Vision Transformer for Multi-Domain Phase Retrieval in Coherent Diffraction Imaging

1. DETAILS OF MULTI-SCALE FOURIER ATTENTION

The spectral mixer introduced in Sec. 2.1.2 is defined here. Let $\mathbf{X} \in \mathbb{R}^{B\times E\times H\times W}$ with $H{=}W{=}16$ denote the token feature map inside a transformer block. At scale s we downsample by average pooling, apply a per-channel Fourier gate, and upsample back:

$$\mathbf{Z}_s = \mathcal{U}_s\Big(\mathcal{F}^{-1}\big(\,\mathcal{F}(\mathcal{D}_s\mathbf{X}) \,\odot\, \mathbf{W}_s \,\odot\, \mathbf{M}_s\,\big)\Big), \tag{S1}$$

where $\mathcal{D}_s$ is average pooling with stride s to $h_s{\times}w_s = (H/s){\times}(W/s)$, $\mathcal{U}_s$ upsamples back to (H, W), $\mathcal{F}$ is the channel-wise 2D FFT, $\mathbf{W}_s \in \mathbb{R}^{E\times h_s\times w_s}$ are learned per-channel frequency responses, and $\mathbf{M}_s \in \mathbb{R}^{1\times h_s\times w_s}$ is a shared gate. The multi-scale output is

$$\mathbf{Z} = \sum_s \alpha_s\, \mathbf{Z}_s, \qquad \boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{\theta}), \tag{S2}$$

with scale weights α_s parameterised by $\boldsymbol{\theta} \in \mathbb{R}^{|\{s\}|}$.

By the convolution theorem,

$$\mathbf{Z}_s = \mathcal{U}_s\big(\,\mathcal{D}_s\mathbf{X} \,*\, \mathbf{g}_s\,\big), \quad \mathbf{g}_s = \mathcal{F}^{-1}(\mathbf{W}_s \odot \mathbf{M}_s), \tag{S3}$$

with $*$ denoting circular convolution on the $h_s{\times}w_s$ grid. Each scale applies the learned kernel $\mathbf{g}_s$, upsamples the result to (H, W), and combines the scales using the weights α_s. This gives global token mixing via FFT-based convolution. In frequency space, $\mathbf{W}_s \odot \mathbf{M}_s$ is a multiplicative gate on the Fourier modes.

Let $N = HW$ denote the number of tokens. For each transformer block, the spectral mixer requires $\mathcal{O}(|\{s\}|\, B\, E\, N \log N)$ time and $\mathcal{O}(|\{s\}|\, B\, E\, N)$ memory, compared with $\mathcal{O}(B\, E\, N^2)$ time and $\mathcal{O}(B\, N^2)$ memory for dot-product attention. This follows Fourier-based token mixing (FNet) [1] and spectral parameterisations used in Fourier neural operators [2], but in our model it appears only as a compact multi-scale gate inside each transformer block.

2. AMPLITUDE PRIOR AND NORMALISATION

The decoder predicts a real-space amplitude $\hat{A}(x, y)$ on the 64×64 grid, together with a phase field as described in Sec. 2.1.3. Before forming the complex density we combine $\hat{A}(x, y)$ with a simple prior amplitude $A_{\text{prior}}(x, y)$ defined on the real-space support $S(x, y)$, where $S(x, y) = 1$ inside the support region and $S(x, y) = 0$ outside. From this support we construct one of three priors: (i) a flat prior with $A_{\text{prior}}(x, y) = S(x, y)$; (ii) a blurred prior obtained by repeated local averaging of $S(x, y)$; or (iii) a Gaussian prior centred on the support and truncated by $S(x, y)$.

