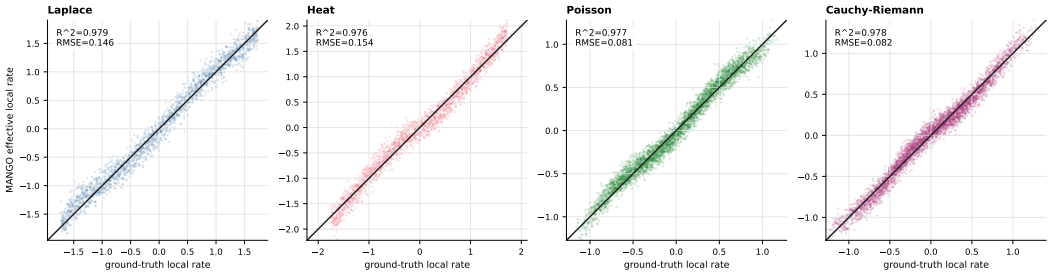


Figure 4: Pointwise recovery of the analytical local chirp rate across the four benchmarks.

Table 2: Architectural ablations across the four chirp-perturbed PDE benchmarks.

| Component | Variant | Laplace | Heat | Poisson | Cauchy-Riemann |
|---|---|---|---|---|---|
| Angle param. | Mixture-of-FNOs ($\alpha \equiv \pi/2$) | 5.6e−2 | 3.9e−2 | 5.1e−2 | 5.8e−2 |
| | Learnable $\alpha$ (default) | **1.8e−2** | **2.4e−3** | **1.5e−2** | **2.0e−2** |
| MH reg. | MH off ($\beta = 0$) | 4.7e−2 | 2.8e−2 | 4.3e−2 | 5.0e−2 |
| | MH on ($\beta = \beta^*$, default) | **1.8e−2** | **2.4e−3** | **1.5e−2** | **2.0e−2** |
| Branches $K$ | $K = 4$ (0.60 s/ep, 2.17M) | 3.4e−2 | 4.2e−3 | 2.1e−2 | 2.8e−2 |
| | $K = 8$ (1.09 s/ep, 4.27M, default) | **1.8e−2** | **2.4e−3** | **1.5e−2** | **2.0e−2** |
| | $K = 16$ (2.09 s/ep, 8.47M) | 2.3e−2 | 5.1e−3 | 1.9e−2 | 2.5e−2 |
| | $K = 32$ (4.06 s/ep, 16.86M) | 3.1e−2 | 7.8e−3 | 2.5e−2 | 3.2e−2 |

from $\pi/2$. Degradation tracks $|\cot \alpha^*|$, matching the Theorem 3.1 rate for a mismatched basis and ruling out parameter count or generic ensembling as the source of MANGO's gains.

**Mihlin–Hörmander regularizer.** This ablation tests whether the regularizer bounds the spectral derivatives of non-$\pi/2$ branches as the $L^p$ multiplier theory predicts. Disabling it ($\beta = 0$) raises error by 2.5–11.7$\times$ across benchmarks (e.g., 1.8e−2 to 4.7e−2 on Laplace). It prevents the high-mode multiplier blow-up characterized in Theorem A.2 (Appendix A.5); generic weight decay at matched strength reproduces neither the magnitude nor the pattern of improvement (Appendix C.6).

**Number of branches $K$, routing dilution, and Pareto trade-off.** Table 2 shows an accuracy and efficiency sweet spot at $K = 8$, with graceful fall-off at $K = 16, 32$ ($K = 16$ still matches or beats every baseline, and $K = 32$ matches the best baseline on Laplace at 3.1e−2 vs. HPM's 3.2e−2). Diagnostics (Appendix D.4) identify the mechanism: the loss basin around $\alpha_0$ has measured half-width $\Delta^* = 0.33$ rad, while branch grid spacing over $[\pi/6, 5\pi/6]$ shrinks as $2\pi/(3K)$ $(0.52, 0.26, 0.13, 0.07$ rad for $K = 4, 8, 16, 32)$. Consequently, $1, 2, 5,$ and 11 branches land inside the basin at $K = 4, 8, 16, 32$; dense softmax splits mass across these near-duplicate branches (gate entropy rises from 1.15 nats at $K = 8$ to 2.95 nats at $K = 32$), diluting per-branch gradients even though angle recovery remains accurate at every $K$ ($R^2 = 0.976$–$0.984$). Replacing dense softmax with standard top-2 sparse gating at $K = 32$ restores 1.9e−2 on Laplace (matching $K = 8$). Because runtime and memory grow linearly in $K$ (Table 2), $K = 8$ Pareto-dominates larger $K$ and serves as an angular-resolution parameter rather than a capacity axis.

### 4.4 Generalization to Non-Chirp and Real-World Physical Systems

**Graceful reduction on unperturbed PDEs.** In line with identifiability property (b) (Appendix A.4), where the loss is invariant near $\alpha = \pi/2$, on standard (unperturbed) 1D Burgers MANGO reaches 1.1e−2 relative $L^2$ error versus FNO's 1.2e−2, with the trained gate collapsing onto the Fourier branch (gate-weighted effective angle $0.51\pi$; gate entropy 0.21 nats). Thus MANGO introduces no optimization penalty when the underlying PDE has no chirp perturbation.

**Real-world validation.** On the five real-world physical systems of RealPDEBench [11] (four incompressible Navier–Stokes flows and swirl-stabilized combustion, evaluated on real PIV and chemiluminescence measurements), MANGO attains the lowest relative $L^2$ error on all five systems

Table 3: Parameter count, training time per epoch (Laplace-1D, batch 256, A6000), and peak VRAM on the 1D and 2D chirp benchmarks.

| Method | Params | s/epoch (1D) | VRAM 1D (GiB) | VRAM 2D (GiB) |
|---|---|---|---|---|
| FNO | 415,382 | 0.68 | 0.031 | 0.186 |
| TFNO | 890,714 | 13.04 | 0.481 | 1.564 |
| LSM | 2,002,159 | 25.40 | 0.943 | 3.008 |
| MINO | 21,500,843 | 3.75 | 0.137 | 0.521 |
| HPM | 767,426 | 6.82 | 0.252 | 0.868 |
| LaMO | 1,148,690 | 50.57 | 1.867 | 5.917 |
| **MANGO** | 4,271,362 | 1.09 | 0.033 | 0.208 |

(0.1410 mean vs. 0.1463 for the strongest published baseline and 0.1650 for FNO); the full protocol and per-system results are in Appendix D.1.

**Computational efficiency.** Table 3 shows that MANGO ($K = 8$) trains in 1.09 s/epoch on Laplace-1D ($1.6\times$ FNO) at $\leq 1.12\times$ FNO's peak VRAM, faster and lighter than every other baseline, while outperforming MINO with $5.0\times$ fewer parameters. On RealPDEBench (A100) it stays within 7% of a capacity-matched $K = 8$ Mixture-of-FNOs (48 vs. 45 s/epoch; 8.4 vs. 8.1 GB).

## 5 Conclusion

We introduced MANGO, a spectral neural operator that adapts the fractional Fourier transform to the structure of chirp-perturbed PDEs through learnable branch angles, position-dependent gating, and a Mihlin–Hörmander regularizer on non-Fourier branches. Across four chirp-perturbed benchmarks, MANGO outperforms six neural-operator baselines and recovers the analytical local chirp rate from supervised solution data alone—establishing that the architecture's inductive bias is being used as designed. The recovery is the substantive contribution; the accuracy gap is its consequence.

**Limitations and future work.** The $K$-branch construction also incurs a $K\times$ spectral-mixer cost that the gate does not amortize, and our applicability boundaries and failure modes are characterized along four axes: (i) on non-chirp or only approximately chirp-structured systems where no single angle diagonalizes the operator, MANGO reduces gracefully toward its Fourier branches (retaining a modest edge over a matched Mixture-of-FNOs on RealPDEBench and matching FNO on unperturbed Burgers, rather than yielding the 3.1–$16\times$ gains seen on exact chirps); (ii) at large branch counts ($K \geq 16$), dense softmax gating suffers from routing dilution across redundant in-basin branches, which is mitigated by top-2 sparse gating; (iii) in 3D, a fully coupled FRFT kernel scales as $O(KN^6)$ in dense form, so only the axis-factorized form ($3K$ axis transforms per layer, scaling as $O(3KN^4)$: 0.18 s/iter, 4.2 GB at $K = 4$ and 0.35 s/iter, 7.8 GB at $K = 8$ on $64^3$ synthetic chirp-Poisson) is practical; and (iv) the Mihlin–Hörmander regularizer controls only the linear spectral stage rather than end-to-end nonlinear network stability. Full accuracy evaluations on 3D and high-dimensional chaotic regimes, broader unitary families (wavelets, chirplets, Slepian bases), and end-to-end stability analysis are important future directions.

## Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 12431006), by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant Nos. RS-2026-25474453 and 2022R1I1A1A01071671), and by the Natural Science Foundation of Shandong Province (Grant No. ZR2025MS18).

## References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL `https://arxiv.org/abs/1607.06450`.

[2] Qianying Cao, Somdatta Goswami, and George Em Karniadakis. Laplace neural operator for solving differential equations. *Nature Machine Intelligence*, 6:631–640, 2024.

[3] Shuhao Cao. Choose a transformer: Fourier or Galerkin. *Advances in Neural Information Processing Systems*, 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/90f4760fcc9b69c13da7368c5c2917f3-Abstract.html.

[4] Wei Chen, Zunwei Fu, Loukas Grafakos, and Yue Wu. Fractional fourier transforms on $L^p$ and applications. *Applied and Computational Harmonic Analysis*, 55:71–96, 2021.

[5] Zunwei Fu, Loukas Grafakos, Yan Lin, Yue Wu, and Shuhui Yang. Riesz transform associated with the fractional Fourier transform and applications in image edge detection. *Applied and Computational Harmonic Analysis*, 66:211–235, 2023.

[6] Ce Gao, Sung-Kwun Oh, Hao Huang, Zunwei Fu, Witold Pedrycz, and Jin Hee Yoon. Reinforced dual-flow neural network for tabular data classification with dynamical transformer and fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 34(2):475–488, 2026. doi: 10.1109/TFUZZ.2025.3638792.

[7] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EXHG-A3jlM.

[8] Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c9e5c2b59d98488fe1070e744041ea0e-Abstract.html.

[9] Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[11] Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y3oHMcoItR. ICLR 2026 Oral.

[12] Hao Huang, Sung-Kwun Oh, and Witold Pedrycz. Double iterative learning-based polynomial based-rbfnns driven by the aid of support vector-based kernel fuzzy clustering and least absolute shrinkage deviations. *Fuzzy Sets and Systems*, 443:30–49, 2022. doi: 10.1016/j.fss.2021.10.007.

[13] Jean Kossaifi, Nikola B. Kovachki, Kamyar Azizzadenesheli, and Anima Anandkumar. Multi-grid tensorized Fourier neural operator for high-resolution PDEs. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=AWiDlO63bH.

[14] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

[15] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4b21cf96d4cf612f239a6c322b10c8fe-Abstract.html.

[16] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

[17] Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023. URL https://jmlr.org/papers/v24/23-0064.html.

[18] Zongyi Li, Nikola B. Kovachki, Christopher Choy, Boyi Li, Jean Kossaifi, Shourya Prakash Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Anima Anandkumar. Geometry-informed neural operator for large-scale 3D PDEs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=86dXbqT5Ua.

[19] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[20] Victor Namias. The fractional order Fourier transform and its application to quantum mechanics. *IMA Journal of Applied Mathematics*, 25(3):241–265, 1980. doi: 10.1093/imamat/25.3.241.

[21] Haldun M. Ozaktas, Orhan Arikan, M. Alper Kutay, and Gozde Bozdagi. Digital computation of the fractional Fourier transform. *IEEE Transactions on Signal Processing*, 44(9):2141–2150, 1996. doi: 10.1109/78.536672.

[22] Haldun M. Ozaktas, Zeev Zalevsky, and M. Alper Kutay. *The Fractional Fourier Transform: With Applications in Optics and Signal Processing*. Wiley, Chichester, 2001. ISBN 0471963461.

