Research Article Vol. 34, No. 14 / 13 Jul 2026 / Optics Express 26196

Optics EXPRESS

Vision transformer for multi-domain phase retrieval in coherent diffraction imaging

JIALUN LIU,[1,*] DAVID YANG,[2] AND IAN ROBINSON[1,2]

[1] London Centre for Nanotechnology, University College London, London WC1E 6BT, United Kingdom
[2] Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, New York 11973, USA
[*] jialun.liu.17@ucl.ac.uk

Abstract: Bragg coherent diffraction imaging (BCDI) phase retrieval becomes difficult in the strong-phase regime, where a crystal contains distortions beyond half a lattice spacing. An important special case is the phase domain problem, where blocks of a crystal are displaced with sharp jumps at domain walls. This generates split Bragg peaks and dense fringe structures for which classical iterative solvers often stagnate or return different solutions from different initialisations. We introduce an unsupervised vision transformer (ViT) using Fourier attention to solve this block-phase, multi-domain phase retrieval problem directly from measured 2D Bragg diffraction intensities. Through token mixing and multiscale Fourier attention, Fourier ViT couples reciprocal-space information globally to improve convergence in the strong-phase regime. We have validated the approach on synthetic Voronoi multi-domain crystals with strong-phase contrast under realistic noise corruptions. We further tested it on experimental diffraction from weak-phase $SrTiO_3$, Fe_3O_4 across its phase transition, and strong-phase $La_{0.5}Ca_{0.5}MnO_3$ nanocrystals. For the cases studied here, we have achieved reciprocal-space mismatch (χ^2) comparable to or lower than the compared baselines. Our method preserves domain-resolved phase reconstructions on synthetic data as the number of domains increases. On experimental data, the iterative method, Fourier ViT, and the complex convolutional neural network baseline reach similar χ^2 under the same fixed support.

1. Introduction

Bragg coherent diffraction imaging (BCDI) has transformed the way we characterise nanoscale materials by revealing the internal structure and lattice distortions of single crystals in three dimensions [1–4]. This lensless X-ray technique can invert Bragg peaks into real-space by iteratively applying reciprocal and real-space constraints via fast Fourier transforms [5,6]. BCDI reconstructions have revealed how edge strain in single Pt nanocatalysts influences methane-oxidation activity [7], how nanostrain accumulation drives structural degradation and voltage fade in Li-rich layered oxide cathodes [8], and how electric fields can affect domain wall dynamics in $BaTiO_3$ nanoparticles [9]. However, its core challenge remains in phase retrieval: detectors record only diffraction intensities, so the phase required for the real-space reconstruction is missing [5,10].

Recovering phase from intensities alone is a non-convex, ill-conditioned inverse problem whose difficulty grows sharply with internal strain and domain complexity. Classical algorithms such as Gerchberg–Saxton error reduction [11], hybrid input–output (HIO) [5], difference map [12] and relaxed averaged alternating reflections (RAAR) [13] alternate between real and reciprocal-space, enforcing constraints such as support and measured intensities. For "weak-phase" crystals, where phase shifts stay below about $\pi/2$, these methods are often reliable and converge in a few hundred iterations [2]. Once phase shifts exceed this regime, interference between differently phased

#597254 https://doi.org/10.1364/OE.597254
Journal © 2026 Received 5 Mar 2026; revised 25 Jun 2026; accepted 25 Jun 2026; published 8 Jul 2026

regions leads to peak splitting and complex fringe patterns, the so called "strong-phase" problem [14–16]. The central Bragg maximum can split into multiple lobes, producing separated intensity maxima and dense fringes around the Bragg peak, so the measured diffraction becomes extremely sensitive to the internal phase distribution. In this strong-phase, multi-domain regime, iterative solvers stagnate or converge to different solutions under different random seeds, reflecting an apparent non-uniqueness of the reconstruction [17–19]. This slow and fragile convergence limits routine use of BCDI and makes real-time feedback for in situ or operando experiments difficult, especially at X-ray free-electron laser (XFEL) and modern synchrotron sources [20,21].

Under idealised conditions, theoretical work on the uniqueness of the phase problem shows that oversampling in two or more dimensions is sufficient to guarantee a unique solution up to trivial symmetries such as real-space translation, conjugate inversion, a global phase shift and Fresnel propagation [17,22–24]. In contrast, the one-dimensional problem generally admits many non-equivalent solutions that share the same Fourier magnitudes [25,26]. Motivated by this, we focus on a two-dimensional formulation, considering both purely 2D objects and 2D projections of three-dimensional nanocrystals, thereby preserving essential Fourier-phase structure and serving as a controlled proof-of-principle benchmark for strong-phase, multi-domain optimisation. The extension to fully 3D BCDI is an important direction for future work.

We focus on the phase-domain limit relevant to multi-domain crystals. In BCDI, the real-space phase $\phi(\mathbf{r}) = \mathbf{Q}\cdot\mathbf{u}(\mathbf{r})$ is the projection of the lattice displacement field $\mathbf{u}(\mathbf{r})$ onto the reciprocal-lattice vector $\mathbf{Q}$ of the measured Bragg peak [2,3]. We define a phase domain as a connected block-shaped region in which this projected phase is approximately constant, separated from neighbouring regions by sharp phase discontinuities across narrow domain walls. As the number of domains increases, coherent interference between domains produces stronger Bragg-peak splitting and richer fringe structure, and the phase-retrieval landscape becomes more complicated [14]. Such phase domain textures occur in real quantum materials, for example, BCDI has been used to image the formation and scaling of low-temperature orthorhombic domains in the cuprate superconductor $La_{1.875}Ba_{0.125}CuO_4$ [27] and monoclinic domains in magnetite Fe_3O_4 [28].

In practice, experimental data can be noisy and constraints such as support or non-negativity are only approximate [29]. For noisy data the phase problem no longer has a single exact solution. Instead, one obtains a family of near-solutions whose spread reflects both noise and algorithmic trapping [30]. This has motivated ensemble strategies, in which many independent reconstructions are run and then either selected or averaged according to a figure of merit. Examples include guided HIO procedures [18,31] and the common-mode (eigen-solution) analysis [32] implemented in PyNX, which combines multiple BCDI reconstructions using a free log-likelihood metric [33,34]. Alternative projection schemes, including Fourier-weighted projections [35], improve robustness for truncated diffraction data but remain iterative projection algorithms and can still be sensitive to initialisation. Overall, phase retrieval for multi-domain crystals remains slow, fragile, and difficult to automate.

In parallel, deep learning is increasingly used to solve coherent imaging problems. Supervised neural networks learn the inverse mapping from diffraction intensities to real-space structure, enabling near real-time reconstructions compared with iterative phase retrieval [36–38]. Convolutional neural networks (CNNs) with U-Net architectures have been applied to optical holography [39], X-ray ptychography [40], and coherent X-ray diffraction imaging (CXDI) [41], including BCDI [15,37,42]. After training, these networks can return reconstructions orders of magnitude faster than classical iterative solvers [36,40,42]. The main limitation of supervised approaches is that they learn only what has been given in the training set, and therefore do not guarantee the correct solution for a new object outside the training distribution, particularly when ground-truth labels are unavailable for experimental patterns [36,37,43]. Since label generation and verification are most straightforward in simple regimes, many published demonstrations and benchmarks focus on single- to few-domain systems [37]. Recent work has shown that supervised CNNs can be trained to recover reciprocal-space phase for highly strained BCDI [44], and experimental methods such as Bragg coherent modulation imaging have also been introduced to address challenges in highly strained diffraction phase retrieval [45].

Unsupervised, physics-aware models are trained by minimising a forward-model loss between measured and predicted diffraction intensities, allowing diffraction patterns to be inverted without any training based on ground-truth real-space amplitude or phase labels. This setting is particularly natural for BCDI, where real-space phase labels are unavailable and strong-phase, multi-domain structures make the optimisation landscape highly non-convex and sensitive to initialisation [37,43]. Although physics-aware CNNs such as AutoPhaseNN [43] and complex-valued CNNs [42] can perform unsupervised phase retrieval without ground-truth labels, their convolutional architectures rely on local feature extraction and do not include mechanisms that directly integrate information across the entire reciprocal space. Vision transformers (ViTs), through attention mechanisms that allow each token to attend to all others, inherently provide a global receptive field. ViTs have recently been applied in coherent X-ray imaging, for example in ptychography [46], but have not yet been integrated into a fully unsupervised, physics-informed BCDI framework.

Here, we introduce a simple block-phase domain model to access the strong-phase, multi-domain regime, and solve the resulting phase-retrieval problem using an unsupervised Fourier ViT with global reciprocal-space token mixing. The proposed Fourier ViT combines convolutional feature extraction with spectral token mixing, providing global coupling at $O(N \log N)$ complexity, in contrast to the $O(N^2)$ cost of full dot-product self-attention in standard transformers [47]. Fourier token mixing has been explored across modalities, from unparameterised Fourier mixing (FNet) to learnable global frequency filters (GFNet) and adaptive Fourier operators (AFNO/FNO) [48–51]. In this work, we adapt these ideas to oversampled Bragg diffraction through a multiscale spectral design tailored to the structure of BCDI patterns. This architectural choice is motivated by the structure of strong-phase BCDI diffraction itself: split Bragg peaks, weak high-q fringes, and domain-interference features are distributed across the detector and should not be treated as isolated local patches. Our multiscale Fourier mixing is intended to couple these non-local reciprocal-space correlations throughout the entire calculation more directly than a purely local convolutional backbone.

We first validated Fourier ViT on synthetic datasets of Voronoi multi-domain crystals. Over the range of conditions studied, Fourier ViT achieves reciprocal-space mismatch (χ^2) comparable to or lower than that of the compared methods, and preserves domain-resolved phase reconstructions on synthetic data under realistic noise degradations. We further apply the method to experimental diffraction from $SrTiO_3$ (STO), Fe_3O_4 and $La_{0.5}Ca_{0.5}MnO_3$ (LCMO) nanocrystals [28,52]. In these experimental cases, Fourier ViT achieves χ^2 values competitive with the iterative benchmark and the C-CNN baseline [42].

2. Methods

2.1. Fourier ViT model architecture

2.1.1. Tokenisation and positional embedding

As shown in Fig. 1 (a), the input is a 2D diffraction pattern magnitude with 64×64 pixels, max-normalised to $[0, 1]$. In panel (b), a shallow CNN extracts features at the original 64×64 resolution and outputs a 128-channel feature map. We retain this as a full-resolution skip map $\mathbf{S} \in \mathbb{R}^{B\times128\times64\times64}$, where B denotes the batch size. Patch embedding in Fig. 1(c) uses a single 4×4 convolution with stride 4 applied to $\mathbf{S}$, partitioning the image into non-overlapping 4×4 patches and projecting each patch to a token of embedding dimension E=128. This produces a 16×16 token grid, thus N=16×16=256 tokens in total. The typical speckle/fringe spacing in our datasets is about two pixels after binning, so a 4×4 patch spans at least one full fringe period and allows each token to encode a physically meaningful local piece of the diffraction pattern. Following the vision transformer formulation [47,53], we then flatten the 16×16 grid into a token sequence and refine it with two 1×1 convolutions before adding a learned positional embedding. The resulting sequence $\mathbf{X} \in \mathbb{R}^{B \times N \times E}$ is passed to the stack of Fourier transformer blocks described in Sec. 2.1.2.

