Propagation Is Not Decoration: When Physical Diffraction Carries Generative Computation

Anonymous authors
Paper under double-blind review

Abstract

We introduce the conditioned Diffraction Field Model (cDFM), a generative architecture built from Wave-Aperture-Diffraction (WAD) blocks. Each block conditions a complex field with a learned, class-conditioned aperture and propagates it with a fixed Angular Spectrum Method (ASM) operator. One random field is carried through four resolutions and read out directly as intensity, with no learned decoder. A learned aperture can form images by itself and bypass propagation, so the presence of propagation in such a model does not show that it does any work. We test this by retraining eight controls from scratch with matched parameter counts and update budgets. In an earlier image-fitting model, restricting the aperture to 16×16 spatial resolution lets ASM remove 30.7% of the excess error of a retrained bypass, compared with 2.1% at 64×64. On MNIST, the full generator reaches 83.1% class accuracy from fresh sources across three seeds. Replacing ASM with the identity, $\mathcal{E}(x) = x$, drops this to 20.3%, and rebuilding the source at each scale drops it to 16.4%. An identity control whose aperture can amplify light reaches 78.5%, so the identity drop depends on the aperture being passive, as a physical aperture is. A nonphysical operator matched to ASM's local reach reaches 77.0%, and a global permuted spectral operator 55.7%. A wider version (994k parameters) trained for 30 epochs reaches 95.9% on one seed. Retrained at that budget, the passive identity control reaches 90.3% and the permuted spectral control 94.1%, but both fit held-out digits worse. In a separate strict-locality task, identity evolution leaves distant prediction at chance (25%), ASM at propagation distance 20 reaches $\sim 55\%$, and a global random unitary and a permuted spectral operator reach $\sim 86\%$ and $\sim 88\%$. Propagation matters most in this generator when the aperture is passive and the learned path is small or briefly trained, and less as the learned path grows. In the locality task, diffraction is one of several global operators that can carry distant information. On CIFAR-10 the same design produced mostly blobs with a coarse lattice texture, so the generative results are limited to MNIST.

1 Introduction

Diffraction moves information across a field without learning a spatial mixing matrix. A model can include physical propagation and still learn to bypass it.

We introduce the conditioned Diffraction Field Model (cDFM), built from Wave-Aperture-Diffraction (WAD) blocks. Each block conditions a complex field with a learned, class-conditioned aperture, applies fixed angular-spectrum propagation, and reads out the resulting intensity. In our generator, a single random field is carried through every stage (Figure 1). The model must develop that field into an image; it has no learned image decoder after propagation and does not rebuild its source at each stage.

A learned aperture can form images by itself. Given enough spatial freedom, it takes over work we meant propagation to do. Deleting propagation from a trained model does not test this, because the other components were trained with propagation present. We instead retrain every control from scratch and limit the spatial resolution of the aperture. In earlier class-conditioned image-fitting experiments, lowering aperture resolution raised the benefit of ASM over a retrained bypass, although

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.