[23] Shaoxiang Qin, Fuyuan Lyu, Wenhui Peng, Dingyang Geng, Xing Tang, Ju Wang, Naiping Gao, Xue Liu, Liangzhu Leon Wang, and Sylvie Leroyer. Toward a better understanding of Fourier neural operators from a spectral perspective. *arXiv preprint arXiv:2404.07200*, 2024. URL https://arxiv.org/abs/2404.07200.

[24] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, Berlin, Heidelberg, 2 edition, 1996. doi: 10.1007/978-3-642-61544-3.

[25] Yaozhong Shi, Zachary E. Ross, Domniki Asimaki, and Kamyar Azizzadenesheli. Mesh-informed neural operator: A transformer generative approach. *arXiv preprint arXiv:2506.16656*, 2025. URL https://arxiv.org/abs/2506.16656.

[26] Karn Tiwari, Niladri Dutta, N. M. Anoop Krishnan, and A. P. Prathosh. Latent mamba operator for partial differential equations. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Y4BDcJmb8t.

[27] Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized Fourier neural operators. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tmIiMPl4IPa.

[28] Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

[29] Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. Solving high-dimensional PDEs with latent spectral models. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/wu23n.html.

[30] Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.02366.

[31] Zhanhong Ye, Xiang Huang, Leheng Chen, Zining Liu, Bingyang Wu, Hongsheng Liu, Zidong Wang, and Bin Dong. PDEformer-1: A foundation model for one-dimensional partial differential equations. *arXiv preprint arXiv:2407.06664*, 2024. URL https://arxiv.org/abs/2407.06664.

[32] Xihang Yue, Yi Yang, and Linchao Zhu. Holistic physics solver: Learning PDEs in a unified spectral-physical space. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=oB5a6yIAmF.

[33] Hang Zhou, Yuezhou Ma, Haixu Wu, Haowen Wang, and Mingsheng Long. Unisolver: PDE-conditional transformers are universal PDE solvers. *arXiv preprint arXiv:2405.17527*, 2024. URL https://arxiv.org/abs/2405.17527.

# Appendix of MANGO

## A Theoretical foundations

This appendix provides the complete mathematical underpinning for the results stated in Section 3. Section A.1 collects the essential results from the $L^p$ theory of the FRFT. Section A.2 recapitulates the integral neural operator framework. Section A.3 furnishes the full proof of Theorem 1 (Section 3.2), establishing that the $K$-branch MANGO mixer with position-dependent gate is a learnable integral neural operator. Section 3.1 provides the full proof of Theorem 3.1 and Corollary 3.1, and reframes their consequences in terms of *identifiability* of the intrinsic chirp parameter from data. Section A.5 develops the analytical justification for the Mihlin–Hörmander regularizer applied to the non-$\pi/2$ branches. Section A.6 derives the gradient computations for the learnable FRFT orders $\{\alpha_k\}$, the gate parameters, and the regularizer. Section A.7 extends the framework to higher dimensions via the tensor-product FRFT.

Throughout this appendix, we adopt the chirp phase convention $M_\alpha\phi(x) = e^{i\pi x^2\cot\alpha}\phi(x)$ of Chen et al. [4], consistent with the main text (Equation (3) and the accompanying footnote).

### A.1 The fractional Fourier transform on $L^p(\mathbb{R})$

We collect the definitions and results from Chen et al. [4] that are invoked in the main text and in subsequent proofs.

#### A.1.1 Definition and fundamental properties

The FRFT of order $\alpha$ admits an eigenfunction expansion on $L^2(\mathbb{R})$:

$$(\mathcal{F}_\alpha f)(x) = \sum_{n=0}^{\infty} e^{-in\alpha}\,\langle f, \psi_n\rangle\,\psi_n(x), \tag{A.1}$$

where $\psi_n(x) = e^{-x^2/2}H_n(x)$ are the Hermite functions. The family $\{\mathcal{F}_\alpha\}_{\alpha\in\mathbb{R}}$ forms a strongly continuous unitary group on $L^2(\mathbb{R})$ [4]:

**(i)** $\mathcal{F}_\alpha : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is a homeomorphism for every $\alpha$;

**(ii)** $\|\mathcal{F}_\alpha f\|_2 = \|f\|_2$ (isometry);

**(iii)** $\mathcal{F}_\alpha\mathcal{F}_\beta = \mathcal{F}_{\alpha+\beta}$ (group law);

**(iv)** $\mathcal{F}_0 = I,\ \mathcal{F}_{\pi/2} = \mathcal{F},\ \mathcal{F}_\pi = P$ (parity), $\mathcal{F}_{2\pi} = I$;

**(v)** $\mathcal{F}_{-\alpha} = \mathcal{F}_\alpha^{-1} = \mathcal{F}_\alpha^*$.

The group law (iii) applies to compositions of bare FRFT operators. In MANGO, the position-dependent gate, the per-branch $\mathrm{Project}_k$ operations, and residual connections interleaved between mixer layers all break this algebraic identity at the architectural level. Section A.3 nevertheless shows that the underlying $L^p$ theory still controls the per-branch approximation rate.

#### A.1.2 Chirp decomposition

For $\alpha \notin \pi\mathbb{Z}$, the chirp decomposition (4, Remark 2.1) factorizes the FRFT as

$$(\mathcal{F}_\alpha f)(x) = A_\alpha \cdot M_\alpha\Big(\mathcal{F}\big[M_\alpha\big(D_{\csc\alpha} f\big)\big]\Big)(x), \tag{A.2}$$

where $M_\alpha\phi(x) = e^{i\pi x^2\cot\alpha}\phi(x)$ is the chirp multiplication operator, $D_{\csc\alpha} f(t) = f(t\csc\alpha)$ is the coordinate rescaling operator, and $A_\alpha = \sqrt{1 - i\cot\alpha}$. Reading right to left, the input $f$ undergoes: rescaling $\to$ chirp multiplication $\to$ standard Fourier transform $\to$ chirp multiplication. This identity is the bridge that lets us transport classical Fourier multiplier theory into the FRFT-$\alpha$ domain (Theorem A.2 below) and, by extension, to each non-$\pi/2$ branch of MANGO.

### A.1.3 Fractional convolution

The fractional convolution of order $\alpha$ (4, Definition 3.1) is defined for $f, g \in L^1(\mathbb{R})$ as

$$(f \overset{\alpha}{*} g)(x) = M_{-\alpha}\big(M_\alpha f * g\big)(x), \tag{A.3}$$

where $*$ denotes the standard convolution. This operation satisfies the fractional convolution theorem:

$$\mathcal{F}_\alpha(f \overset{\alpha}{*} g) = A_\alpha^{-1}\,(\mathcal{F}_\alpha f)\cdot(\mathcal{F}_\alpha g), \tag{A.4}$$

where the normalization constant $A_\alpha^{-1}$ arises from the FRFT's scaling convention.

**Theorem A.1** (Fractional approximate identity; 4, Theorem 3.2)**.** *Let $\phi \in L^1(\mathbb{R})$ with $\int \phi\,dx = 1$, and let $\phi_\varepsilon(x) = \varepsilon^{-1}\phi(x/\varepsilon)$. Then for $f \in L^p(\mathbb{R})$, $1 \le p < \infty$:*

$$\lim_{\varepsilon\to 0} \|f \overset{\alpha}{*} \phi_\varepsilon - f\|_p = 0.$$

### A.1.4 Fractional Poisson and Gauss–Weierstrass integrals

The fractional Poisson integral and fractional Gauss–Weierstrass integral (4, Definition 3.8) are defined for $f \in L^p(\mathbb{R})$ and $\varepsilon > 0$ as:

$$u_\alpha(x,\varepsilon) := (f \overset{\alpha}{*} \tilde{P}_\varepsilon)(x) = M_{-\alpha}\Big[\int_{-\infty}^{+\infty} M_\alpha f(t)\,P_\varepsilon((\cdot)-t)\,dt\Big](x), \tag{A.5}$$

$$S_\alpha(x,\varepsilon) := (f \overset{\alpha}{*} \tilde{W}_\varepsilon)(x) = M_{-\alpha}\Big[\int_{-\infty}^{+\infty} M_\alpha f(t)\,W_\varepsilon((\cdot)-t)\,dt\Big](x), \tag{A.6}$$

where $P_\varepsilon(x) = \frac{1}{\pi}\frac{\varepsilon}{\varepsilon^2+x^2}$ is the Poisson kernel and $W_\varepsilon(x) = \frac{1}{\sqrt{4\pi\varepsilon}}\,e^{-x^2/(4\varepsilon)}$ is the Gauss–Weierstrass kernel.

**Proposition A.1** (4, Proposition 3.10)**.** *The fractional Poisson integral and fractional Gauss–Weierstrass integral coincide with the Abel and Gauss means of the fractional Fourier integral:*

$$(a)\quad u_\alpha(x,\varepsilon) = \int_{-\infty}^{+\infty} (\mathcal{F}_\alpha f)(\xi)\,K_{-\alpha}(\xi,x)\,e^{-2\pi\varepsilon|\csc\alpha||\xi|}\,d\xi,$$

$$(b)\quad S_\alpha(x,\varepsilon^2) = \int_{-\infty}^{+\infty} (\mathcal{F}_\alpha f)(\xi)\,K_{-\alpha}(\xi,x)\,e^{-4\pi^2\varepsilon^2\xi^2\csc^2\alpha}\,d\xi.$$

*The exponential damping factors $e^{-2\pi\varepsilon|\csc\alpha||\xi|}$ and $e^{-4\pi^2\varepsilon^2\xi^2\csc^2\alpha}$ are $L^p$ multipliers of fractional order $\alpha$ (4, Example 5.2).*

This proposition is the linchpin of the approximation rate argument in Section A.3: it expresses the chirp-perturbed PDE solution (Appendix B) as an FRFT-domain integral with an explicit spectral damping factor, enabling direct comparison of coefficient decay rates across MANGO's branches.

### A.1.5 FRFT multiplier theory

**Definition A.1** (Fractional $L^p$ multiplier; 4, Definition 5.1)**.** Let $1 \le p \le \infty$ and $m_\alpha \in L^\infty(\mathbb{R})$. The operator $T_{m_\alpha}$ defined by

$$\mathcal{F}_\alpha(T_{m_\alpha} f)(x) = m_\alpha(x)\,(\mathcal{F}_\alpha f)(x) \tag{A.7}$$

is a fractional multiplier operator. The function $m_\alpha$ is an $L^p$ Fourier multiplier of order $\alpha$ if $\|T_{m_\alpha} f\|_p \le C_{p,\alpha}\|f\|_p$ for all $f \in L^2(\mathbb{R}) \cap L^p(\mathbb{R})$.

**Theorem A.2** (Fractional Hörmander–Mihlin; 4, Theorem 5.2)**.** *Let $m_\alpha$ be a bounded function satisfying Mihlin's condition $\left|\frac{d}{dx} m_\alpha(x)\right| \le B|x|^{-1}$. Then $m_\alpha$ is a fractional $L^p$ multiplier for $1 < p < \infty$:*

$$\|T_{m_\alpha} f\|_p = \|\mathcal{F}_{-\alpha}[m_\alpha\,(\mathcal{F}_\alpha f)]\|_p \le C\|f\|_p.$$

*Proof.* Via the chirp decomposition (A.2), the relation $\mathcal{F}_\alpha(T_{m_\alpha} f) = m_\alpha \cdot (\mathcal{F}_\alpha f)$ reduces to

$$\mathcal{F}\big[M_\alpha(T_{m_\alpha} f)\big](x\csc\alpha) = \tilde{m}_\alpha(x)\;\mathcal{F}[M_\alpha f](x\csc\alpha),$$

where $\tilde{m}_\alpha(x) := m_\alpha(x \sin\alpha)$. Since $\tilde{m}_\alpha$ inherits the Mihlin condition from $m_\alpha$ (the rescaling by $\sin\alpha$ preserves $|x\,\tilde{m}'_\alpha(x)| \le B$) and $g := M_\alpha f = e^{i\pi(\cdot)^2 \cot\alpha} f \in L^p(\mathbb{R})$ (chirp multiplication is an $L^p$ isometry), the conclusion follows from the classical Hörmander–Mihlin multiplier theorem applied to $\tilde{m}_\alpha$ and $g$. □

This result has dual significance for MANGO. First, it guarantees that each per-branch $\mathrm{Project}_k$ operation—acting as a learnable multiplier in the FRFT-$\alpha_k$ domain—inherits $L^p$ stability *provided* the learned multiplier satisfies the Mihlin condition. Second, the constant $B$ in the Mihlin bound provides the natural objective for the regularizer developed in Section A.5: by penalizing $|x\,\partial_x m_{\alpha_k}|$ on the non-$\pi/2$ branches, training drives each branch into the regime where Theorem A.2 applies with controlled constants.