During training we blend the prior and predicted amplitudes via a scalar weight $\alpha(t)$:

$$A_{\text{blend}}(x, y; t) = \alpha(t)\, A_{\text{prior}}(x, y) + \big(1 - \alpha(t)\big)\, \hat{A}(x, y). \tag{S4}$$

This keeps early epochs guided by the support-based prior and later epochs dominated by the network prediction. We use a cosine schedule with a lock epoch t_{lock} and a decay-end epoch t_{end}:

$$\alpha(t) = \begin{cases} 1, & t < t_{\text{lock}}, \\ \dfrac{1}{2}\Big(1 + \cos\dfrac{\pi\,(t - t_{\text{lock}})}{t_{\text{end}} - t_{\text{lock}}}\Big), & t_{\text{lock}} \le t < t_{\text{end}}, \\ 0, & t \ge t_{\text{end}}. \end{cases} \tag{S5}$$

For the smooth-amplitude synthetic dataset runs, we use a slower schedule ($t_{\text{lock}} = 50$, $t_{\text{end}} = 200$) to reduce early amplitude drift in joint amplitude-phase recovery. For the STO weak-phase support-refinement run, we use $t_{\text{lock}} = 20$ and $t_{\text{end}} = 100$. For the synthetic 2D projection data and the other experimental single-pattern runs, the prior is used only for amplitude initialisation and the blending schedule is not applied. In all cases where the schedule is used, the prior is removed once $\alpha(t) = 0$.

To stabilise the overall scale of the reconstructed amplitude, we apply a soft energy-matching step inside the support. Let $A_{\text{blend}}(x, y; t)$ denote the blended amplitude at epoch t. We compute the total amplitude "energy" inside the support for the blended and prior amplitudes as

$$E_{\text{cur}}(t) = \sqrt{\sum_{x,y} \big(A_{\text{blend}}(x, y; t)\, S(x, y)\big)^2}, \qquad E_{\text{tgt}} = \sqrt{\sum_{x,y} \big(A_{\text{prior}}(x, y)\, S(x, y)\big)^2}. \tag{S6}$$

where E_{cur} is the current energy, E_{tgt} is the target energy. We then rescale

$$\tilde{A}(x, y; t) = \gamma(t)\, A_{\text{blend}}(x, y; t), \qquad \gamma(t) = \text{clip}\big(E_{\text{tgt}}/E_{\text{cur}}(t),\, \gamma_{\min}, \gamma_{\max}\big), \tag{S7}$$

with $\gamma_{\min}$=0.8 and $\gamma_{\max}$=1.2 for the smooth-amplitude synthetic data. This constrains the support energy to remain close to the prior value. In practice, it prevents global amplitude drift while allowing local deviations from A_{prior}. For the projected 2D synthetic data and experimental reconstructions, this energy-matching step was not used; the reconstructed amplitude was max-normalised directly inside the support.

3. LOSS WEIGHTS AND EXPONENTS

The hybrid loss combines Pearson correlation coefficient (PCC), standard χ^2, and a total variation (TV) regulariser,

$$L_{\text{total}}(t) = w_{\text{PCC}}(t)\, L_{\text{PCC}} + w_{\chi^2}(t)\, L_{\chi^2} + \lambda_{\text{TV}}(t)\, L_{\text{TV}}, \tag{S8}$$

with epoch index $t \geq 0$. We set the raw weight schedules as

$$\text{raw}_{\text{PCC}}(t) = 200\, \exp(-t/15) + 1, \quad \text{raw}_{\chi^2}(t) = 1, \tag{S9}$$

We then normalise by the sum of raw weights at epoch t to obtain $w_{\text{PCC}}(t)$ and $w_{\chi^2}(t)$. For the smooth-amplitude synthetic data, we used

$$\lambda_{\text{TV}}(t) = \begin{cases} 0.2, & t < 200 \\ 0, & t \geq 200 \end{cases} \tag{S10}$$

For the projected 2D synthetic data and the experimental data, we set $\lambda_{\text{TV}}(t) = 0$ for all epochs.