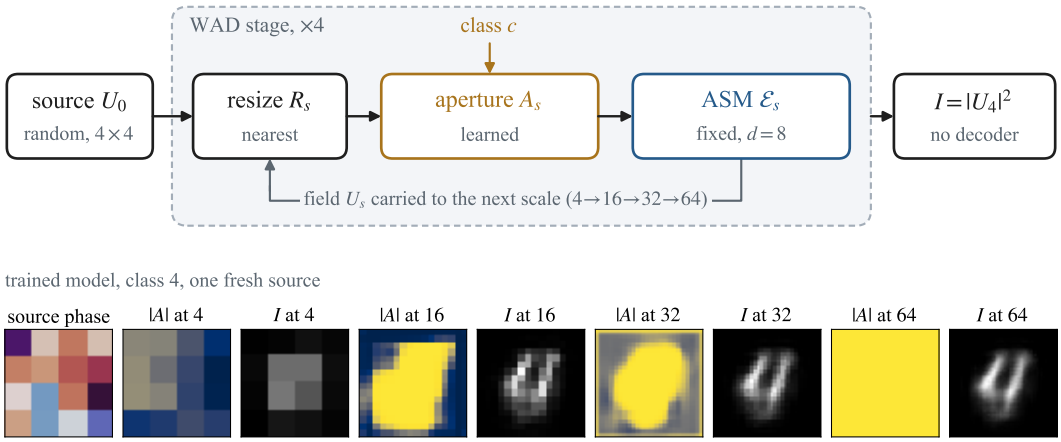

Figure 1: cDFM. Top: one 4×4 random field passes through four WAD stages, each a nearest-neighbor resize, a learned class-conditioned aperture, and fixed ASM propagation. The final intensity is the image. Bottom: the trained full model (seed 17) on one fresh source for class 4, showing aperture amplitude $|A|$ and intensity I after each stage. Across all 320 fresh sources, the 64×64 aperture amplitude stays between 0.997 and 1.

the benefit did not grow steadily as the aperture shrank. When every learned component is strictly local, distant spatial prediction stays at chance without global evolution.

Our main generative experiment carries this field through $4 \to 16 \to 32 \to 64$ grids, with an aperture cap of $m = 32$ and a shared multiscale assignment loss. Across three training seeds, the full model reaches 87.3% class accuracy from fixed sources and 83.1% from fresh sources. Its mean class accuracy is higher than that of every retrained control we test, each of which removes or replaces propagation, spatial variation in the aperture, the carried field, or the multiscale schedule. The strongest replacement, a random operator matched to the local reach of ASM, reaches 77.0% from fresh sources, and a permuted spectral operator that shuffles the ASM transfer values reaches 55.7%. Removing either propagation or the persistent field drops fresh class accuracy below 21% at this budget. With an aperture that can amplify light, the model reaches 78.5% without propagation, so the drop from removing propagation depends on the aperture being passive. Trained for 30 epochs with wider stage networks, the same architecture reaches 95.9% from fresh sources on one seed. Retrained at that budget, the passive identity and permuted spectral controls reach 90.3% and 94.1%, so propagation helps most when the learned path is small or briefly trained.

Across these experiments, propagation carries the computation when the learned path is constrained: when the aperture is passive, its resolution is limited, or the model is small or briefly trained. In the separate strict-locality task, nonphysical global operators outperform ASM. The evidence does not yet establish natural-image generation at scale.

2 Related work

Learned diffraction and optical generation. Diffractive neural networks already use learned optical modulation and propagation for computation (Lin et al., 2018). Optical generative models also precede this work: Chen et al. (2025) combine digitally prepared phase seeds with an optical decoder, and Oguz et al. (2024) study optical denoising within a diffusion process. Adversarially trained optical generators include scattering-based noise sampling (Zhan et al., 2024), optoelectronic generation with an electronic discriminator (Qiu et al., 2025), and simulated diffractive generators and discriminators (Hu et al., 2026). Stochastic optical sources and fields carried between optical layers therefore already exist, and we claim neither. cDFM is a numerical architecture in which the learned masks depend on both the requested class and the current state of the field, while a fixed propagation law does all spatial mixing between them.

Fixed mixing and architectural context. FNet demonstrates that fixed Fourier mixing can replace learned attention in a language encoder (Lee-Thorp et al., 2022). MetaFormer studies how much performance comes from the surrounding architecture rather than a particular token mixer (Yu et al., 2022). We therefore compare against controls with matched parameter counts, including generic global mixing operators. Our final-generator replacements test two such transformations and one local operator matched to ASM, while the locality task shows a setting in which the global ones are stronger than ASM.

Matching generated and real samples. Distribution matching through transport costs is established in generative modeling (Genevay et al., 2018). We use exact, finite-batch assignment within each class, with one assignment shared across image scales. We solve an exact permutation problem with no entropic regularization and propose no new transport algorithm.