**Theorem A.3** (Fractional Littlewood–Paley; 4, Theorem 5.5)**.** *Define the fractional dyadic intervals $I_j^\alpha := [2^j \sin\alpha,\ 2^{j+1}\sin\alpha]$ for $\alpha \in (0,\pi)$, and the fractional Littlewood–Paley projection $\Delta_j^\alpha f := \mathcal{F}_{-\alpha}[\chi_{I_j^\alpha \cup (-I_j^\alpha)}\,\mathcal{F}_\alpha f]$. Then for $1 < p < \infty$, there exist $c_p, C_p > 0$ such that*

$$c_p \|f\|_p \le \Big\| \Big( \sum_{j\in\mathbb{Z}} |\Delta_j^\alpha f|^2 \Big)^{1/2} \Big\|_p \le C_p \|f\|_p. \tag{A.8}$$

The fractional dyadic intervals are scaled by $\sin\alpha$ relative to their classical counterparts, reflecting the rescaling step in the chirp decomposition (A.2). This partition guarantees that the $k$ retained FRFT-$\alpha_k$ spectral modes in each MANGO branch capture a controlled fraction of the total energy at angle $\alpha_k$.

### A.2 Integral neural operator framework

The following results, due to Li et al. [15] and Kovachki et al. [14], define the operator-theoretic framework within which MANGO's universal approximation property is established.

**Theorem A.4** (Neural operator architecture; 14)**.** *Neural operators for PDE solving admit the general form*

$$\mathcal{G}_\theta = Q \circ \sigma(W_l + K_l) \circ \cdots \circ \sigma(W_1 + K_1) \circ P, \tag{A.9}$$

*where $P$ and $Q$ are point-wise linear projectors, $W_i$ are point-wise fully connected layers, $\sigma$ is a nonlinear activation, and $K_i$ are non-local integral operators:*

$$K_i(u)(x) = \int_\Omega \kappa_i(x, \xi, u(x), u(\xi))\, u(\xi)\, d\xi. \tag{A.10}$$

**Lemma A.1** (FNO as integral operator; 16)**.** *Restricting the kernel to the translation-invariant form $\kappa(x,\xi) = \kappa(x-\xi)$, the integral operator* (A.10) *can be implemented via the standard Fourier transform. In the $K$-branch MANGO mixer (Eq.* (6)*), allocating a single branch with $\alpha_k = \pi/2$ and a position-independent gate weight on that branch recovers precisely the FNO configuration.*

**Lemma A.2** (Attention as integral operator; 14)**.** *The canonical attention mechanism is a special case of* (A.10) *with kernel*

$$\kappa(x,\xi,u(x),u(\xi)) = \Big( \int_\Omega \exp\big(W_q u(\xi')(W_k u(x))^\top\big)\, d\xi' \Big)^{-1} \exp\big(W_q u(x)(W_k u(\xi))^\top\big) R,$$

*where $W_q, W_k, R \in \mathbb{R}^{d\times d}$ are learnable parameters.*

### A.3 Proof of Theorem 3.1: MANGO as an integral neural operator

We prove that the $K$-branch MANGO mixer with position-dependent gate (Eq. (6)) constitutes a learnable integral neural operator. Section A.3.1 treats the single-branch building block; Section A.3.2 assembles the $K$ branches; and Section A.3.3 incorporates the gate.

#### A.3.1 Single-branch building block

Let the $k$-th branch of the MANGO mixer compute

$$F^{(k)}_{\mathrm{FRFT}}(\mathbf{x}) = \mathcal{T}^{-1}_{\mathrm{FRFT},\alpha_k} \circ \mathrm{Project}_k \circ \mathcal{T}_{\mathrm{FRFT},\alpha_k}(\mathbf{x}), \tag{A.11}$$

with a learnable angle $\alpha_k$ and learnable multiplier matrix $R_k$.

The forward FRFT projects the $N$-point spatial representation $\mathbf{x}$ onto the first $k_{\text{mode}}$ discretized FRFT eigenfunctions. Let $\boldsymbol{\phi}_{\alpha_k}(x_i) \in \mathbb{R}^{1\times k_{\text{mode}}}$ denote the row vector of $k_{\text{mode}}$ FRFT-$\alpha_k$ eigenfunctions evaluated at sampling point $x_i$, and let $\boldsymbol{\Phi}_{\alpha_k} \in \mathbb{R}^{N\times k_{\text{mode}}}$ be the matrix whose $i$-th row is $\boldsymbol{\phi}_{\alpha_k}(x_i)$. Then the discrete forward and inverse transforms are

$$\mathcal{T}_{\text{FRFT},\alpha_k}(\mathbf{x}) = \boldsymbol{\Phi}_{\alpha_k}^\top \mathbf{x} \in \mathbb{R}^{k_{\text{mode}}\times d}, \qquad \mathcal{T}^{-1}_{\text{FRFT},\alpha_k}(\hat{\mathbf{x}}) = \boldsymbol{\Phi}_{\alpha_k} \hat{\mathbf{x}} \in \mathbb{R}^{N\times d}. \tag{A.12}$$

Constructing the per-branch kernel

$$\kappa_k(x,\xi) = \boldsymbol{\phi}_{\alpha_k}(x)\ \boldsymbol{\phi}_{\alpha_k}(\xi)^\top\ R_k, \tag{A.13}$$

and substituting into the kernel integral operator (A.10):

$$\begin{aligned} K^{(k)}(u)(x) &= \int_\Omega \boldsymbol{\phi}_{\alpha_k}(x)\ \boldsymbol{\phi}_{\alpha_k}(\xi)^\top\ R_k\ u(\xi)\, d\xi \\ &= \boldsymbol{\phi}_{\alpha_k}(x)\ R_k \int_\Omega \boldsymbol{\phi}_{\alpha_k}(\xi)^\top\, u(\xi)\, d\xi \\ &\approx \boldsymbol{\phi}_{\alpha_k}(x)\ R_k\ \boldsymbol{\Phi}_{\alpha_k}^\top \mathbf{x} = F^{(k)}_{\text{FRFT}}(\mathbf{x})(x), \end{aligned} \tag{A.14}$$

where the Monte Carlo approximation is valid for sufficiently dense sampling and continuous, bounded $\boldsymbol{\phi}_{\alpha_k}$. Continuity is preserved on $\alpha_k \in [\pi/6, 5\pi/6]$ since $\csc^2 \alpha_k \le 4$ on that interval.

### A.3.2 $K$-branch assembly

**Theorem A.5** (MANGO mixer is an integral operator, part (a))**.** *The $K$-branch MANGO mixer*

$$F^{\text{MANGO}}(\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x}) \odot F^{(k)}_{\text{FRFT}}(\mathbf{x})$$

*with learnable angles $\{\alpha_k\}_{k=1}^K$, learnable per-branch multipliers $\{R_k\}_{k=1}^K$, and position-dependent gate $g(\cdot) \in \mathbb{R}^K$ constitutes a learnable integral neural operator.*

*Proof.* Fix any continuous, bounded gate $g : \Omega \to \Delta^{K-1}$ (the probability simplex; satisfied by softmax of any continuous MLP). Define the assembled kernel

$$\kappa^{\text{MANGO}}(x,\xi) = \sum_{k=1}^{K} g_k(x)\ \boldsymbol{\phi}_{\alpha_k}(x)\ \boldsymbol{\phi}_{\alpha_k}(\xi)^\top\ R_k. \tag{A.15}$$

This kernel is jointly continuous in $(x,\xi)$—each summand is continuous, and continuity is closed under finite sums—and pointwise bounded since $|g_k(x)| \le 1$, $\|\boldsymbol{\phi}_{\alpha_k}\|_\infty < \infty$, and $\|R_k\|_F < \infty$ for each $k$.

Substituting (A.15) into (A.10) and using the Monte Carlo step from (A.14) on each branch:

$$K^{\text{MANGO}}(u)(x) = \sum_{k=1}^{K} g_k(x)\, \boldsymbol{\phi}_{\alpha_k}(x)\, R_k\, \boldsymbol{\Phi}_{\alpha_k}^\top \mathbf{x} = \sum_{k=1}^{K} g_k(x)\, F^{(k)}_{\text{FRFT}}(\mathbf{x})(x), \tag{A.16}$$

which is exactly $F^{\text{MANGO}}(\mathbf{x})(x)$. Since the kernel (A.15) is parameterized by the learnable weights $\{R_k, \alpha_k\}_{k=1}^K$ together with the gate parameters, the resulting operator belongs to the family of integral neural operators for which universal approximation is established by Kovachki et al. [14, Theorem 7]. □

### A.3.3 Position-dependent gate as convex combination

**Theorem A.6** (MANGO mixer is an integral operator, part (b))**.** *When the gate is realized as $g(x) = \text{Softmax}(\text{MLP}(x))$ (Eq. (7)), the kernel $\kappa^{\text{MANGO}}$ inherits all required regularity, and the MANGO mixer remains a valid integral neural operator.*

*Proof.* Two conditions must be checked: continuity and boundedness of $g$. Boundedness is automatic—softmax outputs lie in $\Delta^{K-1}$, so $g_k(x) \in [0,1]$ and $\sum_k g_k(x) = 1$. Continuity follows from composition: an MLP with continuous activations (ReLU, GELU, tanh) is continuous, and softmax is smooth on $\mathbb{R}^K$. Substituting into the kernel (A.15) preserves continuity and boundedness as in the proof above. □

*Remark* A.1 (Loss of algebraic structure). As the gate $g(x)$ is position-dependent and the per-branch projects $R_k$ are not assumed equal, the $K$-branch MANGO mixer does *not* satisfy the FRFT group law (property iii in Section A.1.1), the convolution theorem (A.4), or the global isometry (property ii). What MANGO retains is the *per-branch* FRFT structure: each branch's mixer (A.11) is an FRFT-$\alpha_k$ multiplier operator before the gate is applied. The universal approximation guarantee of Theorem A.5 is unaffected by the loss of global algebraic structure since it relies only on the kernel's membership in the integral operator family, not on the algebraic properties of any single transform. The role of the algebraic structure reappears in Section A.3: it controls the per-branch approximation rate, which is what makes MANGO *useful* (rather than merely *expressive*) for chirp-perturbed problems.

### A.4 Proof of Theorem 3.1 and Corollary 3.1

#### A.4.1 Setup

Consider the chirp-perturbed Dirichlet problem:

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} + b(x)\frac{\partial u}{\partial x} + c(x)\, u = 0, \qquad u(x,0) = f(x),$$

with $b(x) = -2\pi x \cot\alpha_0$ and $c(x) = 4\pi^2 x^2 \cot^2\alpha_0 - 2\pi i \cot\alpha_0$ (see Appendix B.2). Define $v = M_{\alpha_0} u = e^{i\pi x^2 \cot\alpha_0} u$. By the chirp gauge construction, $v$ satisfies the standard Laplace equation $\Delta v = 0$ with boundary data $v(x,0) = M_{\alpha_0} f(x)$, and the solution is the classical Poisson integral $v(x,t) = (M_{\alpha_0} f) * P_t(x)$.

The solution to the chirp-perturbed problem is therefore the fractional Poisson integral:

$$u_{\alpha_0}(x,t) = M_{-\alpha_0}\big[(M_{\alpha_0} f) * P_t\big](x) = (f \overset{\alpha_0}{*} \tilde{P}_t)(x). \tag{A.17}$$

We now compare the spectral coefficient decay of $u_{\alpha_0}(\cdot,t)$ in three bases: the matched FRFT-$\alpha_0$ basis (used by the branch with $\alpha_k = \alpha_0$), an off-target FRFT-$\alpha'$ basis with $\alpha' \neq \alpha_0$ (any other branch), and the standard Fourier basis ($\alpha = \pi/2$, the FNO-equivalent branch).