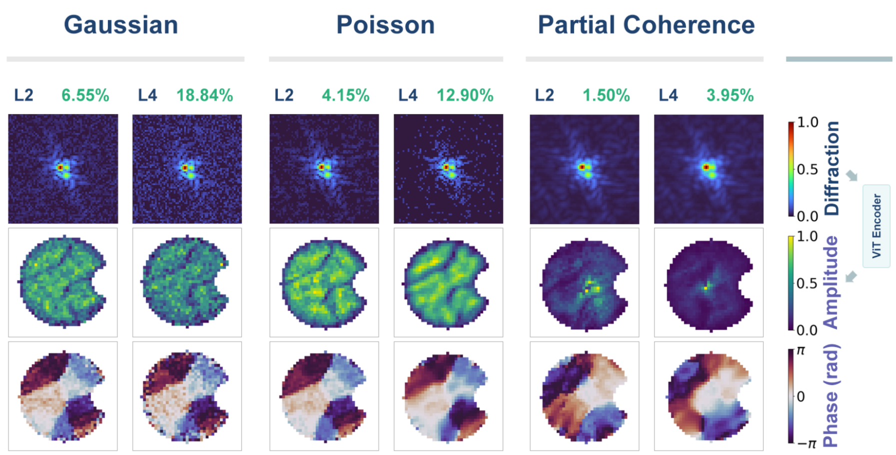

Fig. S1. Additional noise robustness visualisations for the Fourier ViT at Levels 2 and 4 (Levels 1, 3 and 5 are shown in Fig. 3). Columns show Gaussian noise, Poisson noise, and partial coherence applied to the same reference diffraction pattern. For each noise model, the green percentages report χ^2_{n} (in %), the input diffraction error relative to the clean ground truth. The top row shows the noisy input diffraction magnitude (64×64 pixels). The middle and bottom rows show the reconstructed crystal amplitude and phase (40×40 pixels), respectively; both are masked to the support and the phase is shown in radians.

