HOLOFNO: A PHYSICS-SUPERVISED NEURAL OPERATOR FOR LOW-LATENCY HOLOGRAPHIC PHASE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fast synthesis of phase-only holograms is critical for programmable optical trapping, for which classic iterative phase-retrieval algorithms need repeated forward-backward propagation and constraint-projection steps. We formulate holographic phase synthesis as a physics-supervised inverse-design problem: the network predicts a spatial light modulator (or SLM) phase mask and is optimized through a differentiable Fourier-optics model against the reconstructed target-plane intensity, without requiring Gerchberg-Saxton (GS) generated phase labels. We investigate this problem at a fixed native discretization resolution of 128×128 pixels across both the SLM aperture and the focal plane, and introduce HoloFNO, a Fourier neural operator augmented with a harmonic-oscillator-inspired multi-frequency sinusoidal bypass. To uncover the internal representation mechanics and validate that the operator learns genuine diffraction physics rather than superficial heuristics, we perform fixed-resolution mechanistic evaluations consisting of branch-scrubbing interventions, low-order Zernike-residual decomposition, and mechanistic diagnostics across 3 independent training runs and 5 shared evaluation draws per run. Diagnostic probes demonstrate that the selected topology exhibits a reproducible phase content. At its 60-iteration quality-matched point, HoloFNO leads GS on reconstruction fidelity within the region of interest, quantified as the relative root-mean-square error (Rel-RMSE) between the reconstructed and ideal target intensity patterns ($Rel-RMSE_{ROI} = 0.1621 \pm 0.0253$ vs. 0.1494 ± 0.0213 for HoloFNO), while HoloFNO's fixed one-shot cost eliminates the per-sample runtime variability inherent to GS's convergence-dependent iteration count.

1 INTRODUCTION

Ultracold atoms in optical lattices provide a platform for quantum simulation with single-atom-resolved detection (Bakr et al., 2009; Gross & Bloch, 2017). In the blue-detuned setting considered here, optical dipole forces repel atoms from bright regions and can provide confinement near intensity minima (Grimm et al., 2000). A practical challenge is the rapid synthesis of a phase-only spatial-light modulator (SLM) pattern whose reconstructed intensity directly translates to a desired trapping configuration.

In our setting, the post-measurement atomic state ψ_B defines a probability density $\rho_B = |\psi_B|^2$, from which we construct a bounded blueprint intensity I_{bp}. The inverse-design problem can therefore be written as follows. $I_{\mathrm{bp}} \longmapsto \phi_{\mathrm{SLM}} \longmapsto I_\phi$, where I_ϕ is obtained through a differentiable Fourier-optics propagation model. The formulation separates the physical target from the particular phase representative used to realize it.

The problem is directly relevant to fast optical control and holographic generation in neutral-atom experiments, including learned hologram synthesis for tweezer arrays (Ren et al., 2024; Lin et al., 2025; Yan et al., 2026). We study a controlled family of targets generated from superpositions of two-dimensional quantum-harmonic-oscillator eigenstates.

For a phase-only SLM illuminated by a known input field, the focal-plane complex field is related to the SLM-plane field through the Fourier-optics propagation model used during training (Goodman,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

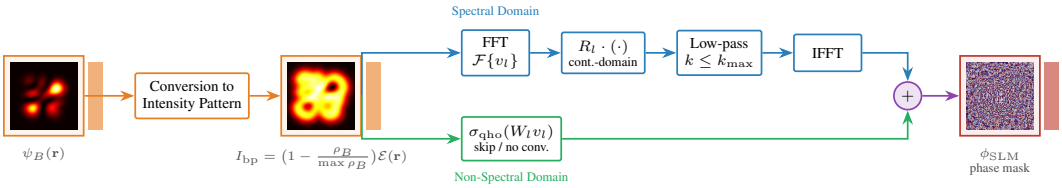

Figure 1: Forward pipeline from the measured atomic state to the retrieved SLM phase mask. The atomic wavefunction ψ_B (shown as its measured $|\psi_B|^2$ density plot) yields the blueprint intensity I_{bp} (Eq. 1), the atom density inverted and modulated by the envelope $\mathcal{E}(\mathbf{r})$. FNO produces the unwrapped phase, which is wrapped to give the final SLM mask ϕ_{SLM}

2005). Thus, the reconstructed trap depends on the spatial-frequency content of the phase mask. This global coupling motivates a spectral operator as a natural starting point. It also makes the effects of spectral truncation, positional information, and pointwise activations relevant to the learned representation.

The Computation. A classical approach to phase-mask generation is the Gerchberg–Saxton (GS) algorithm (Gerchberg & Saxton, 1972). GS alternates between propagation and constraint projection, so its runtime depends on the number of iterations. Moreover, phase retrieval is non-unique: different phase masks can produce similar reconstructed intensities (Fienup, 1982). A regression model trained against one GS solution therefore learns a particular numerical representative of a non-unique inverse problem. We instead optimize reconstructed intensity through the differentiable forward model.

Neural operators learn mappings between discretized functions and provide a natural framework for global interactions and inverse design (Kovachki et al., 2023). The Fourier Neural Operator (FNO) represents non-local interactions in Fourier space and shares its learned spectral parameters across the spatial domain (Li et al., 2021). However, neural-operator terminology does not by itself guarantee that a model learns a discretization-independent representation (Sakarvadia et al., 2026). We therefore ask a narrower question: *how does a spectral operator represent and route spatial structure when it is trained at a single native resolution?*

We study this question through three related objectives: (1) evaluate whether a physics-supervised FNO can generate useful phase masks with low latency; (2) determine whether an explicitly structured sinusoidal pathway contributes measurably to the learned representation; and (3) compare the resulting behavior with convolutional and U-Net baselines under in-distribution and out-of-distribution target families.

Our Contributions.

1. **Physics-supervised spectral hologram synthesis** We formulate phase-mask generation as a differentiable inverse-design problem in which the network is optimized against reconstructed optical intensity rather than GS-generated phase labels. At the native 128×128 resolution, the resulting model produces phase masks in 5.35 ± 0.19 ms on the reported hardware.

2. **Fixed-resolution analysis of learned spectral representations.** We evaluate the pure spectral FNO using coordinate-frame perturbations and spectral-mode diagnostics to characterize its dependence on scale, positional information, and retained spectral content.

3. **A structured sinusoidal bypass and branch intervention analysis.** We introduce a harmonic-oscillator-inspired multi-frequency sinusoidal pathway and compare alternative activation topologies. Fixed-weight branch interventions show that the single-saturation topology exhibits substantially stronger causal dependence of the non-low-order phase representation on the designated sinusoidal pathway.

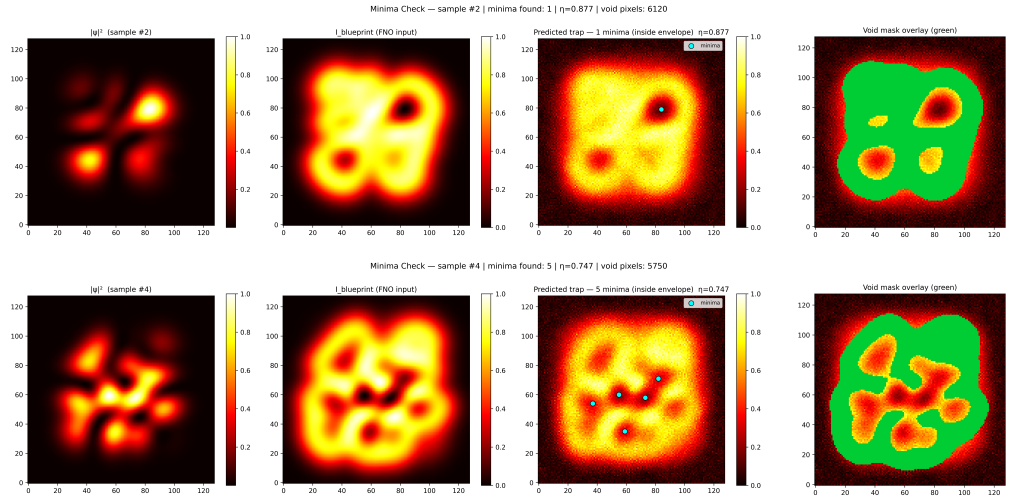

Figure 2: Examples of held-out targets and reconstructions for test samples #2 (the top one) and #4 (the bottom one). From left to right in each example: atomic probability density $|\psi|^2$, blueprint intensity I_{bp}, predicted SLM phase mask $\phi(x, y)$, reconstructed trap intensity with detected local minima, and the void mask used for confinement scoring.

2 Problem Formulation and Physics-Supervised Objective

We use (x, y) for coordinates in the SLM plane and (u, v) for coordinates in the target (focal) plane. Given the post-measurement wavefunction $\psi_B(u, v)$, the atomic probability density is

$$\rho_B(u, v) = |\psi_B(u, v)|^2.$$

We construct the normalized blueprint intensity as

$$I_{\mathrm{bp}}(u, v) = \left(1 - \frac{\rho_B(u, v)}{\max_{u,v} \rho_B(u, v)}\right) \mathcal{E}(u, v), \tag{1}$$

where $\mathcal{E}(\mathbf{r})$ is a smooth Gaussian envelope obtained by convolving a soft-thresholded version of ρ_B with a Gaussian kernel of width σ_{blur}. We also define the normalized density

$$\tilde{\rho}_B(u, v) = \frac{\rho_B(u, v)}{\max_{u,v} \rho_B(u, v)}.$$

The envelope confines the target structure to the physical region of the atomic state and suppresses spurious far-field content. The resulting I_{bp} is normalized to $[0, 1]$ and is used as the network input and target observable.

The network predicts a phase mask whose reconstructed intensity is evaluated by the differentiable optical model. This avoids matching a particular GS phase representative while retaining the non-uniqueness of the inverse problem. We therefore optimize reconstructed intensity rather than an algorithm-dependent phase realization.

We use a composite objective with five terms: density-weighted intensity fidelity ($\mathcal{L}_{fid}$), Sobel edge consistency ($\mathcal{L}_{edge}$), wrap-aware phase total variation ($\mathcal{L}_{TV}$), quadratic void confinement ($\mathcal{L}_{void}$), and boundary gradient-direction alignment ($\mathcal{L}_{dir}$):

$$\mathcal{L} = \lambda_{\mathrm{fid}}\mathcal{L}_{\mathrm{fid}} + \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}} + \lambda_{\mathrm{TV}}\mathcal{L}_{\mathrm{TV}} + \lambda_{\mathrm{void}}\mathcal{L}_{\mathrm{void}} + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}}. \tag{2}$$

We fix the loss weights across the principal architecture and activation ablations. We do not use test-set quantities to select them; we evaluate their sensitivity separately in Appendix B. The individual terms penalize residual intensity near the atomic density, encourage edge agreement, regularize phase variation, suppress intensity in the confinement region, and align the predicted and target intensity-gradient directions near the void boundary. Full definitions are given in Appendix B.

3 SELECTION OF ACTIVATION FUNCTION FOR THE SPECTRAL BLOCK

Before introducing the sinusoidal bypass (Section 5), we first determine the activation function for the spectral block itself by comparing the pure spectral baseline (Eq. 3) underGELU (Hendrycks & Gimpel, 2016), ReLU, and sigmoid. All three use the same architecture, spectral truncation, training protocol, and hyperparameters; only the pointwise activation differs.

Table 1: Trap-quality diagnostics for the pure spectral baseline under GELU, ReLU, and Sigmoid activations, evaluated across all 5 datasets (500 samples total).

Metric	**GELU**	**ReLU**	**Sigmoid**
$\mathrm{RMSE}_{\mathrm{ROI}}$	0.0847 ± 0.0107	0.0987 ± 0.0129	0.1038 ± 0.0127
S_{grad}	0.475 ± 0.025	0.445 ± 0.021	0.446 ± 0.020
κ_{geom} (target $= 1$)	0.821 ± 0.120	0.774 ± 0.144	0.780 ± 0.203
η_{conf}	0.405 ± 0.033	0.398 ± 0.028	0.406 ± 0.026

A complementary analysis of the late-layer spectral response is provided in Appendix G.2.

Among the three activations, GELU achieves the lowest $\mathrm{RMSE} * \mathrm{ROI}$ and the highest $S * \mathrm{grad}$ and κ_{geom} (Table 1), while η_{conf} remains comparable across the three variants. We therefore use GELU as the pointwise activation for the subsequent spectral and bypass-topology experiments.

4 THE PURE SPECTRAL FNO BASELINE AND SPECTRAL TRUNCATION

We deliberately begin with a pure spectral FNO baseline with no non-spectral bypass branch. At the native grid resolution $N = 128$, the model retains $k_{\max} = 60$ Fourier modes per spatial dimension. This value was selected empirically from a truncation sweep over $k_{\max} \in \{20, 40, 60\}$ (Appendix C.1), where $k_{\max} = 60$ gave the best trap-quality metrics before returns diminished with additional modes; we treat this truncation as an architectural hyperparameter of the native-resolution model rather than as a universal physical threshold. The terminal-frequency diagnostics and spectral-truncation analysis below assess how much of the learned representation occupies the retained frequency range.

4.1 ARCHITECTURE

Our baseline uses a spectral-only Fourier neural operator (FNO) backbone derived from the Fourier-domain operator formulation of Li et al. (2021). The input $I_{\mathrm{bp}} \in \mathbb{R}^{N \times N}$ is augmented with normalized Cartesian coordinates $\mathbf{c}_{ij} = (i/(N-1),\, j/(N-1))$, $i, j \in \{0, \ldots, N-1\}$, lifted to a latent dimension d by a pointwise linear layer, and processed through $L = 4$ spectral blocks:

$$v_{l+1}(\mathbf{x}) = \mathrm{GELU}\big(\mathcal{F}^{-1}\{R_l \cdot \mathcal{F}\{v_l\}\}(\mathbf{x})\big), \qquad l = 0, \ldots, L-1, \tag{3}$$

where $R_l \in \mathbb{C}^{d \times d \times k_{\max} \times k_{\max}}$ are learned complex-valued spectral weights and $k_{\max}$ is the number of retained Fourier modes per spatial dimension. The latent representation is projected to a scalar output by two pointwise linear layers with an intermediate GELU activation, producing the continuous unwrapped phase $\tilde{\phi}(\mathbf{x})$. The final SLM mask is

$$\phi(\mathbf{x}) = \mathrm{rem}\Big(\tilde{\phi}(\mathbf{x}) + \pi,\, 2\pi\Big) - \pi, \tag{4}$$

where $\mathrm{rem}(\cdot, 2\pi)$ wraps the phase to $[-\pi, \pi]$. A trainable spatial phase bias $\phi_{\mathrm{bias}} \in \mathbb{R}^{1 \times 1 \times N \times N}$, initialized as a diverging parabolic lens, is added before phase wrapping and optimized jointly with the network weights. Since this bias is spatially indexed but input-independent, it provides an additional fixed positional prior.