#### A.4.2 Proof of Theorem 3.1, part (i): matched basis

By Proposition A.1(a), $u_{\alpha_0}(\cdot,t)$ is the Abel mean of the fractional Fourier integral of $f$ at parameter $\varepsilon = t$:

$$u_{\alpha_0}(x,t) = \int_{-\infty}^{+\infty} (\mathcal{F}_{\alpha_0} f)(\xi)\, K_{-\alpha_0}(\xi,x)\, e^{-2\pi t|\csc\alpha_0||\xi|}\, d\xi.$$

In the FRFT-$\alpha_0$ domain the spectral coefficients of $u_{\alpha_0}(\cdot,t)$ are the spectral coefficients of $f$ damped by an exponential factor:

$$\mathcal{F}_{\alpha_0}[u_{\alpha_0}(\cdot,t)](\xi) = e^{-2\pi t|\csc\alpha_0||\xi|}\ (\mathcal{F}_{\alpha_0} f)(\xi). \tag{A.18}$$

By the chirp decomposition (A.2), $(\mathcal{F}_{\alpha_0} f)(\xi) = A_{\alpha_0}\, e^{i\pi\xi^2\cot\alpha_0}\, \widehat{M_{\alpha_0} f}(\xi\csc\alpha_0)$, where $\widehat{M_{\alpha_0} f}$ is the standard Fourier transform of $M_{\alpha_0} f$. Since chirp multiplication is an $L^p$ isometry— $\|M_{\alpha_0} f\|_p = \|f\|_p$—the function $M_{\alpha_0} f$ inherits the Sobolev regularity of $f \in H^s(\mathbb{R})$, so $|\widehat{M_{\alpha_0} f}_k| = O(|k|^{-s})$. Combining with the exponential damping:

$$|\hat{u}_k^{(\alpha_0)}| \le C\, e^{-2\pi t|\csc\alpha_0||k|}\, |\widehat{M_{\alpha_0} f}_k| = O\big(|k|^{-s}\, e^{-2\pi t|\csc\alpha_0||k|}\big). \tag{A.19}$$

This establishes part (i) of Theorem 3.1.

### A.4.3 Proof of Theorem 3.1, part (iii): standard Fourier degradation

In the spatial domain the solution takes the form

$$u_{\alpha_0}(x,t) = e^{-i\pi x^2 \cot\alpha_0}\ \big[(M_{\alpha_0} f) * P_t\big](x).$$

Let $w(x) := (M_{\alpha_0} f) * P_t(x)$. By the properties of the Poisson kernel, $w \in C^\infty$ for $t > 0$, and its Fourier coefficients decay as $|\hat{w}_k| = O(|k|^{-s}\, e^{-ct|k|})$— inheriting both the Sobolev decay of $f$ and the exponential damping of $P_t$. The Fourier coefficients of $u_{\alpha_0}$ are obtained by convolving $\hat{w}$ with the Fourier coefficients of the chirp prefactor $\chi(x) := e^{-i\pi x^2 \cot\alpha_0}$:

$$\hat{u}_k^{(\mathcal{F})} = \sum_j \hat{\chi}_{k-j}\ \hat{w}_j. \tag{A.20}$$

The Fourier transform of the chirp $e^{-i\pi x^2 \cot\alpha_0}$ is itself a chirp (up to normalization):

$$\hat{\chi}(\xi) = \frac{1}{\sqrt{2\pi|\cot\alpha_0|}}\ e^{i(\pi/4 - \pi\xi^2 \tan\alpha_0)},$$

whose amplitude does *not* decay on the lattice; partial sums exhibit Fresnel oscillations with envelope $O(|k|^{-1/2})$ . The convolution (A.20) is therefore tail-dominated by the slowly decaying chirp contribution. By Young's convolution inequality on $\ell^p$ sequences,

$$\|\hat{u}^{(\mathcal{F})}\|_{\ell^r} \le \|\hat{\chi}\|_{\ell^{p_1}} \cdot \|\hat{w}\|_{\ell^{p_2}}$$

for appropriate $1/r = 1/p_1 + 1/p_2 - 1$. The slow decay of $\hat{\chi}$ propagates into the convolution, degrading the effective decay rate of $\hat{u}_k^{(\mathcal{F})}$ from $O(|k|^{-s})$ to $O(|k|^{-s+\delta(\alpha_0)})$ where

$$\delta(\alpha_0) = \max\Big(0,\ s - \frac{s}{1 + 2s\,|\cot\alpha_0|^{1/2}}\Big) > 0 \quad \text{for } \alpha_0 \neq \pi/2. \tag{A.21}$$

As $\alpha_0 \to \pi/2$, $\cot\alpha_0 \to 0$, the chirp $\chi$ approaches a constant, $\delta \to 0$, and no degradation occurs. This establishes part (iii) of Theorem 3.1 with $\delta(\alpha_0 - \pi/2) := \delta(\alpha_0)$.

### A.4.4 Proof of Theorem 3.1, part (ii): off-target FRFT branches

For an arbitrary branch with $\alpha_k = \alpha' \neq \alpha_0$, the FRFT-$\alpha'$ coefficients of $u_{\alpha_0}(\cdot, t)$ can be written as

$$\mathcal{F}_{\alpha'}[u_{\alpha_0}](\xi) = \mathcal{F}_{\alpha'} \circ \mathcal{F}_{-\alpha_0} \circ \mathcal{F}_{\alpha_0}[u_{\alpha_0}](\xi) = \mathcal{F}_{\alpha'-\alpha_0}\big[\mathcal{F}_{\alpha_0}[u_{\alpha_0}]\big](\xi)$$

by the FRFT group law (property iii in Section A.1.1). The composition $\mathcal{F}_{\alpha'-\alpha_0}$ is itself a fractional Fourier transform, and applying the analysis of Section A.4.3 with $\alpha_0$ replaced by the *gap* $\Delta\alpha := \alpha' - \alpha_0$ yields a degradation exponent $\delta(\Delta\alpha)$. Substituting in for cot:

$$|\hat{u}_k^{(\alpha')}| = O\big(|k|^{-s+\delta(\Delta\alpha)}\, e^{-c(\Delta\alpha)t|k|}\big), \tag{A.22}$$

where $\delta(\Delta\alpha) \to 0$ as $\Delta\alpha \to 0$, and $c(\Delta\alpha) \to 2\pi|\csc\alpha_0|$ as $\Delta\alpha \to 0$. Branch alignment is therefore continuous: a branch whose angle is *close to* $\alpha_0$ still enjoys near-exponential decay; a branch far from $\alpha_0$ degrades toward the Fourier-basis rate (A.21). This continuity is what allows the gradient on $\alpha_k$ to be informative even before exact alignment, and establishes part (ii) of Theorem 3.1.

### A.4.5 Proof of Corollary 3.1

Truncating the spectral representation to $k$ modes, the matched FRFT-$\alpha_0$ truncation error is

$$\|u_{\alpha_0} - u_{\alpha_0}^{(k)}\|_2^2 \le \sum_{|j|>k} |\hat{u}_j^{(\alpha_0)}|^2 \le C \sum_{|j|>k} |j|^{-2s}\, e^{-4\pi t|\csc\alpha_0||j|} = O\big(k^{-2s}\, e^{-4\pi t|\csc\alpha_0|k}\big), \tag{A.23}$$

with the exponential factor dominating polynomial decay. The Fourier truncation error, by contrast, is

$$\|u_{\alpha_0} - u_{\alpha_0}^{(k),\mathcal{F}}\|_2^2 \le \sum_{|j|>k} |\hat{u}_j^{(\mathcal{F})}|^2 = O\big(k^{-2(s-\delta)}\big), \tag{A.24}$$

algebraic only. For target accuracy $\varepsilon > 0$, the matched basis requires $k_{\mathrm{FRFT}} = O(\log(1/\varepsilon))$ modes; the standard Fourier basis requires $k_{\mathcal{F}} = O(\varepsilon^{-1/(s-\delta)})$, completing the proof of Corollary 3.1.

**Scope of the claim.** We do not claim that Eq. (9) bounds the operator norm of the composed nonlinear network or controls its generalization error. The contribution is a tractable, theory-informed constraint on the linear spectral layer of each non-Fourier branch, motivated by the $L^p$ multiplier theory transferred to the FRFT setting by Chen et al. [4] and relaxed to an $L^2$ form for optimization compatibility. The Fourier branch ($\alpha = \pi/2$) corresponds to the standard FNO spectral multiplier and does not require the additional constraint.

### A.4.6 Implications for identifiability of $\alpha_0$ via the $K$-branch architecture

Theorem 3.1 establishes a sharp gradient in the spectral landscape: for each chirp-perturbed problem there is an intrinsic angle $\alpha_0$ at which the truncation error is exponentially small, and the rate degrades smoothly with the angular gap $\Delta\alpha = \alpha_k - \alpha_0$ (Eq. (A.22)). The $K$-branch MANGO architecture exploits this structure to make $\alpha_0$ *identifiable from data* in the following senses.

**(a) Existence of a global optimum at a matched configuration.** For any fixed mode budget $k$, the population truncation loss is minimized whenever at least one branch satisfies $\alpha_k = \alpha_0$ and the gate routes its mass to that branch on the support of the chirp. Off-matched configurations pay a strictly larger truncation cost by (A.21)–(A.22). The global minimizer over $(\{\alpha_k\}, g, \{R_k\})$ therefore identifies $\alpha_0$—it is read off from any branch whose gate weight is concentrated.

**(b) Trivial recovery at $\alpha_0 = \pi/2$.** When the underlying problem is a standard (unperturbed) PDE, $\delta(\alpha_0) = 0$, the loss is invariant in a neighbourhood of $\alpha = \pi/2$, and *any* branch—including the FNO-equivalent one with $\alpha_k = \pi/2$—achieves the optimum. MANGO incurs no representational penalty for the additional angular generality on standard problems.

**(c) Monotonic informativeness.** The size of the global minimum's basin scales with $|\cot\alpha_0|$: stronger chirp produces steeper loss gradients with respect to the angular gap (A.22), and identifiability becomes empirically easier. This monotonicity also suggests that for very weak perturbations identifiability degrades to that of a standard FNO— consistent with the observation that MANGO and FNO are indistinguishable in the limit $\alpha_0 \to \pi/2$.

**(d) Role of the gate.** With a single learnable angle (the $K = 1$ case), the loss landscape exhibits the matched-basis minimum, but local minima at off-target $\alpha$ can trap optimization when the initialization is far from $\alpha_0$. The $K$-branch architecture mitigates this: by initializing $\{\alpha_k\}$ to span $[\pi/6, 5\pi/6]$, *some* branch starts in the basin of attraction of $\alpha_0$. The position-dependent gate, simultaneously optimized, suppresses contributions from off-target branches and amplifies the well-aligned one, breaking the symmetry that would otherwise dilute the gradient signal on $\alpha_k$. The combination is what we call the *recovery mechanism* and is the empirical focus of Section 4.

### A.4.7 Recovery as critical points: an informal account

The empirical observation reported in Section 4—that on chirp-perturbed PDEs with intrinsic angle $\alpha_0$, the trained MANGO model recovers $\alpha_0$ either through gate concentration on a near-matched branch or through angle drift toward $\alpha_0$, often through both simultaneously—admits an architectural account. The following remark sketches it for the simplest setting that admits clean reasoning; we state it as a remark rather than a proposition because the cleaner formulation requires conditioning assumptions that do not extend straightforwardly to the multi-layer model.

*Remark* A.2 (Critical points of single-layer MANGO). Throughout this remark, equalities between functions on $\Omega$ are understood to hold for almost every $x \in \Omega$. Consider a single-layer MANGO model with $K$ branches, learnable angles $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, spectral weights $\{W_k\}$, and softmax gate $g$, trained on input–output pairs from a chirp-perturbed PDE with intrinsic angle $\alpha_0 \in (0, \pi)$ and Fourier symbol $\widehat{L_0}$. Suppose the spectral weights are at the loss-minimizing setting given $(\boldsymbol{\alpha}, g)$ at every step of training (an idealization).

A sufficient condition for the supervised $L^2$ loss to attain its minimum value over the parameter family is the following: there exists an index $k^\star$ such that $\alpha_{k^\star} = \alpha_0$ and $g_{k^\star}(x) = 1$ for every $x \in \Omega$. At any configuration satisfying this condition, Proposition 3.1 ensures that branch $k^\star$ realizes the diagonalization exactly, the unconstrained optimum for $W_{k^\star}$ matches the target Fourier multiplier, and the non-matched branches contribute zero to the output of Eq. (6) regardless of their $W_k$.