3 The conditioned Diffraction Field Model

3.1 A persistent Wave-Aperture-Diffraction path

Let c be the requested class and $z \in \mathbb{R}^{4\times4}$ have independent standard normal entries. We initialize one unit-amplitude field,

$$U_0 = \exp(i\pi \tanh z)\,. \tag{1}$$

The four stages operate at $n_s \in \{4, 16, 32, 64\}$. At stage s, a fixed nearest-neighbor resize R_s expands the real and imaginary parts of the preceding field to $n_s \times n_s$. The first resize is the identity. A class embedding $e(c)$ is broadcast over the grid and concatenated with the two field components. A stage-specific network f_{θ_s} produces amplitude and phase logits:

$$\widetilde{U}_s = R_s(U_{s-1}), \qquad (a_s, p_s) = P_{m_s} f_{\theta_s}\big[\Re\widetilde{U}_s, \Im\widetilde{U}_s, e(c)\big], \tag{2}$$

$$A_s = \sigma(a_s)\exp(i\pi\tanh p_s)\,, \qquad U_s = \mathcal{E}_s(A_s \odot \widetilde{U}_s), \qquad I_s = |U_s|^2. \tag{3}$$

Here P_{m_s} average-pools the logits to $m_s \times m_s$ and nearest-expands them back before the nonlinearities, with $m_s = \min(m, n_s)$. The generator uses $m = 32$, giving effective aperture grids $4, 16, 32, 32$. The aperture network still sees the full stage grid; the cap only limits how finely its output mask can vary.

Each stage uses two 3×3 convolutions of width 96, each followed by GELU, then a 1×1 convolution producing two logits. The class embedding has width eight. With four separate stage networks, the model has 367,960 trainable parameters. The final image is I_4 itself. There is no learned upsampler, convolutional readout, target-correlated input, or source reinjection. Intermediate intensities are supervised during training but are not fed back as new source images.

3.2 Fixed angular-spectrum evolution

We use scalar angular-spectrum propagation,

$$\mathcal{E}_s(U) = \mathcal{F}^{-1}[H_s \odot \mathcal{F}(U)]\,, \qquad H_s(f_x, f_y) = \exp\left(id\sqrt{(2\pi/\lambda)^2 - (2\pi f_x)^2 - (2\pi f_y)^2}\right). \tag{4}$$

The implementation uses orthonormal FFTs, wavelength $\lambda = 0.5$, distance $d = 8$, and grid pitch $\Delta x = 1$ in consistent simulation units. The square root is evaluated in complex arithmetic. No propagation parameter is trained. At these settings, the sampled modes are propagating modes.

The FFT operates on an unpadded periodic grid. This is a discrete angular-spectrum model, and we have not tested how closely it matches free-space optics. Numerical sampling matters in angular-spectrum calculations (Matsushima & Shimobaba, 2009); our implementation is not the band-limited method of that paper. Since pitch is fixed as the grid grows, the multiscale path changes computational extent as well as resolution. Digital resizing and neural, state-dependent aperture prediction are part of cDFM. We make no all-optical implementation or energy-efficiency claim.

3.3 Shared multiscale Hungarian training

Each update draws $B = 32$ target images from one class and 32 independent sources. Let $T_s(y)$ resize the target to stage s, and define $D_s(x, y) = \|x - y\|_2^2 / n_s^2$. We form one cost matrix,

$$C_{ij} = \frac{1}{4} \sum_{s=1}^{4} D_s(I_s(z_i, c), T_s(y_j)), \qquad \pi^* = \arg\min_{\pi \in S_B} \sum_{i=1}^{B} C_{i,\pi(i)},$$
$$\mathcal{L} = \frac{1}{4B} \sum_{i=1}^{B} \sum_{s=1}^{4} D_s(I_s(z_i, c), T_s(y_{\pi^*(i)})). \tag{5}$$

We solve the assignment with the Hungarian method (Kuhn, 1955) on detached costs, then differentiate the selected pairs. Every scale is assigned to the same target instance for that update. Targets enter only this loss. At inference the model takes a class and a new random field and makes one four-stage pass.