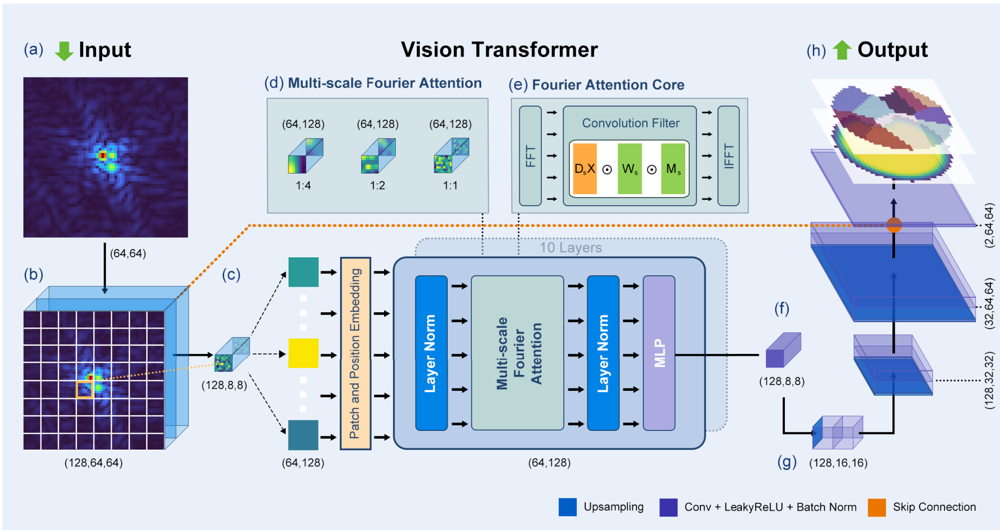

Fig. 1. Architecture of the proposed Fourier ViT model for BCDI phase retrieval. (a) The input diffraction magnitude (64×64 pixels) is passed through (b) a shallow convolutional feature extractor to produce 128 feature channels ($128 \times 64 \times 64$). (c) The feature map is partitioned into non-overlapping patches and embedded as tokens. For visual clarity, the schematic shows an 8×8 token grid, corresponding to 64 tokens and a token sequence of shape 64×128. This sequence is processed by (d) a multi-layer Vision Transformer with multi-scale Fourier attention, where feature mixing is performed at 1:4, 1:2, and 1:1 resolution scales, and by (e) an FFT-based global convolution block. (f) The transformer output is reshaped into a latent feature map ($128 \times 8 \times 8$) and decoded by (g) a convolutional upsampling path with a skip connection to reconstruct the complex crystal field, yielding (h) 64×64 real-space amplitude and phase maps. In the actual implementation used for all experiments, the model uses 4×4 pixel patches on the 64×64 feature map, producing a 16×16 token grid, 256 tokens, a transformer input sequence of shape 256×128, and a corresponding latent feature map of size $128 \times 16 \times 16$.

2.1.2. Multi-scale Fourier attention

Inspired by Fourier-based transformers and spectral gating mechanisms [48,54], we replace dot product self-attention [47] with multi-scale Fourier attention, indicated by the green boxes in Fig. 1(d,e). The tokens are reshaped to a feature map $\mathbf{X} \in \mathbb{R}^{B \times E \times H \times W}$ with H=W=16 and processed at three spatial scales $s \in \{1, 2, 4\}$:

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

here $\mathcal{D}_s$ denotes average pooling with stride s, which reduces the spatial size to $h_s \times w_s = (H/s) \times (W/s)$, and $\mathcal{U}_s$ upsamples back to (H, W). The operator $\mathcal{F}$ is the channel-wise 2D FFT, and $\odot$ denotes element-wise multiplication in frequency space. In practice, we use three mixers at resolutions (H, W), $(H/2, W/2)$ and $(H/4, W/4)$ and learn their relative importance via the scale weights α_s, giving $\mathbf{Z} = \sum_s \alpha_s \mathbf{Z}_s$ with $\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{\theta})$. At each scale, we learn per-channel frequency responses $\mathbf{W}_s \in \mathbb{R}^{E\times h_s\times w_s}$ and a channel-shared spectral gate $\mathbf{M}_s \in \mathbb{R}^{1\times h_s\times w_s}$. The $s=1$ branch acts as the finest-resolution branch and retains high-frequency fringe information. The $s=4$ branch provides a coarser representation of broader reciprocal-space features, including the central Bragg peak region. The different spectral branches are then recombined with learned scale weights, rather than fixed by hand. This design is well suited to strong-phase multi-domain BCDI, where both the global peak structure and fine fringe pattern carry information about the object.

We use pre-normalised residual connections in each Fourier transformer block:

$$\mathbf{X}' = \mathbf{X} + \mathrm{Mixer}\big(\mathrm{LN}(\mathbf{X})\big), \qquad \mathbf{X}_{\mathrm{out}} = \mathbf{X}' + \mathrm{MLP}\big(\mathrm{LN}(\mathbf{X}')\big), \tag{2}$$

where LN represents layer normalisation [53,55], Mixer denotes the multi-scale Fourier spectral mixer defined above, and the multilayer perceptron (MLP) is a feed-forward projection with hidden dimension d_{mlp} and ReLU nonlinearity. The residual paths improve gradient flow, while pre-normalisation stabilises optimisation for deeper stacks of blocks. Additional details, including the convolution theorem and complexity analysis, are given in Supplement 1, Sec. S1.

2.1.3. Decoder

After the final transformer block, the output sequence $\mathbf{X} \in \mathbb{R}^{B\times N\times E}$ (with N=16 × 16 and E=128) is reshaped to an $E \times 16 \times 16$ feature map as shown in Fig. 1(f) and upsampled to 64×64 by a small CNN decoder [Fig. 1(g)]. At full resolution the decoder fuses three inputs: (i) the upsampled token features, (ii) the encoder skip map $\mathbf{S} \in \mathbb{R}^{B\times 128\times 64\times 64}$, which carries local diffraction pattern context from the CNN feature extractor, and (iii) a frequency–space summary of $\mathbf{S}$ obtained by taking its channel-wise Fourier magnitude, applying $\log\big(1+|\mathcal{F}\{\mathbf{S}\}|\big)$ to compress the dynamic range, and normalising per channel (see Supplement 1, Sec. S2 for details). The three streams are concatenated along the channel dimension and passed through 3×3 convolutions for channel mixing, followed by a short refinement stack. This fusion supplies the decoder with complementary information: (i) non-local diffraction correlations from the transformer, (ii) fine-scale speckle features from the encoder, and (iii) an explicit spectral summary of the learned diffraction features.

The decoder heads output a real-valued crystal amplitude map $\hat{A}(x,y)$ and two channels $(\hat{c}(x,y), \hat{s}(x,y))$ that parametrise the crystal phase. The amplitude head uses a softplus nonlinearity to enforce non-negativity, and both amplitude and phase are multiplied by a fixed real-space support mask $S(x,y)$, in line with standard CXDI support constraints [5,56]. For experimental data, Fourier ViT uses the same support as the MATLAB iterative reconstruction (see Sec. 2.6) for a fair comparison. Inside the support, the predicted amplitude is blended with a simple prior via a scalar schedule $\alpha(t)$ and then softly rescaled so that its ℓ_2 norm matches that of the prior amplitude; the precise blending and normalisation rules are given in Supplement 1, Sec. S2. For the phase, the two decoder outputs are first normalised onto the unit circle, $(\tilde{c},\tilde{s}) = (\hat{c},\hat{s})/\sqrt{\hat{c}^2+\hat{s}^2}$, and the phase is recovered as $\phi(x,y) = \mathrm{atan2}(\tilde{s}(x,y), \tilde{c}(x,y))$.

Given an input diffraction pattern magnitude $M(\mathbf{q})$, the network therefore outputs a complex real-space density $\rho(x,y)$ on a 64×64 grid, with amplitude and phase constrained by the fixed support.

2.1.4. Loss function

For a given input diffraction pattern, the network predicts a complex real-space density $\rho(x,y)$ from the crystal's amplitude and phase. This estimate is propagated through the BCDI forward model to obtain a predicted diffraction magnitude $M_{\mathrm{pred}}(\mathbf{q})$ (Sec. 2.1.3), which is compared with the measured magnitude $M_{\mathrm{meas}}(\mathbf{q})$ via a hybrid Fourier-space loss. The loss combines Pearson correlation coefficient (PCC), a root mean square (RMS)-normalised χ^2, and a small

Optics EXPRESS

total variation (TV) regulariser, with epoch-dependent weights that shift the emphasis from global pattern correlation to fine scale intensity agreement as training proceeds.

For the χ^2-type terms we work with an RMS-normalised magnitude $\hat{M}(\mathbf{q}) = \frac{M(\mathbf{q})}{\sqrt{\langle M(\mathbf{q})^2\rangle_{\mathbf{q}}}}$, applied separately to M_{meas} and M_{pred} so that any global intensity scale is removed. At training epoch t, we minimise

$$L_{\mathrm{total}}(t) = w_{\mathrm{PCC}}(t)\, L_{\mathrm{PCC}} + w_{\chi^2}(t)\, L_{\chi^2} + \lambda_{\mathrm{TV}}(t)\, L_{\mathrm{TV}}, \tag{3}$$

where the data-fidelity weights $w_{\mathrm{PCC}}(t)$, $w_{\chi^2}(t)$, and the TV weight $\lambda_{\mathrm{TV}}(t)$ follow simple epoch-dependent schedules defined in Supplement 1, Sec. S3. They are designed so that PCC dominates early training to stabilise the global reconstruction, before its weight decays towards an approximately balanced PCC/χ^2 contribution during later refinement. The amplitude TV term was used only for the smooth-amplitude synthetic data, where it acts as an edge-preserving smoothing prior on the reconstructed modulus. We used $\lambda_{\mathrm{TV}}(t) = 0.2$ for $t<200$ and $\lambda_{\mathrm{TV}}(t) = 0$ for $t \geq 200$. Ablations without TV showed that it mainly suppresses small modulus artefacts and improves amplitude smoothness in this idealised synthetic case. It was not used for the projected 2D synthetic data or the experimental data, because projection produces physically meaningful amplitude variation that should be retained.

(i) PCC loss. To enforce global similarity of the diffraction patterns we use

$$L_{\mathrm{PCC}} = 1 - \rho\big(M_{\mathrm{meas}}, M_{\mathrm{pred}}\big), \tag{4}$$

where, for a single pattern,

$$\rho\big(M_{\mathrm{meas}}, M_{\mathrm{pred}}\big) = \frac{\sum_{\mathbf{q}} \big(M_{\mathrm{meas}}(\mathbf{q}) - \mu_{\mathrm{meas}}\big)\big(M_{\mathrm{pred}}(\mathbf{q}) - \mu_{\mathrm{pred}}\big)}{\sqrt{\sum_{\mathbf{q}} \big(M_{\mathrm{meas}}(\mathbf{q}) - \mu_{\mathrm{meas}}\big)^2}\sqrt{\sum_{\mathbf{q}} \big(M_{\mathrm{pred}}(\mathbf{q}) - \mu_{\mathrm{pred}}\big)^2}}, \tag{5}$$

where $\mu_{\mathrm{meas}} = \langle M_{\mathrm{meas}}(\mathbf{q})\rangle_{\mathbf{q}}$, $\mu_{\mathrm{pred}} = \langle M_{\mathrm{pred}}(\mathbf{q})\rangle_{\mathbf{q}}$.

(ii) RMS-normalised χ^2 loss. To penalise absolute mismatches while remaining insensitive to any residual scale factor, we define

$$L_{\chi^2} = \frac{\sum_{\mathbf{q}} \big[\hat{M}_{\mathrm{meas}}(\mathbf{q}) - \hat{M}_{\mathrm{pred}}(\mathbf{q})\big]^2}{\sum_{\mathbf{q}} \hat{M}_{\mathrm{meas}}(\mathbf{q})^2}, \tag{6}$$

where $\hat{M}_{\mathrm{meas}}$ and $\hat{M}_{\mathrm{pred}}$ are the RMS-normalised magnitudes defined above.

(iii) Total-variation amplitude regularization. Let $S(x, y) \in \{0, 1\}$ be the binary support mask and $N_S = \sum_{x,y} S(x, y)$ the number of support pixels. We define an anisotropic TV penalty

$$L_{\mathrm{TV}} = \frac{1}{N_S} \sum_{x,y} S(x, y)\, \big(|\nabla_x A(x, y)| + |\nabla_y A(x, y)|\big), \tag{7}$$

where $\nabla_x A(x, y) = A(x + 1, y) - A(x, y)$ and $\nabla_y A(x, y) = A(x, y + 1) - A(x, y)$. In implementation, we mask the amplitude by $S(x, y)$ before taking finite differences, so the gradients across the support boundary do not contribute.

2.2. Synthetic data and noise

2.2.1. Voronoi multi-domain crystals

To simulate the heterogeneous structure of functional nanocrystals, we constructed two-dimensional crystals made up of multiple phase domains with sharp, straight boundaries.

In Fig. 2(a), seed points were randomly placed inside a soft-edged circular support with a crescent cutout. This geometry represents a central cross section of a finite 3D particle. Each seed was then assigned a random phase in $[-\pi, \pi]$. The soft amplitude edges suppress artificial high-frequency components in Fourier space that would arise from box-shaped boundaries, while the crescent cut breaks symmetry and helps to avoid twin solutions. A minimum spacing of 2 pixels was enforced between seeds to prevent clustering. Each seed then generated a Voronoi cell whose pixels take the phase of the closest seed, producing a tessellation of non-overlapping domains with straight boundaries. Physically meaningful examples of sharp domain walls occur in thin films. This Voronoi-based construction mimics the abrupt domain walls typically observed in ferroelectric and ferroelastic materials [57].