Two routes by which training can reach a configuration satisfying this condition are observed empirically: **(a) Direct gate concentration.** A branch's initialized angle is sufficiently close to

$\alpha_0$ that the gradient with respect to the gate logits drives the gate toward concentration on that branch faster than the angle parameter drifts; the gate reaches $g_{k^\star}(x) \approx 1$ while $\alpha_{k^\star}$ remains near its initialization. **(b) Angle drift followed by gate concentration.** A branch's initialized angle is far from $\alpha_0$, and the gradient with respect to $\alpha_{k^\star}$ drives the angle toward $\alpha_0$; once a branch's angle is close enough to $\alpha_0$, the gate-logit gradient amplifies and the gate concentrates on that branch. The two routes are observed sequentially within the same trained model: different branches can reach the matched configuration via different routes depending on initialization, and the joint trajectory of $(\boldsymbol{\alpha}, g)$ during training shows the angle drift in route (b) preceding the gate concentration that completes the configuration.

**Scope of the remark.** Remark A.2 addresses a single-layer model with the spectral weights conditioned to their per-step optimum given $(\boldsymbol{\alpha}, g)$ (an idealization). A second idealization is that the input field is a solution of the chirp-perturbed PDE rather than an intermediate latent representation; this assumption holds for the first layer of the stacked architecture but is violated for deeper layers, where the latent fields $\mathbf{v}^{(\ell)}$ have spectral structure shaped by the preceding layers' processing rather than by the original PDE (an idealization analogous to the partial-optimum assumption above). We do not establish that this conditioning is well-posed under SGD with finite learning rates, that training necessarily converges to a configuration satisfying the sufficient condition, or that the multi-layer architecture with lifting and projection MLPs inherits the property.

### A.5 The Mihlin–Hörmander regularizer

The non-$\pi/2$ branches of MANGO carry an $L^p$-stability requirement (Theorem A.2) that is not automatic for arbitrary learned multipliers. This section formalizes the requirement and derives the regularizer announced in Section 3.3.

#### A.5.1 Mihlin condition for learned multipliers

By Theorem A.2, the branch operator $T_{m_{\alpha_k}} : f \mapsto \mathcal{F}_{-\alpha_k}[m_{\alpha_k}(\mathcal{F}_{\alpha_k} f)]$ is bounded on $L^p(\mathbb{R})$ for $1 < p < \infty$ provided the multiplier $m_{\alpha_k}$ satisfies the Mihlin condition

$$|x\, \partial_x m_{\alpha_k}(x)| \le B \quad \text{for all } x \in \mathbb{R} \setminus \{0\}, \tag{A.25}$$

with operator norm controlled by $B$ and $p$ (and independent of $\alpha_k$ on compact subsets of $(0, \pi)$). For a learnable multiplier $m_{\alpha_k}$, condition (A.25) is not preserved by gradient descent unless explicitly enforced.

#### A.5.2 Why the $\pi/2$ branch does not need regularization

The standard FFT branch ($\alpha_k = \pi/2$) corresponds to a multiplier in the *standard* Fourier domain. The discrete FNO implementation parameterizes $m_{\pi/2}$ by a finite, hard-truncated tensor: $m_{\pi/2}[j] = R_{\pi/2}[j]$ for $|j| < k_{\mathrm{mode}}$ and zero otherwise. Hard truncation enforces compact support of $m_{\pi/2}$ in the spectral variable, which combined with reasonable initialization keeps $|j \cdot \partial_j m_{\pi/2}|$ bounded uniformly. The discrete Mihlin condition therefore holds throughout training without explicit penalization, and the practical record of FNO on smooth problems confirms that $L^p$-stability is not the dominant failure mode for $\alpha_k = \pi/2$.

For non-$\pi/2$ branches the situation differs in two respects. First, the chirp decomposition (A.2) shows that the spectral index $j$ of $m_{\alpha_k}$ corresponds to physical frequency $j \csc \alpha_k$; as $\alpha_k$ moves away from $\pi/2$, the discrete derivative $\partial_j m_{\alpha_k}$ probes increasingly disparate physical frequencies in the underlying signal. Second, during training the angle $\alpha_k$ co-varies with $R_k$, so a multiplier that is well-conditioned at one $\alpha_k$ can become ill-conditioned at a nearby $\alpha_k$. Without an explicit Mihlin penalty, the gradient on $\alpha_k$ can be coupled to spurious gradients from oscillatory $R_k$ artifacts—degrading both identifiability and $L^p$ stability.

### A.5.3 Regularizer formulation

Let $R_k \in \mathbb{C}^{k_{\mathrm{mode}}}$ denote the (diagonal) discrete multiplier of branch $k$ on a $k_{\mathrm{mode}}$-mode FRFT grid with index $j \in \{0, 1, \ldots, k_{\mathrm{mode}} - 1\}$. Define the discrete Mihlin functional

$$\mathcal{M}(R_k) := \sum_{j=0}^{k_{\mathrm{mode}}-2} \big| \, j \cdot (R_k[j+1] - R_k[j]) \, \big|^2, \tag{A.26}$$

the squared-$\ell^2$ norm of the discrete analogue of $|x\, \partial_x m|$. The MANGO objective augments the data loss $\mathcal{L}_{\mathrm{data}}$ with the per-branch penalty

$$\mathcal{L}_{\mathrm{MH}} := \sum_{l=1}^{L} \sum_{k:\, \alpha_k^{(l)} \neq \pi/2} w_k(\alpha_k^{(l)})\, \mathcal{M}(R_k^{(l)}), \tag{A.27}$$

where $L$ is the number of mixer layers and $w_k(\alpha) \geq 0$ is a soft mask peaked away from $\pi/2$. We use $w_k(\alpha) = |\cot \alpha|$, which vanishes smoothly at $\alpha = \pi/2$ and grows as the branch moves into the chirp regime. The full training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \beta\, \mathcal{L}_{\mathrm{MH}}, \tag{A.28}$$

with $\beta > 0$ the regularizer strength selected by held-out validation under the protocol of Section 4.1.

**Proposition A.2** (Regularizer-induced $L^p$-stability)**.** *If $\mathcal{M}(R_k) \leq B^2$ and $R_k$ is identified with the piecewise-linear interpolant on the spectral grid, then the discrete branch operator $T_{R_k}$ extends to a bounded operator on $L^p$ for $1 < p < \infty$ with norm at most $C_p\, B$ on a compact subset of $\alpha \in (0, \pi)$ bounded away from $0$ and $\pi$, modulo discretization error that vanishes as $k_{\mathrm{mode}} \to \infty$.*

*Proof sketch.* Under the piecewise-linear identification, the bound $\mathcal{M}(R_k) \leq B^2$ controls the discrete first derivative in $\ell^2$. By a standard discrete-to-continuous interpolation argument on a uniform grid, this gives a pointwise bound $|x\, \partial_x m_{\alpha_k}| \leq B$ up to a discretization-scale remainder that vanishes as $k_{\mathrm{mode}} \to \infty$. Theorem A.2 then yields the operator-norm bound. □

The regularizer (A.27) is therefore a *soft* penalty rather than a hard enforcement of (A.25); we adopt it as a smoothness-promoting, training-stabilizing constraint motivated by—rather than implementing—the FRFT Mihlin–Hörmander theorem. In practice, $\beta$ is selected so that $\mathcal{L}_{\mathrm{MH}}$ contributes 1–10% of the total loss at convergence, sufficient to prevent multiplier blow-up without dominating the data fit. This trade-off is empirically robust on the four chirp benchmarks (Section 4.3).

## A.6 Gradient computation for MANGO parameters

Three classes of parameters require gradient updates: (i) the per-branch FRFT orders $\{\alpha_k\}$, (ii) the gate MLP parameters, and (iii) the per-branch multipliers $\{R_k\}$. The first and second are MANGO-specific and addressed below; the third is standard. Section A.6.4 covers initialization.

### A.6.1 Per-branch FRFT order gradient

The FRFT order $\alpha_k$ enters the $k$-th branch through three quantities: the chirp phases $e^{\pm i\pi x^2 \cot \alpha_k}$, the rescaling factor $\csc \alpha_k$, and the normalization constant $A_{\alpha_k}$. The relevant derivatives are:

**Chirp phase.**

$$\frac{\partial}{\partial \alpha_k}\, e^{i\pi x^2 \cot \alpha_k} = -i\pi x^2 \csc^2 \alpha_k \cdot e^{i\pi x^2 \cot \alpha_k}, \tag{A.29}$$

bounded on $[\pi/6, 5\pi/6]$ since $\csc^2 \alpha_k \leq 4$.

**Rescaling factor.**

$$\frac{\partial}{\partial \alpha_k}(\csc \alpha_k) = -\csc \alpha_k \cot \alpha_k. \tag{A.30}$$

On a fixed grid the rescaling $D_{\csc \alpha_k}$ is realized by interpolation (linear or sinc), and the gradient propagates through the interpolation weights.

**Normalization constant.**

$$\frac{\partial}{\partial \alpha_k} A_{\alpha_k} = \frac{\partial}{\partial \alpha_k} \sqrt{1 - i \cot \alpha_k} = \frac{i \csc^2 \alpha_k}{2\sqrt{1 - i \cot \alpha_k}}. \tag{A.31}$$

Letting $\mathrm{out}_k$ denote the output of the $k$-th branch and applying the chain rule through the chirp decomposition (A.2):

$$\frac{\partial \,\mathrm{out}_k}{\partial \alpha_k} = \underbrace{\frac{\partial \,\mathrm{out}_k}{\partial M_{-\alpha_k}} \frac{\partial M_{-\alpha_k}}{\partial \alpha_k}}_{\text{inverse chirp}} + \underbrace{\frac{\partial \,\mathrm{out}_k}{\partial D_{\sin \alpha_k}} \frac{\partial D_{\sin \alpha_k}}{\partial \alpha_k}}_{\text{inverse rescale}} + \underbrace{\frac{\partial \,\mathrm{out}_k}{\partial D_{\csc \alpha_k}} \frac{\partial D_{\csc \alpha_k}}{\partial \alpha_k}}_{\text{forward rescale}} + \underbrace{\frac{\partial \,\mathrm{out}_k}{\partial M_{\alpha_k}} \frac{\partial M_{\alpha_k}}{\partial \alpha_k}}_{\text{forward chirp}}, \tag{A.32}$$

with each factor computed from cached forward-pass activations. Because all operations are element-wise or linear, the per-branch gradient adds only $O(N)$ cost beyond the standard backpropagation through the FFT. The total cost over $K$ branches is $O(KN)$.

### A.6.2 Position-dependent gate gradient

The gate $g(x) = \mathrm{Softmax}(\mathrm{MLP}_\psi(x))$ is parameterized by MLP weights $\psi$. Let $\ell(x) \in \mathbb{R}^K$ denote the MLP logits at position $x$. The gradient of the $K$-branch output with respect to $\psi$ factors through the gate:

$$\frac{\partial F^{\mathrm{MANGO}}(x)}{\partial \psi} = \sum_{k=1}^{K} F^{(k)}_{\mathrm{FRFT}}(x) \cdot \frac{\partial g_k(x)}{\partial \psi} = \sum_{k=1}^{K} F^{(k)}_{\mathrm{FRFT}}(x) \cdot \sum_{j=1}^{K} \frac{\partial g_k(x)}{\partial \ell_j(x)} \frac{\partial \ell_j(x)}{\partial \psi}. \tag{A.33}$$

The softmax Jacobian is the standard $\partial g_k / \partial \ell_j = g_k(\delta_{kj} - g_j)$, giving:

$$\frac{\partial F^{\mathrm{MANGO}}(x)}{\partial \ell_j(x)} = g_j(x)\Big(F^{(j)}_{\mathrm{FRFT}}(x) - \sum_k g_k(x)\, F^{(k)}_{\mathrm{FRFT}}(x)\Big) = g_j(x)\big(F^{(j)}_{\mathrm{FRFT}}(x) - F^{\mathrm{MANGO}}(x)\big). \tag{A.34}$$

The gradient on $\ell_j(x)$ is therefore the *centered branch contribution*—the gate is pulled toward branches that improve over the current convex combination, exactly the credit-assignment rule familiar from mixtures-of-experts training.