4 Experimental design

Data and optimization. We split the 60,000 MNIST (LeCun et al., 1998) training examples into 55,000 training and 5,000 validation examples with a fixed permutation. Pixel values are divided by 255 and bilinearly resized from 28×28 to 64×64; lower-scale targets are area-resized from that image. This preprocessing smooths the target distribution and is shared by all models. Each model receives 6,000 Adam (Kingma & Ba, 2014) updates at learning rate 10^{-3}, with class labels cycling through $0, \ldots, 9$. Assignment uses the full batch of 32; gradients accumulate over microbatches of eight without recomputing separate assignments. This corresponds to 192,000 training-image presentations. We compare seeds 17, 29, and 43 at the final update rather than selecting a best intermediate checkpoint.

Aperture selection and evaluation. An exploratory sweep compared $m \in \{4, 8, 16, 32, 64\}$ using the same three training seeds. The fixed-source class-accuracy criterion selected $m = 32$. The final controls reuse these baseline runs, so they are not an independent post-selection replication. For each seed, every model uses identical saved fixed and fresh source banks, each containing 32 draws per class. The fresh bank was excluded from the aperture ranking. We report both.

Class accuracy is the fraction of generated samples that a frozen MNIST classifier assigns to the requested class. The classifier is an evaluation proxy; it reaches 97.3% real-digit accuracy on a 1,000-example test sanity check under the same preprocessing. Generator matching uses 32 held-out validation targets per class. We report final-scale matched MSE, within-class pairwise pixel MSE V, and the same quantity restricted to classifier-correct samples, V_{correct}. These distances measure pixel-level output variation and give no estimate of distribution coverage or semantic diversity, so we inspect unfiltered sample sheets alongside them. Reported sample standard deviations are taken across training seeds.

Retrained controls. All controls retain the same nominal parameter count, optimizer, sources, target streams, and update budget. *Identity* removes propagation. *Identity with gain* also replaces the sigmoid amplitude with an unbounded softplus, so the aperture can add light. *Uniform aperture* retains a state- and class-conditioned scalar amplitude and phase at each stage but averages away spatial mask variation. *PermutedSpectral* shuffles the ASM transfer values over frequencies. *Local random spectral* replaces ASM with a random phase-only transfer function matched to the width of the ASM impulse response. *Structured RandomUnitary* applies fixed random spatial and spectral phase masks around Fourier mixing.

Same-source reinjection resets the field at each later stage to a resized copy of the original random source. It preserves the scale ladder and shared assignment but removes the carried field. *Single-scale* instead runs all four stages on the 64×64 grid. The latter changes both computation and the supervision schedule, so it does not isolate persistence. Equal parameter and update budgets do not imply equal FLOPs or equal effective capacity. Appendix B gives the exact interventions.

Configuration	Fixed class acc. (%)	Fresh class acc. (%)	Fresh MSE	Fresh V_{correct}
Full ASM	87.29 ± 1.26	83.12 ± 4.23	0.03803	0.01455
Identity	20.62 ± 9.21	20.31 ± 9.00	0.06765	0.00389
Identity with gain	79.27 ± 8.31	78.54 ± 7.33	0.03744	0.01277
Uniform aperture	11.04 ± 0.95	10.42 ± 0.48	0.09285	0.00393
Structured random unitary	18.96 ± 1.26	19.79 ± 0.18	0.07315	0.00542
Local random spectral	76.77 ± 5.63	76.98 ± 3.08	0.04275	0.00923
Permuted spectral	59.48 ± 13.30	55.73 ± 10.00	0.05020	0.00686
Same-source reinjection	16.88 ± 2.71	16.35 ± 3.96	0.07810	0.00347
Single-scale	16.46 ± 2.53	17.92 ± 2.84	0.06519	0.00304

Table 1: Final generator and retrained controls, three paired seeds and 6,000 updates. Class accuracy is mean ± sample SD. MSE and V_{correct} are seed means; lower matched MSE is better, whereas larger variation is not necessarily better. All models have 367,960 nominal trainable parameters.