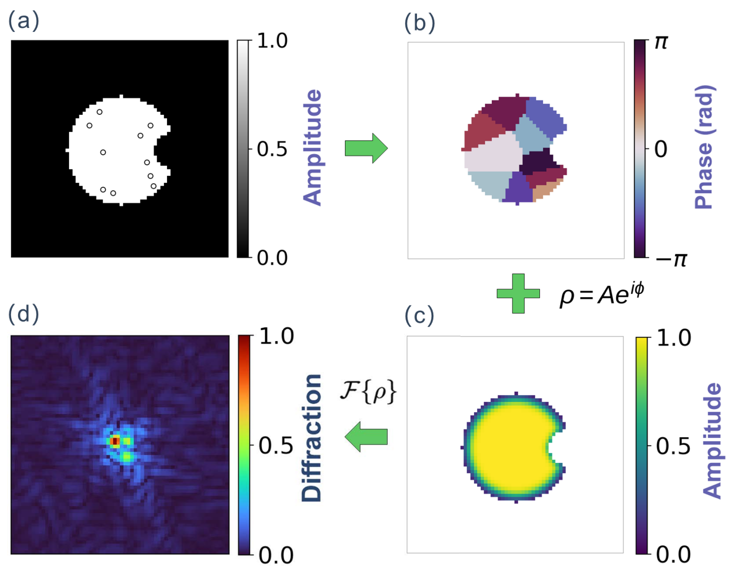

Fig. 2. Simulation pipeline for synthetic crystals and diffraction patterns. (a) Random seed placement within the support. (b) Voronoi phase domain generation from seeds, with phases wrapped to $[-\pi, \pi]$. (c) Construction of the crystal amplitude mask, max-normalised to $[0, 1]$. (d) Corresponding diffraction magnitude computed by fast Fourier transform (FFT), max-normalised to $[0, 1]$. All images are 64×64 pixels.

In Fig. 2(b), we set the central domain to $\phi = 0$ to remove the global phase offset. We required $\Delta\phi > 0.5$ rad between neighbouring domains to enforce strong-phase contrast and avoid inversion-symmetric cases. Figure 2(c) shows the crystal amplitude $A(x, y)$, chosen to represent an idealised 2D cross-section of a finite 3D crystal. We use a flat-core radial profile with Gaussian edge smoothing, modulated by the crescent mask, to produce a confined but soft-edged support. The final real-space object is defined as the complex density

$$\rho(x, y) = A(x, y)\, e^{i\phi(x,y)}, \tag{8}$$

with $A(x, y)$ fixed and $\phi(x, y)$ set by the Voronoi partition.

In Fig. 2(d), the diffraction intensity is obtained as the modulus squared of the Fourier transform of the complex density,

$$I_{\text{pred}}(\mathbf{q}) = \left|\mathcal{F}\{\rho(x, y)\}(\mathbf{q})\right|^2, \tag{9}$$

on a 64×64 grid containing a 32-pixel-diameter support, corresponding to an oversampling ratio of 2 in reciprocal space. Outside the support, $\rho(x, y)$ is set to zero so that only physically allowed regions contribute to the scattering. The crystal amplitude, phase and diffraction pattern are stored as 16-bit TIFF images.

We extended the Voronoi construction to projected 3D objects. A spherical 3D support with a cylindrical side cut produced a non-centrosymmetric particle shape in Fig. 3(a), and seeds within this volume defined Voronoi phase domains. Figure 3(b) shows orthogonal xy, xz and yz amplitude and phase slices through the object. The projected 2D object was obtained by summing the complex density along the z direction:

$$\rho_{\text{proj}}(x, y) = \sum_{z} \rho(x, y, z). \tag{10}$$

This differs from prescribing a 2D amplitude and phase independently, because complex values with different phases interfere through the depth of the particle. As a result, the projected amplitude becomes spatially non-uniform and the apparent phase boundaries are softened compared with a single 2D slice, as shown in Fig. 3(d). The corresponding 2D diffraction pattern was calculated from $\mathcal{F}\{\rho_{\text{proj}}(x, y)\}$, giving the projected diffraction pattern in Fig. 3(e). This is consistent with the central $q_x q_y$ slice of the 3D diffraction volume shown in Fig. 3(c) because of the Fourier slice theorem [58].

We also generated an irregular projected-object dataset by perturbing the particle boundary with a smooth random field before Gaussian edge smoothing. Voronoi phase domains were assigned inside this support and projected to 2D using the same procedure, providing a less regular geometry for testing reconstruction stability.

This Voronoi-based construction is intended as a controlled benchmark for abrupt domain walls, rather than as a full model of real domain microstructure. It is physically motivated by materials in which relatively sharp domain boundaries are observed [57], but does not capture diffuse walls, local strain gradients, defects, measurement background errors, or other experimental complications. Overall, this pipeline produces synthetic data with controlled domain number, sharp phase boundaries, and structured reciprocal-space intensity. The first workflow isolates the sharp-boundary phase-retrieval problem in a clean 2D cross-section, and the projected workflow introduces amplitude–phase coupling closer to the experimental setting considered later in this work.

2.2.2. Noise tests

We assess robustness in the regime where the network predicts both crystal amplitude and phase with noise added to the diffraction pattern. Starting from a max-normalised diffraction pattern intensity $I(q_x, q_y) \in [0, 1]$, we construct three families of corrupted data: Poisson shot noise, Gaussian read noise and partial coherence.

For the two detector-noise models we work on the exposure (photon-count) scale. In this setting, read noise from the sensor chain is well approximated by additive, zero-mean Gaussian noise in counts, whereas photon shot noise is described by Poisson statistics [59–61]. Given a conversion factor $S>0$ from normalised intensity to expected counts, the ideal noiseless counts are $C_0(q_x, q_y) = S\,I(q_x, q_y)$. Gaussian read noise is then modelled as

$$C_g(q_x, q_y) = C_0(q_x, q_y) + \eta(q_x, q_y), \tag{11}$$

where the random variable $\eta(q_x, q_y)$ is drawn independently at each pixel from a normal distribution with zero mean and variance σ_c^2. Poisson shot noise is modelled as

$$C_p(q_x, q_y) \sim \text{Poisson}\big(C_0(q_x, q_y)\big), \tag{12}$$

reflecting the discrete, independent arrival of photons at each pixel. The function of Poisson is generated using the Poisson random number generator in the NumPy array library [62].

Optics EXPRESS

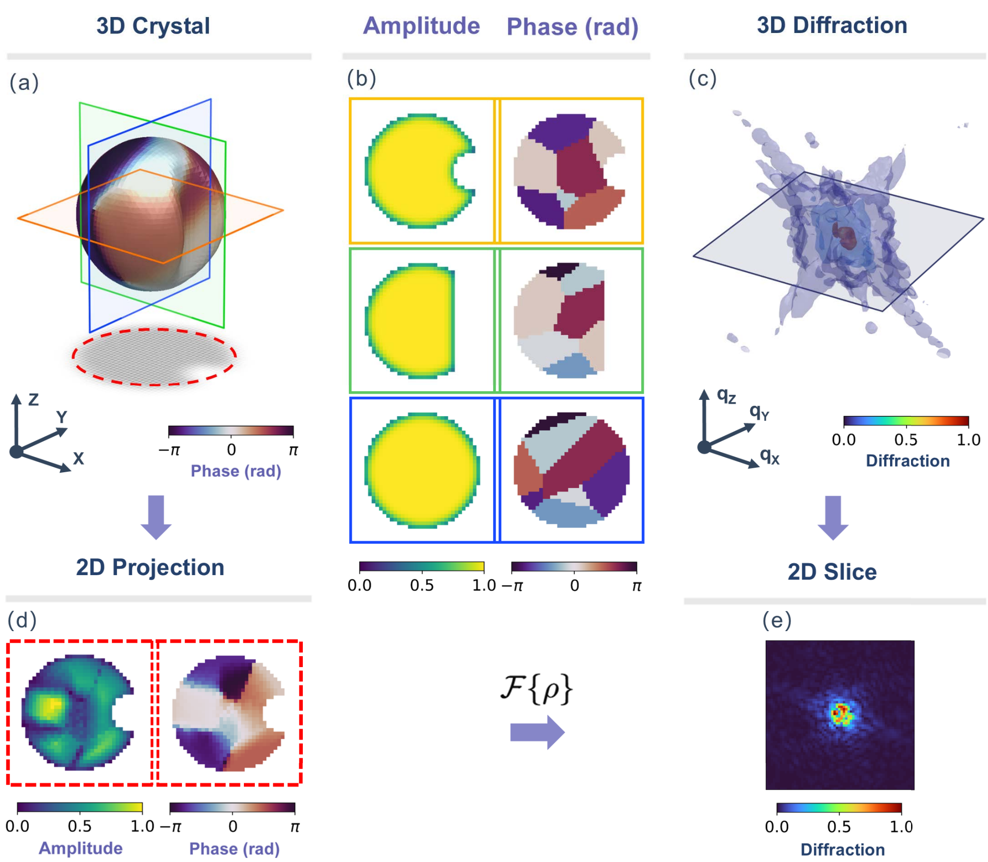

Fig. 3. Simulation pipeline for projected multi-domain crystals and diffraction patterns. (a) 3D Voronoi crystal with a compact support and crescent cutout. The surface colour represents the phase in $[-\pi, \pi]$, and the three orthogonal planes indicate the central slices used for visualisation. (b) Orthogonal amplitude and phase slices through the 3D object. The amplitude is max-normalised to $[0, 1]$, and the phase is wrapped to $[-\pi, \pi]$. The orange, green and blue frames correspond to the xy, xz and yz planes shown in (a), respectively. (c) 3D diffraction magnitude calculated from the complex crystal density, shown in reciprocal space. (d) 2D projected amplitude and phase of the 3D crystal in the xy plane, marked by dashed red boxes. (e) Corresponding 2D diffraction magnitude obtained from the Fourier transform of the projected complex density in (d). This 2D diffraction corresponds to the central $q_x q_y$ slice of the 3D diffraction volume in (c). The 2D diffraction patterns are 64×64 pixels; the 2D crystal amplitude and phase are shown as 40×40-pixel zoomed-in regions.

For the Gaussian read-noise model, negative values are truncated to zero to enforce physical non-negativity of the recorded counts. Poisson counts are non-negative by construction. For analysis, we return to an intensity scale via $\tilde{I}(q_x, q_y) = C(q_x, q_y)/S$, and use the corresponding diffraction magnitude $\tilde{M}(q_x, q_y) = \sqrt{\tilde{I}(q_x, q_y)}$ for all reported image space metrics and for the reconstruction loss χ^2_{rec} in Table 1.

Partial coherence is treated as a degradation of the diffraction intensity rather than a stochastic detector noise. We model it as a contrast-reducing blur

$$I_{\text{pc}}(q_x, q_y) = \mathcal{G}_\sigma * I(q_x, q_y), \tag{13}$$

Table 1. Noise robustness of the Fourier ViT (magnitude domain). All values are max-normalised χ^2 errors in %. χ^2_{n} compares the noisy input to the clean ground truth; $\chi^2_{\mathrm{rec,n}}$ and $\chi^2_{\mathrm{rec,c}}$ compare the reconstruction to the noisy input and to the clean ground truth, respectively. 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	3.28	1.76	1.54	1.86	1.18	0.92	0.31	0.13	0.25
Level 2	6.55	3.48	2.71	4.15	2.67	1.92	1.50	0.25	1.94
Level 3	10.07	5.16	4.14	7.68	4.69	3.65	2.30	0.23	2.78
Level 4	18.84	9.03	8.21	12.90	8.07	5.82	3.95	0.17	4.66
Level 5	28.32	12.49	14.36	20.61	11.91	10.53	5.62	0.11	6.37

where $\mathcal{G}_\sigma$ approximates the far-field mutual-coherence function as a Gaussian in reciprocal space, so that partial coherence is modelled as a convolution of the coherent intensity [63,64]. For simulated data, we convolve the coherent diffraction with $\mathcal{G}_\sigma$ and then renormalise I_{pc} to match the clean pattern's overall scaling, so that changes in reconstruction quality primarily reflect the loss of high-q fringe visibility rather than intensity rescaling. For experimental data (Sec. 3.4), we apply the same Gaussian model in reverse using a regularised deconvolution to partially undo the blurring and obtain a more nearly coherent effective diffraction pattern.