4.2 Training Setup and Native-Grid Performance

The training dataset is constructed from the 16 low-lying two-dimensional quantum harmonic-oscillator (QHO) eigenstates with $n_x, n_y \in \{0, 1, 2, 3\}$ and random normalized complex superpositions over the complete 16-state basis. Each random state is generated as

$$\psi_B(x, y) = \frac{1}{\mathcal{N}} \sum_{n_x=0}^{3} \sum_{n_y=0}^{3} c_{n_x,n_y}\, \psi_{n_x,n_y}(x, y), \tag{5}$$

where $c_{n_x,n_y} = a_{n_x,n_y} + i\, b_{n_x,n_y}$, a_{n_x,n_y}, b_{n_x,n_y} independently, and $\mathcal{N}$ normalizes the resulting wavefunction to unit norm. The generated dataset contains the 16 individual QHO eigenstates together with 2000 accepted random superpositions. Random samples are retained only when their Gerchberg–Saxton reconstruction satisfies the dataset-generation quality criteria of reconstruction error ≤ 0.15 and void depth ≥ 0.85.

For network training, the stored dataset is divided into 1516 training, 300 validation, and 200 held-out test samples. For each wavefunction, the blueprint intensity I_{bp} and normalized atomic density are constructed dynamically during training using the procedure described in Section 2. The phase labels stored during dataset generation are not used as supervision; optimization is performed entirely through the differentiable physics-based objective in Eq. equation 2.

The models are optimized using AdamW with batch size 32 for 750 epochs. For training from scratch, the learning rate is controlled by a OneCycleLR schedule with maximum learning rate 10^{-3} and weight decay 10^{-4}. Gradient norms are clipped to a maximum value of 1.0. The checkpoint with the lowest validation intensity-fidelity loss is retained for subsequent evaluation. Trap-quality metrics used in the evaluation are defined in the Appendix D. The 5 different cases evaluated are briefly explained in the Appendix E.

5 Harmonic-oscillator-inspired sinusoidal bypass

To support the Fourier-domain global branch, a pointwise multi-frequency sinusoidal bypass is utilised, where it applies

$$q_{\mathrm{QHO}}(z) = g_l \sum_{n=0}^{N_h} a_n \sin(\omega_n z), \qquad z = W_l v_l + b_l, \tag{6}$$

where W_l and b_l define a learnable 1×1 pointwise projection, g_l is a learnable sigmoid-parameterized gate, and a_n and ω_n are fixed amplitude and frequency coefficients, respectively. In the reported models, $N_h = 6$, $a_n = 1/(N_h + 1)$, and $\omega_n = \omega_0\sqrt{2n+1}$ with $\omega_0 = 4$. During training, the bypass contribution is linearly ramped from zero over the first 50 epochs.

The sinusoidal bypass is motivated by two considerations. First, periodic representations and Fourier-feature mappings have been shown to facilitate representing oscillatory and high-frequency signals in coordinate-based neural models (Tancik et al., 2020; Sitzmann et al., 2020). Second, we generate the target family from low-lying 2-D Quantum Harmonic Oscillator (QHO) eigenstates, which provides a natural, problem-specific set of characteristic scales. We therefore use a fixed multi-frequency sinusoidal bank whose frequency schedule follows the QHO-inspired $\sqrt{2n+1}$ scaling described in Appendix C. The frequency bank is therefore treated as a problem-specific inductive bias rather than as a direct correspondence between QHO spatial frequencies and latent-space frequencies.

The bypass acts pointwise on latent features and is evaluated in parallel with the Fourier-domain global branch. The two pathways are combined according to the topology-specific equations defined below. Because ω_n is a frequency with respect to the projected latent feature value z, rather than directly with respect to the spatial coordinate, the form of Eq. equation 6 alone does not imply injection of fixed spatial Fourier carriers. We therefore evaluate the computational role of the bypass empirically through fixed-weight branch interventions, spectral-mode comparisons, and downstream reconstruction and trap-proxy metrics.

Table 2: Full no-scrub trap-quality metrics over five held-out evaluation sets. Cases 3 and 5 pool three independently trained checkpoints evaluated on the same 500 examples; their sample-level standard deviations therefore include both sample and seed variation.

Case	S_{grad}	η_{conf}	$\mathrm{RMSE}_{\mathrm{ROI}}$
1 (linear)	0.360 ± 0.014	0.435 ± 0.034	0.1254 ± 0.0137
2 (sine, outside)	0.453 ± 0.020	0.413 ± 0.033	0.0959 ± 0.0120
3 (sine, double)	0.449 ± 0.026	0.414 ± 0.032	0.1019 ± 0.0160
5 (sine, single)	0.474 ± 0.024	0.404 ± 0.031	0.0883 ± 0.0113

6 Mechanistic Analysis: Causal Role of the Sinusoidal Bypass

End-task performance alone cannot determine whether the sinusoidal bypass performs the intended computational function. An improvement in trap quality could instead result from the additional capacity introduced by the branch. We therefore perform a fixed-weight branch-scrubbing intervention: the bypass activation is set to zero at its point of injection while all other network parameters and the input are held fixed. For each held-out sample, we compare the standard forward pass (ϕ_{full}) with the corresponding scrubbed pass (ϕ_{scrub}).

With weights and inputs fixed, the full-versus-scrubbed difference estimates the effect of removing the bypass at the chosen intervention site. This is a within-model intervention; it does not establish that the bypass is uniquely necessary, because the scrubbed activation may be off the training manifold and other pathways may compensate.

6.1 Zernike Residual as a Representation-Level Diagnostic

To assess whether bypass removal changes the structure of the predicted phase masks, we use a fixed low-order Zernike basis over the circular pupil. Let $\mathcal{Z}_{15}$ denote orthogonal projection onto the first 15 Noll-indexed Zernike modes. For an unwrapped phase mask ϕ, we define the residual

$$R_{\mathrm{Z}}(\phi) = \|\phi - \mathcal{Z}_{15}\phi\|_2 , \tag{7}$$

This measures phase structure outside the selected low-order subspace and should not be interpreted as a direct measure of Fourier high-frequency energy. We quantify bypass removal using

$$\Delta_{\mathrm{Z}} = R_{\mathrm{Z}}(\phi_{\mathrm{full}}) - R_{\mathrm{Z}}(\phi_{\mathrm{scrub}}), \tag{8}$$

and the residual-retention ratio;

$$\rho_{\mathrm{Z}} = \frac{R_{\mathrm{Z}}(\phi_{\text{spectral-only}})}{R_{\mathrm{Z}}(\phi_{\mathrm{full}})}. \tag{9}$$

Here, smaller ρ_{Z} indicates a larger reduction in the measured non-low-order residual after bypass removal.

The Zernike intervention produces a pronounced topology-dependent effect. For the selected Case 5 topology, the full model has $R_{\mathrm{Z}} = 1109.4 \pm 244.7$, whereas suppressing the sinusoidal bypass reduces this value to 18.40 ± 0.14, giving $\rho_{\mathrm{Z}} = 0.0166$. Thus, only approximately 1.7% of the measured residual remains after bypass removal. In contrast, Cases 1–3 show substantially weaker changes, with $\rho_{\mathrm{Z}} = 1.226$, 1.074, and 0.916, respectively (Appendix F). The strong reduction observed for Case 5 is therefore specific to the selected topology rather than a generic consequence of introducing a bypass.

6.2 Multi-Seed Causal Results

We repeat the branch-scrubbing intervention across five independently drawn evaluation sets, with 100 held-out samples per set (500 samples total). These evaluation sets change the held-out data draw while keeping the trained weights fixed for each case. They therefore test consistency across evaluation data rather than robustness to training initialization. Independent training variability is evaluated separately for Cases 3 and 5.

Case 5 shows the largest and most consistently signed intervention effects among the evaluated sinusoidal topologies, with $\Delta S_{\mathrm{grad}} = 0.457 \pm 0.012$, $\Delta\kappa_{\mathrm{geom}} = 0.803 \pm 0.005$, and $\Delta\eta_{\mathrm{conf}} =$

Table 3: Causal branch-scrubbing and trap-quality effects evaluated across 5 independent held-out evaluation sets (100 samples/set, 500 samples total). Cases 3 and 5 report mean ± std across 3 independently trained seeds. $\mathrm{RMSE}_{\mathrm{ROI}}$ is evaluated strictly inside the optical trapping envelope.

Case	ΔS_{grad}	$\Delta\kappa_{\mathrm{geom}}$	$\Delta\eta_{\mathrm{conf}}$	Full-model $\mathrm{RMSE}_{\mathrm{ROI}}\downarrow$
1 (linear)	0.360 ± 0.014	0.366 ± 0.377	-0.467 ± 0.036	0.1254 ± 0.0137
2 (sine, outside)	0.048 ± 0.010	-0.030 ± 0.149	-0.031 ± 0.017	0.0959 ± 0.0120
3 (sine, double)	0.204 ± 0.114	0.621 ± 0.144	-0.483 ± 0.102	0.1019 ± 0.012
5 (sine, single)	$\mathbf{0.457 \pm 0.012}$	$\mathbf{0.803 \pm 0.005}$	$\mathbf{-0.596 \pm 0.002}$	$\mathbf{0.0883 \pm 0.001}$

-0.596 ± 0.002, while also achieving the lowest full-model $\mathrm{RMSE}_{\mathrm{ROI}} = 0.0883 \pm 0.001$. Case 3 exhibits the same general pattern, but with smaller intervention effects and higher $\mathrm{RMSE}_{\mathrm{ROI}} = 0.1019 \pm 0.012$.

Cases 2 and 1 behave differently. Case 2 shows only weak intervention effects, with $\Delta S_{\mathrm{grad}} = 0.048 \pm 0.010$, $\Delta\kappa_{\mathrm{geom}} = -0.030 \pm 0.149$, and $\Delta\eta_{\mathrm{conf}} = -0.031 \pm 0.017$, despite a comparatively low full-model $\mathrm{RMSE}_{\mathrm{ROI}} = 0.0959 \pm 0.0120$. Case 1 shows substantially larger changes in S_{grad} and η_{conf}, but has the highest full-model $\mathrm{RMSE}_{\mathrm{ROI}} = 0.1254 \pm 0.0137$ among the four cases.

Taken together, the tabulated results distinguish Case 5 by combining the lowest full-model reconstruction error with the largest intervention-induced changes in S_{grad}, κ_{geom}, and η_{conf}. This indicates that the bypass has a substantially stronger effect on the final trap-quality metrics in Case 5 than in the other evaluated topologies.

6.3 Final Architecture Selection

Case 3 applies GELU to the spectral branch before adding the bypass and then applies a second GELU:

$$v_{l+1} = \mathrm{GELU}[\mathrm{GELU}(S_l(v_l)) + \sigma_{\mathrm{qho}}(v_l)]\,. \tag{10}$$

Case 5 combines the two pathways before a single GELU:

$$v_{l+1} = \mathrm{GELU}[S_l(v_l) + \sigma_{\mathrm{qho}}(v_l)]\,. \tag{11}$$

The two topologies have similar end-task performance, so trap-quality metrics alone do not clearly distinguish them. We therefore use branch-intervention consistency as the architecture-selection criterion. Case 5 exhibits a large, consistently signed intervention effect in physical diagnostics and lower variation in the Zernike intervention effect than Case 3. We consequently use Case 5 as HoloFNO in the remaining experiments. This selection criterion identifies the topology with the most reproducible branch dependence among the evaluated cases.

7 Related Work and Connection to Learned Holography

Table 4: Comparison with representative learned hologram-generation approaches.

Method	**Training signal**	**Architecture**	**Target regime**	**Analysis beyond end-task metrics**
Lin et al. (2025)	GS-supervised	CNN	Static/dynamic arrays	No
Ren et al. (2024)	GS-supervised	U-Net	Static tweezer arrays	No
Yan et al. (2026)	Self-supervised NPCC	FNO bottleneck in U-Net	Static/dynamic arrays	No
HoloFNO	Physics-supervised	Spectral FNO + sinusoidal bypass	State-dependent QHO densities	Branch scrubbing, Zernike and coordinate diagnostics

Recent work has demonstrated learned hologram synthesis using both phase-supervised and differentiable propagation objectives. Ren et al. (2024) and Lin et al. (2025) use convolutional architectures trained with Gerchberg–Saxton-derived supervision, whereas Yan et al. (2026) introduces a Fourier U-Net in which an FNO-style spectral operator is embedded within an encoder–decoder bottleneck and trained using an NPCC objective. HoloFNO differs in both architecture and supervision. It is trained directly through a differentiable Fourier-optics objective without GS phase labels and performs spectral mixing at the native spatial resolution rather than inside a compressed U-Net bottleneck. Because our experiments are intentionally restricted to a fixed 128×128 discretization, we do not claim zero-shot resolution transfer; instead, the analysis focuses on how spectral and

non-spectral pathways contribute to the learned phase representation through fixed-weight branch interventions, Zernike-residual decomposition, and coordinate-frame perturbations. A comprehensive comparison of existing works is summarized in Table 4.

8 BASELINE COMPARISON RESULTS

We compare HoloFNO with U-Net and convolutional baselines on the same held-out target sets. The architectures used for comparison are explained in detail in the Appendix H. The primary metrics are envelope-restricted reconstruction error, gradient-direction alignment, confinement curvature, and confinement purity. Trap-quality metrics are defined in the Appendix D.

Table 5: In-distribution benchmark over all five held-out evaluation datasets, with envelope-restricted $\mathrm{RMSE}_{\mathrm{ROI}}$ and robust κ_{geom}. HoloFNO is averaged across all 3 trained seeds.

Architecture	$\mathrm{RMSE}_{\mathrm{ROI}} \downarrow$	$S_{\mathrm{grad}} \uparrow$	κ_{geom}	η_{conf}
U-Net (Phys Loss)	0.1043 ± 0.0111	0.433 ± 0.021	0.774 ± 0.134	0.404 ± 0.035
U-Net (NPCC)	0.0912 ± 0.0139	0.424 ± 0.033	0.792 ± 0.176	0.419 ± 0.031
U-Net (Yang)	0.3814 ± 0.0195	0.113 ± 0.005	0.971 ± 0.681	0.736 ± 0.016
HDC CNN	0.2050 ± 0.0186	$\mathbf{0.502 \pm 0.049}$	0.157 ± 0.118	0.207 ± 0.028
HoloFNO (Ours)	$\mathbf{0.0883 \pm 0.001}$	0.474 ± 0.002	$\mathbf{0.804 \pm 0.005}$	0.404 ± 0.002

Table 5 compares all architectures on the in-distribution held-out evaluation sets. HoloFNO achieves the lowest $\mathrm{RMSE}_{\mathrm{ROI}}$ at 0.0883 ± 0.001, while HDC CNN achieves the highest S_{grad} at 0.502 ± 0.049. All architectures have positive mean κ_{geom}, although HDC CNN has the lowest value (0.157 ± 0.118). U-Net (Yang) attains a relatively high κ_{geom} (0.971 ± 0.681) but with substantially higher variability, reconstruction error (0.3814 ± 0.0195), and lower gradient-direction alignment (0.113 ± 0.005). The corresponding OOD results of the anharmonic oscillator are reported separately in the Appendix I.