5 Results

5.1 The full generator uses propagation when the aperture is passive

Table 1 compares the full generator with its retrained controls. On fresh sources, replacing ASM with Identity reduces class accuracy from 83.12% to 20.31% and increases matched MSE from 0.03803 to 0.06765. Under this training budget and with a passive aperture, the learned aperture path does not recover the full generator's result without propagation.

The Identity outputs in Figure 2 are black. A mask can only remove light, and without propagation the light cannot move, so no output pixel can be brighter than the 4×4 stage at that position. The multiscale loss pushes that stage toward a dim, blurred digit, and in the seed-17 run its mean amplitude falls below 10^{-3} within 500 updates. A dim, blocky digit was still possible, so this is partly an optimization collapse.

To test whether the collapse comes from the amplitude bound, we retrained Identity with an unbounded softplus amplitude that starts from the same initial mask (Appendix B). This Identity-with-gain control reaches $78.54 \pm 7.33\%$ fresh class accuracy. ASM's paired fresh advantage over it is 4.58 ± 6.93 points, within seed variation, and the control's fresh matched MSE, 0.03744, is slightly lower than ASM's 0.03803. Once the mask can add light, it forms images without propagation, which is the bypass described in the introduction. Most of Identity's failure therefore comes from the passive aperture. A physical aperture cannot add light, so propagation carries the computation under that physical constraint.

The uniform-aperture control stays at chance, 10.4% fresh-source class accuracy. Conditioning alone, without spatial variation in the aperture, does not reproduce the result.

We also retrained the full model at propagation distances $d \in \{0, 1, 2, 4, 8, 16, 32\}$ (seed 17, 6,000 updates; $d = 8$ was chosen before this sweep). At $d = 0$ the ASM transfer function is exactly one, the model is Identity, and fresh class accuracy is 10.0%. It rises to 60.6% at $d = 1$, 81.3% at $d = 4$, and 86.6% at $d = 16$, then falls to 79.7% at $d = 32$ (Appendix I). The $d = 0$ and $d = 8$ runs reproduce the seed-17 Identity and full-model rows of Table 3 exactly. The physical distance setting changes what the generator can learn.

5.2 ASM outperforms the tested nonphysical propagation operators

PermutedSpectral is the strongest global control, reaching 55.73% class accuracy from fresh sources. Its paired deficit relative to ASM is 27.40 ± 14.06 percentage points across seeds. Structured RandomUnitary reaches 19.79%, a deficit of 63.33 ± 4.08 points. Both also have worse matched MSE. PermutedSpectral varies widely across seeds (Table 1), but ASM wins on every paired seed, on both banks.

PermutedSpectral keeps the ASM transfer values but scrambles their frequency arrangement, and the structured unitary control supplies global mixing without that arrangement. Both mix globally and still lose to ASM, so mixing by itself does not explain the gap. Translation equivariance does not either, since PermutedSpectral is still a convolution. The clearer difference is reach: at $d = 8$, ASM