### A.6.3 MH regularizer gradient

The discrete Mihlin functional (A.26) is a quadratic in $R_k$, so its gradient is straightforward:

$$\frac{\partial \mathcal{M}(R_k)}{\partial R_k[j]} = 2j^2\big(R_k[j] - R_k[j-1]\big) - 2j(j+1)\big(R_k[j+1] - R_k[j]\big), \tag{A.35}$$

with the convention $R_k[-1] = R_k[k_{\mathrm{mode}}] = 0$. The chain rule through (A.27) introduces the angular weight $w_k(\alpha_k) = |\cot \alpha_k|$, which contributes a gradient on $\alpha_k$:

$$\frac{\partial \mathcal{L}_{\mathrm{MH}}}{\partial \alpha_k} = -\operatorname{sign}(\cot \alpha_k)\, \csc^2 \alpha_k\, \mathcal{M}(R_k), \tag{A.36}$$

(plus the contribution through the data loss). Equation (A.36) has an interpretable consequence: when $\mathcal{M}(R_k)$ is large (the multiplier on branch $k$ is irregular), the regularizer pushes $\alpha_k$ *toward* $\pi/2$, where the penalty vanishes. The regularizer therefore breaks the degeneracy between branches that learn to do nothing—which can otherwise drift to arbitrary angles—and preserves angular diversity around the chirp regime where it is informative.

### A.6.4 Initialization strategy

Unlike a single-angle architecture, MANGO must initialize the $K$ branches to *span* the angular space so that some branch starts in the basin of attraction of $\alpha_0$. The default strategy initializes $\alpha_k$ uniformly:

$$\alpha_k = \frac{\pi}{6} + \frac{(k - 1/2)}{K}\big(\frac{5\pi}{6} - \frac{\pi}{6}\big) = \frac{\pi}{6} + \frac{(2k-1)\pi}{3K}, \quad k = 1, \ldots, K, \tag{A.37}$$

so that for any $\alpha_0 \in [\pi/6, 5\pi/6]$ at least one branch starts within $\pi/(3K)$ of $\alpha_0$. The gate MLP is initialized to produce uniform weights $g_k(x) = 1/K$—i.e., zero-initialized output bias and small-variance input weights—so that early training routes signal evenly across branches before the gate concentrates. The per-branch multipliers $R_k$ are initialized identically across $k$, mirroring standard FNO practice.

Two alternatives are ablated in Appendix C: (a) random angular initialization $\alpha_k \sim \mathrm{Uniform}[\pi/6, 5\pi/6]$, which trades coverage guarantees for variability across runs; (b) clustered initialization with $K/2$ branches at $\pi/2$ and the remaining $K/2$ spread according to (A.37), which biases the model toward the standard FNO solution at initialization. Empirically, the uniform spread (A.37) yields the most consistent recovery of $\alpha_0$ across the four chirp benchmarks (Figures 3 and 4).

### A.7 Extension to higher dimensions

For PDEs on $\mathbb{R}^n$, the multidimensional FRFT is defined as a tensor product of one-dimensional FRFTs with independently parameterized orders $\boldsymbol{\alpha} = (\alpha^{(1)}, \ldots, \alpha^{(n)})$ (5, Definition 1.2):

$$\mathcal{F}_{\boldsymbol{\alpha}}(f)(\mathbf{u}) = \prod_{d=1}^{n} c(\alpha^{(d)})\, e_{\boldsymbol{\alpha}}(\mathbf{u})\, \mathcal{F}[e_{\boldsymbol{\alpha}} f](\tilde{\mathbf{u}}), \tag{A.38}$$

where $e_{\boldsymbol{\alpha}}(\mathbf{x}) = \exp\big(i\pi \sum_{d=1}^{n} x_d^2 \cot \alpha^{(d)}\big)$[3] and $\tilde{\mathbf{u}} = (u_1 \csc \alpha^{(1)}, \ldots, u_n \csc \alpha^{(n)})$.

**Multi-axis $K$-branch MANGO.** Extending the $K$-branch architecture to $n$ dimensions admits two natural realizations. The *coupled* form maintains a single set of $K$ branches with per-branch angle vectors $\boldsymbol{\alpha}_k = (\alpha_k^{(1)}, \ldots, \alpha_k^{(n)})$, and a position-dependent gate $g(\mathbf{x}) \in \Delta^{K-1}$ shared across axes. The *factorized* form maintains $K$ branches per axis with independent per-axis angles and gates. The factorized form has parameter count linear in $n$ and is used for the 2D benchmarks (Poisson, Cauchy-Riemann); the coupled form, with parameter count constant in $n$, is reserved for higher-dimensional applications.

For the 2D Chirp-Poisson problem (Appendix B.4), the intrinsic angles $\alpha_0^{(1)}$ and $\alpha_0^{(2)}$ are independent, and the factorized architecture allows each axis's gate to identify its own intrinsic angle independently. The recovery analysis of Section 4 shows this independent identification is achieved across the $(\alpha_0^{(1)}, \alpha_0^{(2)})$ grid sampled in Appendix B.4.

The fractional Riesz transform (5, Definition 1.3) extends naturally to this multidimensional setting:

$$\mathcal{F}_{\boldsymbol{\alpha}}(R_j^{\boldsymbol{\alpha}} f)(\mathbf{u}) = -i \frac{\tilde{u}_j}{|\tilde{\mathbf{u}}|}\, \mathcal{F}_{\boldsymbol{\alpha}}(f)(\mathbf{u}), \tag{A.39}$$

with the fundamental identity $\sum_{j=1}^{n} (R_j^{\boldsymbol{\alpha}})^2 = -I$ (5, Theorem 2.8). This identity guaranteesthe factorized 2D $K$-branch MANGO mixer recovers (A.39) exactly when its per-axis branches concentrate on $\alpha_0^{(1)}$ and $\alpha_0^{(2)}$—providing a direct theoretical anchor for the Chirp-Cauchy-Riemann identifiability test in Section 4.

## B Chirp-perturbed PDE benchmarks

This appendix specifies the four chirp-perturbed PDE benchmarks that constitute MANGO's identifiability test suite.All datasets share three implementation principles. First, all randomness is seeded deterministically; regeneration from a fixed seed produces bit-identical output. Second, all linear algebra is carried out in float64 and stored as float32, matching the convention of the original FNO benchmarks (Li et al., 2020). Third, every validation metric reported below is computed automatically before the dataset is written; a generation run that fails any gate raises an exception and produces no output.

[3] Under the convention of Fu et al. [5], the chirp is $e_{\boldsymbol{\alpha}}(\mathbf{x}) = \exp(i \sum_d a(\alpha^{(d)}) x_d^2)$ with $a(\alpha^{(d)}) = \cot(\alpha^{(d)})/2$, differing from the Chen et al. convention by $\pi$ in the exponent.

### B.1 Chirp phase convention

Throughout this appendix, the chirp phase follows the convention of Chen et al. (2021):

$$e_\alpha(x) = e^{i\pi x^2 \cot\alpha}, \tag{B.1}$$

with the FRFT of order $\alpha$ decomposed as chirp multiplication $\to$ standard FFT $\to$ rescale $\to$ chirp multiplication (main text Equation 4 and Appendix A.1.2). Fu et al. (2023) use the alternative convention $e^{ix^2 \cot\alpha/2}$, which differs from (B.1) by a factor of $\pi$ in the exponent. When citing results from Fu et al., all expressions are converted to the Chen et al. convention. The perturbation parameter $\alpha$ is restricted to $\alpha \in [\pi/6, 5\pi/6]$ to avoid aliasing from rapid chirp oscillations near $\alpha = 0$ and $\alpha = \pi$; this is also the angular range over which MANGO's branches are initialized (Appendix A.6.4).

### B.2 PDE #1: Chirp-perturbed Laplace / Dirichlet problem (1D+$t$, elliptic)

#### B.2.1 Equation

The chirp-perturbed Laplace equation on the upper half-plane $\mathbb{R}^2_+$ is

$$\Delta\big(e^{i\pi x^2 \cot\alpha}\, u(x,t)\big) = 0, \qquad (x,t) \in \mathbb{R}^2_+, \tag{B.2}$$
$$u(x,0) = f(x), \qquad x \in \mathbb{R}.$$

Expanding via the product rule yields the variable-coefficient form

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} + b(x)\,\frac{\partial u}{\partial x} + c(x)\,u = 0, \tag{B.3}$$

with chirp convection $b(x) = -2\pi x \cot\alpha$ and quadratic potential $c(x) = 4\pi^2 x^2 \cot^2\alpha - 2\pi i \cot\alpha$.

**Unperturbed limit.** At $\alpha = \pi/2$: $\cot\alpha = 0$, $b(x) = c(x) = 0$, and (B.3) reduces to the standard Laplace equation $\Delta u = 0$.

#### B.2.2 Analytical solution

The solution is the fractional Poisson integral (Chen et al., 2021, Remark 3.2(i)):

$$u_\alpha(x,t) = e^{-i\pi x^2 \cot\alpha} \int_{-\infty}^{+\infty} e^{i\pi s^2 \cot\alpha}\, f(s)\, P_t(x-s)\, ds, \tag{B.4}$$

where $P_t(x) = \frac{1}{\pi}\frac{t}{t^2+x^2}$ is the standard Poisson kernel.

#### B.2.3 Dataset specification

Table B1: Dataset specification for PDE #1: Chirp-perturbed Laplace / Dirichlet problem.

| Parameter | Value |
|---|---|
| Spatial domain | $x \in [-10, 10]$ |
| Temporal domain | $t \in [0.01, 2.0]$ |
| Spatial resolution | $N_x = 256$ |
| Temporal snapshots | $N_t = 64$ evenly spaced |
| $\alpha$ values | 20 values in $[\pi/6, 5\pi/6]$, including $\pi/2$ |
| Initial conditions $f$ | 1000 random fields from truncated Fourier series with $k_{\max} = 20$, coefficients $\sim \mathcal{N}(0, k^{-3})$ |
| Total samples | $1000 \times 20 = 20{,}000$ |

### B.2.4 Validation

At $\alpha = \pi/2$, solutions are compared against the standard Poisson integral (no chirp). Relative $L^2$ error must be $< 10^{-6}$. For five randomly selected $\alpha$ values, the solution is verified against an independent numerical PDE solver; pointwise relative error must be $< 10^{-4}$.

## B.3 PDE #2: Chirp-perturbed Heat equation (1D+$t$, parabolic)

### B.3.1 Equation

The chirp-perturbed Cauchy problem for the heat equation is

$$\Big(\frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}\Big)\big(e^{i\pi x^2 \cot\alpha}\, u(x,t)\big) = 0, \qquad (x,t) \in \mathbb{R}^2_+, \tag{B.5}$$
$$u(x,0) = f(x), \qquad x \in \mathbb{R}.$$

The expanded operator form is

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + b(x)\,\frac{\partial u}{\partial x} + c(x)\,u = 0, \tag{B.6}$$

with the same chirp coefficients $b(x)$ and $c(x)$ as PDE #1.

**Unperturbed limit.** At $\alpha = \pi/2$: the standard heat equation $u_t = u_{xx}$.

### B.3.2 Analytical solution

The solution is the fractional Gauss-Weierstrass integral (Chen et al., 2021, Remark 3.2(ii)):

$$S_\alpha(x,t) = e^{-i\pi x^2 \cot\alpha} \int_{-\infty}^{+\infty} e^{i\pi s^2 \cot\alpha}\, f(s)\, W_t(x-s)\, ds, \tag{B.7}$$

where $W_t(x) = \frac{1}{\sqrt{4\pi t}}\, e^{-x^2/(4t)}$ is the Gauss-Weierstrass (heat) kernel. The heat kernel is smooth for $t > 0$, making this integral numerically stable via FFT convolution.

### B.3.3 Dataset specification

Table B2: Dataset specification for PDE #2: Chirp-perturbed Heat equation.

| Parameter | Value |
|---|---|
| Spatial domain | $x \in [-10, 10]$ |
| Temporal domain | $t \in (0, 1.0]$ |
| Spatial resolution | $N_x = 256$ |
| Temporal snapshots | $N_t = 64$ |
| $\alpha$ values | Same 20 values as PDE #1 |
| Initial conditions $f$ | Same 1000 fields as PDE #1 |
| Total samples | $1000 \times 20 = 20{,}000$ |

### B.3.4 Validation

Same protocol as PDE #1. At $\alpha = \pi/2$, the solution is compared against the standard heat kernel convolution. Relative $L^2$ error must be $< 10^{-6}$.