HDC CNN's results show that strong local gradients do not necessarily imply accurate global trap geometry: despite its highest S_{grad}, it has comparatively poor reconstruction accuracy ($\mathrm{RMSE}_{\mathrm{ROI}} = 0.2050 \pm 0.0186$) and the lowest κ_{geom}. In contrast, U-Net (Yang) exhibits the largest reconstruction error and lowest S_{grad}, together with substantial variability in κ_{geom}. Its encoder–bottleneck–decoder structure performs global processing after spatial compression and reconstructs the full-resolution phase through the decoder. These observations are consistent with effects from spatial compression, bottleneck representation, decoder reconstruction, and possible aliasing, but the present experiments do not isolate any single mechanism as causal.

HoloFNO performs global spectral mixing directly at the native 128×128 resolution, avoiding spatial compression before the Fourier blocks. It achieves the lowest reconstruction error while maintaining high S_{grad} and κ_{geom}. For the selected topology, the branch-intervention experiments further show that neither the spectral-only nor bypass-only counterfactual reproduces the full-model non-low-order phase residual, supporting a contribution from the interaction between the globally mixed spectral pathway and the structured pointwise bypass. Overall, these results reflect differences in global context, spatial resolution, information routing, and optimization bias, rather than a claim that CNN or U-Net architectures are fundamentally incapable of holographic inverse design.

Single-Pass Hologram Synthesis with Predictable Latency. Across 500 held-out test instances, HoloFNO produces a phase mask in a deterministic single forward pass of 5.219 ± 0.19 ms on the NVIDIA RTX 4090 GPU. Quality-matched iterative Gerchberg–Saxton (GS) reaches comparable mean reconstruction error at a median of 60 iterations (3.368 ms), but only on 89.6% of evaluated samples (448/500), with the remainder requiring additional iterations or failing to match within the tested range; HoloFNO's fixed one-shot cost avoids this per-sample convergence variability entirely. (Table 12) At the quality-matched operating point, HoloFNO achieves a region-of-interest relative reconstruction error of $Rel\text{-}RMSE_{\mathrm{ROI}} = 0.1494 \pm 0.0213$, compared with 0.1621 ± 0.0253 for GS at 60 iterations, though GS attains marginally higher simulated confinement ($\kappa_{\mathrm{geom}} = 0.890 \pm 0.123$ versus 0.787 ± 0.113 for HoloFNO).

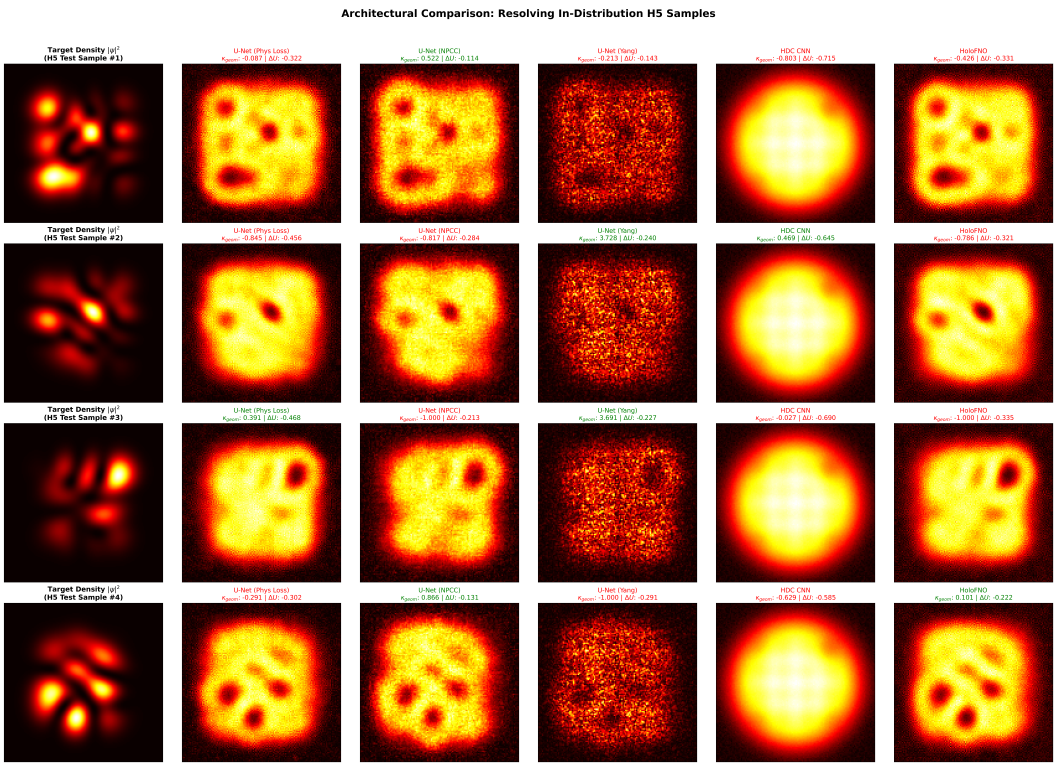

Figure 3: Qualitative comparison on four held-out H5 target states. Each row corresponds to a target density $|\psi|^2$ and each subsequent column shows the reconstructed trap intensity produced by a different learned hologram generator.

9 Conclusion

We introduced HoloFNO, a physics-supervised Fourier neural operator for phase-only hologram synthesis that is trained through differentiable Fourier-optics reconstruction rather than regression to a single Gerchberg–Saxton phase solution. HoloFNO combines a global Fourier-domain branch with a harmonic-oscillator-inspired multi-frequency sinusoidal pathway. Through fixed-weight branch-scrubbing interventions and representation-level diagnostics, we find that the selected single-saturation topology exhibits reproducible dependence on the sinusoidal pathway, accompanied by corresponding changes in the trap-quality metrics, providing evidence that the structured pathway contributes to the learned representation beyond simply increasing model capacity.

Scaling the native grid size while holding RMSE fixed shows GS's fixed-iteration runtime diverging from HoloFNO's near-flat latency (Fig. 11), suggesting the advantage widens at higher resolutions, though this trend is extrapolated from limited tested grid sizes rather than validated at full SLM resolution.

References

Waseem S. Bakr, Jonathon I. Gillen, Amy Peng, Simon Fölling, and Markus Greiner. A quantum gas microscope for detecting single atoms in a hubbard-regime optical lattice. *Nature*, 462:74–77, 2009. doi: 10.1038/nature08482.

James R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982. doi: 10.1364/AO.21.002758.

R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35(2):237–246, 1972.

Dennis C. Ghiglia and Mark D. Pritt. *Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software*. Wiley-Interscience, 1998. ISBN 0471249351.

Joseph W. Goodman. *Introduction to Fourier Optics*. Roberts and Company Publishers, 3 edition, 2005.

Rudolf Grimm, Matthias Weidemüller, and Yurii B. Ovchinnikov. Optical dipole traps for neutral atoms. In *Advances in Atomic, Molecular, and Optical Physics*, volume 42, pp. 95–170. Academic Press, 2000. URL https://arxiv.org/abs/physics/9902072.

Philipp Grohs, Sarah Koppensteiner, and Martin Rathmair. Phase retrieval: Uniqueness and stability. *SIAM Review*, 62(2):301–350, 2020. doi: 10.1137/19M1256865.

Christian Gross and Immanuel Bloch. Quantum simulations with ultracold atoms in optical lattices. *Science*, 357(6355):995–1001, 2017. doi: 10.1126/science.aal3837.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016. URL https://arxiv.org/abs/1606.08415.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.08895.

Rui Lin, Han-Sen Zhong, You Li, Zhang-Rui Zhao, Le-Tian Zheng, Tai-Ran Hu, Hong-Ming Wu, Zhan Wu, Wei-Jie Ma, Yan Gao, Yi-Kang Zhu, Zhao-Feng Su, Wan-Li Ouyang, Yu-Chen Zhang, Jun Rui, Ming-Cheng Chen, Chao-Yang Lu, and Jian-Wei Pan. AI-enabled parallel assembly of thousands of defect-free neutral atom arrays. *Physical Review Letters*, 135(6):060602, 2025. doi: 10.1103/2ym8-vs82.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4898–4906, 2016.

Robert J. Noll. Zernike polynomials and atmospheric turbulence. *Journal of the Optical Society of America*, 66(3):207–211, 1976. doi: 10.1364/JOSA.66.000207.

Zejian Ren, Xu Yan, Kai Wen, Huijin Chen, Elnur Hajiyev, Chengdong He, and Gyu-Boong Jo. Creation of a tweezer array for cold atoms utilizing a generative neural network. *APL Quantum*, 1(4):046111, 2024. doi: 10.1063/5.0219170.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992. doi: 10.1016/0167-2789(92)90242-F.

Mansi Sakarvadia, Kareem Hegazy, Amin Totounferoush, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W. Mahoney. The false promise of zero-shot super-resolution in machine-learned operators. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.06646.

Vincent Sitzmann, Julien N. P. Martel, Alexander Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7462–7473. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html.

Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. Unpublished presentation, Stanford Artificial Intelligence Project, 1968.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451–1460, 2018. doi: 10.1109/WACV.2018.00163. URL https://arxiv.org/abs/1702.08502.

Ge Yan, Yuchen Wang, and Junchi Yan. Towards real-time neutral atom array assembly via unsupervised hologram generation and path optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(32):27486–27493, 2026. doi: 10.1609/aaai.v40i32.39967.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.07122.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7324–7334. PMLR, 2019.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020. doi: 10.1016/j.acha.2019.06.004.

APPENDIX

A OPTICAL FORWARD MODEL AND TARGET CONSTRUCTION

Field at the SLM Plane. A continuous-wave monochromatic field of wavelength λ illuminates the phase-only SLM. At the SLM plane (x, y), the field is

$$E_{\mathrm{SLM}}(x, y) = A_0(x, y)e^{i\phi(x,y)}, \tag{12}$$

where $A_0(x, y)$ is the known incident amplitude profile and $\phi(x, y)$ is the phase applied by the SLM. In simulation, the incident field is modeled as a Gaussian,

$$A_0(x, y) = A_{\max} \exp\left[-\frac{x^2 + y^2}{w_0^2}\right], \tag{13}$$

with beam waist w_0 chosen relative to the simulated SLM aperture.

Focal-Plane Intensity. Under the scalar paraxial approximation, a thin lens maps the SLM field to the focal plane (u, v) through a Fourier transform (Goodman, 2005). The reconstructed intensity is therefore

$$I_\phi(u, v) = \frac{1}{\lambda^2 f^2} \left|\mathcal{F}_{2D}\Big\{A_0(x, y)e^{i\phi(x,y)}\Big\}\right|^2, \tag{14}$$

where f is the focal length. Numerically, Eq. equation 14 is evaluated using discrete fast fourier transforms (FFTs).

Table 6: Notation used throughout the manuscript.

Notation	Meaning
Optical field and target	
$\mathbf{r} = (x, y)$	Transverse spatial coordinate in the SLM or focal plane.
$A_0(\mathbf{r})$	Known amplitude profile of the incident field at the SLM.
$\phi(\mathbf{r})$	Phase mask applied by the phase-only SLM.
$\tilde{\phi}(\mathbf{r})$	Continuous, unwrapped phase predicted by the network before 2π wrapping.
$E_{\mathrm{SLM}}(\mathbf{r})$	Complex optical field at the SLM, $A_0(\mathbf{r})e^{i\phi(\mathbf{r})}$.
$E_f(u, v)$	Complex field in the focal plane after Fourier propagation.
$I_\phi(u, v)$	Predicted focal-plane intensity generated by the phase mask ϕ.
$I_{\mathrm{bp}}(\mathbf{r})$	Blueprint target intensity constructed from the atomic probability density.
$\rho_B(\mathbf{r})$	Atomic probability density, $\rho_B = \lvert\psi_B\rvert^2$.
$\mathcal{E}(\mathbf{r})$	Smooth spatial envelope used to confine the target trap structure.
Neural operator	
v_l	Latent feature representation at spectral block l.
L	Number of spectral blocks in the FNO.
d	Latent channel dimension of the neural operator.
R_l	Learned complex Fourier-domain weight tensor in spectral block l.
$\mathcal{F}, \mathcal{F}^{-1}$	Forward and inverse two-dimensional Fourier transforms.
$k_{\max}$	Number of retained Fourier modes per spatial dimension.
σ	Pointwise activation function used after the spectral transformation.
GELU	Gaussian Error Linear Unit used as the block activation in the main models.
Sinusoidal bypass	
σ_{qho}	Multi-frequency sinusoidal bypass activation inspired by the QHO state family.
a_n	Amplitude coefficient of the nth sinusoidal component.
ω_n	Frequency parameter of the nth sinusoidal component.
$N_{\max}$	Maximum harmonic index used in the sinusoidal frequency bank.
$S_l(v_l)$	Spectral transformation applied to the latent representation at block l.
Causal and representation diagnostics	
$\mathcal{Z}_{15}$	Orthogonal projection onto the first 15 Noll Zernike modes.
$R_{\mathrm{Z}}(\phi)$	ℓ_2 norm of the phase component outside the selected Zernike subspace.
R_{full}	Zernike residual of the complete network.
R_{scrub}	Zernike residual after causal removal of the non-spectral bypass.
Δ_{Z}	Causal intervention effect, $R_{\mathrm{full}} - R_{\mathrm{scrub}}$.
α	Optional bypass-scaling coefficient in controlled branch interventions.
Evaluation and indexing	
N	Native spatial resolution of the square simulation grid.
K	Number of held-out samples used in an adaptation or fine-tuning experiment.
$\mathbf{k} = (k_x, k_y)$	Discrete spatial-frequency index in the Fourier domain.
k	Radial spatial-frequency magnitude used in spectral diagnostics.

Blue-Detuned Dipole Potential. For a far-detuned two-level atom, the optical dipole potential is proportional to the reconstructed intensity,

$$U_{\mathrm{dip}}(u,v) = \frac{\hbar\Gamma^2}{8\Delta I_{\mathrm{sat}}} I_\phi(u,v), \tag{15}$$

where Γ is the natural linewidth, I_{sat} is the saturation intensity, and $\Delta = \omega_L - \omega_0$. For blue detuning ($\Delta > 0$), atoms are repelled from high-intensity regions and confined near local intensity minima. The transverse dipole force is

$$\mathbf{F}_{\mathrm{dip}}(u,v) = -\nabla U_{\mathrm{dip}}(u,v) = -\frac{\hbar\Gamma^2}{8\Delta I_{\mathrm{sat}}} \nabla I_\phi(u,v). \tag{16}$$

Accordingly, a useful reconstructed trap should produce low intensity in the desired confinement regions together with surrounding intensity gradients that provide restoring confinement.