2.3. Fourier ViT-network training

Fourier ViT operates on 64×64 max-normalised diffraction magnitudes. With a patch size of 4, this gives 16×16 tokens and an embedding dimension of 128 processed through 10 Fourier transformer blocks. The networks are trained for 1000 epochs with batch size 1 using AdamW [65] (initial learning rate 5×10^{-4}, weight decay 10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with a cosine-annealed learning rate schedule down to 10^{-6}. Training uses automatic mixed precision on an NVIDIA Quadro RTX 5000 (Max-Q). For a full 1000-epoch optimisation of a single 64×64 diffraction pattern, the mean wall-clock time was 94.42 s, corresponding to 1.57 min per reconstruction on this GPU.

The hybrid loss function of Sec. 2.1.4 is used throughout. A fixed binary support is imposed; the amplitude is initialised from a Gaussian prior within the support and kept fixed for the first 10 epochs, before being learned jointly with the phase via the blending schedule $\alpha(t)$ (Supplement 1, Sec. S2). Ablation tests in which the amplitude-prior blending schedule was turned off, with the prior used only for initialisation, produced qualitatively reasonable reconstructions with low χ^2 for the synthetic 2D projection data and most experimental data. For the smooth-amplitude synthetic cases, and for the STO weak-phase support-refinement run, the blending schedule improved convergence and reduced amplitude artefacts. Dropout is set to 0, and we apply gradient clipping with a maximum gradient norm of 1.0 to stabilise optimisation.

For each domain count, we perform 100 independent runs with different random seeds and report statistics over these runs. We select the best run across random initialisations by minimising the diffraction mismatch χ^2 between the predicted and measured diffraction magnitudes, and report the corresponding reconstruction metrics, including χ^2 against the clean target when available. In the experimental case, χ^2 is reported as a convenient reciprocal-space mismatch measure under identical support constraints for direct comparison with the simulations. A free log-likelihood evaluation [33] would be more appropriate for noisy experimental data and therefore more valuable as a future extension.

Optics EXPRESS

2.4. Experimental support refinement

For experimental data without a known real-space support, we refine the support using an amplitude-based procedure within the Fourier ViT workflow. The initial support was chosen as a square mask that loosely enclosed the object. A first batch of 100 Fourier ViT reconstructions was then performed using this neutral support. The reconstructed amplitude was max-normalised and thresholded at 0.06 to suppress weak diffuse background. The thresholded image was binarised and cleaned by connected-component filtering, retaining only the principal connected object region and removing small isolated features. This cleaned binary mask was used as the refined support for a second batch of Fourier ViT reconstructions.

This procedure follows the same physical logic as amplitude-based shrink-wrap support refinement in coherent diffraction imaging [56], but is implemented here within the Fourier ViT workflow. Adaptive support updates beyond this scheme are left for future work.

2.5. BCDI experiment

We measured BCDI data from a crystal of $La_{0.5}Ca_{0.5}MnO_3$ (LCMO). LCMO is an important complex transition-metal oxide showing colossal magnetoresistance (CMR) close to room temperature, in which large changes in electrical resistance are induced by small changes in magnetic field. We are interested in whether the phase domains play a role in the CMR behaviour. LCMO was prepared by chemical synthesis in the form of nanocrystals with sizes in the range of 300 nm, aggregated into clusters [52]. These were attached to silicon wafer substrates for measurement by BCDI using the tetraethyl orthosilicate (TEOS) bonding method [66]: a dilute solution of TEOS was used to coat the dry powder, and it was then calcined at 500 °C in air using a furnace for several hours.

X-ray measurements were made at beamline 34-ID-C of the advanced photon source (APS) [52]. The monochromatic X-ray beam, with wavelength $\lambda = 0.138$ nm, was made coherent with a 50 μm (vertical) × 20 μm (horizontal) slit, 48 m from the source, then focused with Kirkpatrick–Baez (KB) mirrors to a beam size of 600 nm × 800 nm at the centre of a precise air-bearing rotation axis, θ. A single nanocrystal was selected by rotating θ until an isolated peak appeared on the Medipix area detector, with 55 μm pixels, positioned at the Bragg angle of the (110) Bragg peak, at a distance $D = 2.2$ m away from the sample. A single frame from that 3D data set, LCMO, was selected for the study. The corresponding image is a phased projection of the 3D nanocrystal structure, whose Fourier transform magnitude squared (plus noise) should match exactly the measured frame, according to the Fourier slice theorem [58] and from the projected-density definition in Eq. (10), as discussed in Sec. 2.2.1.

2.6. Iterative method

We implement a conventional 2D iterative phase retrieval algorithm in MATLAB following standard BCDI practice [2,64]. The initial support is defined relative to the reconstruction grid, with $s_x = 0.5\,N_x$ and $s_y = 0.5\,N_y$. We then use a predefined switching schedule that alternates between 20 iterations of error-reduction (ER) and 80 iterations of hybrid input-output (HIO) algorithms, with HIO relaxation parameter $\beta = 0.9$ [5]. The support was refined using shrink-wrap updates [56] with a max-normalised amplitude threshold of 0.15. During ER, we enforce the crystal's amplitude to zero outside of the support and keep phase within $[-\pi, \pi]$. Reconstructions of LCMO used 500 iterations, with shrink-wrap updates applied from iteration 200, to generate support masks from independent runs. As these runs did not converge to a unique final support, we selected the largest support among the main solutions and used it as a fixed common real-space constraint for the iterative, Fourier ViT and C-CNN comparisons, rather than as an independent ground-truth morphology.

2.7. Complex convolutional neural network

For experimental comparison with the proposed Fourier ViT, we implement a complex-valued CNN baseline following the complex-network recipe used for BCDI phase retrieval [42]. The model has a U-Net encoder-decoder architecture. In our implementation, the encoder uses complex channel widths of 32, 64, 128, 256 and 512 (with a decoder of 256, 128, 64 and 32), followed by a final 1×1 complex convolution that outputs a single complex channel. We used LeakyReLU activations with a 0.2 negative slope throughout. Feature maps are represented as real and imaginary channels, and all convolutions are performed via complex operations by mixing the two channels according to complex multiplication. Skip connections bridge the encoder and decoder stages to preserve fine spatial detail. For the LCMO measurement, we initialised the network input as two channels, with $\Re\{x\}$ set to the max-normalised measured diffraction magnitude and $\Im\{x\} = 0$, corresponding to a zero initial phase. The network outputs a complex object $\rho(x) = \rho_{\Re}(x) + i\rho_{\Im}(x)$, from which amplitude and phase are obtained by $|\rho|$ and $\arg(\rho)$. Training, optimisation, support constraint, and best-run selection for the C-CNN baseline follow the same procedure and loss function as Sec. 2.3, so that the comparison isolates the effect of network architecture. The C-CNN therefore serves as a controlled baseline under matched conditions, rather than as an exhaustively tuned model.

3. Results

3.1. Performance on synthetic data

On synthetic multi-domain crystals, the Fourier ViT recovers the real-space object from reciprocal-space diffraction magnitude, without real-space inputs beyond a support mask. At 64×64 diffraction resolution with known crystal amplitude, the method resolves up to 19 domains with runs that achieve very low reciprocal-space mismatch at $\chi^2 \leq 10^{-5}$. Those reconstructions occur 38 times out of 100 runs for 10 domains, 18 times for 15 domains, and 4 times for 19 domains. When crystal amplitude and phase are recovered jointly, convergence is harder but still feasible: for a 10-domain crystal, 13 out of 100 runs reach $\chi^2 \leq 10^{-5}$ after 1000 epochs.

To illustrate object amplitude-phase retrieval, we show a representative 10-domain case in Fig. 4. Figure 4(a,b) compare the ground truth and Fourier ViT predicted diffraction magnitudes, while Fig. 4(c) shows the radially averaged log-intensity profile of the diffraction pattern as a function of momentum transfer q (radial coordinate in reciprocal-space). The profile matches the ground truth across the full q range, including the high-q tails produced by strong-phase contrasts and sharp domain walls. Instead of a single central peak, familiar in the weak-phase case, this pattern has multiple centres because the object has highly varying phase values. Figure 4(d,e) demonstrate that the recovered crystal amplitude matches the ground truth smoothly, and Fig. 4(g,h) show that the phase solution recovers the real-space domain configurations with sharp, correctly placed boundaries and no spurious domains. The ground truth panels are used only for quantitative assessment and visualisation, as the network is trained unsupervised from diffraction data without real-space labels. For visual comparison, we fix the global phase offset by setting the phase at the central pixel to zero in all reconstructions.

Figure 4(f) shows a bimodal distribution over 100 random initialisations for the phase-only prediction case. Many runs converge to near-global-minimum solutions with $\chi^2 \leq 10^{-5}$, while the remaining runs settle into higher-error local minima, forming a second cluster at $\chi^2 \approx 1.5$–2.7×10^{-2}. This behaviour reflects the non-convex strong-phase retrieval landscape, where small changes in domain-wall position or topology can produce a structured mismatch in the high-q fringe field. Similar sensitivity to initialisation is also common in iterative phase retrieval. In this controlled synthetic case, the fixed amplitude/support constraint and available ground truth confirm that the minimum-χ^2 solution gives the most accurate real-space reconstruction. The higher-χ^2 cluster corresponds to runs that remain in local minima within the fixed optimisation

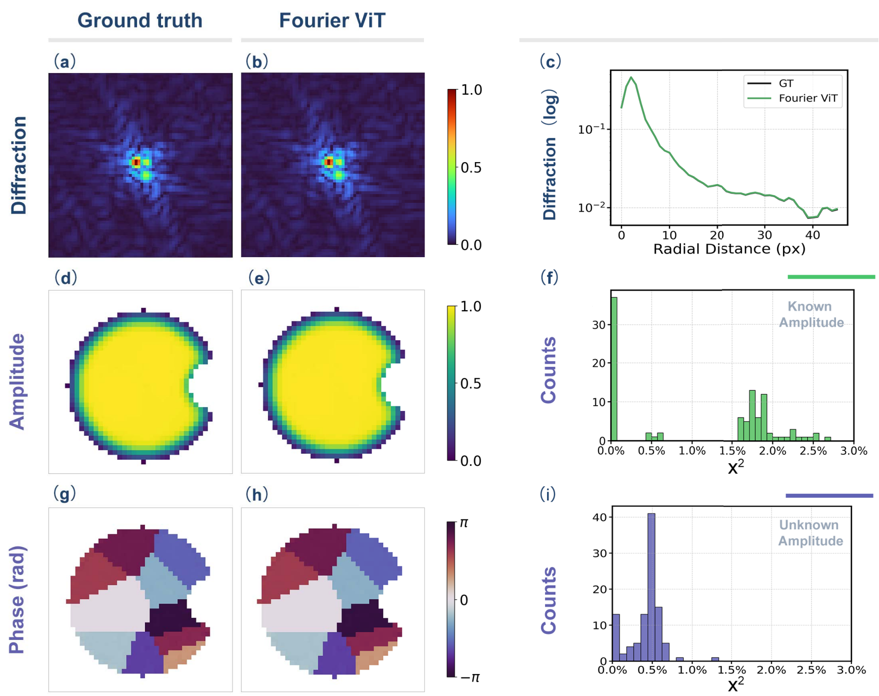

Fig. 4. Fourier ViT reconstruction of a synthetic 10-domain crystal from diffraction only. (a,d,g) Ground truth diffraction magnitude (64×64 pixels), crystal amplitude and phase (40×40 pixels); (b,e,h) corresponding Fourier ViT reconstructions. (c) Radial log intensity against q profiles of ground truth and reconstructed diffraction, showing agreement across the full q-range. (f,i) Histograms of diffraction χ^2 over 100 random initialisations for known crystal amplitude (f) and unknown amplitude (i) phase retrieval.

time; in a practical multi-start protocol, such runs would be discarded. With fixed amplitude, the optimisation has limited freedom to compensate for phase errors, so these solutions retain stable non-zero residuals. In the joint amplitude–phase recovery case in Fig. 4(i), the additional degrees of freedom from amplitude prediction reduce the number of near-perfectly converged runs but concentrate the remaining outcomes into a narrower band of lower χ^2. This indicates that small phase deviations can be partially offset by amplitude reweighting within the support, improving agreement with the measured diffraction magnitude.

These results show that reciprocal-space modelling with Fourier attention reproduces the diffraction structure and reconstructs the multi-domain phase topology without real-space labels. Unless otherwise stated, all Fourier ViT models were trained in an unsupervised manner using the optimisation and loss schedule described in Sec. 2.3.