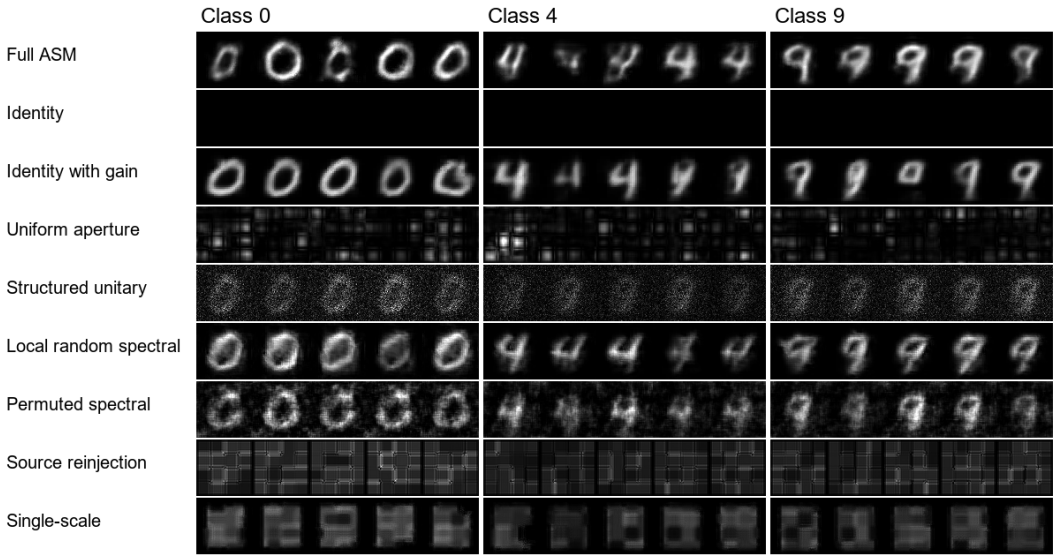

Figure 2: Fresh-source outputs from the first training seed. For each configuration, the first five draws of classes 0, 4, and 9 are shown without filtering. The full ASM model, Identity with gain, and the local control form recognizable digits; PermutedSpectral retains some digit structure, while the other interventions mainly produce coarse patterns. Display range is fixed to $[0, 1]$.

keeps about 95% of an impulse's energy within 3 pixels at every stage, while both controls spread it across the grid.

To test whether reach alone explains the gap, we retrained a nonphysical local control: a smooth random phase-only transfer function, calibrated at each stage so its impulse response has the same RMS width as ASM (Appendix B). It keeps ASM's reach, energy conservation, and translation equivariance, but not its dispersion law. It reaches $76.98 \pm 3.08\%$ fresh class accuracy, far above PermutedSpectral but below ASM on every paired seed and both banks (paired fresh deficit 6.15 ± 2.30 points), with higher matched MSE (0.04275 against 0.03803). Local reach accounts for most of ASM's advantage over the global controls, and ASM keeps a smaller, consistent edge beyond it.

5.3 Persistence and recognizable variation

Rebuilding the same source at each scale reduces fresh class accuracy to 16.35%. Of all the controls, this one changes the least: the source distribution, scale ladder, aperture budget, and assignment objective stay the same. Later stages no longer receive the field that earlier stages shaped. Reinjection also cuts the gradient path from the final loss back through earlier fields, so the result reflects the forward computation and training together. Stage-by-stage outputs (Appendix G) show the full model forming a pixelated digit at 16×16 that becomes smooth by 64×64, whereas the reinjection control ends in a lattice-like pattern.

The single-scale configuration reaches 17.92%. It also changes the scale schedule, computation, and target resolutions. The single-scale comparison alone does not show whether later stages specialize by scale.

Figure 3 shows fresh samples for every class. Sources produce different stroke shapes and loops as well as malformed examples. The full model's fresh pairwise MSE is 0.01641; restricting to classifier-correct outputs gives 0.01455. The samples cannot tell us how much of the digit distribution the model covers or whether it copies training examples. None of the 27 trained models in Table 1 meets the internal screen of at least 90% overall class accuracy and at least 75% for every class on either bank; at the longer budget of Section 5.4, ASM meets it from epoch 10 on, and both retrained controls meet it at some evaluations.