Mapping the Atomic State to Blueprint Intensity. Given the transverse atomic wavefunction $\psi_B(u,v)$, we define

$$\rho_B(u,v) = |\psi_B(u,v)|^2, \qquad \tilde{\rho}_B(u,v) = \frac{\rho_B(u,v)}{\max_{u,v} \rho_B(u,v)}. \tag{17}$$

The target blueprint used throughout the experiments is

$$I_{\mathrm{bp}}(u,v) = [1 - \tilde{\rho}_B(u,v)]\, \mathcal{E}(u,v), \tag{18}$$

where $\mathcal{E}(u,v)$ is a smooth Gaussian envelope constructed from a soft-thresholded version of the atomic density. The envelope confines the target structure to the physically relevant region and suppresses far-field background content. The resulting blueprint is normalized to $[0,1]$ and serves as both the network input and the target observable in the physics-supervised objective.

The atomic-state family is generated from normalized superpositions of the 16 two-dimensional harmonic-oscillator eigenstates with $n_x, n_y \in \{0,1,2,3\}$, as described in the main text.

B The Significance of the Five Different Physics Loss

Density-Weighted Intensity Fidelity ($\mathcal{L}_{fid}$)

$$\mathcal{L}_{fid} = \mathbb{E}_r\Big[\big(1 + 10\,\tilde{\rho}_B(r)\big)\big(I_\phi(r) - I_{bp}(r)\big)^2\Big]. \tag{19}$$

This is a weighted mean-squared error between the reconstructed focal-plane intensity I_ϕ produced by the predicted phase mask and the target blueprint intensity I_{bp}. The multiplicative weight $\big(1 + 10\tilde{\rho}_B(r)\big)$ scales the squared error by the local normalized atomic density $\tilde{\rho}_B(r) \in [0,1]$, so error occurring exactly at an atom's location is penalized up to $11\times$ more heavily than the same-magnitude error occurring in open background. This penalizing is important because, as pixels away from the voids are probed, their $\mathcal{L}_{fid}$ value is required to be negligible. A residual intensity error of a given size is far more consequential when it lands on an atom (where it directly disturbs confinement and can eject or heat the atom) than when it lands in empty background space the atom never visits.

Sobel Edge-Consistency Term ($\mathcal{L}_{edge}$)

$$\mathcal{L}_{edge} = \mathbb{E}_r\Big[\big\|\nabla_S I_\phi(r) - \nabla_S I_{bp}(r)\big\|_2^2\Big], \qquad \nabla_S I = \sqrt{(G_x * I)^2 + (G_y * I)^2}, \tag{20}$$

where G_x and G_y are the 3×3 Sobel kernels(Sobel & Feldman, 1968)

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \qquad G_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}. \tag{21}$$

This term penalizes the mismatch between the Sobel-filtered gradient *magnitude* of the predicted and target intensity fields, computed via 2D convolution of each image with the two kernels above.

$\mathcal{L}_{fid}$ can be minimized by a solution that gets the average intensity approximately right while still blurring, softening, or slightly mislocating the sharp walls that separate the dark trap region from the bright confining boundary. Because $\mathcal{L}_{fid}$ operates pointwise, it has no mechanism for distinguishing a blurred solution from a genuinely accurate reconstruction (Sobel & Feldman, 1968). $\mathcal{L}_{edge}$ closes this gap by comparing gradient-magnitude fields directly, so reproduction of location and sharpness of boundaries is penalized rather than only correct raw pixel values.

Wrapping-Aware Phase Total Variation ($\mathcal{L}_{TV}$)

$$\mathcal{L}_{TV} = \mathbb{E}_r\Big[\big|\,\mathrm{wrap}\big(\partial_x\phi(r)\big)\,\big| + \big|\,\mathrm{wrap}\big(\partial_y\phi(r)\big)\,\big|\Big], \qquad \mathrm{wrap}(\Delta) = \mathrm{rem}(\Delta+\pi, 2\pi) - \pi. \quad (22)$$

This regularizes the *predicted phase mask* ϕ toward piecewise smoothness by penalizing the total variation (sum of absolute gradient magnitudes) of the phase field (Rudin et al., 1992), using a wrap-consistent finite difference ((Ghiglia & Pritt, 1998)). The predicted phase $\phi(r)$ is only physically meaningful in 2π. A finite difference computed across a wrap boundary between $\phi = -\pi + \epsilon$ and an adjacent pixel at $\phi = \pi - \epsilon$, would register a sudden spike gradient of magnitude $\approx 2\pi$, even though the true underlying phase is physically smooth at that point. The $\mathrm{wrap}(\cdot)$ operator re-maps every finite difference back into $(-\pi, \pi]$ before computing its magnitude, so the penalty only activates on genuine phase discontinuities, never on wrapping artifacts (Ghiglia & Pritt, 1998).

Quadratic Void-Confinement Penalty ($\mathcal{L}_{void}$)

$$\mathcal{L}_{void} = \mathbb{E}_r\Big[\tilde{\rho}_B(r)^4\, I_\phi(r)^2\Big]. \quad (23)$$

This term penalizes residual reconstructed intensity specifically at the atom's exact location, using a quartic density weight $\tilde{\rho}_B(r)^4$ that concentrates the penalty almost entirely within the small region where $\tilde{\rho}_B \approx 1$, contributing negligibly everywhere else. For a far-detuned two-level atom under blue detuning ($\Delta > 0$), the optical dipole potential $U_{dip}(r) \propto I_\phi(r)$ is positive, so the atom is repelled from bright regions and confined only where $I_\phi(r) \to 0$. A moderately dark but non-zero residual could still achieve acceptably low $\mathcal{L}_{fid}$. $\mathcal{L}_{void}$'s aggressive, spatially concentrated weighting directly supplies the missing incentive to drive the void all the way to the physically required near-zero intensity.

Boundary Gradient-Direction Alignment ($\mathcal{L}_{dir}$)

$$\mathcal{L}_{dir} = \mathbb{E}_{r\in\partial\Omega}\Big[1 - \frac{\nabla I_\phi(r)\cdot\nabla I_{bp}(r)}{\|\nabla I_\phi(r)\|\,\|\nabla I_{bp}(r)\| + \epsilon}\Big], \quad (24)$$

evaluated only on the boundary ring $\partial\Omega$ obtained via erosion of the physical envelope mask. This is a cosine-similarity penalty between the predicted and target intensity *gradient vectors* (not just their magnitudes), evaluated along the confining wall's boundary ring.

B.1 Ablation Analysis and Optimization Dynamics

To examine the contribution of each loss component, we perform a 300-epoch leave-one-out ablation in which each term is removed individually from the full objective. Figure 4 shows the corresponding optimization trajectories, while Table 7 reports the resulting reconstruction and trap-quality metrics.

The ablations reveal distinct contributions of the loss terms. Removing $\mathcal{L}_{\mathrm{fid}}$ produces by far the largest degradation, increasing RMSE from 0.0928 to 0.5165, reducing S_{grad} from 0.447 to 0.352, driving κ_{geom} negative, and increasing η_{conf} to 0.996. This identifies the density-weighted fidelity term as the principal constraint on reconstruction accuracy.

Removing $\mathcal{L}_{\mathrm{edge}}$ has the strongest effect on boundary-gradient quality, reducing S_{grad} from 0.447 to 0.347, while also increasing RMSE to 0.0988 and η_{conf} to 0.437. This is consistent with its intended role in preserving the location and sharpness of trap boundaries.

The effects of removing $\mathcal{L}_{\mathrm{TV}}$, $\mathcal{L}_{\mathrm{void}}$, or $\mathcal{L}_{\mathrm{dir}}$ are smaller under the reported metrics. Removing $\mathcal{L}_{\mathrm{TV}}$ or $\mathcal{L}_{\mathrm{dir}}$ leaves RMSE and S_{grad} close to the full model, whereas removing $\mathcal{L}_{\mathrm{void}}$ produces a modest increase in RMSE and reduction in S_{grad}. The large standard deviations of κ_{geom} for

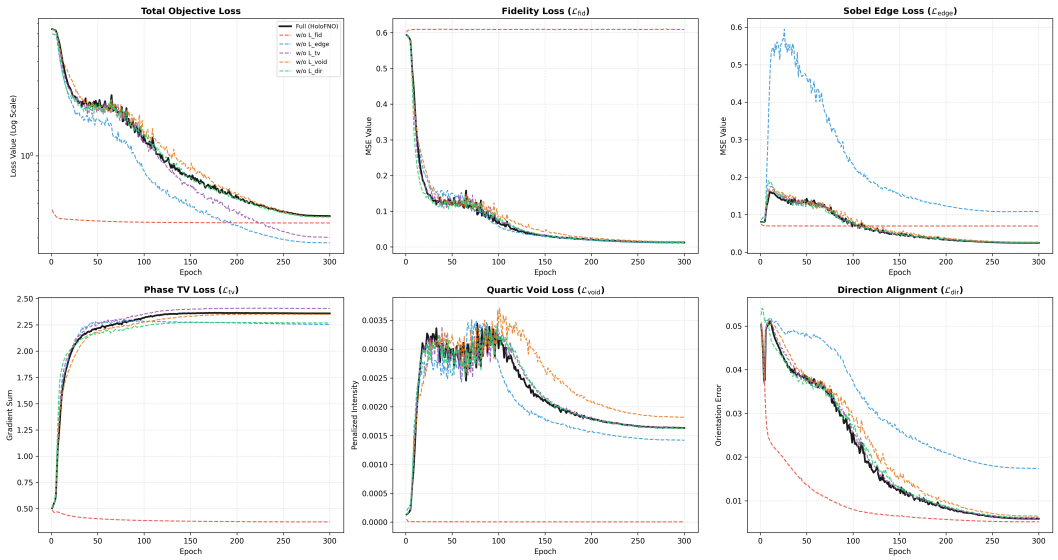

Figure 4: **Optimization trajectories across 300-epoch loss ablations.** Evolution of the total objective and individual loss components under leave-one-out training configurations.

Table 7: **Performance across 300-epoch loss-ablation variants.** Values report mean ± standard deviation over the evaluation set.

Variant	RMSE ↓	S_{grad} ↑	κ_{geom}	η_{conf} ↓
Full (HoloFNO)	**0.0928** ± 0.010	0.447 ± 0.022	0.169 ± 0.844	0.400 ± 0.032
w/o $\mathcal{L}_{\text{fid}}$	0.5165 ± 0.023	0.352 ± 0.022	-0.740 ± 0.232	0.996 ± 0.002
w/o $\mathcal{L}_{\text{edge}}$	0.0988 ± 0.013	0.347 ± 0.020	0.807 ± 1.981	0.437 ± 0.044
w/o $\mathcal{L}_{\text{TV}}$	0.0949 ± 0.010	0.447 ± 0.024	0.302 ± 1.147	0.399 ± 0.034
w/o $\mathcal{L}_{\text{void}}$	0.0984 ± 0.010	0.425 ± 0.022	0.206 ± 0.859	0.391 ± 0.031
w/o $\mathcal{L}_{\text{dir}}$	0.0934 ± 0.011	0.448 ± 0.021	0.274 ± 1.252	0.401 ± 0.033

several ablations indicate that curvature is comparatively unstable in this 300-epoch study and should therefore be interpreted cautiously.

Overall, the ablation supports $\mathcal{L}_{\text{fid}}$ as the dominant reconstruction term and $\mathcal{L}_{\text{edge}}$ as an important contributor to gradient fidelity, while the remaining auxiliary terms produce more modest changes in the four reported metrics.

C Physics-Inspired Construction of the HO-Ladder Sinusoidal Bank

The sinusoidal bypass is motivated by the hierarchy of characteristic scales in the two-dimensional quantum harmonic oscillator (QHO). The 2-D oscillator eigenstate ψ_{n_x,n_y} has energy

$$E_{n_x,n_y} = \hbar\omega(n_x + n_y + 1). \tag{25}$$

Defining the total quantum number $n = n_x + n_y$, the corresponding energy level is

$$E_n = \hbar\omega(n + 1), \tag{26}$$

with degeneracy $g_n = n + 1$. Thus, increasing n corresponds to progressively more spatially extended and oscillatory state structure.

For the one-dimensional oscillator, the classical turning-point scale obeys

$$r_n \propto \sqrt{2n + 1}. \tag{27}$$

The same $\sqrt{2n+1}$ progression therefore provides a natural hierarchy of characteristic scales associated with increasing oscillator order. We use this progression only as a *physics-inspired spacing rule* for the latent sinusoidal frequencies,

$$\omega_n = \omega_0\sqrt{2n+1}, \qquad n = 0, \ldots, N_h. \tag{28}$$

For the training family used in this work, $n_x, n_y \in \{0, 1, 2, 3\}$, so the largest total quantum number is $n_x + n_y = 6$. We consequently use

$$N_h = 6, \tag{29}$$

giving seven sinusoidal components. The base frequency is set to $\omega_0 = 4$.

Importantly, the implemented model does *not* weight the components according to the QHO energy or degeneracy. Instead, all components are assigned the same fixed amplitude,

$$a_n = \frac{1}{N_h + 1}, \qquad n = 0, \ldots, N_h. \tag{30}$$

The energy-level and turning-point arguments above therefore motivate the spacing of the frequency bank, but do not constitute a derivation of the component amplitudes.

The complete bypass at spectral block l is

$$q_{\mathrm{QHO}}(v_l) = g_l \sum_{n=0}^{N_h} a_n \sin(\omega_n z_l), \qquad z_l = W_l v_l + b_l, \tag{31}$$

where W_l and b_l define a learned 1×1 pointwise projection. The scalar gate is

$$g_l = \sigma(\gamma_l)\, \alpha_{\mathrm{warm}}, \tag{32}$$

where γ_l is learned through a sigmoid parameterization and α_{warm} is an epoch-dependent warm-up factor. During training, α_{warm} is increased linearly from zero to one over the first 50 epochs, allowing the spectral pathway to establish a representation before the sinusoidal contribution reaches its full scale.

The frequencies ω_n in Eq. equation 28 act on the projected latent feature value z_l, not directly on the spatial coordinate (x, y). Consequently, the bypass should not be interpreted as injecting spatial Fourier carriers at the QHO turning-point frequencies. Rather, the QHO hierarchy provides a structured set of nonlinear latent frequencies that acts as a problem-specific inductive bias.

This distinction is important because the physical QHO spectrum does not uniquely determine the optimal nonlinear frequency bank of a neural network. The present construction therefore makes the more limited claim that the $\sqrt{2n+1}$ progression provides a physically motivated ordering of latent scales. Whether this particular schedule is superior to generic matched-frequency banks remains an empirical question.