For the same simulated diffraction pattern of a 10-domain crystal and fixed support, iterative methods combining Error Reduction (ER) and Hybrid Input-Output (HIO) reach χ^2 values between 10^{-4} and 10^{-3} for approximately half of reconstructions after 500 iterations. No runs reach $\chi^2 \leq 10^{-5}$ with the iterative method, but increasing to 2000 iterations raises the fraction of runs with $10^{-4} \leq \chi^2 \leq 10^{-3}$ to 90%. Under matched initialisation and support constraints, the Complex Convolutional Neural Network (C-CNN) with skip connections achieves χ^2 values of 3.1×10^{-3}–1.1×10^{-2}. Among all three methods, the χ^2 outcomes over random initialisations are bimodal; however, only the Fourier ViT reaches the lowest-χ^2 value ($\chi^2 \leq 10^{-5}$), whereas the C-CNN remains confined to a higher-error mode.

Optics EXPRESS

3.2. Projected 3D crystals and support robustness

The previous synthetic tests in Fig. 4 used a flat-core amplitude profile with sharp Voronoi phase domains. We next considered 2D projected objects generated from simulated 3D complex crystals using the workflow demonstrated in Fig. 3. After projection, the amplitude becomes spatially non-uniform and the phase variations are smoother than in the idealised Voronoi case.

Figure 5 first considers Crystal A, a projected sphere with a crescent cutout. With the ground-truth support, Fourier ViT recovers the main non-uniform amplitude contrast and the dominant projected phase structure in Figs. 5(a,b). Over 100 independent random initialisations, the χ^2 histogram in Fig. 5(g) is bimodal, with values ranging from 6.6×10^{-4} to 2.8×10^{-2} after 1000 epochs.

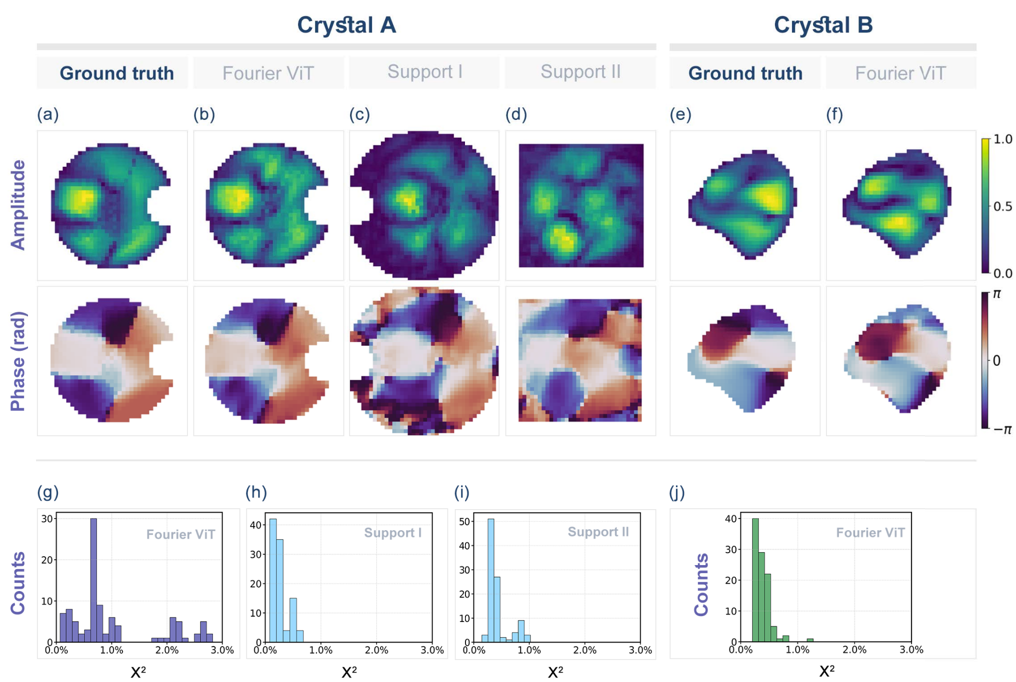

Fig. 5. Fourier ViT reconstructions from 2D projections of simulated 3D crystal models. All images are 40×40 pixel crops of the projected object amplitude and corresponding phase. (a) Ground-truth projected amplitude (upper) and phase (lower) for Crystal A. (b) Fourier ViT reconstruction of Crystal A using the ground-truth support. (c,d) Reconstructions of the same diffraction pattern with two alternative support constraints, denoted Support I and Support II. (e) Ground-truth projected amplitude and phase for Crystal B. (f) Fourier ViT reconstruction of Crystal B with the ground-truth support. (g–j) Histograms of diffraction χ^2 over 100 independent random initialisations for the reconstructions in (b), (c), (d), and (f), respectively.

We then reconstructed the same diffraction pattern using two deliberately mismatched supports. Support I is an enlarged circular mask with a left-side notch that does not encode the true right-side crescent cutout of Crystal A. Even under this mismatch, the reconstruction in Fig. 5(c) does not simply conform to the imposed support boundary. Instead, the recovered amplitude develops a reduced-intensity region on the right-hand side and retains the main internal contrast of the projected object, while the phase map preserves the dominant large-scale phase variation. The corresponding χ^2 values in Fig. 5(h) lie between 8.0×10^{-4} and 6.8×10^{-3}, remaining concentrated in the low-error regime. Using the geometrically mismatched 32×32 square

Optics EXPRESS

Support II in Fig. 5(d) leads to larger local deviations in both amplitude and phase, but low-error solutions are still obtained, with χ^2 values between 1.5×10^{-3} and 1.0×10^{-2} in Fig. 5(i).

To test whether the result depends on the crescent geometry, we also generated Crystal B with a less ideal shape, projected to 2D by the same workflow. In Figs. 5(e,f), Fourier ViT still recovers the main amplitude lobes and the large-scale projected phase variation, although the local deviations are stronger than for Crystal A. The χ^2 histogram in Fig. 5(j) remains concentrated below $\sim 7.0 \times 10^{-3}$, with values between 2.1×10^{-3} and 1.3×10^{-2}, and the best run reaches $\chi^2 = 2.0 \times 10^{-3}$.

Taken together, Fig. 5 shows that the method is not limited to the idealised Voronoi benchmark of Fig. 4. When the object is a 2D projection of a 3D complex crystal, Fourier ViT still matches the target diffraction magnitude and recovers the main projected amplitude–phase morphology. Solutions that match the measured diffraction magnitude with low χ^2 remain accessible even when the support is enlarged or geometrically mismatched.

3.3. Noise robustness

To assess robustness under realistic experimental degradations, the same noise-free diffraction pattern used in Fig. 4 was corrupted by three noise models: additive Gaussian noise, accounting for detector background; Poisson counting noise representing counting statistics; and partial coherence of the incident beam as demonstrated in Fig. 6. Strictly, partial coherence is not the same as the statistical noise models because it can be corrected [64], but its effects on convergence are worth identifying so they can be recognised and corrected in practice. Each noise model was applied at five strengths, from Level 1 to Level 5, according to its effect on χ^2. For visual clarity, Fig. 6 shows Levels 1, 3 and 5; Levels 2 and 4 are provided in Supplement 1, Fig. S1. We applied the same noise-addition procedure to the 2D projection data for crystal A in Fig. 5; the corresponding results are provided in Supplement 1, Fig. S2 and Table S1.

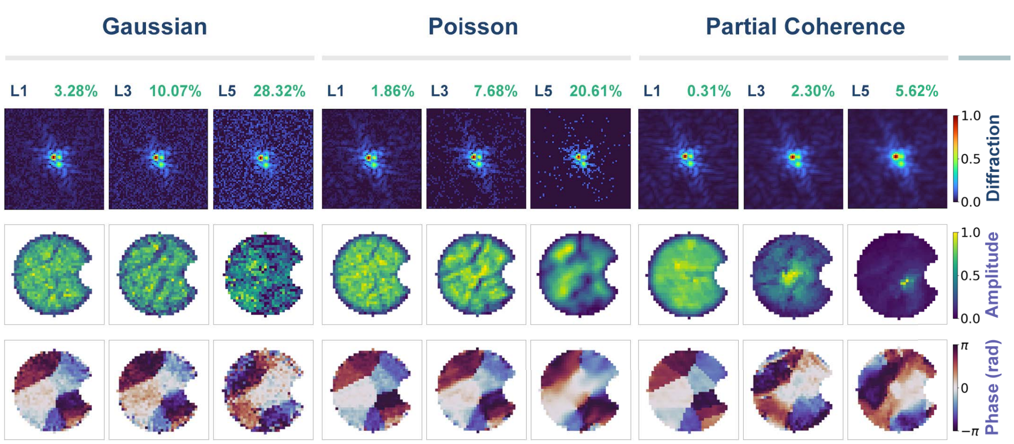

Fig. 6. Noise robustness of the Fourier ViT. Columns show three types of noise applied to the same diffraction pattern: Gaussian noise, Poisson noise, and partial coherence. For each model, L1, L3 and L5 denote increasing noise strengths (levels 1, 3 and 5); green percentages give χ^2_{n} (in %), the input diffraction error relative to the clean ground truth. Top row shows the noisy input diffraction amplitudes (64×64 pixels). Middle and bottom rows show the reconstructed crystal amplitude and phase zoomed in 40×40 pixels, respectively; both are masked to the support and phase is in radians.

For each noise type, Table 1 reports three normalised diffraction errors: (i) the noisy input relative to the clean ground truth, χ^2_{n}; (ii) the reconstruction mismatch to the noisy input, $\chi^2_{\mathrm{rec,n}}$; and (iii) the reconstruction error relative to the clean ground truth, $\chi^2_{\mathrm{rec,c}}$. In our unsupervised setting, we select the best run across random initialisations by minimising $\chi^2_{\mathrm{rec,n}}$, and report the corresponding $\chi^2_{\mathrm{rec,c}}$ as evaluation against the clean ground truth.

For Gaussian noise, χ^2_{n} increases from 3.28% to 28.32% with noise strength. The Fourier ViT reduces the error relative to the clean diffraction to $\chi^2_{\mathrm{rec,c}} = 1.54$–14.36%. For Poisson noise, χ^2_{n} increases from 1.86% to 20.61%, and the model reduces this to $\chi^2_{\mathrm{rec,c}} = 0.92$–10.53%. Across both stochastic noise processes, the improvements are likely due to oversampling and the use of a support.

Figure 6 illustrates the effects of Gaussian and Poisson noise on the real-space images. At low noise levels, the fringes of the diffraction remain sharp and the reconstructions show clean and domain-resolved crystal phase maps. As noise increases, pixel-scale fluctuations appear and progressively corrupt the phase, eventually blurring or erasing the domain boundaries. Under Gaussian noise, pixel-scale intensity fluctuations developed in both the reconstructed crystal's amplitude and phase, while under Poisson statistics the amplitude develops grain boundary contrast that mirrors the underlying domain structure. In both cases, the Fourier ViT suppresses unstructured speckle and background fluctuations, and preserves the physically meaningful fringe pattern around the centre of the Bragg peak, thereby enabling robust phase retrieval even under strongly degraded conditions. Consistent with Table 1, the Fourier ViT therefore acts like a filter, reducing the diffraction domain error by roughly a factor of two in χ^2 relative to the noisy input.

For the partial coherence condition, χ^2_{n} increases from 0.31% to 5.62% with increasing blur. The reconstruction matches the blurred measurement extremely closely as $\chi^2_{\mathrm{rec,n}} = 0.11$–0.25%, but the error relative to the clean diffraction lies in the range $\chi^2_{\mathrm{rec,c}} = 0.25$–6.37%. The ratio $\chi^2_{\mathrm{rec,c}}/\chi^2_{\mathrm{n}}$ is ≈ 0.81 at the weakest level, but increases to ≈ 1.13–1.29 at stronger blur. The diffraction fringes are progressively blurred and power is concentrated near the central Bragg peak. The reconstructed amplitude shows an over-localised central intensity ("hot spot" region) and a smoothed phase, reflecting the loss of high-q information. Partial coherence has been reported to produce an over-localised central amplitude "hot spot" [67]. We therefore consider it a possible mechanism for similar artefacts in the experimental reconstructions discussed in Sec. 3.4.