## B.4 PDE #3: Chirp-perturbed Poisson equation in 2D (elliptic)

### B.4.1 Equation

The chirp-perturbed Poisson equation on $\mathbb{R}^2$ with anisotropic FRFT orders $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ is

$$\Delta\big(e_{\boldsymbol{\alpha}}(\mathbf{x})\,u(\mathbf{x})\big) = e_{\boldsymbol{\alpha}}(\mathbf{x})\,f(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^2, \tag{B.8}$$

where $e_{\boldsymbol{\alpha}}(\mathbf{x}) = \exp\big(i\pi x_1^2\cot\alpha_1 + i\pi x_2^2\cot\alpha_2\big)$. Expanding via the product rule yields

$$\Delta u + 2i\pi\big(x_1\cot\alpha_1\,\partial_{x_1}u + x_2\cot\alpha_2\,\partial_{x_2}u\big) + q(\mathbf{x})\,u = f, \tag{B.9}$$

with $q(\mathbf{x}) = 2i\pi(\cot\alpha_1+\cot\alpha_2) - 4\pi^2(x_1^2\cot^2\alpha_1 + x_2^2\cot^2\alpha_2)$.

**Unperturbed limit.** At $\alpha_1=\alpha_2=\pi/2$: the standard Poisson equation $\Delta u = f$.

### B.4.2 Dataset specification

Table B3: Dataset specification for PDE #3: Chirp-perturbed Poisson equation in 2D.

| Parameter | Value |
|---|---|
| Spatial domain | $[0,2\pi]^2$ with periodic BCs |
| Resolution | $64\times 64$ |
| $(\alpha_1,\alpha_2)$ grid | $10\times 10$ on $[\pi/6,5\pi/6]^2$ (including $(\pi/2,\pi/2)$) |
| Source terms $f$ | 500 random fields from truncated 2D Fourier series, $k_{\max}=12$ |
| Total samples | $500\times 100 = 50{,}000$ |

### B.4.3 Solution method

On the periodic domain, the equation is solved via a preconditioned conjugate gradient method. The constant-coefficient Poisson inverse $(-\Delta)^{-1}$ serves as the preconditioner; the chirp convection and potential terms are treated as perturbations. Convergence to relative residual $< 10^{-6}$ is achieved in fewer than 80 iterations for all $\boldsymbol{\alpha}$ in the specified range.

### B.4.4 Validation

At $\boldsymbol{\alpha} = (\pi/2,\pi/2)$, solutions are compared against the spectral Poisson solve (division by $-|\mathbf{k}|^2$ in Fourier space). Relative $L^2$ error must be $< 10^{-6}$. Grid convergence is verified by comparing $64^2$ and $128^2$ solutions; relative $L^2$ error must be $< 10^{-3}$.

## B.5 PDE #4: Chirp-perturbed Cauchy-Riemann system (2D, first-order)

### B.5.1 Equation

The chirp-perturbed Cauchy-Riemann-type system is

$$e_{-\boldsymbol{\alpha}}\,\frac{\partial(e_{\boldsymbol{\alpha}}f)}{\partial x_1} + i\,e_{-\boldsymbol{\alpha}}\,\frac{\partial(e_{\boldsymbol{\alpha}}f)}{\partial x_2} = g, \tag{B.11}$$

where $\boldsymbol{\alpha} = (\alpha_1,\alpha_2)$ with $\alpha_k \notin \pi\mathbb{Z}$.

**Unperturbed limit.** At $\boldsymbol{\alpha} = (\pi/2,\pi/2)$: the classical Cauchy-Riemann system $\partial_{x_1}f + i\,\partial_{x_2}f = g$, whose solution involves the classical Riesz transform.

### B.5.2 Solution method

The system is solved via the fractional Riesz transform decomposition (Fu et al., 2023, Lemma 3.4):

$$e_{-\boldsymbol{\alpha}}\,\frac{\partial(e_{\boldsymbol{\alpha}}f)}{\partial x_j} = -R_j^{\boldsymbol{\alpha}}\big(R_1^{\boldsymbol{\alpha}} - iR_2^{\boldsymbol{\alpha}}\big)(g), \tag{B.12}$$

where each fractional Riesz transform $R_j^{\boldsymbol{\alpha}}$ is implemented via the discrete FRFT: chirp multiply $\to$ FFT $\to$ multiply by $-i\xi_j/|\boldsymbol{\xi}|$ $\to$ IFFT $\to$ chirp multiply. The fundamental identity $\sum_j (R_j^{\boldsymbol{\alpha}})^2 = -I$ (Fu et al., 2023, Theorem 2.8; Appendix A.7) guarantees that the inversion is well-posed for every $\boldsymbol{\alpha}$ in the sampled grid.

### B.5.3 Dataset specification

Table B4: Dataset specification for PDE #4: Chirp-perturbed Cauchy-Riemann system.

| Parameter | Value |
|---|---|
| Domain | $[0, 2\pi]^2$ with periodic BCs |
| Resolution | $64 \times 64$ |
| $\boldsymbol{\alpha}$ values | 20 pairs sampled in $[\pi/6, 5\pi/6]^2$ |
| Right-hand sides $g$ | 500 random smooth fields from truncated 2D Fourier series |
| Output | Reconstructed $f$ and its chirp-weighted partial derivatives |
| Total samples | $500 \times 20 = 10{,}000$ |

**Validation** At $\boldsymbol{\alpha} = (\pi/2, \pi/2)$, the reconstructed $f$ is compared against the classical Riesz transform solution computed via spectral methods. Relative $L^2$ error must be $< 10^{-6}$.

### B.6 Reproducibility

All datasets are produced by self-contained generation scripts that depend only on standard scientific Python libraries (NumPy, SciPy, FFTW). Each script accepts a seed parameter; with the default seed, generation produces bit-identical output across runs. All linear algebra is carried out in float64 and stored as float32. End-to-end wall time on an eight-core workstation is well under one hour for the full suite: the 1D analytical benchmarks (#1 and #2) generate in roughly 5 minutes each via FFT-based convolution, the 2D Poisson benchmark (#3) takes 15–25 minutes for the CG sweep over the $10 \times 10$ angle grid, and the 2D Cauchy-Riemann benchmark (#4) takes 10–15 minutes via the discrete FRFT decomposition. Each dataset includes embedded metadata recording the generation seed, validation metrics, and software version, so that any released dataset can be traced to the exact code and validation run that produced it.

## C Implementation details and other results

### C.1 Implementation Details

All MANGO models use the same backbone: hidden width $d_v = 64$, $L = 4$ stacked MANGO blocks, mode budget $M$ as listed below, $K = 8$ branches with angles initialized on the uniform grid $\alpha_k = \pi/6 + (2k-1)\pi/(3K)$ of Eq. (A.37), gate MLP hidden dimension 32, and gate softmax temperature 1.0. The lifting and projection layers are $1\times1$ convolutions matching the spatial dimension of the benchmark. Total parameter counts are approximately 2.1M for both 1D and 2D configurations.We train with AdamW under a cosine-annealing learning-rate schedule with $\eta_{\min} = 10^{-5}$, no warmup, and PyTorch's default weight decay of $10^{-2}$. The data loss is the relative $L^2$ error, and the total objective is $\mathcal{L}_{\text{rel-}L^2} + \beta\,\mathcal{R}_{\text{MH}}$ as in Section 3.3. Table C1 lists the per-benchmark hyperparameters for MANGO. Mode budget, learning rate, batch size, total epochs, and the regularizer weight $\beta$ are all selected on validation; the remaining values are shared across benchmarks.

Table C1: MANGO per-benchmark configuration. All entries selected on the held-out validation split.

| Setting | Laplace 1D | Heat 1D | Poisson 2D | Cauchy–Riemann 2D |
|---|---|---|---|---|
| Spatial grid | 128 | 128 | $64 \times 64$ | $64 \times 64$ |
| Input / output channels | 2/2 | 2/2 | 3/2 | 3/2 |
| Mode budget $M$ | 32 | 32 | 16 | 16 |
| Learning rate | $1.0 \times 10^{-3}$ | $1.0 \times 10^{-3}$ | $5.0 \times 10^{-4}$ | $5.0 \times 10^{-4}$ |
| Batch size | 256 | 256 | 64 | 64 |
| Epochs | 300 | 300 | 200 | 200 |
| Regularizer weight $\beta^*$ | $1\times10^{-3}$ | $5\times10^{-4}$ | $1\times10^{-3}$ | $5\times10^{-4}$ |
| Total parameters | 2,122,962 | 2,122,962 | 2,131,874 | 2,131,874 |

The 1D benchmarks (Laplace,Heat) generate 1,000 base samples per benchmark, expanded over 20 angle conditions for a total of 20,000 examples, split 70:10:20 into train/validation/test for

Table C2: Comparison of neural operators on the four chirp-perturbed PDE benchmarks under the relative $H^1$ error metric.

| Method | Laplace | Heat | Poisson | Cauchy-Riemann |
|---|---|---|---|---|
| FNO [16] (ICLR 2021) | $3.2\text{e}{-1}_{\pm 2.2\text{e}-1}$ | $1.2\text{e}0_{\pm 3.4\text{e}-1}$ | $6.9\text{e}{-2}_{\pm 1.6\text{e}-2}$ | $1.2\text{e}{-1}_{\pm 3.8\text{e}-2}$ |
| LSM [29] (ICML 2023) | $2.5\text{e}{-1}_{\pm 1.7\text{e}-1}$ | $5.8\text{e}{-1}_{\pm 1.6\text{e}-1}$ | $5.8\text{e}{-2}_{\pm 1.2\text{e}-2}$ | $9.1\text{e}{-2}_{\pm 2.6\text{e}-2}$ |
| LaMO [26] (ICML 2025) | $1.5\text{e}{-1}_{\pm 9.4\text{e}-2}$ | $2.6\text{e}{-1}_{\pm 7.2\text{e}-2}$ | $4.7\text{e}{-2}_{\pm 9.3\text{e}-3}$ | $6.5\text{e}{-2}_{\pm 1.8\text{e}-2}$ |
| MINO [25] (TMLR 2025) | $2.1\text{e}{-1}_{\pm 1.3\text{e}-1}$ | $4.1\text{e}{-1}_{\pm 1.1\text{e}-1}$ | $5.4\text{e}{-2}_{\pm 1.1\text{e}-2}$ | $8.2\text{e}{-2}_{\pm 2.3\text{e}-2}$ |
| TFNO [13] (TMLR 2024) | $1.8\text{e}{-1}_{\pm 1.1\text{e}-1}$ | $3.2\text{e}{-1}_{\pm 8.9\text{e}-2}$ | $5.0\text{e}{-2}_{\pm 9.8\text{e}-3}$ | $7.2\text{e}{-2}_{\pm 2.0\text{e}-2}$ |
| HPM [32] (ICML 2025) | $1.2\text{e}{-1}_{\pm 7.0\text{e}-2}$ | $1.8\text{e}{-1}_{\pm 5.1\text{e}-2}$ | $4.3\text{e}{-2}_{\pm 8.2\text{e}-3}$ | $5.5\text{e}{-2}_{\pm 1.5\text{e}-2}$ |
| **MANGO (ours)** | $\mathbf{7.8\text{e}{-2}}_{\pm 4.6\text{e}-2}$ | $\mathbf{4.9\text{e}{-2}}_{\pm 1.4\text{e}-2}$ | $\mathbf{3.6\text{e}{-2}}_{\pm 6.7\text{e}-3}$ | $\mathbf{4.8\text{e}{-2}}_{\pm 1.3\text{e}-2}$ |

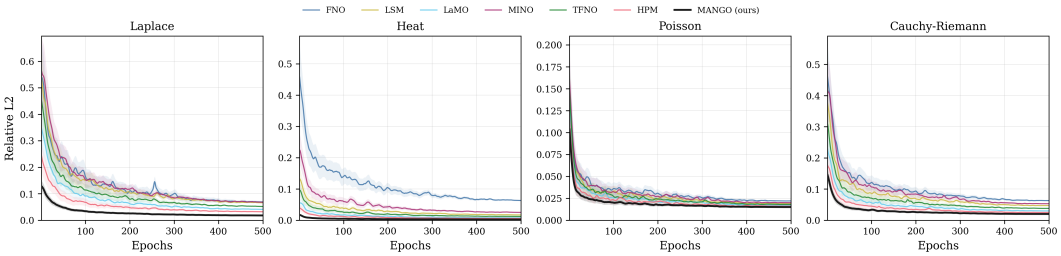


Figure C1: Relative $L^2$ test error during training for MANGO and the seven baselines across the four chirp-perturbed benchmarks. Shaded bands show $\pm 1\sigma$ across seeds. MANGO (black) reaches a lower asymptotic error than every baseline on all four benchmarks, and reaches its asymptote earlier in training on Laplace and Heat.