C.1 Spectral-Energy Diagnostic and Fourier-Mode Budget

The Fourier-mode count $k_{\max}$ of the spectral branch is treated as an architectural hyperparameter rather than as a quantity analytically derived from the target phase spectrum. Nevertheless, the spectral distribution of predicted phase masks provides a useful diagnostic for understanding how much phase structure occupies different spatial-frequency ranges.

For an unwrapped predicted phase $\phi(\mathbf{x})$, let

$$\hat{\phi}(\mathbf{k}) = \mathcal{F}\{\phi\}(\mathbf{k}) \tag{33}$$

denote its two-dimensional discrete Fourier transform. We define the radial power spectrum as

$$S(k) = \sum_{\|\mathbf{k}'\| \in [k, k+1)} \left|\hat{\phi}(\mathbf{k}')\right|^2. \tag{34}$$

The cumulative fraction of spectral energy contained below radial wavenumber K is

$$\mathcal{E}(K) = \frac{\sum_{k=0}^{K} S(k)}{\sum_{k} S(k)}. \tag{35}$$

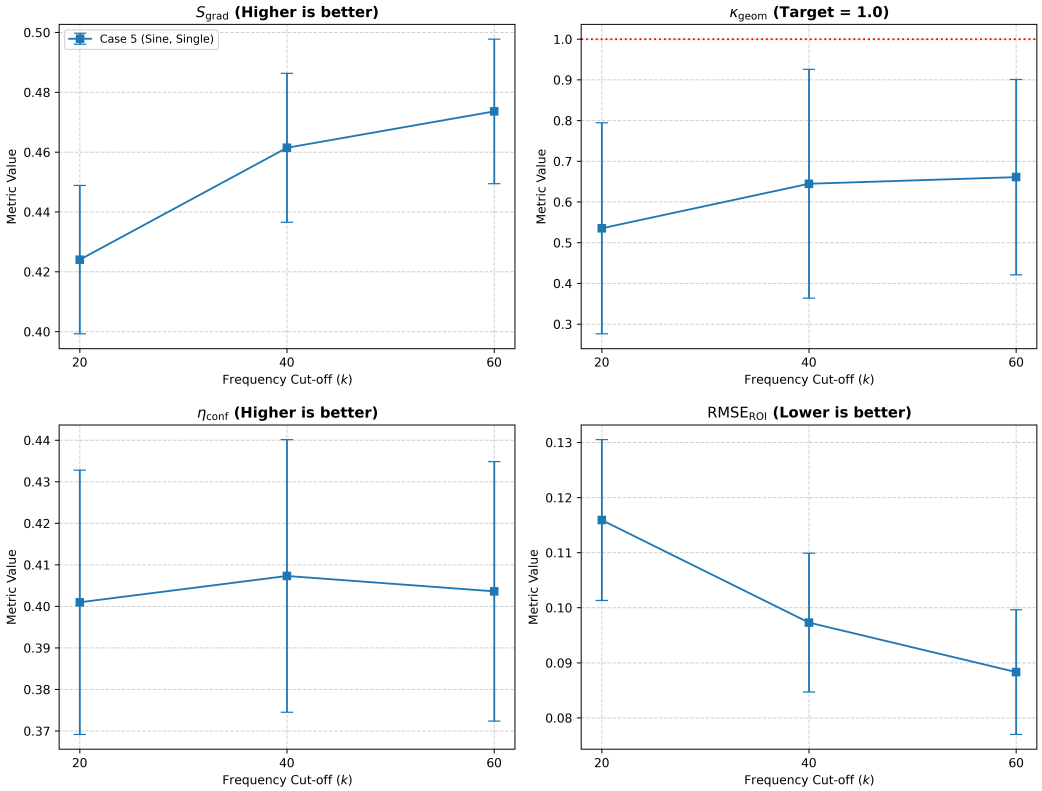

Figure 5: Effect of the Fourier-mode budget on the selected topology. Each $k_{\max}$ corresponds to a separately trained model. The normalized residual is the fraction of the unwrapped phase inside the unit disk that remains after least-squares projection onto the first 15 Zernike modes (Noll, 1976). Error bars denote the sample standard deviation over $n = 20$ held-out inputs. The full-model residual remains approximately 0.61–0.63 across $k_{\max} = 20, 40, 60$, indicating that increasing the retained spectral budget over this range does not substantially change the fraction of phase structure lying outside the selected low-order Zernike subspace. Scrubbing either pathway reduces this residual, with the largest separation observed for $k_{\max} = 60$, consistent with the non-low-order phase structure depending on interaction between the spectral and sinusoidal pathways.

For descriptive purposes, one may define

$$K_\eta = \min \{K : \mathcal{E}(K) \geq \eta\}, \tag{36}$$

for a chosen energy fraction η. This quantity characterizes the spectral concentration of a predicted phase mask, but is *not* used in this work to analytically determine $k_{\max}$.

Instead, we evaluate the retained Fourier-mode budget empirically by independently training the selected topology with

$$k_{\max} \in \{20, 40, 60\}. \tag{37}$$

Each value therefore corresponds to a separately optimized network rather than to inference-time truncation of a common checkpoint.

The relatively stable full-model residual across the three spectral budgets indicates that the measured non-low-order phase structure is not simply a consequence of retaining more Fourier modes. The stronger full-versus-scrubbed separation at $k_{\max} = 60$ motivates its use in the reported architecture, while this experiment does not establish $k_{\max} = 60$ as a universal or analytically optimal cutoff.

C.2 Specificity of the QHO-Inspired Frequency Schedule

The QHO-inspired frequency bank should be interpreted as a physics-motivated inductive bias rather than as an established optimal construction. Periodic representations and Fourier-feature mappings are known to facilitate learning of oscillatory and fine-scale functions (Tancik et al., 2020; Sitzmann et al., 2020), but these results do not establish that the particular $\sqrt{2n+1}$ schedule used here is uniquely advantageous for holographic inverse design.

A definitive specificity test would compare the implemented bank against generic fixed-frequency banks matched in component count, frequency range, amplitude norm, topology, and training protocol. Because such matched controls are not part of the present reported experiments, we do not attribute the observed causal role of the bypass specifically to the QHO frequency schedule. Instead, the current results establish the computational role of the structured sinusoidal pathway as implemented; isolating the contribution of its QHO-inspired frequency placement remains a direction for future work.

D Trap-Quality Metric Definitions

We evaluate reconstructed traps using four complementary metrics: reconstruction fidelity ($\mathrm{RMSE}_{\mathrm{ROI}}$), boundary-gradient agreement (S_{grad}), local confinement curvature (κ_{geom}), and confinement-region intensity (η_{conf}). These metrics probe different aspects of the reconstructed intensity and are therefore interpreted jointly rather than combined into a single score.

Region-of-Interest Reconstruction Error ($\mathrm{RMSE}_{\mathrm{ROI}}$). $\mathrm{RMSE}_{\mathrm{ROI}}$ measures the root-mean-square intensity difference between the reconstructed field I_ϕ and blueprint I_{bp}, restricted to the physical trapping envelope. Lower values indicate greater reconstruction fidelity. Restricting the calculation to the trapping region prevents large inactive background areas from dominating the reported error.

Boundary Gradient Agreement (S_{grad}). S_{grad} measures agreement between the reconstructed and target intensity-gradient directions along the confining boundary ring. It is computed from the cosine similarity between the corresponding gradient vectors and therefore assesses whether the reconstructed intensity increases in the intended direction around the confinement region. Larger values indicate better directional agreement.

Geometric Confinement Curvature (κ_{geom}). κ_{geom} compares the local curvature of the reconstructed intensity minimum with that of the target trap, with values near unity indicating agreement in confinement scale. Positive curvature corresponds to a locally confining intensity minimum, whereas negative curvature indicates locally anti-confining geometry.

The naive curvature estimate obtained directly from the discrete Laplacian at the target density maximum is sensitive to small spatial misalignment and pixel-scale intensity texture. We therefore use a robust estimator. First, I_{pred} and I_{targ} are smoothed with a Gaussian kernel of width $\sigma_\kappa = 2.0$ pixels. For each target location $\mathbf{r}_0$, the reconstructed local minimum is then identified within a search radius of $r_{\mathrm{search}} = 4$ pixels. Finally, the local second-derivative estimates are averaged over a $(2w+1) \times (2w+1)$ neighborhood with $w = 2$ around this refined location before computing the curvature ratio. This procedure reduces sensitivity to sub-pixel displacement and high-frequency numerical texture.

Confinement-Region Intensity (η_{conf}). η_{conf} quantifies residual reconstructed intensity within the intended low-intensity confinement region. Because blue-detuned trapping requires atoms to occupy dark regions of the optical field, lower values indicate cleaner confinement regions and are therefore preferred.(Appendix A)

These four quantities are reported separately because good holographic reconstruction requires simultaneous agreement in global intensity structure, boundary orientation, local confinement geometry, and residual intensity within the trapping region.

E BYPASS-TOPOLOGY VARIANT REFERENCE

For completeness, the five architectural variants compared in the topology study are summarized in Table 8, where $S_l(v_l)$ denotes the spectral transformation and $\sigma_{\mathrm{qho}}(v_l)$ the sinusoidal bypass.

Table 8: Definition of the five spectral/bypass topology variants.

Case	Block update	Description
1	$\mathrm{GELU}[S_l(v_l) + W_l v_l]$	Linear 1×1 bypass
2	$\mathrm{GELU}[S_l(v_l)] + \sigma_{\mathrm{qho}}(v_l)$	Sinusoidal bypass outside GELU
3	$\mathrm{GELU}[\mathrm{GELU}(S_l(v_l)) + \sigma_{\mathrm{qho}}(v_l)]$	Double-GELU topology
4	$\mathrm{GELU}[S_l(v_l)]$	Pure spectral baseline
5	$\mathrm{GELU}[S_l(v_l) + \sigma_{\mathrm{qho}}(v_l)]$	Single-GELU topology (HoloFNO)

F ZERNIKE DECOMPOSITION

Why use a Zernike decomposition? The first 15 Noll-indexed Zernike modes provide a physically interpretable low-order basis for optical wavefront structure, including piston, tilt, defocus, astigmatism, coma, trefoil, and spherical-aberration-like terms. Let $Z_{15}\phi$ denote the least-squares projection of the unwrapped phase ϕ onto this subspace. We define the corresponding residual as given in Eq. 7 Thus, $Z_{15}\phi$ captures the component explained by the selected low-order optical modes, while R_Z measures phase structure lying outside that subspace. A large residual should therefore be interpreted as non-low-order phase structure, rather than being identified with a particular spatial-frequency band or physical mechanism.

Fixed-weight branch interventions. Comparing independently trained models with and without a bypass would confound the architectural intervention with differences in initialization, optimization trajectory, effective capacity, and convergence. Instead, we evaluate counterfactual forward passes using the same trained checkpoint and the same input. In the spectral-only intervention, the sinusoidal bypass is forced to zero at every block; in the bypass-only intervention, the spectral contribution is suppressed instead. The resulting phases are denoted ϕ_{full}, $\phi_{\mathrm{spectral\text{-}only}}$, and $\phi_{\mathrm{bypass\text{-}only}}$.

Because the network weights and physical input remain fixed, differences between these forward passes provide a direct estimate of the sensitivity of the trained model to each pathway. However, branch removal can produce internal activation states not encountered during training, and the remaining pathway may partially compensate for the removed pathway. Consequently, these interventions support causal statements about the effect of removing a branch from the fixed trained model, but do not by themselves establish a unique internal mechanism.

We quantify bypass removal using the residual-retention ratio

$$\rho_Z = \frac{R_Z(\phi_{\mathrm{spectral\text{-}only}})}{R_Z(\phi_{\mathrm{full}})}. \tag{38}$$

Values near one indicate that bypass removal leaves the measured residual largely unchanged, values below one indicate loss of non-low-order phase structure after bypass removal, and values above one indicate that removing the bypass increases the residual.

F.1 ZERNIKE BRANCH-INTERVENTION RESULTS

Topology-dependent residual decomposition. Cases 1 and 2 have $\rho_Z > 1$, meaning that suppressing the bypass increases rather than decreases the residual outside the selected low-order Zernike subspace. In these topologies, the bypass is therefore consistent with suppressing or reshaping part of the non-low-order phase structure generated by the rest of the network rather than independently supplying that structure.

Case 3 gives $\rho_Z = 0.916$, so approximately 91.6% of the full-model residual remains after bypass removal. Its Zernike intervention effect is therefore comparatively modest.

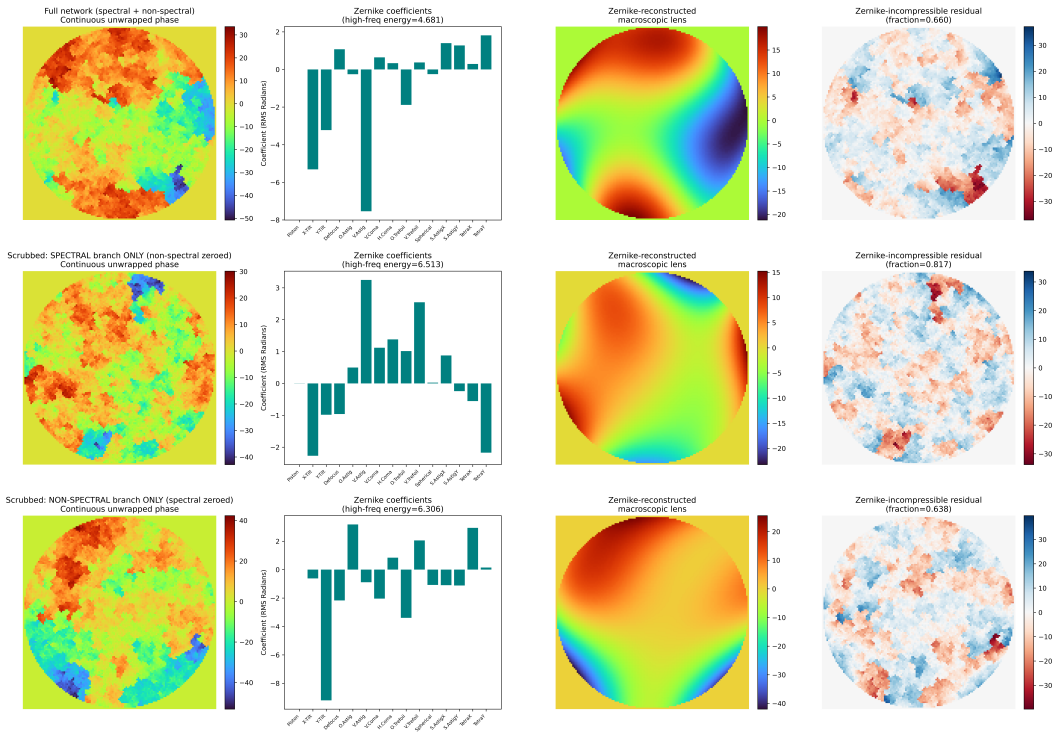

Figure 6: Fixed-weight branch intervention and Zernike decomposition for Case 1 (linear bypass).