3.4. Experimental verification

3.4.1. Weak-phase $SrTiO_3$

As a preliminary experimental check, we tested the Fourier ViT on a 2D diffraction pattern from a weak-phase, single-domain $SrTiO_3$ (STO) nanocrystal with a cross-section size of $\sim 460 \times 450$ nm^2. Following the support-refinement procedure in Sec. 2.4, we first reconstructed the object using an initial 25×25 pixel square support, as shown in Fig. 7(b,f). This first reconstruction gives a confined single-crystal morphology, but the amplitude contains a central hot spot and residual intensity near the support boundary, and the phase still shows weak texture inside the support.

The amplitude from this first Fourier ViT reconstruction was then thresholded at 0.06 to define a refined support, and the reconstruction was repeated using this support, as shown in Fig. 7(c,g). The refined-support reconstruction gives a more filled amplitude distribution and suppresses diffuse background outside the crystal. Furthermore, the flatter phase observed within the support is consistent with the expected weak-phase character of the STO nanocrystal. Using the same refined support, the iterative reconstruction provides a conventional reference under the same real-space constraint [Fig. 7(d,h)]. For Fourier ViT, these improvements arise from the combination of the tighter real-space support and the amplitude-prior blending schedule described in Sec. 2.3, which stabilises the modulus during early optimisation before being relaxed.

Optics EXPRESS

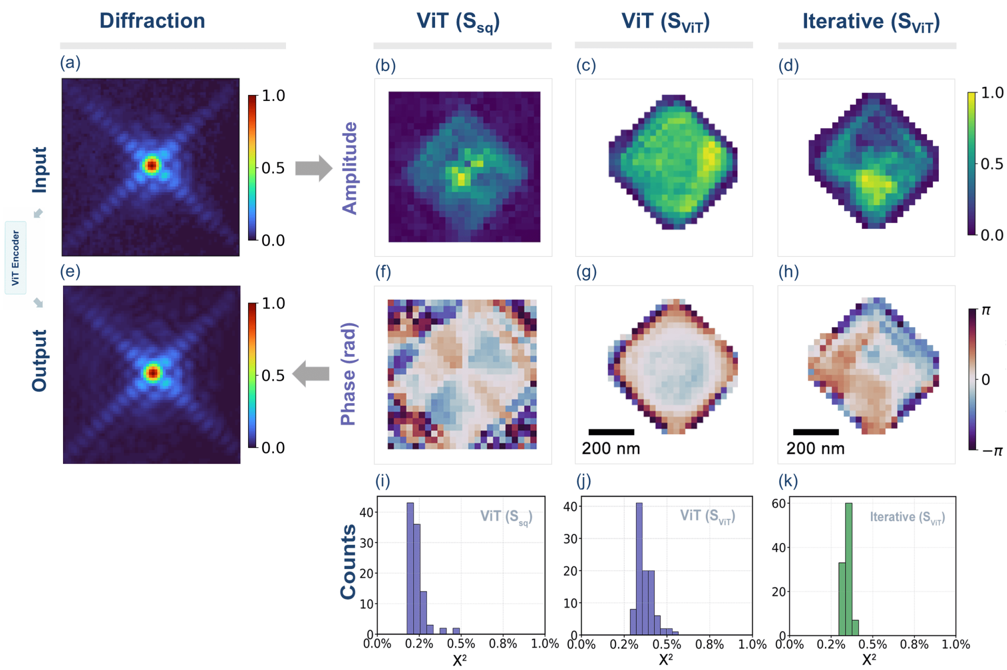

Fig. 7. Single-domain $SrTiO_3$ (STO) nanocrystal with a cross-section size of $\sim 460 \times 450$ nm^2, reconstructed from the same measured diffraction pattern using Fourier ViT with an initial square support (S_{sq}), Fourier ViT with a refined support (S_{ViT}), and an iterative method with S_{ViT}. (a,e) Measured and reconstructed diffraction magnitudes, shown as 64×64 pixel crops centred on the Bragg peak and max-normalised to [0, 1]; (e) is calculated from the lowest-χ^2 Fourier ViT reconstruction using S_{ViT}. (b–d) Reconstructed real-space amplitudes for Fourier ViT (S_{sq}), Fourier ViT (S_{ViT}), and the iterative method (S_{ViT}), respectively. (f–h) Corresponding phases wrapped to $[-\pi, \pi]$; scale bars represent 200 nm. (i–k) Corresponding χ^2 histograms over 100 random initialisations, shown over the range 0.0%–1.0%.

Across 100 independent random initialisations, Fourier ViT with the square support S_{sq} gives $\chi^2 = 0.18$–0.49% [Fig. 7(i)], Fourier ViT with the refined support S_{ViT} gives $\chi^2 = 0.29$–0.57% [Fig. 7(j)], and the iterative reconstruction using S_{ViT} gives $\chi^2 = 0.30$–0.42% [Fig. 7(k)]. These values remain below 1.0%, consistent with stable weak-phase reconstruction. The main effect of support refinement in this case is therefore not simply to reduce the reciprocal-space error, but to give a cleaner real-space object with a better-defined boundary and a flatter weak-phase distribution.

3.4.2. Strong-phase Fe_3O_4 with support transfer

To address the support uncertainty raised for strong-phase experimental reconstructions more directly, we next consider BCDI data from a micron-sized Fe_3O_4 crystal measured above and below its Verwey transition temperature (T_c) [28]. Domain formation occurs in the monoclinic phase of Fe_3O_4 below T_c, requiring a strong phase reconstruction. This example is particularly useful because the same crystal was measured both above and below T_c, and its overall morphology remained approximately block-like.

Above T_c, the diffraction pattern is comparatively simple [Fig. 8(a)] and the reconstructed crystal phase is weak. This allowed us to derive a tight support from the Fourier ViT amplitude without recourse to iterative reconstruction. Starting from a neutral square support, we first reconstructed the high-temperature amplitude [Fig. 8(b)]. Thresholding this amplitude at 0.06,

Optics EXPRESS

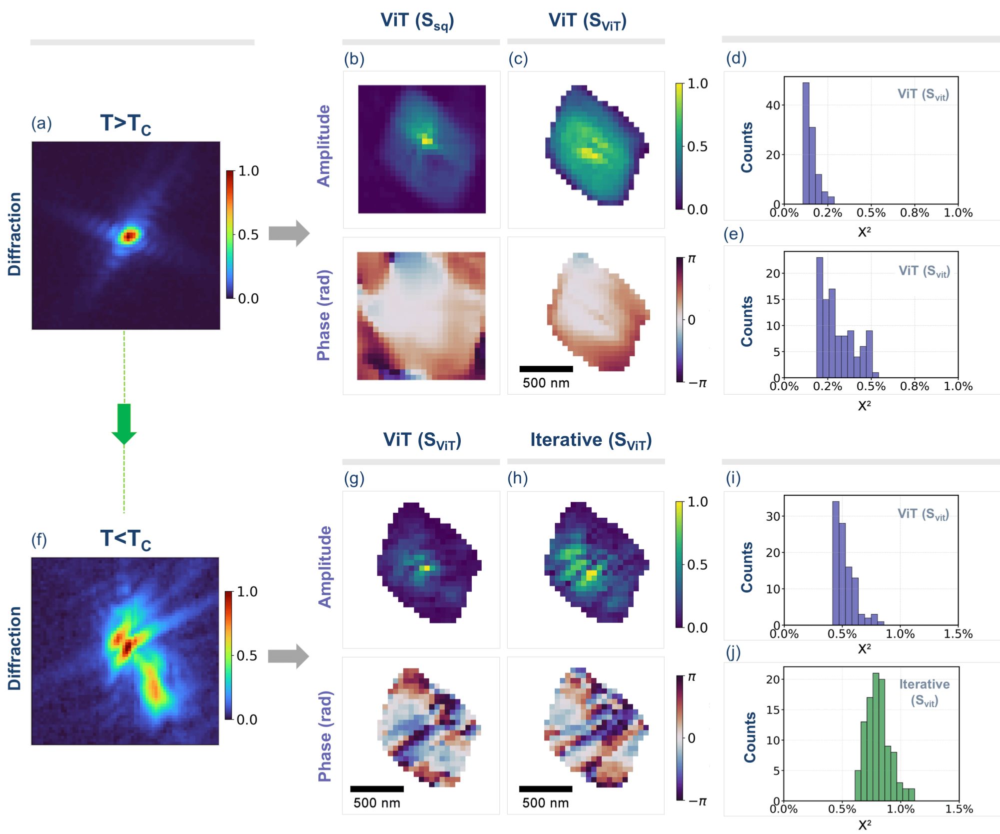

Fig. 8. Support-transfer study of phase retrieval on experimental BCDI data from a Fe_3O_4 crystal measured above and below the Verwey transition. (a,f) Measured diffraction magnitudes (64 × 64 pixels) for $T>T_c$ and $T<T_c$, respectively. (b,c) Fourier ViT reconstructions above T_c using an initial square support (S_{sq}) and a refined Fourier ViT support (S_{ViT}). (d,e) Corresponding χ^2 histograms over 100 independent random initialisations. (g,h) Fourier ViT and iterative reconstructions below T_c, both using the same refined support S_{ViT} transferred from the above-T_c reconstruction. (i,j) Corresponding χ^2 histograms over 100 independent random initialisations. Amplitudes are independently normalised to their maximum value and phases are shown in radians over $[-\pi, \pi]$. Scale bars represent 500 nm.

followed by connected-component cleaning, produced the refined particle support used for the subsequent high- and low-temperature reconstructions shown in Figs. 8(c,g,h). For $T>T_c$, the first reconstruction from the square support already converges reliably, with χ^2 values between 0.13% and 0.27% over 100 random initialisations, as shown by the χ^2 histogram in Fig. 8(d). Using the refined support, the $T>T_c$ reconstruction remains consistent with a weak-phase, single-crystal state and yields χ^2 values between 0.19% and 0.54% [Fig. 8(e)]. The refined support produces a more compact, filled amplitude within the particle envelope in Fig. 8(c), and the corresponding phase remains weak and comparatively smooth.

The low-temperature diffraction pattern at $T<T_c$ in Fig. 8(f) shows strong splitting due to domain formation. Using the refined support from the high-temperature state, the Fourier ViT recovers a stripe-like projected phase contrast below T_c [Fig. 8(g)], with χ^2 values between 0.42% and 0.86% over 100 independent runs [Fig. 8(i)]. The iterative reconstruction using the same support gives a similar low-temperature phase contrast, but with a higher and broader χ^2 distribution of 0.61%–1.12% over 100 independent runs [Figs. 8(h,j)]. The key change between

Optics EXPRESS

high- and low-temperature reconstructions lies in the crystal's phase: it is comparatively weak and smooth above T_c [Fig. 8(c)] and develops a stripe-like projected phase contrast below T_c [Figs. 8(g,h)].

Since reproducibility is a central difficulty in strong-phase BCDI, we further tested the Fourier ViT result using the 50 lowest-χ^2 reconstructions below T_c. Before comparison, each reconstruction was corrected for the conjugate-twin ambiguity, integer-pixel shifts and global phase offset. The complex-object average preserves the particle support and the main stripe-like projected phase contrast [Supplement 1, Fig. S3(a)], and the three lowest-χ^2 runs show similar dominant phase features [Supplement 1, Fig. S3(b–d)]. To check that this contrast is not only caused by the cyclic wrapped-phase colour scale, we also unwrapped the averaged reconstruction and the lowest-χ^2 run and plotted them with a linear, non-periodic colour scale [Supplement 1, Fig. S4(a,b)]. The stripe-like contrast remains visible after unwrapping and is consistent with monoclinic-domain contrast below (T_c) [28]. We therefore interpret the dominant stripe-like projected phase contrast as reproducible within the best-fitting ensemble.

3.4.3. Strong-phase $La_{0.5}Ca_{0.5}MnO_3$

We then consider the strong-phase regime using an experimental BCDI pattern from a strongly distorted, multi-domain LCMO nanocrystal (Sec. 2.5). Figure 9 compares reconstructions from a traditional iterative method (ER/HIO), the Fourier ViT, and a C-CNN baseline, using the same real-space support from the iterative shrink-wrap reconstruction described in Sec. 2.6. For each method, we report the lowest χ^2 over 100 random initialisations. The support is used as a common reconstruction constraint, not as an independent ground-truth morphology.