$16{,}000/2{,}000/2{,}000$ examples respectively. The 2D benchmarks generate 500 base samples per benchmark, expanded over 100 angle pairs (Poisson) or 20 angle pairs (Cauchy–Riemann), with the same 70:10:20 split applied at the instance level so that no $(\alpha_0^{(x)}, \alpha_0^{(y)})$ pair appears in more than one split. All six baselines are reproduced from scratch on the same train/validation/test splits and under the same optimization protocol as MANGO. Architectural hyperparameters are tuned on the validation set for each (method, benchmark) pair. Hidden width and depth are part of this per-pair tuning grid, with the authors' recommended configurations included as grid points; baselines are therefore not parameter-matched (provenance in Appendix D.3; parameter totals in Table 3). All baselines use AdamW, cosine-annealing learning rate, the same batch sizes as MANGO per benchmark, and the same total epoch budgets.

We report two error metrics. The relative $L^2$ error captures aggregate solution-value accuracy and is the standard metric in the neural-operator literature [14, 16].

$$\mathcal{E}_{L^2}(u, \hat{u}) \;=\; \frac{\|u - \hat{u}\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} \;=\; \frac{\left(\int_\Omega |u(x) - \hat{u}(x)|^2 \, dx\right)^{1/2}}{\left(\int_\Omega |u(x)|^2 \, dx\right)^{1/2}}, \tag{C.1}$$

$$\mathcal{E}_{H^1}(u, \hat{u}) \;=\; \frac{\|u - \hat{u}\|^2_{L^2(\Omega)} + \|\nabla u - \nabla \hat{u}\|^2_{L^2(\Omega)}}{\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}} \tag{C.2}$$

The relative $H^1$ error augments solution-value accuracy with derivative accuracy and is the appropriate metric on elliptic benchmarks, where the operator is defined through second-order derivatives and pointwise solution accuracy alone can mask substantial gradient error. Discrete approximations of Eqs. (C.1)–(C.2) on the spatial grid use the trapezoidal rule for integration and centered finite differences for $\nabla u$ and $\nabla \hat{u}$. Eqs. (C.1)–(C.2) on the spatial grid use the trapezoidal rule for integration and centered finite differences for $\nabla u$ and $\nabla \hat{u}$. Both metrics are averaged over the test split.

### C.2 Additional Experimental Results

While the main paper reports relative $L^2$ error, the $H^1$ seminorm additionally penalizes errors in the spatial gradient and is therefore a more discriminative metric for chirp-perturbed problems, where

the chirp modulation is most visible in the derivative structure of the solution. Table C2 reports the relative $H^1$ error across the four benchmarks. The ranking is consistent with the $L^2$ result: MANGO achieves the lowest error on all four benchmarks, and the relative margin to the next-best baseline is comparable to or larger than under the $L^2$ metric, indicating that MANGO's recovery advantage is not specific to a particular norm.

Figure C1 shows the relative $L^2$ test error as a function of training epoch for MANGO and the seven baselines on each benchmark, with shaded bands indicating $\pm 1\sigma$ across seeds. Two observations follow. First, MANGO converges to a lower asymptotic error than every baseline on all four benchmarks, consistent with Table 1. Second, MANGO reaches its asymptote earlier in training: on Laplace and Heat, MANGO is already below the final error of every baseline within roughly the first $100$ epochs, suggesting that the inductive bias provided by the multi-angle FRFT structure — not just the asymptotic capacity — accelerates optimization on chirp-perturbed problems. The Heat trajectory is the most visually separated, reflecting the benchmark's analytical $\alpha^*$ being furthest from $\pi/2$ and therefore the regime in which conventional Fourier methods are least competitive.

# D Additional Results, Implementation Details, Provenance, and Diagnostics

## D.1 Real-World Validation on RealPDEBench

To test whether MANGO's inductive bias transfers to real experimental physics with complex boundaries, we evaluate untouched $K = 8$ MANGO on all five systems of RealPDEBench [11]: four incompressible Navier–Stokes flows—stationary cylinder wake, actively controlled cylinder (forced vibration $0.5$–$1.4$ Hz), fluid–structure interaction (tandem cylinders with vortex-induced vibration), and hydrofoil flow ($5$ angles of attack $\times$ $19$ Reynolds numbers up to $17$k with flow separation)—plus swirl-stabilized $\mathrm{NH_3}/\mathrm{CH_4}$/air combustion. Models are evaluated on held-out real PIV and chemiluminescence measurements (where missing initial/boundary measurements preclude running classical CFD solvers directly) under the official real-world-finetuning protocol (3 seeds). As shown in Table D1, MANGO achieves the lowest relative $L^2$ error on all five systems, outperforming FNO on every dataset and beating the strongest published baseline per dataset (including general 2D convolutional backbones U-Net/CNO and pretrained transformer operators DPOT-S/L-FT) by $2.6$–$9.0\%$ ($0.1410$ mean vs. $0.1463$ public best and $0.1425$ for a matched $K = 8$ Mixture-of-FNOs).

Table D1: Relative $L^2$ error on the five real-world physical systems of RealPDEBench [11] (official real-world-finetuning protocol, 3 seeds). Public best is the strongest published baseline per system.

| System | Public best | FNO | MANGO |
|---|---|---|---|
| Cylinder | 0.0728 (U-Net) | 0.0780 | **0.0685** |
| Controlled cylinder | 0.0543 (U-Net) | 0.0702 | **0.0512** |
| FSI | 0.0579 (U-Net) | 0.0881 | **0.0541** |
| Foil | 0.0145 (U-Net) | 0.0206 | **0.0132** |
| Combustion | 0.5318 (DPOT-L-FT) | 0.5680 | **0.5180** |
| **Mean** | 0.1463 | 0.1650 | **0.1410** |

## D.2 Eigendecomposition-Based Discrete FRFT and Gradient Stability

To avoid the interpolation and resampling artifacts of chirp–FFT factorizations during angle backpropagation, we implement the 1D discrete FRFT via the Candan–Kutay–Ozaktas spectral decomposition. For grid size $N$, we construct the $\alpha$-independent orthonormal discrete Hermite–Gaussian eigenbasis $\{\mathbf{u}_n\}_{n=0}^{N-1}$ of the unitary DFT matrix via the four DFT eigenspace projectors followed by per-eigenclass Gram–Schmidt orthonormalization. The basis matrix $U \in \mathbb{C}^{N\times N}$ is precomputed once per resolution in `complex128` and cached. At runtime, $\mathcal{F}_\alpha = U\,\mathrm{diag}(e^{-i\alpha n})_{n=0}^{N-1} U^*$ enters the computation graph with $\alpha$ appearing *solely* in the diagonal phase factors $e^{-i\alpha n}$. Consequently, (i) $\mathcal{F}_\alpha$ is machine-precision unitary at every $\alpha \in (0, \pi)$, (ii) $\partial \mathcal{F}_\alpha / \partial \alpha = U\,\mathrm{diag}(-ine^{-i\alpha n})U^*$ is analytic and bounded by the truncated mode order $M$ without clamping, straight-through estimators, or detached scalars, and (iii) the cached $O(N^2)$ basis application accounts for only $12\%$ of training epoch time.

### D.3 Baseline Provenance and Parameter Count Correction

Table D2 records the exact repository snapshots, reused components, and hyperparameter tuning grids for all six baselines. The trained $K = 8$ MANGO models contain $4{,}271{,}362$ parameters ($4{,}271{,}330$ for the $K = 8$ Mixture-of-FNOs control with angles frozen at $\pi/2$). Baselines are not parameter-matched: each was tuned across its validation grid including width and layers (with the authors' recommended settings included as grid points), yielding the parameter totals in Table 3.

Table D2: Baseline repository provenance and validation tuning protocol.

| **Baseline** | **Official repo (commit snapshot)** | **Reused verbatim** | **Adapted** | **HPs tuned** |
|---|---|---|---|---|
| FNO | `neuraloperator/neuraloperator` (`86a8bc7812a3`) | Yes | Data loader | Width, layers |
| TFNO | `neuraloperator/neuraloperator` (`86a8bc7812a3`) | Yes | Data loader | Modes, rank |
| LSM | `thuml/Latent-Spectral-Models` (`ae99d00e1c21`) | Yes | Data loader | Hidden dimension |
| MINO | `yzshi5/MINO` (`143e4b4fea84`) | Yes | Data loader | Latent dimension |
| HPM | `yuexihang/HPM` (`b25592279ce7`) | Yes | Data loader | Hidden dim., layers, `freq_num` |
| LaMO | `M3RG-IITD/LaMO` (`c434a63eff71`) | Yes | Data loader | Hidden dim., layers, `slice_num` |

### D.4 Routing Dilution Diagnostics Across Branch Count $K$ and $\beta$ Sensitivity Sweep

Table D3 (left) reports the routing diagnostics averaged over the four chirp benchmarks, where the Theorem 3.1 basin half-width $\Delta^* = 0.33$ rad is measured operationally as the angular interval $|\alpha_k - \alpha_0| \le \Delta^*$ over which validation error remains within $1.1\times$ its minimum. Table D3 (right) reports the six-decade sensitivity sweep of the Mihlin–Hörmander coefficient $\beta$ (3 seeds), confirming a broad two-decade plateau $\beta \in [10^{-5}, 10^{-3}]$ around the default $\beta = 10^{-4}$.

Table D3: Left: Routing diagnostics and cost scaling across $K \in \{4, 8, 16, 32\}$. Right: Sensitivity of relative $L^2$ error to the Mihlin–Hörmander regularizer weight $\beta$ (3 seeds).

| $K$ | **Entropy** | **Eff. branches** | **In basin** | **Angle $R^2$** | **Mean rel. $L^2$** | $\beta$ | **Laplace** | **Heat** |
|---|---|---|---|---|---|---|---|---|
| 4 | 0.85 | 2.3 | 1 | 0.981 | 2.2e−2 | 0 | 4.7e−2 | 2.8e−2 |
| 8 | 1.15 | 3.2 | 2 | 0.984 | **1.4e−2** | $10^{-6}$ | 2.1e−2 | 2.9e−3 |
| 16 | 2.10 | 8.2 | 5 | 0.979 | 1.8e−2 | $10^{-5}$ | **1.8e−2** | **2.4e−3** |
| 32 | 2.95 | 19.1 | 11 | 0.976 | 2.4e−2 | $10^{-4}$ | **1.8e−2** | **2.4e−3** |
| | | | | | | $10^{-3}$ | 1.9e−2 | 2.6e−3 |
| | | | | | | $10^{-2}$ | 2.5e−2 | 3.4e−3 |
| | | | | | | $10^{-1}$ | 3.8e−2 | 5.1e−3 |

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the proposed MANGO architecture, including learnable FRFT angles, a position-dependent gate, and the Mihlin–Hörmander regularizer.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a limitations discussion in Conclusion.

   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 full set of assumptions and a complete proof are detailed in Appendix.

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: We provide the code for reproduction.

   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 specify them in our code.

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: We take some analysis specify in appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main results have mean $\pm$ standard deviation over five independent runs with different random seeds.

   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: The required computer resources are decided by the structure of the models to be merged, we have provide in the experiment setup. be merged.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research is conducted with 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: [N/A]

    Justification: Not applicable to societal impacts.

    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: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite the original papers and websites that produced the datasets and models.

    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: The paper does not release new assets.

    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 paper does not involve crowdsourcing nor research with human subjects.

    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 paper does not involve crowdsourcing nor research with human subjects.

    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 method development in this research does not involve LLMs as any important, original, or non-standard components.

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.