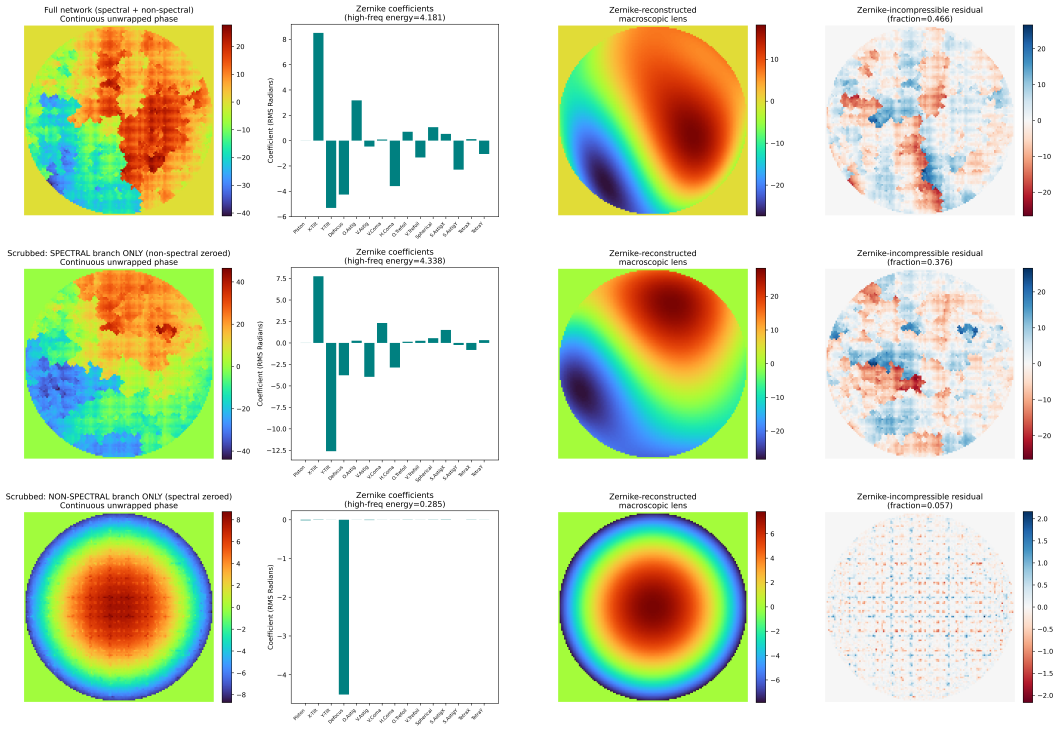

Figure 7: Fixed-weight branch intervention and Zernike decomposition for Case 2 (sinusoidal bypass outside GELU).

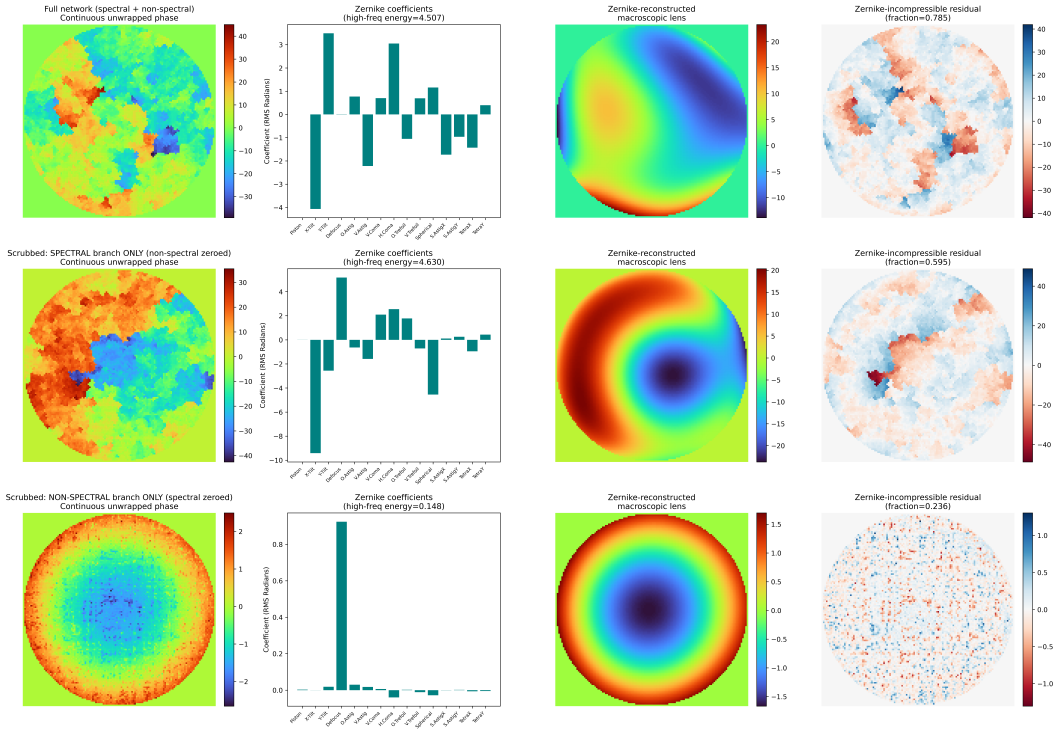

Figure 8: Fixed-weight branch intervention and Zernike decomposition for Case 3 (double-GELU topology).

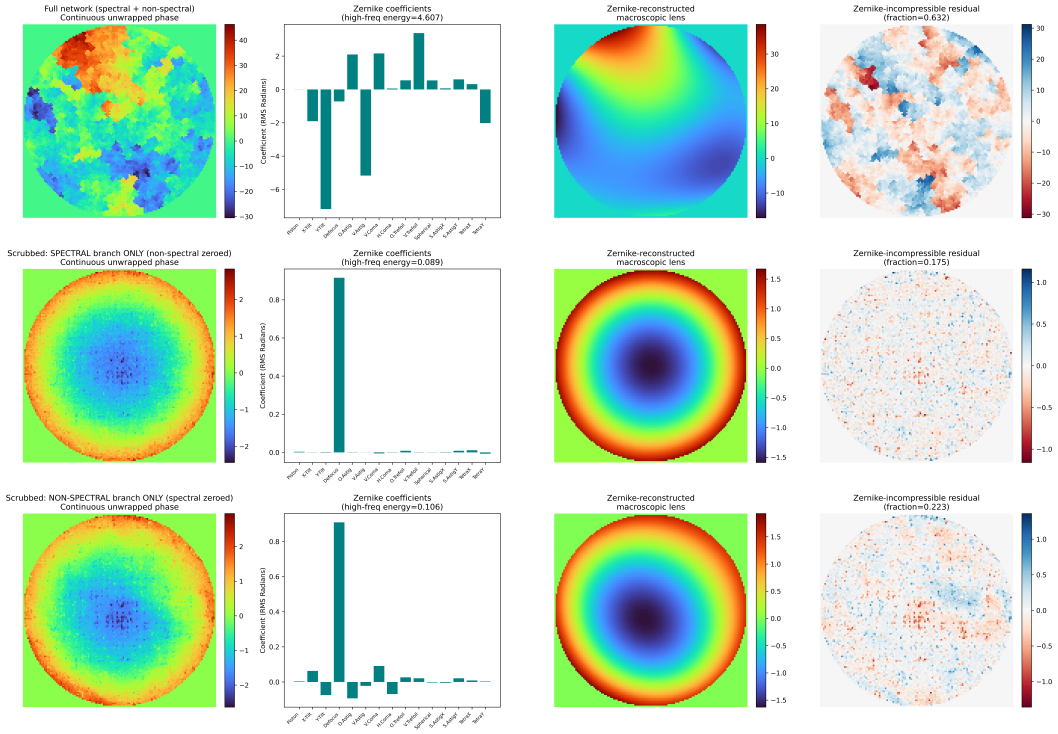

Figure 9: Fixed-weight branch intervention and Zernike decomposition for Case 5 (single-GELU HoloFNO topology).

Table 9: Zernike residual decomposition for the four bypass-equipped topologies over five independently generated evaluation sets ($K = 5$, 100 samples per set; 500 samples total per case). R_{full} denotes the complete model, $R_{\text{spectral-only}}$ the model with the bypass suppressed, and $R_{\text{bypass-only}}$ the model with the spectral pathway suppressed. $\rho_Z = R_{\text{spectral-only}}/R_{\text{full}}$ measures retention of the non-low-order residual after bypass removal. Case 4 is omitted because it contains no bypass.

Metric	Case 1	Case 2	Case 3	Case 5
R_{full}	1008.0 ± 173.4	757.5 ± 181.6	1026.0 ± 199.5	1109.4 ± 244.7
$R_{\text{spectral-only}}$	1235.5 ± 224.0	813.3 ± 186.1	939.8 ± 180.6	18.40 ± 0.14
$R_{\text{bypass-only}}$	1186.6 ± 141.5	29.10 ± 0.00	25.29 ± 0.05	23.66 ± 0.26
ρ_Z	1.226	1.074	0.916	**0.0166**

Case 5 behaves qualitatively differently. Its full-model residual is $R_{\text{full}} = 1109.4 \pm 244.7$, whereas the spectral-only and bypass-only counterfactuals fall to 18.40 ± 0.14 and 23.66 ± 0.26, respectively. The corresponding $\rho_Z = 0.0166$ means that only approximately 1.7% of the measured full-model residual remains when the sinusoidal pathway is removed.

Neither isolated pathway therefore reproduces the non-low-order residual present in the complete Case 5 network. This pattern is consistent with the measured phase structure arising through nonlinear interaction between the spectral and sinusoidal pathways rather than from either pathway acting independently. Together with the corresponding changes in the physical trap-quality metrics reported in Table 3, the result supports a substantial functional role for the bypass in the selected topology.

The Zernike analysis does not establish that the bypass represents a particular spatial Fourier band, that the QHO-inspired frequency schedule is uniquely responsible for the effect, or that the intervention identifies a unique internal mechanism. Those interpretations would require additional matched-frequency and intervention controls.

G Mechanistic Diagnostics: Coordinate-Frame Sensitivity and Terminal Spectral Response

We use two complementary diagnostics that probe properties of the learned representation without treating either as a direct measure of end-task quality. The coordinate-frame intervention tests whether the learned mapping depends on the explicit positional channels, whereas the terminal high-frequency response measures how much spatial high-frequency content remains in a late latent representation. These diagnostics complement, but are distinct from, the fixed-weight branch interventions and Zernike decomposition described in Section F.

G.1 Coordinate-Frame Perturbation

The FNO input contains two explicit Cartesian coordinate channels in addition to the physical blueprint intensity. We therefore test whether the trained mapping depends on the coordinate frame by perturbing only these coordinate channels while holding the physical input, network weights, and all other quantities fixed.

This experiment is not a conventional rotation- or reflection-equivariance test. The physical blueprint remains unchanged while its coordinate representation is modified. The resulting change therefore measures *coordinate-frame sensitivity*: how strongly the prediction depends on the orientation encoded by the explicit positional channels.

Reference coordinate field. For an $N \times N$ grid, the coordinate channels used during training are

$$\mathbf{c}_{ij} = \left(\frac{i}{N-1}, \frac{j}{N-1} \right), \qquad i, j \in \{0, \ldots, N-1\}. \tag{39}$$

Thus, both coordinate channels span the interval $[0, 1]$.

For each held-out blueprint I_{bp}, the reference prediction is

$$\phi_{\mathrm{normal}} = \Phi(I_{\mathrm{bp}}, \mathbf{c}), \tag{40}$$

where Φ denotes the trained hologram generator. The same physical input is then evaluated after replacing only $\mathbf{c}$ by a transformed coordinate field.

Centered rotations. Coordinates are centered about $(0.5, 0.5)$,

$$\tilde{\mathbf{c}}_{ij} = \begin{bmatrix} x_{ij} - 0.5 \\ y_{ij} - 0.5 \end{bmatrix}, \tag{41}$$

and transformed according to

$$\mathbf{c}_{ij}^{(\theta)} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tilde{\mathbf{c}}_{ij} + \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \tag{42}$$

for $\theta \in \{45^\circ, 90^\circ, 180^\circ\}$.

The transformed coordinates are used directly without clipping. In particular, a 45° rotation does not preserve the square coordinate domain and can produce values outside $[0, 1]$. It is therefore interpreted as a stronger off-training-coordinate perturbation than the grid-preserving 90° and 180° rotations.

Diagonal reflection. The reflection condition exchanges the two coordinate channels,

$$\mathbf{c}_{ij}^{(\mathrm{diag})} = (y_{ij}, x_{ij}), \tag{43}$$

corresponding to reflection about the diagonal $x = y$ while leaving I_{bp} unchanged.

Coordinate-perturbation RMSE. For a coordinate transformation T, let

$$\phi_T = \Phi(I_{\mathrm{bp}}, T(\mathbf{c})). \tag{44}$$

The resulting phase change is measured using

$$\mathrm{RMSE}_{\mathrm{coord}}(T) = \sqrt{\frac{1}{HW} \|\phi_{\mathrm{normal}} - \phi_T\|_2^2}, \tag{45}$$

where $H = W = 128$. No independent standardization of the two phase predictions is applied before computing the metric.

A value near zero indicates relative insensitivity to the corresponding coordinate transformation, while a larger value indicates stronger dependence on the explicit coordinate frame.

Table 10: Phase RMSE after perturbing only the coordinate channels while holding the physical blueprint fixed. For Cases 3 and 5, values are aggregated across three independently trained seeds at $k_{\max} = 60$. Cases 1 and 2 use their corresponding single trained checkpoints. The 45° transformation additionally probes coordinates outside the training-domain square and should therefore be interpreted separately from the grid-preserving transformations.

Case	Rotate 45°	Rotate 90°	Rotate 180°	Diagonal mirror
1	1.576 ± 0.029	2.032 ± 0.035	1.982 ± 0.027	2.003 ± 0.036
2	1.307 ± 0.028	2.492 ± 0.025	2.656 ± 0.016	2.470 ± 0.025
3	1.654 ± 0.224	2.625 ± 0.037	2.693 ± 0.039	2.656 ± 0.056
5	1.899 ± 0.225	2.619 ± 0.074	2.678 ± 0.025	2.619 ± 0.060

Interpretation. All four bypass-equipped topologies exhibit substantial prediction changes under coordinate-frame rotations and reflection. The effect is particularly large for the 90°, 180°, and diagonal-reflection conditions in Cases 2, 3, and 5, where the phase RMSE is approximately 2.5–2.7. This shows that the explicit positional channels contribute materially to the learned mapping.