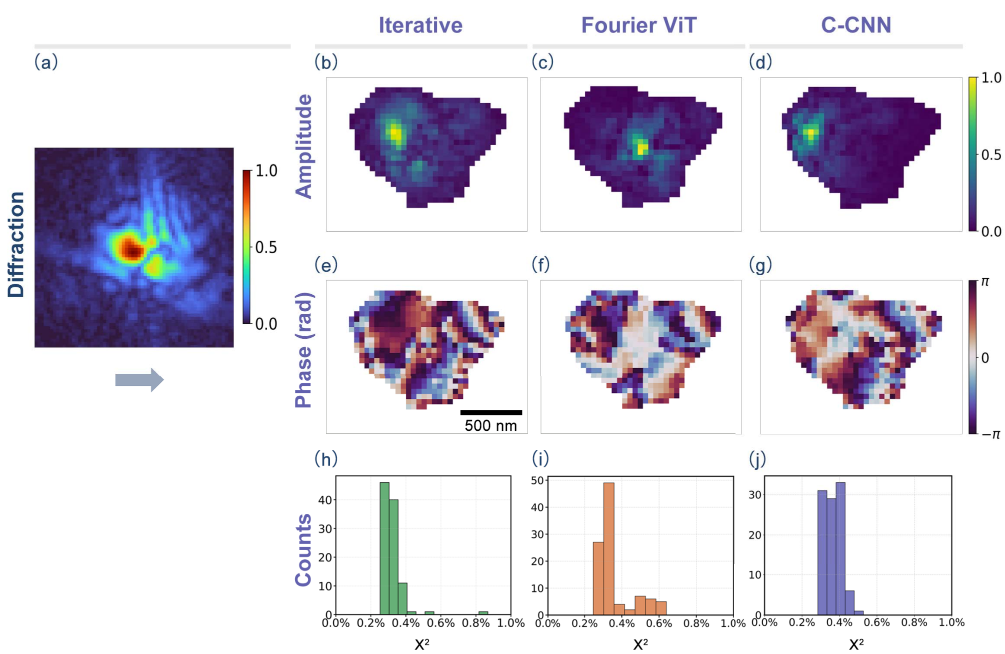

Fig. 9. Comparison study of phase retrieval on experimental data from a multi-domain LCMO nanocrystal. (a) Measured BCDI diffraction magnitude (64 × 64 pixels). (b–d) Reconstructed amplitude and (e–g) phase from the iterative method, Fourier ViT, and C-CNN using the same fixed real-space support for controlled comparison; scale bar represents 500 nm. (h–j) Corresponding χ^2 histograms over 100 independent random initialisations per method. The χ^2 range is shown between 0.0% and 1.0%.

Trained unsupervised on this single diffraction pattern, the Fourier ViT returns the real-space amplitude and phase in Fig. 9(c,f). The amplitude reconstructions obtained with the three methods in Fig. 9(b–d) all contain localised high-amplitude features. These hot spots may reflect the remaining real-space freedom allowed by the fixed support, together with detector noise and possible partial-coherence effects. In addition, the max-pixel normalisation used during reconstruction and visualisation can make localised maxima appear more prominent in the amplitude maps. We therefore avoid assigning the small differences in amplitude localisation to a unique reconstruction advantage. In the phase maps [Fig. 9(e–g)], the detailed phase contrast differs between methods, with the most reliable phase interpretation coming from the higher-amplitude regions of the reconstructed objects.

To reduce the hot spot in the crystal amplitude, we tested a mild partial-coherence correction by Gaussian deconvolution of the measured diffraction magnitude [67]. While the deconvolution visibly sharpens the fringes, it also amplifies detector noise, and the reconstruction χ^2 increases relative to the raw-data case: the iterative method gives $\chi^2 = 0.61\%$, the Fourier ViT gives $\chi^2 = 0.55\%$, and the C-CNN gives $\chi^2 = 0.58\%$. We therefore use the raw measured pattern as the primary basis for comparison in Fig. 9.

The χ^2 histograms in Fig. 9(h–j), plotted on a common axis, summarise run-to-run variability over 100 independent initialisations. For the iterative method [Fig. 9(h)], most reconstructions fall between $\chi^2 = 0.25$–0.40%, with a few higher-error runs extending to $\chi^2 = 0.87\%$. Fourier ViT [Fig. 9(i)] is mainly distributed between $\chi^2 = 0.25$–0.35%, with a tail extending to $\chi^2 = 0.64\%$, while C-CNN [Fig. 9(j)] lies within $\chi^2 = 0.28$–0.53%. These overlapping ranges show that all three methods can fit the measured diffraction magnitude to a similar level, but their real-space amplitude and phase solutions are not identical. Visualization 1, Visualization 2 and Visualization 3 show the per-epoch evolution of the three methods and highlight distinct optimisation trajectories. In ER/HIO, the negative feedback step helps the iterations escape local minima and reach low-χ^2 solutions. The C-CNN tends to build the phase map from smooth initial fields towards sharper domain boundaries, adding spatial detail gradually during training; in our tests, it more often converges to solutions with edge-localised amplitude and a phase topology that differs from the other methods. By contrast, the Fourier ViT couples the diffraction-pattern features globally through token mixing and produces multi-domain configurations early in training. This reduces time spent in over-smoothed intermediate states and moves the optimisation quickly into the relevant strong-phase solution class.

Overall, on this experimental dataset, the Fourier ViT reaches reciprocal-space agreement comparable to the iterative method and the C-CNN baseline. The LCMO result should therefore be read as a controlled support-uncertain comparison under the same fixed real-space constraint, rather than as proof of a unique real-space reconstruction.

4. Conclusions

We developed a Vision Transformer with Fourier attention (Fourier ViT) for multi-domain phase retrieval in BCDI. On synthetic 64×64 diffraction patterns, Fourier ViT recovers strong-phase domain structure directly from diffraction in an unsupervised setting. We first tested this in a smooth-amplitude, sharp-boundary synthetic benchmark, considering both the phase-only limit with known amplitude and joint amplitude-phase reconstruction. For phase-only prediction, runs can reach very close agreement with the target diffraction and resolve complex domain structure up to 19 domains [Fig. 4(f)]. For a representative 10-domain case, the model reproduces the ground-truth diffraction magnitude over the full q range, including the high-q tails, and recovers the same phase-domain layout with sharp domain walls [Fig. 4(c,h)]. Under the same fixed support on the same pattern, ER/HIO can also reach low reciprocal-space error when sufficient iterations are used, whereas the complex convolutional neural network (C-CNN) remains at higher error under the same initial setup. Across all three methods, random initialisation in the strong-phase regime separates runs into low- and higher-χ^2 outcomes under the same constraints.

The smooth-amplitude, sharp-boundary benchmark in Fig. 4 provides a controlled thin-film-like 2D limit for testing strong-phase retrieval. We then extended the test to simulated 3D complex crystals projected into 2D using the workflow in Fig. 3, where amplitude becomes non-uniform and phase variations are smoother. This setting is closer to projected multi-domain nanocrystals such as LCMO and magnetite Fe_3O_4. Fourier ViT still recovers the main projected amplitude–phase morphology and reaches low reciprocal-space error (Fig. 5). For Crystal A, the reconstruction with the ground-truth support recovers the main amplitude and phase features, and low-error solutions remain accessible with mismatched supports. For the less ideal Crystal B, Fourier ViT recovers the main amplitude lobes and large-scale phase variation. These results show that Fourier ViT can recover low-error solutions beyond the controlled thin-film-like 2D limit, although local amplitude and phase details remain sensitive to support accuracy.

Robustness tests (Fig. 6, Table 1) show that, under Gaussian and Poisson noise, the reconstructions remain closer to the clean diffraction than the noisy input across all levels ($\chi^2_{\mathrm{rec,c}} < \chi^2_{\mathrm{n}}$). The same trend is observed for the projected Crystal A test in Supplement 1, Fig. S2 and Table S1, despite the non-uniform projected amplitude. By contrast, partial coherence blurs the diffraction and reduces fringe visibility. Although the model can fit the blurred measurement extremely closely, the reconstructed diffraction departs from the clean target as blur increases. Partial coherence can also produce an over-localised central amplitude "hot spot", a behaviour documented previously [67]. This may contribute to similar artefacts in the experimental reconstructions, particularly because simple deconvolution methods intended to address this blur can simultaneously amplify existing noise. Therefore, an explicit coherence model is needed when the blur is significant. All three distortions appear differently in the predicted images: Gaussian noise introduces noisy local features, Poisson noise produces grainy fluctuations, and partial coherence produces less well-defined domain boundaries.

We next tested the support-refinement strategy on experimental weak- and strong-phase data. For the weak-phase STO nanocrystal, Fourier ViT reconstructs a confined single-crystal object from an initial square support [Fig. 7(d,f)]. Thresholding this first amplitude reconstruction gives a refined support, which suppresses diffuse intensity outside the crystal and produces a flatter phase within the support [Fig. 7(e,g)]. Therefore, in the weak-phase STO case, support refinement mainly improves the real-space support shape rather than simply lowering the reciprocal-space error.

The Fe_3O_4 experiment provides a more direct test of support transfer for strong-phase reconstruction. Above T_c, the same crystal is in the weak-phase state and approximately block-like, allowing a refined support to be obtained from Fourier ViT without using an iterative shrink-wrap reconstruction [Figs. 8(a,e,i)]. This support is then transferred to the low-temperature diffraction pattern, where domain formation produces strong splitting in reciprocal space [Fig. 8(c)]. Using the high-temperature support, Fourier ViT recovers a striped low-temperature phase-domain structure [Fig. 8(k)]. The key point is that the support is derived from the weak-phase state of the same crystal and then used for the strong-phase state, reducing the ambiguity associated with choosing a support directly from a distorted multi-domain reconstruction.

We finally tested a more challenging experimental strong-phase case using measured diffraction from the LCMO nanocrystal. Fourier ViT was compared with ER/HIO and the C-CNN under the same support constraint and FFT-based forward model in the loss. In Fig. 9, Fourier ViT reaches χ^2 values comparable to both the iterative reconstruction and the C-CNN baseline. The overlapping χ^2 ranges mean that the LCMO comparison should not be read as a strict ranking of the three methods. Instead, it shows that Fourier ViT remains competitive in an experimental strong-phase case, using a support taken from an iterative shrink-wrap reconstruction and applied equally to all methods.

The reconstruction error χ^2 histograms in Fig. 9(h–j) also highlight a practical aspect of the strong-phase regime. With realistic noise and unavoidable model mismatch under a magnitude-only data constraint, many near-degenerate minima can fit the measured diffraction to a similar level. As a result, reconstructions may have similar reciprocal-space errors while remaining meaningfully different in real space. Therefore, reducing χ^2 is necessary but not sufficient to identify a unique real-space domain solution. The broader spread of Fourier ViT outcomes may be consistent with access to multiple acceptable strong-phase solutions through global token mixing, but may also indicate run-to-run optimisation sensitivity. Practically, we report the best-χ^2 reconstructions and check that key domain-level features are reproducible across independent runs when the measured intensity and imposed constraints do not by themselves enforce a unique solution.

Computationally, ER/HIO is dominated by repeated FFT projections, typically one forward and one inverse FFT per iteration [5,68]. In the unsupervised learning condition, the FFT forward model is embedded in the loss alongside network forward/backward propagation for both the C-CNN and Fourier ViT. Within the Fourier ViT, Fourier token mixing enables global coupling while avoiding the quadratic scaling of dot-product self-attention [47,48]. Scaling the approach to larger grids and fully 3D BCDI volumes will therefore require additional memory and computational power.

Overall, Fourier ViT offers a practical route to strong-phase, multi-domain BCDI reconstructions. In near-ideal conditions, all methods improve with additional iterations or epochs. However, for the optimisation settings tested here, Fourier ViT reaches structured multi-domain solutions within early training epochs, while remaining comparable to the best ER/HIO and C-CNN baseline reconstructions on experimental data. Future work will incorporate explicit noise and partial-coherence models in the forward operator, adapt shrink-wrap support updating within the Fourier ViT optimisation, use free log-likelihood together with multi-start reconstruction ensembles to assess solution reliability, and extend the approach to fully 3D experimental datasets.

Funding. United States Department of Energy (DE-SC0012704, DE-AC02-06CH11357); U.S. National Science Foundation (DMR-9724294); Engineering and Physical Sciences Research Council (EP/I022562/1); European Research Council (ERC) (227711).