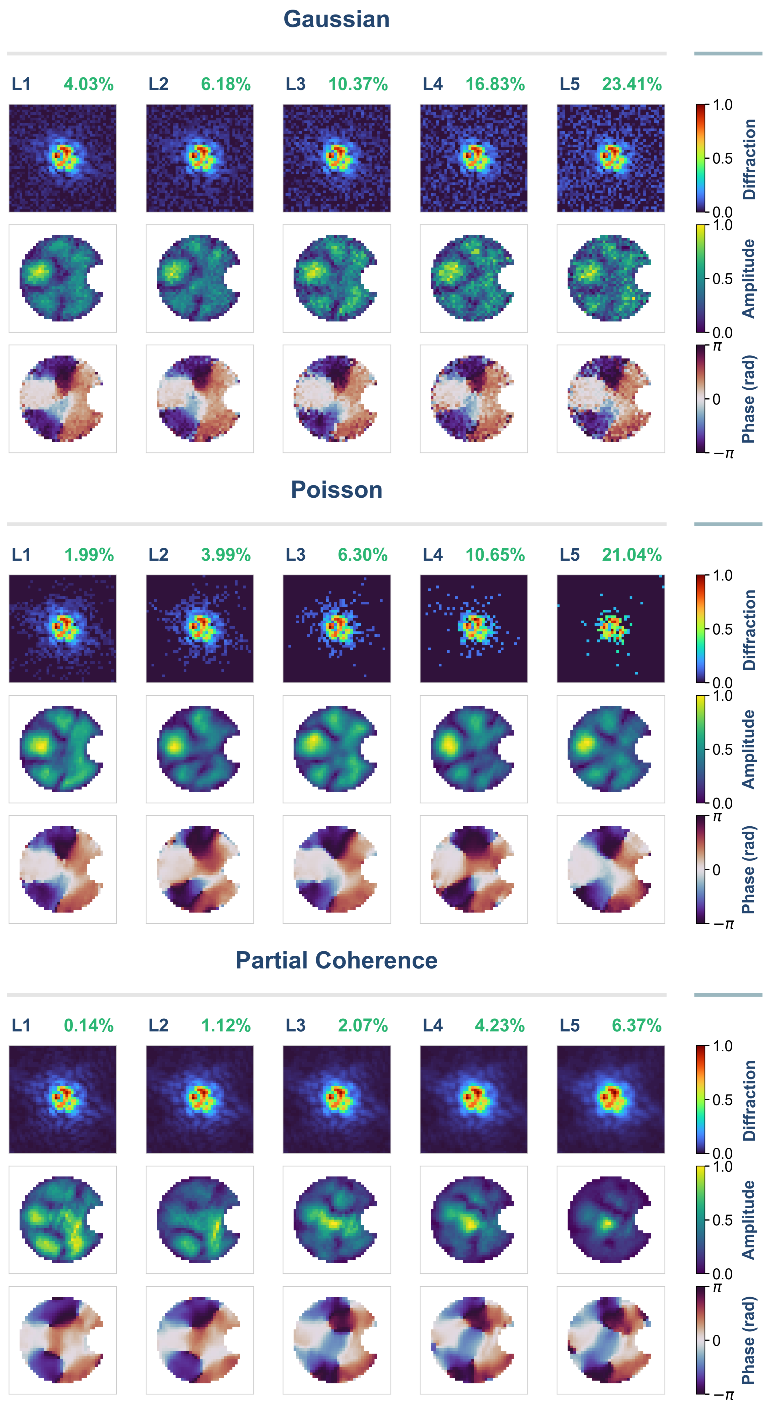

Fig. S2. Noise robustness of the Fourier ViT for the 2D projected synthetic crystal A from Fig. 5. The three blocks show Gaussian noise, Poisson noise, and partial coherence applied to the same clean diffraction pattern. Within each block, columns L1–L5 denote increasing degradation strength; the green percentages give χ^2_n (in %), the input diffraction error relative to the clean ground truth. For each degradation level, the top row shows the degraded input diffraction amplitude (64×64 pixels). The middle and bottom rows show the reconstructed crystal amplitude and phase zoomed to 40×40 pixels, respectively; both are masked to the support and the phase is shown in radians.