These results should not be interpreted as evidence that the physical holographic mapping itself lacks rotational or reflection equivariance. Because only the coordinate channels are transformed while

I_{bp} remains unchanged, the intervention deliberately creates a mismatch between physical content and positional encoding. The experiment therefore establishes coordinate-frame dependence of the trained operator, not geometric generalization under transformed physical targets.

The comparable responses of Cases 3 and 5 under the grid-preserving transformations also indicate that the stronger branch-intervention effect observed for Case 5 is not simply explained by an unusually strong dependence on the Cartesian coordinate channels.

G.2 Terminal High-Frequency Response

The coordinate intervention probes explicit positional dependence, but does not reveal how different pointwise activations affect the spatial-frequency content propagated through the network. We therefore additionally examine the spectral distribution of a late latent representation.

Procedure. Let $v_3(\mathbf{x})$ denote the latent activation map at the third spectral block for a batch of held-out inputs. For each latent channel, we compute the two-dimensional discrete Fourier transform

$$\hat{v}_3(\mathbf{k}) = \mathcal{F}_{2D}\{v_3(\mathbf{x})\}\,, \tag{46}$$

and form a radially aggregated power spectrum

$$S_3(k) = \sum_{\|\mathbf{k}'\|\in[k,k+1)} |\hat{v}_3(\mathbf{k}')|^2\,. \tag{47}$$

We define the terminal high-frequency response as the fraction of spectral energy lying in the upper third of the retained spectral range,

$$H_{\mathrm{term}} = \frac{\sum_{k\geq k_{\mathrm{hi}}}^{k_{\max}} S_3(k)}{\sum_{k=0}^{k_{\max}} S_3(k)}, \qquad k_{\mathrm{hi}} = \frac{2}{3}k_{\max}. \tag{48}$$

This quantity is not an end-task performance metric. Rather, it measures how much high-spatial-frequency energy remains in the late latent representation under otherwise matched architectures.

Table 11: Terminal high-frequency response for the pure spectral baseline under different pointwise activation functions. Higher values indicate a larger fraction of late-layer latent spectral energy in the upper third of the retained band.

Activation	H_{term}
GELU	0.182
ReLU	0.198
Sigmoid	0.054

Interpretation. The Sigmoid model exhibits a substantially lower terminal high-frequency fraction (0.054) than either GELU (0.182) or ReLU (0.198). This is consistent with the stronger saturation of Sigmoid: once its pre-activations enter a saturated regime, variations in the input produce smaller changes in the output, which can suppress fine-scale variation propagated through later layers.

GELU and ReLU both preserve considerably more late-layer high-frequency content. ReLU yields the largest value under this diagnostic, while GELU retains a similarly high fraction without relying on this metric alone for activation selection. The activation choice in the main text is therefore based on the reported reconstruction and trap-quality metrics, with H_{term} serving only as a complementary representation-level diagnostic.

Importantly, a larger H_{term} is not inherently better. High spectral energy may represent useful fine-scale phase structure, but it may also arise from unnecessary texture or noise. The metric therefore characterizes spectral behavior rather than ranking hologram quality, and it should not be interpreted as evidence that a particular activation or branch uniquely generates the physically relevant high-frequency structure.

H BASELINE ARCHITECTURES

We compare HoloFNO against four neural baselines chosen to separate, where possible, the effects of backbone architecture and training objective.

U-Net (Phys Loss). This baseline uses a standard convolutional U-Net (Ronneberger et al., 2015) with an encoder–decoder structure, skip connections, and a purely convolutional bottleneck. It contains no spectral operator and is trained using the same five-term physics-supervised objective as HoloFNO (Eq. 2). It therefore provides the most direct comparison between a conventional convolutional encoder–decoder and the proposed native-resolution spectral architecture under the same training objective.

U-Net (NPCC). This model uses the same convolutional U-Net architecture as U-Net (Phys Loss), but replaces the five-term physics objective with the Normalized Pearson Correlation Coefficient (NPCC) objective used in related self-supervised holographic learning (Yan et al., 2026). Comparing the two plain U-Nets therefore isolates the effect of the training objective while holding the architecture fixed.

HDC CNN. The HDC baseline is a dilated convolutional network that maintains the native spatial resolution throughout the model and expands its receptive field through dilated convolutions rather than spatial downsampling (Yu & Koltun, 2016; Wang et al., 2018). Unlike the U-Net baselines, it therefore avoids an encoder–decoder bottleneck while retaining purely convolutional information propagation.

For a unit-stride dilated network, the theoretical receptive field is

$$RF_{\mathrm{HDC}} = 1 + \sum_{l=1}^{D} (k_l - 1) d_l, \tag{49}$$

where k_l and d_l denote the kernel size and dilation of layer l. A strict locality-based limitation applies only if the dilation schedule used in the evaluated model satisfies $RF_{\mathrm{HDC}} < 128$. If $RF_{\mathrm{HDC}} \geq 128$, the architecture has theoretical full-domain coverage, and differences in performance must instead be interpreted in terms of effective receptive field (Luo et al., 2016), sparse information propagation, dilation sampling structure (Wang et al., 2018), and optimization bias rather than an inability to represent global dependencies.

U-Net (Yang). This baseline follows the Fourier U-Net formulation of Yan et al. (2026). A convolutional encoder first compresses the spatial representation, after which an FNO-style spectral operator is applied at the bottleneck together with a parallel 1×1 convolutional pathway. The representation is then decoded back to the full spatial resolution, and the model is trained using NPCC.

Unlike a purely convolutional U-Net, this architecture possesses explicit global spectral interaction. Its main architectural distinction from HoloFNO is therefore not the presence or absence of global coupling, but the resolution at which that coupling occurs: the Fourier U-Net performs spectral mixing after spatial compression, whereas HoloFNO performs Fourier-domain mixing directly at the native 128×128 resolution.

HoloFNO. HoloFNO uses a spectral Fourier neural operator backbone (Li et al., 2021) without a convolutional encoder–decoder or spatial downsampling. Each block performs global Fourier-domain mixing directly at the native 128×128 resolution and combines the spectral pathway with the harmonic-oscillator-inspired sinusoidal bypass defined in Eq. equation 6. The selected Case 5 topology combines both pathways before a single GELU activation,

$$v_{l+1} = \mathrm{GELU}[S_l(v_l) + \sigma_{\mathrm{qho}}(v_l)]\,. \tag{50}$$

HoloFNO is trained using the same five-term physics-supervised objective as U-Net (Phys Loss).

Together, these baselines provide complementary comparisons: U-Net (Phys Loss) tests backbone architecture under matched physics supervision; U-Net (NPCC) tests the effect of the objective within a fixed convolutional architecture; U-Net (Yang) tests spectral mixing within a compressed encoder–decoder representation; and HDC CNN tests native-resolution convolutional propagation without spatial compression.

Table 12: Quality-latency comparison between HoloFNO (Case 5, single deterministic forward pass) and fixed-iteration Gerchberg–Saxton (GS) at native resolution 128×128, evaluated on 500 held-out test instances. Runtime is reported as average in milliseconds; all other metrics are reported as mean $\pm$ std. GS reaches HoloFNO's mean full-image relative RMSE at approximately 60 iterations (3.368 ms), which remains faster than HoloFNO's fixed cost at this resolution, though only 89.6% of samples reach that accuracy within 60 iterations (see quality-match summary).

Method	Average Runtime (ms)	Rel-RMSE (Full)	Rel-RMSE (ROI)
HoloFNO (Case 5)	5.219	0.1640 ± 0.0237	0.1494 ± 0.0213
GS (1 iter.)	0.071	0.9993 ± 0.0003	0.9993 ± 0.0003
GS (3 iter.)	0.193	0.6038 ± 0.0588	0.6029 ± 0.0591
GS (5 iter.)	0.315	0.4929 ± 0.0600	0.4922 ± 0.0602
GS (10 iter.)	0.582	0.3708 ± 0.0502	0.3703 ± 0.0503
GS (20 iter.)	1.144	0.2711 ± 0.0363	0.2707 ± 0.0364
GS (40 iter.)	2.252	0.1947 ± 0.0280	0.1944 ± 0.0281
GS (60 iter.)	3.368	0.1623 ± 0.0253	0.1621 ± 0.0253

I Out-of-Distribution Generalization: Anharmonic Lattices

To probe generalization beyond the harmonic-oscillator superpositions used for training, we evaluate the trained models on target densities generated from eigenstates of the quartic-perturbed Hamiltonian

$$H = -\frac{1}{2}\nabla^2 + \frac{1}{2}r^2 + \lambda r^4, \qquad \lambda = 1.0. \tag{51}$$

We consider four state pairs, $(0, 1)$, $(2, 3)$, $(4, 5)$, and $(6, 7)$, which produce increasingly structured anharmonic target densities. Neither the quartic potential nor these resulting target families is present during training.

Table 13: κ_{geom} for the four anharmonic OOD targets shown in Figure 10. Positive values indicate locally confining curvature under the reported metric, whereas non-positive values indicate failure to recover the corresponding local confinement geometry.

Target pair	U-Net (Phys Loss)	U-Net (NPCC)	HDC CNN	HoloFNO
$(0, 1)$	1.435	-1.000	-1.000	-1.000
$(2, 3)$	-1.000	-1.000	-1.000	-1.000
$(4, 5)$	-1.000	-0.274	-1.000	-0.610
$(6, 7)$	-0.759	-0.750	-0.816	0.364

I.1 Per-Case Behavior

States $(0, 1)$. For the simplest anharmonic target, U-Net (Phys Loss) is the only model with positive reported curvature, reaching $\kappa_{\mathrm{geom}} = 1.435$. The other three models have non-positive curvature under this metric. HoloFNO nevertheless produces a visually well-defined annular dark structure around the target region, illustrating that global visual similarity and local curvature need not track one another exactly.

States $(2, 3)$. All four architectures give $\kappa_{\mathrm{geom}} = -1.000$ for the three-lobed target. The reconstructions still differ visually: the U-Net models preserve portions of the coarse target envelope, HDC CNN remains dominated by a smooth low-frequency blob, and HoloFNO retains more visible internal modulation. However, none of the models recovers positive local confinement curvature for this case.

States $(4, 5)$. For the diagonal multi-lobe target, all models again remain non-positive in κ_{geom}. HDC CNN and U-Net (Phys Loss) give -1.000, U-Net (NPCC) gives -0.274, and HoloFNO gives -0.610. Despite the negative local-curvature score, the HoloFNO reconstruction retains visibly

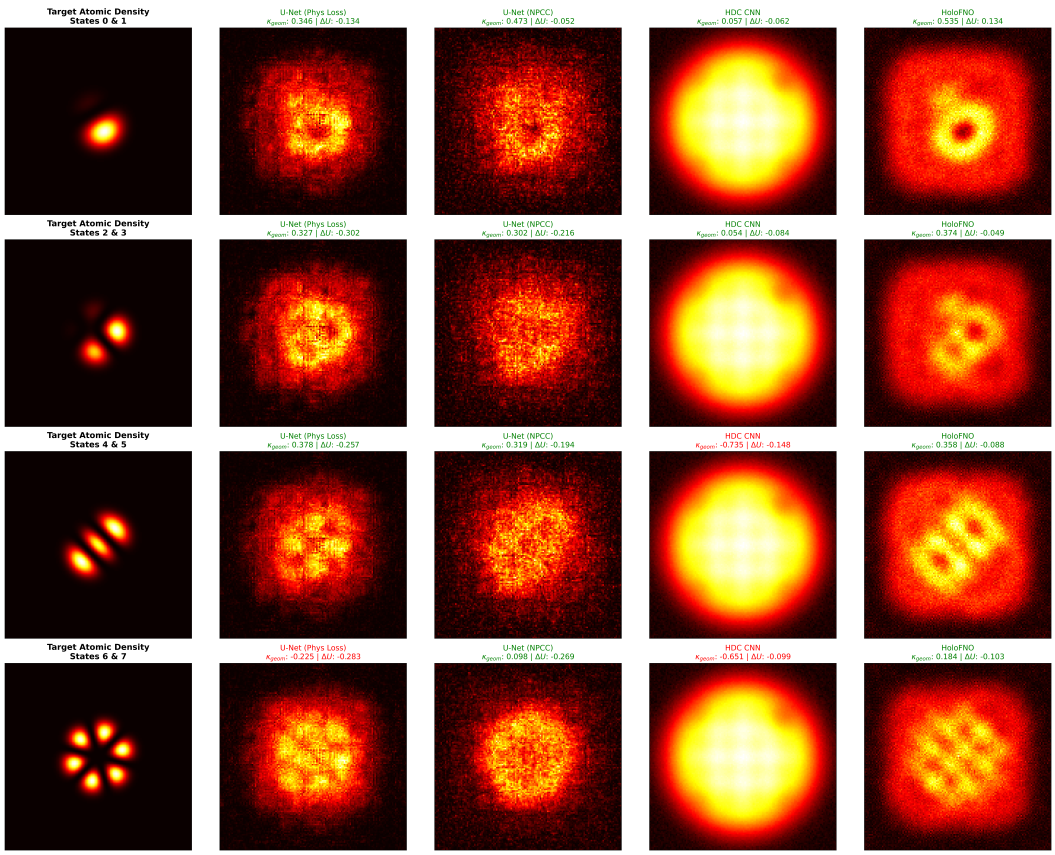

Figure 10: Qualitative OOD comparison on four anharmonic target-state pairs. Rows correspond to $(0, 1)$, $(2, 3)$, $(4, 5)$, and $(6, 7)$, respectively. Columns show the target atomic density followed by reconstructions from U-Net (Phys Loss), U-Net (NPCC), HDC CNN, and HoloFNO. The panel annotations report κ_{geom} together with the auxiliary barrier quantity generated by the diagnostic script; the discussion below uses κ_{geom} and the visible reconstruction structure.

separated internal dark structures that qualitatively follow the multi-lobe organization of the target more closely than the smooth HDC output.

States $(6, 7)$. The most spatially structured target produces a different ordering. HoloFNO is the only evaluated model with positive curvature, $\kappa_{\text{geom}} = 0.364$, while U-Net (Phys Loss), U-Net (NPCC), and HDC CNN give -0.759, -0.750, and -0.816, respectively. The corresponding HoloFNO reconstruction also preserves a clear multi-lobed arrangement resembling the six-fold target geometry.

Interpretation. The anharmonic sweep does not support a claim that any architecture is uniformly robust across all OOD targets. Instead, the results reveal strongly target-dependent behavior. U-Net (Phys Loss) succeeds in local curvature on the simplest case but loses positive confinement curvature as the target structure changes, whereas HoloFNO fails the same curvature test on the first three cases but recovers positive curvature on the most structured $(6, 7)$ target.