Acknowledgments. We would like to thank Xi Yu, Wenxuan Fang, Jack Griffiths, Yue Dong, Ruiyang Duan, Seungjun Lee, Yuewei Lin, Longlong Wu and Shinjae Yoo for helpful discussions, and Weibin Peng for advice on figure design and presentation. We thank Ross Harder, Tadesse Assefa, Yue Cao, Xiaojing Huang and Jesse Clark for help with the measurements of LCMO and Josh Turner and Arup Kumar Raychaudhuri for providing the sample. The measurements were carried out at the Advanced Photon Source (APS) beamline 34-ID-C, which was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. The beamline 34-ID-C was built with U.S. National Science Foundation grant DMR-9724294. Work performed at UCL was supported by EPSRC EP/I022562/1 and ERC Advanced Grant 227711. Work at Brookhaven National Laboratory was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-SC0012704.

Disclosures. The authors declare no conflicts of interest.

Data availability. The source data that support the findings are available from the corresponding author upon reasonable request.

Code availability. The code used to train and evaluate the Fourier ViT and to reproduce all figures is available at [69].

Supplemental document. See Supplement 1 for supporting content.

References

1. J. Miao, P. Charalambous, J. Kirz, *et al.*, “Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens,” Nature **400**(6742), 342–344 (1999).
2. I. K. Robinson and R. Harder, “Coherent x-ray diffraction imaging of strain at the nanoscale,” Nat. Mater. **8**(4), 291–298 (2009).
3. M. A. Pfeifer, G. J. Williams, I. A. Vartanyants, *et al.*, “Three-dimensional mapping of a deformation field inside a nanocrystal,” Nature **442**(7098), 63–66 (2006).
4. J. Gulden, O. M. Yefanov, A. P. Mancuso, *et al.*, “Three-dimensional structure of a single colloidal crystal grain studied by coherent x-ray diffraction,” Opt. Express **20**(4), 4039–4049 (2012).
5. J. R. Fienup, “Phase retrieval algorithms: a comparison,” Appl. Opt. **21**(15), 2758–2769 (1982).

6. S. Marchesini, H. N. Chapman, S. P. Hau-Riege, *et al.*, "Coherent x-ray diffractive imaging: applications and limitations," Opt. Express **11**(19), 2344–2353 (2003).
7. D. Kim, M. Chung, J. Carnis, *et al.*, "Active site localization of methane oxidation on pt nanocrystals," Nat. Commun. **9**(1), 3422 (2018).
8. T. Liu, J. Liu, L. Li, *et al.*, "Origin of structural degradation in li-rich layered oxide cathode," Nature **606**(7913), 305–312 (2022).
9. J. Liu, D. Yang, A. F. Suzana, *et al.*, "Electric field driven domain wall dynamics in batio$_3$ nanoparticles," Phys. Rev. B **111**(5), 054101 (2025).
10. D. Sayre, "Some implications of a theorem due to shannon," Acta Crystallogr. **5**(6), 843 (1952).
11. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
12. V. Elser, "Phase retrieval by iterated projections," J. Opt. Soc. Am. A **20**(1), 40–55 (2003).
13. D. R. Luke, "Relaxed averaged alternating reflections for diffraction imaging," Inverse Problems **21**(1), 37–50 (2005).
14. I. Robinson, T. A. Assefa, Y. Cao, *et al.*, "Domain texture of the orthorhombic phase of $La_{2-x}Ba_xCuO_4$," J. Supercond. Novel Magn. **33**(1), 99–106 (2020).
15. L. Wu, P. Juhas, S. Yoo, *et al.*, "Complex imaging of phase domains by deep neural networks," IUCrJ **8**(1), 12–21 (2021).
16. Y. Gao, X. Huang, H. Yan, *et al.*, "Bragg coherent diffraction imaging by simultaneous reconstruction of multiple diffraction peaks," Phys. Rev. B **103**(1), 014102 (2021).
17. R. H. T. Bates, "Uniqueness of solutions to two-dimensional fourier phase problems for localized and positive images," Computer Vision, Graphics, and Image Processing **25**(2), 205–217 (1984).
18. A. Ulvestad, Y. Nashed, G. Beutier, *et al.*, "Identifying defects with guided algorithms in bragg coherent diffractive imaging," Sci. Rep. **7**(1), 9920 (2017).
19. J. Carnis, L. Gao, S. Labat, *et al.*, "Towards a quantitative determination of strain in bragg coherent x-ray diffraction imaging: artefacts and sign convention in reconstructions," Sci. Rep. **9**(1), 17357 (2019).
20. J. N. Clark, L. Beitra, G. Xiong, *et al.*, "Ultrafast three-dimensional imaging of lattice dynamics in individual gold nanocrystals," Science **341**(6141), 56–59 (2013).
21. R. Harder, "Deep neural networks in real-time coherent diffraction imaging," IUCrJ **8**(1), 1–3 (2021).
22. Y. M. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. **30**(3), 304–308 (1979).
23. Z. Zhuang, D. Yang, F. Hofmann, *et al.*, "Practical phase retrieval using double deep image priors," arXiv (2022).
24. X. Huang, R. Harder, G. Xiong, *et al.*, "Propagation uniqueness in three-dimensional coherent diffractive imaging," Phys. Rev. B **83**(22), 224109 (2011).
25. R. Beinert and G. Plonka, "Ambiguities in one-dimensional discrete phase retrieval from fourier magnitudes," J. Fourier Anal. Appl. **21**(6), 1169–1198 (2015).
26. T. Bendory, R. Beinert, and Y. C. Eldar, "Fourier phase retrieval: Uniqueness and algorithms," in *Compressed Sensing and its Applications*, H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, and R. Mathar, eds. (Springer, 2017), Applied and Numerical Harmonic Analysis, pp. 55–91.
27. T. A. Assefa, Y. Cao, J. Diao, *et al.*, "Scaling behavior of low-temperature orthorhombic domains in the prototypical high-temperature superconductor $la_{1.875}ba_{0.125}cuo_4$," Phys. Rev. B **101**(5), 054104 (2020).
28. Y. Dong, D. Yang, J. Liu, *et al.*, "Symmetry breaking during low-temperature domain formation in micron-sized magnetite crystals," Phys. Scr. **100**(8), 085937 (2025).
29. J. R. Fienup, "Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint," J. Opt. Soc. Am. A **4**(1), 118–123 (1987).
30. R. H. T. Bates and P. J. Napier, "Identification and removal of phase errors in interferometry," Mon. Not. R. Astron. Soc. **158**(4), 405–424 (1972).
31. C. Chen, J. Miao, C. W. Wang, *et al.*, "Application of optimization technique to noncrystalline x-ray diffraction microscopy: Guided hybrid input-output method," Phys. Rev. B **76**(6), 064113 (2007).
32. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature **401**(6755), 788–791 (1999).
33. V. Favre-Nicolin, S. Leake, and Y. Chushkin, "Free log-likelihood as an unbiased metric for coherent diffraction imaging," Sci. Rep. **10**(1), 2664 (2020).
34. V. Favre-Nicolin, G. Girard, S. Leake, *et al.*, "Pynx: high-performance computing toolkit for coherent x-ray imaging based on operators," J. Appl. Crystallogr. **53**(5), 1404–1413 (2020).
35. M. Guizar-Sicairos and J. R. Fienup, "Phase retrieval with fourier-weighted projections," J. Opt. Soc. Am. A **25**(3), 701–709 (2008).
36. M. J. Cherukara, Y. S. G. Nashed, R. J. Harder, *et al.*, "Real-time coherent diffraction inversion using deep generative networks," Sci. Rep. **8**(1), 16520 (2018).
37. L. Wu, S. Yoo, A. F. Suzana, *et al.*, "Three-dimensional coherent x-ray diffraction imaging via deep convolutional neural networks," npj Comput. Mater. **7**(1), 175 (2021).
38. L. Wu, D. Yang, W. Wang, *et al.*, "Unveiling nano-scale crystal deformation using coherent x-ray dynamical diffraction," npj Comput. Mater. **11**(1), 379 (2025).

Optics EXPRESS

39. Y. Rivenson, Y. Zhang, H. Günaydin, *et al.*, "Phase recovery and holographic image reconstruction using deep learning in neural networks," Light: Sci. Appl. **7**(2), 17141 (2018).
40. M. J. Cherukara, T. Zhou, Y. S. G. Nashed, *et al.*, "Ai-enabled high-resolution scanning coherent diffraction imaging," Appl. Phys. Lett. **117**(4), 044103 (2020).
41. T.-S. Vu, M.-Q. Ha, and H.-C. Dam, "Pid3net: a deep learning approach for single-shot coherent x-ray diffraction imaging of dynamic phenomena," npj Comput. Mater. **11**(1), 66 (2025).
42. X. Yu, L. Wu, Y. Lin, *et al.*, "Ultrafast bragg coherent diffraction imaging of epitaxial thin films using deep complex-valued neural networks," npj Comput. Mater. **10**(1), 24 (2024).
43. Y. Yao, H. Chan, S. Sankaranarayanan, *et al.*, "Autophasenn: unsupervised physics-aware deep learning of 3d nanoscale bragg coherent diffraction imaging," npj Comput. Mater. **8**(1), 124 (2022).
44. M. Masto, V. Favre-Nicolin, S. Leake, *et al.*, "Phase retrieval of highly strained bragg coherent diffraction patterns using supervised convolutional neural network," npj Comput. Mater. **12**(1), 164 (2026).
45. J. Zhao, E. Bellec, M.-I. Richard, *et al.*, "Bragg coherent modulation imaging of highly strained nanocrystals," Phys. Rev. Lett. **135**(25), 256101 (2025).
46. W. Gan, Q. Zhai, M. T. McCann, *et al.*, "PtychoDV: Vision transformer-based deep unrolling network for ptychographic image reconstruction," IEEE Open J. Signal Process. **5**, 539–547 (2024).
47. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, (2017).
48. J. Lee-Thorp, J. Ainslie, I. Eckstein, *et al.*, "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.
49. Y. Rao, W. Zhao, Z. Zhu, *et al.*, "Global filter networks for image classification," arXiv (2021).
50. J. Guibas, M. Mardani, Z. Li, *et al.*, "Adaptive fourier neural operators: Efficient token mixers for transformers," arXiv (2021).
51. Z. Li, N. B. Kovachki, K. Azizzadenesheli, *et al.*, "Fourier neural operator for parametric partial differential equations," in *International Conference on Learning Representations*, (2021).
52. I. K. Robinson, T. Assefa, Y. Cao, *et al.*, "Phase domains in crystals," (2026). Manuscript in preparation.
53. A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, (2021).
54. B. N. Patro and V. S. Agneeswaran, "Scattering vision transformer: Spectral mixing matters," arXiv (2023).
55. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv (2016).
56. S. Marchesini, H. He, H. N. Chapman, *et al.*, "X-ray image reconstruction from a diffraction pattern alone," Phys. Rev. B **68**(14), 140101 (2003).
57. N. Setter, D. Damjanovic, L. Eng, *et al.*, "Ferroelectric thin films: Review of materials, properties, and applications," J. Appl. Phys. **100**(5), 051606 (2006).
58. R. N. Bracewell, "Strip integration in radio astronomy," Aust. J. Phys. **9**(2), 198–217 (1956).
59. J. W. Goodman, *Statistical Optics* (John Wiley & Sons, Hoboken, NJ, 2015), 2nd ed.
60. J. R. Janesick, *Scientific Charge-Coupled Devices* (SPIE Press, Bellingham, WA, 2001).
61. J. R. Janesick, *Photon Transfer* (SPIE Press, Bellingham, WA, 2007).
62. C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," Nature **585**(7825), 357–362 (2020).
63. G. J. Williams, H. M. Quiney, A. G. Peele, *et al.*, "Coherent diffractive imaging and partial coherence," Phys. Rev. B **75**(10), 104102 (2007).
64. J. N. Clark, X. Huang, R. Harder, *et al.*, "High-resolution three-dimensional partially coherent diffraction imaging," Nat. Commun. **3**(1), 993 (2012).
65. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, (2019).
66. M. Monteforte, A. K. Estandarte, B. Chen, *et al.*, "Novel silica stabilization method for the analysis of fine nanocrystals using coherent X-ray diffraction imaging," J. Synchrotron Radiat. **23**(4), 953–958 (2016).
67. I. A. Vartanyants and I. K. Robinson, "Partial coherence effects on the imaging of small crystals using coherent x-ray diffraction," J. Phys.: Condens. Matter **13**, 10593–10611 (2001).
68. P. Duhamel and M. Vetterli, "Fast fourier transforms: A tutorial review and a state of the art," Signal Processing **19**(4), 259–299 (1990).
69. J. Liu, "Fast fourier transforms: A tutorial review and a state of the art," v1, Github (2026), https://github.com/JialunSimonLiu/Fourier-Attention_Vision-Transformer