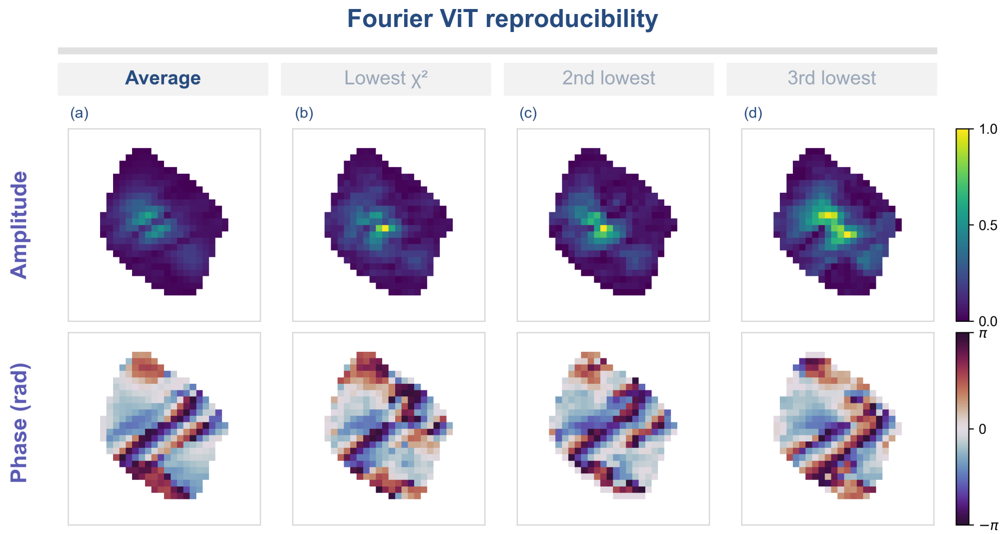

Fig. S3. Fourier ViT reproducibility for the experimental magnetite Fe_3O_4 reconstruction below T_c. All panels show 30×30-pixel crops of the reconstructed amplitude (upper row) and wrapped phase (lower row). (a) Complex-object average of the 50 lowest-χ^2 aligned reconstructions. (b–d) Individual reconstructions ranked by final reciprocal-space χ^2, showing the lowest, second-lowest and third-lowest χ^2 runs. Before averaging, each reconstruction was corrected for the conjugate-twin ambiguity, registered to the same support using integer-pixel shifts, and corrected for a global phase offset.

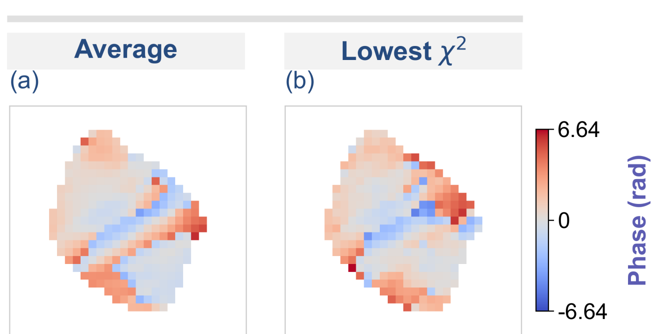

Fig. S4. Unwrapped phase comparison for the experimental magnetite Fe_3O_4 reconstruction below T_c using Fourier ViT. Both panels show 30×30-pixel crops within the same support. (a) Unwrapped phase of the complex-object average from the 50 lowest-χ^2 aligned reconstructions. (b) Unwrapped phase of the reconstruction with the lowest final reciprocal-space χ^2. The phase is plotted using a linear, non-periodic colour scale, showing that the stripe-like contrast is not produced solely by cyclic phase wrapping.

Table S1. Noise robustness of the Fourier ViT for the 2D projected synthetic crystal A from Fig. 5. All values are max-normalised χ^2 errors in %. χ^2_{n} compares the noisy input to the clean ground truth; $\chi^2_{\mathrm{rec,n}}$ compares the reconstruction to the noisy input; and $\chi^2_{\mathrm{rec,c}}$ compares the reconstruction to the clean ground truth. Noise strength increases from level 1 to level 5.

	Gaussian			**Poisson**			**Partial coherence**		
Noise	χ^2_{n}	$\chi^2_{\mathrm{rec,n}}$	$\chi^2_{\mathrm{rec,c}}$	χ^2_{n}	$\chi^2_{\mathrm{rec,n}}$	$\chi^2_{\mathrm{rec,c}}$	χ^2_{n}	$\chi^2_{\mathrm{rec,n}}$	$\chi^2_{\mathrm{rec,c}}$
Level 1	4.03	1.94	2.21	1.99	1.13	0.62	0.14	0.41	0.54
Level 2	6.18	2.74	3.53	3.99	2.72	1.82	1.12	0.42	1.23
Level 3	10.37	4.31	6.01	6.30	3.62	1.78	2.07	0.39	2.22
Level 4	16.83	6.67	10.06	10.65	7.20	3.20	4.23	0.25	3.36
Level 5	23.41	8.84	14.13	21.04	13.74	4.76	6.37	0.15	4.53

VISUALIZATIONS

Visualization 1. (`Visualization1_ER-HIO.mp4`) ER/HIO per-epoch evolution video.
Visualization 2. (`Visualization2_FourierViT.mp4`) Fourier ViT per-epoch evolution video.
Visualization 3. (`Visualization3_C-CNN.mp4`) Complex CNN (C-CNN) per-epoch evolution video.

REFERENCES

1. J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with fourier transforms," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* (Association for Computational Linguistics, Seattle, United States, 2022), pp. 4296–4313.
2. Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier neural operator for parametric partial differential equations," in *International Conference on Learning Representations,* (2021).