The qualitative reconstructions additionally show that κ_{geom} measures only one aspect of OOD behavior. In particular, HoloFNO can preserve recognizable multi-lobe global structure even when the local curvature metric is non-positive. Conversely, visual similarity alone does not establish a valid confining trap. The OOD results should therefore be interpreted jointly from reconstructed structure and the reported physical metrics rather than from either alone.

HDC CNN shows the most consistent qualitative failure mode across the sweep: its outputs remain dominated by broad, smooth intensity profiles with weak reproduction of the target's internal geometry. This is consistent with the architectural-bias discussion in Appendix J, but the OOD experiment alone does not establish a strict receptive-field impossibility.

Overall, the four-case experiment should be regarded as a controlled stress test rather than a comprehensive generalization benchmark. Broader claims would require additional anharmonic strengths, state combinations, independently generated samples, and other families of out-of-distribution potentials.

J Mathematical Analysis of Architectural Expressivity and Bias

The empirical differences between HDC CNN, convolutional U-Nets, the Fourier U-Net, and HoloFNO should not be interpreted as a generic "CNN versus Fourier" expressivity result. The architectures differ more specifically in (i) the spatial range over which information can interact, (ii) how global and fine-resolution information are routed, and (iii) the inductive and optimization biases introduced by their parameterizations.

Sufficiently deep convolutional networks possess universal approximation properties (Zhou, 2020). A strict expressivity restriction therefore applies here only when a particular finite architecture has an insufficient dependency range. When global dependency is structurally available, the comparison concerns architectural and optimization bias rather than representational impossibility.

J.1 Global Coupling of the Fourier-Optics Objective

Under the Fraunhofer model used throughout this work,

$$E_f(u,v) \propto \mathcal{F}_{2D}\left\{A_0(x,y)e^{i\phi(x,y)}\right\}, \qquad I_\phi(u,v) = |E_f(u,v)|^2. \tag{52}$$

For a discrete aperture, let n index SLM pixels and k index focal-plane samples:

$$E_k = \sum_n A_n e^{i\phi_n} e^{-i2\pi k\cdot n/N}, \qquad I_k = |E_k|^2. \tag{53}$$

Differentiating the intensity at focal-plane coordinate k with respect to an arbitrary SLM phase coordinate ϕ_n gives

$$\frac{\partial I_k}{\partial \phi_n} = -2\,\mathrm{Im}\left[E_k^* A_n e^{i\phi_n} e^{-i2\pi k\cdot n/N}\right]. \tag{54}$$

Except at special field configurations for which this derivative vanishes, a phase change at one SLM location can influence intensity throughout the focal plane. The inverse-design problem is therefore intrinsically globally coupled.

For the intensity-dependent portion of the physics-supervised objective, denoted $\mathcal{L}_I$, the gradient with respect to one phase coordinate is

$$\frac{\partial \mathcal{L}_I}{\partial \phi_n} = \sum_k \frac{\partial \mathcal{L}_I}{\partial I_k}\frac{\partial I_k}{\partial \phi_n}. \tag{55}$$

Thus, an update to one predicted phase coordinate can depend on intensity residuals distributed across the complete focal plane. The wrap-aware phase-TV term adds a local phase regularizer but does not remove this global coupling.

This argument concerns dependency structure, not uniqueness. Fourier magnitude phase retrieval remains non-unique and can exhibit ambiguities and stability issues (Grohs et al., 2020; Gerchberg & Saxton, 1972).

J.2 Locality of Finite Convolutional Networks

For a convolutional architecture with kernel size k_l, stride s_l, and dilation d_l at layer l, the theoretical receptive field is

$$RF = 1 + \sum_{l=1}^{D} (k_l - 1) d_l \prod_{j<l} s_j. \tag{56}$$

Let $\mathcal{R}_p$ denote the receptive field associated with output coordinate p. For any fixed convolutional architecture,

$$x|_{\mathcal{R}_p} = x'|_{\mathcal{R}_p} \quad \Longrightarrow \quad f_\theta(x)_p = f_\theta(x')_p, \tag{57}$$

for every parameter choice θ. Equivalently, wherever the network is differentiable,

$$\frac{\partial f_\theta(x)_p}{\partial x_q} = 0, \qquad q \notin \mathcal{R}_p. \tag{58}$$

Consequently, if a selected target-to-phase solution map requires an output coordinate p to depend on an input coordinate q lying outside $\mathcal{R}_p$, that particular finite architecture cannot represent the map exactly. This is the precise condition under which limited receptive field constitutes an expressivity restriction; it should not be generalized to convolutional networks as a class.

Moreover, the effective receptive field of trained CNNs is often smaller and more centrally concentrated than the formal theoretical receptive field (Luo et al., 2016). Thus, theoretical global coverage does not imply that all distant input locations influence the output equally strongly.

J.3 Implications for the Evaluated Baselines

HDC CNN. The HDC baseline preserves native spatial resolution while enlarging its receptive field through dilated convolutions (Wang et al., 2018). As discussed in Section H, a strict locality-based expressivity limitation applies only if the actual dilation schedule gives $RF_{\mathrm{HDC}} < 128$. If the theoretical receptive field spans the full domain, its empirical behavior must instead be interpreted through factors such as effective receptive field, sparse propagation, dilation sampling structure, and optimization bias (Luo et al., 2016).

The poor OOD reconstructions of HDC CNN are therefore evidence that the tested convolutional parameterization is poorly matched to the present inverse-design task, but are not by themselves proof of a generic convolutional expressivity limitation.

Convolutional U-Net. A conventional U-Net obtains broad contextual information through repeated spatial downsampling, followed by reconstruction through an expanding path with skip connections (Ronneberger et al., 2015). Once its theoretical receptive field spans the complete input, there is no corresponding locality-based impossibility result.

Its architectural distinction is instead that global context is represented primarily in spatially compressed deep features, while fine-resolution information is routed through shallower skip pathways. The decoder must combine these information streams during reconstruction.

Downsampling can also introduce aliasing when subsampling is not preceded by sufficient low-pass filtering (Zhang, 2019). This does not imply that U-Nets cannot represent high-frequency outputs; rather, it gives them a different information-routing and optimization bias from architectures that perform global interaction directly at the native spatial resolution.

Fourier U-Net. The Fourier U-Net of Yan et al. (2026) differs importantly from a plain convolutional U-Net because its bottleneck contains an explicit FNO-style spectral operator. It therefore possesses global spectral interaction and should not be characterized as lacking global receptive-field coverage.

Its main distinction from HoloFNO is the spatial resolution at which this interaction occurs. The Fourier U-Net first compresses the input through its encoder, performs global spectral mixing at the bottleneck, and then decodes the result to full resolution. The observed reconstruction behavior is therefore consistent with effects involving spatial compression, bottleneck representation, decoder

reconstruction, and the training objective, but the present experiments do not isolate any one of these as the unique mechanism.

J.4 HoloFNO: Native-Resolution Global Spectral Mixing

HoloFNO retains its principal latent representation at the native 128×128 spatial resolution throughout all four Fourier blocks. For block l, the spectral transformation is

$$S_l(v_l) = \mathcal{F}^{-1}\left[R_l(\mathbf{k})\,\mathcal{F}[v_l](\mathbf{k})\right], \tag{59}$$

where $R_l(\mathbf{k})$ denotes the learned transformation on the retained Fourier modes.

Each retained Fourier basis function has support across the complete spatial domain. Consequently, each spectral block provides direct domain-wide interaction without first reducing the 128×128 spatial grid (Li et al., 2021; Kovachki et al., 2023). This does not imply preservation of all spatial frequencies: the retained mode count $k_{\max}$ remains an explicit architectural bandwidth parameter.

The selected Case 5 topology combines the spectral pathway with the pointwise sinusoidal bypass before a single GELU:

$$v_{l+1} = \mathrm{GELU}\left[S_l(v_l) + B_l(v_l)\right], \tag{60}$$

where B_l is the QHO-inspired multi-frequency bypass defined in Section 5.

The bypass itself should not be interpreted as a spatial Fourier operator. Its sinusoidal frequencies act on projected latent feature values rather than directly on (x, y). Global spatial communication is therefore supplied by S_l, while B_l provides a structured pointwise nonlinear transformation of the globally mixed representation.

The branch-intervention results provide empirical support for interaction between these pathways. For Case 5,

$$R_{\text{full}} \gg R_{\text{spectral-only}}, \qquad R_{\text{full}} \gg R_{\text{bypass-only}}, \tag{61}$$

under the Zernike-residual diagnostic. Neither isolated pathway reproduces the non-low-order residual of the complete network. This is consistent with the measured phase structure arising through nonlinear interaction between native-resolution spectral mixing and the structured pointwise pathway, rather than from either pathway independently.

The explicit coordinate channels provide an additional positional signal. Their contribution is supported by the coordinate-frame perturbation experiment in Appendix G.1. The learned phase bias provides a further spatial prior, but because it is independent of the input blueprint it cannot itself supply input-dependent global coupling.

J.5 Architecture-Specific Summary

Overall, the experiments are most appropriately interpreted as evidence that native-resolution global spectral mixing provides a useful architectural bias for the holographic inverse-design problem studied here. They do not establish that CNNs or U-Nets are fundamentally incapable of representing suitable phase maps.

The only strict architectural statement is conditional: if a specific finite convolutional model has a receptive field smaller than the spatial dependencies required by the selected target-to-phase map, it cannot represent that map exactly. For architectures with global dependency already available, including the evaluated U-Nets and Fourier U-Net, observed performance differences are more appropriately attributed to information routing, architectural inductive bias, and optimization.

K Limitations

This work has three main limitations.

(1) **Fixed-resolution evaluation.** All principal experiments are conducted at a native resolution of 128×128. We therefore do not claim zero-shot resolution transfer or resolution-independent operator behavior. The computational scaling experiment does not establish reconstruction quality or trap quality across resolutions.

Table 14: Architectural mechanisms relevant to the globally coupled holographic inverse-design objective.

Model	Long-range mechanism	Spatial pathway	Interpretation
HDC CNN	Dilated convolutions	Native resolution	Strict locality limitation only if $RF_{\mathrm{HDC}} < 128$; otherwise primarily an effective-RF and optimization issue.
U-Net	Deep multiscale convolution	Downsample–bottleneck–upsample with skips	Can obtain global theoretical RF, but global context and fine detail are routed through different spatial scales.
Fourier U-Net	FNO-style spectral bottleneck	Global mixing after encoder compression followed by decoding	Possesses explicit global interaction; differs from HoloFNO primarily in where spectral mixing occurs.
HoloFNO	Fourier mixing in every block	Native 128×128 grid throughout	Global spectral interaction without progressive spatial compression, combined in Case 5 with a structured pointwise sinusoidal pathway.

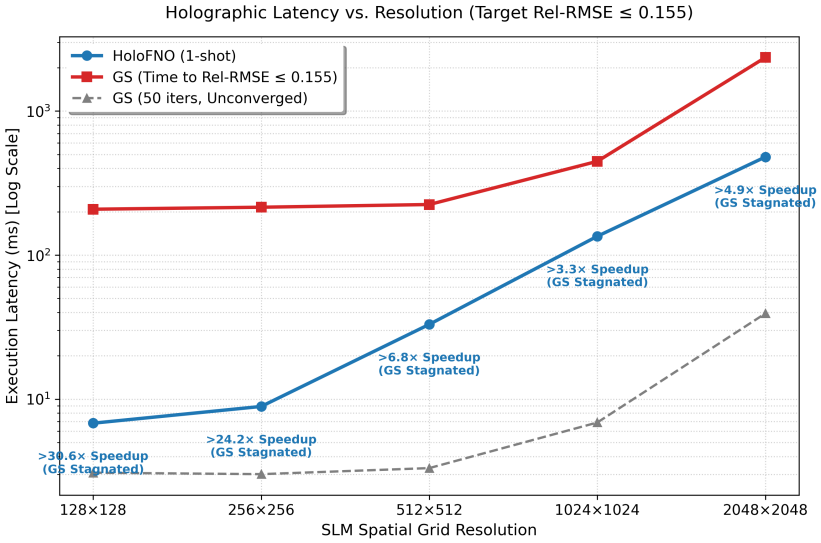

Figure 11: Quality-matched latency versus native grid resolution. For each tested grid size, the fixed-iteration GS iteration count is chosen to match HoloFNO's mean reconstruction RMSE at that resolution, and the resulting GS runtime is plotted against HoloFNO's single-pass runtime. GS latency grows with the per-iteration FFT cost and diverges from HoloFNO's comparatively flat inference cost as resolution increases, indicating that HoloFNO's latency advantage widens at higher native resolutions.

(2) **Simulation-only validation.** The reported results are obtained using a differentiable Fourier-optics model and simulated trap-quality metrics. Effects present in a physical optical system, including SLM calibration error, aberrations, quantization, alignment error, and detector noise, are not included in the reported evaluation. Atomic dynamics, loading efficiency, and atom survival are also not experimentally evaluated. Experimental validation is therefore still required.

(3) **Limited mechanistic and OOD validation.** The branch-scrubbing experiments quantify the effect of removing a pathway within a fixed trained model, but do not uniquely identify the internal mechanism, since the interventions may produce intermediate representations outside the training distribution and the pathways may have co-adapted during training. In addition, the QHO-inspired

frequency schedule has not yet been compared against matched generic frequency-bank controls. Although the two plain U-Net baselines isolate the training objective within a fixed architecture, the comparison between HoloFNO and the Fourier U-Net changes both architecture and objective and therefore does not isolate the effect of spectral-mixing resolution. Finally, the OOD study is restricted to a small anharmonic target family. Broader mechanistic controls, matched architectural comparisons, and OOD evaluations are therefore left for future work.

L Ethics Statement

This work studies computational methods for holographic optical-trap synthesis, motivated by neutral-atom quantum computing and tweezer-array rearrangement. The reported study uses synthetic target densities, including quantum harmonic-oscillator superpositions and anharmonic targets for out-of-distribution evaluation. It does not involve human subjects, personal data, or experimental operation of optical-tweezer hardware. We have not identified specific ethical concerns arising from the reported simulation study beyond those generally associated with quantum computing research. Any future hardware deployment would require experimental validation and compliance with applicable laboratory laser-safety and equipment-operating procedures.

M AI Use and Reproducibility Statement

Generative AI tools were used for writing assistance, language editing, formatting, retrieval and discovery of related work, drafting portions of the manuscript, and assistance with checking internal consistency. The authors reviewed the AI-assisted material and take full responsibility for the final manuscript, including its claims, mathematical formulations, experimental results, references, and other research artifacts.

Generative AI tools were not used to generate the synthetic datasets or target distributions reported in this work; develop the theoretical model or conceptual framework; formulate mathematical claims; provide ingredients for proofs or write proofs. Tasks involving human subjects, surveys, interviews, recordings, and qualitative data were not applicable to this work.

For reproducibility, we provide the source code used for the experiments as supplementary material, including the proposed method, training and evaluation scripts, configuration files, and result-generation scripts, together with environment and dependency specifications.