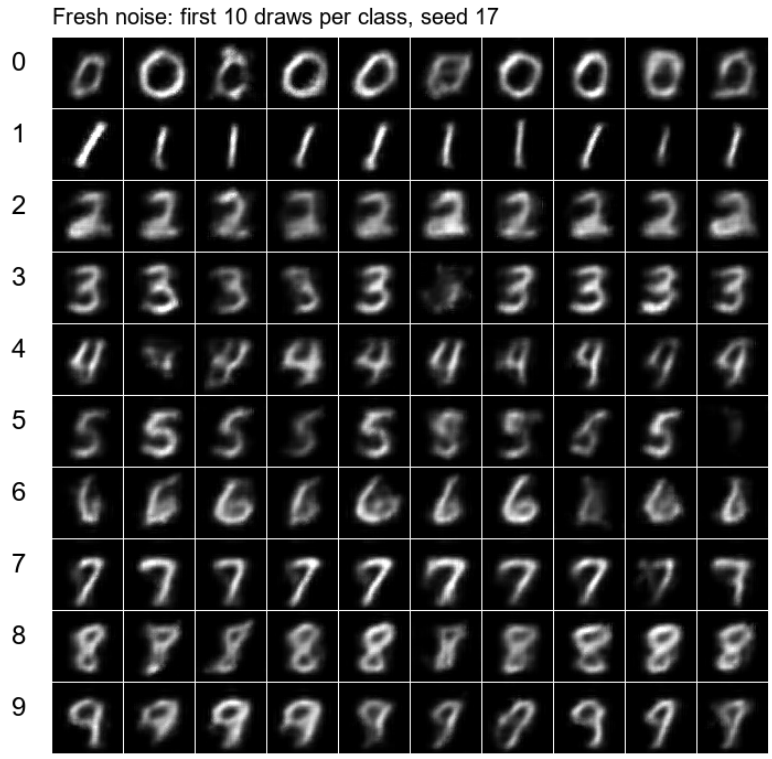

Figure 3: Full persistent ASM generator, fresh source bank, seed 17. Rows are requested classes; columns are the first ten draws, without selection or filtering. The model changes digit morphology, but some outputs are ambiguous. Table 1 gives the three-seed numbers.

5.4 Longer training

We trained one wider version of the generator for 30 passes over the training split. It keeps the persistent $4 \to 16 \to 32 \to 64$ field, ASM, $m = 32$, direct intensity readout, and the shared multiscale assignment. The only architectural change is stage width, from 96 to 161, which gives 993,780 trainable parameters. It uses seed 17 and 51,570 updates of batch 32, or 1,650,240 image presentations, about 8.6 times the budget of Table 1. We evaluated it every five epochs on the same kind of fixed and fresh source banks and did not select a checkpoint. We then retrained Identity, with the passive aperture, and PermutedSpectral at the same width and budget.

	Fresh class acc. (%)			Fresh matched MSE		
Epochs	ASM	Identity	PermSpec	ASM	Identity	PermSpec
5	84.38	63.75	67.50	0.03813	0.04247	0.04698
10	95.31	86.88	87.50	0.03487	0.03975	0.04193
20	95.63	86.56	93.44	0.03384	0.04150	0.03975
30	95.94	90.31	94.06	0.03228	0.03826	0.03769

Table 2: Wider models (993,780 parameters), one seed, 30 epochs, evaluated every five epochs; all evaluations are in Appendix H. Each bank has 32 draws per class.

After 30 epochs the ASM model reaches 97.19% class accuracy on fixed sources and 95.94% on fresh sources, with fresh matched MSE 0.03228 (Table 2). From epoch 10 onward it meets our internal screen on both banks. The weakest fresh class at epoch 30 is 5, at 78.13%. Fixed-bank class accuracy is now close to the classifier's own 97.3% on real test digits, so this proxy has little room left. Fresh pairwise MSE is 0.02386, higher than the 0.01641 of the 6,000-update model. Figure 4 shows unfiltered samples from all three models.

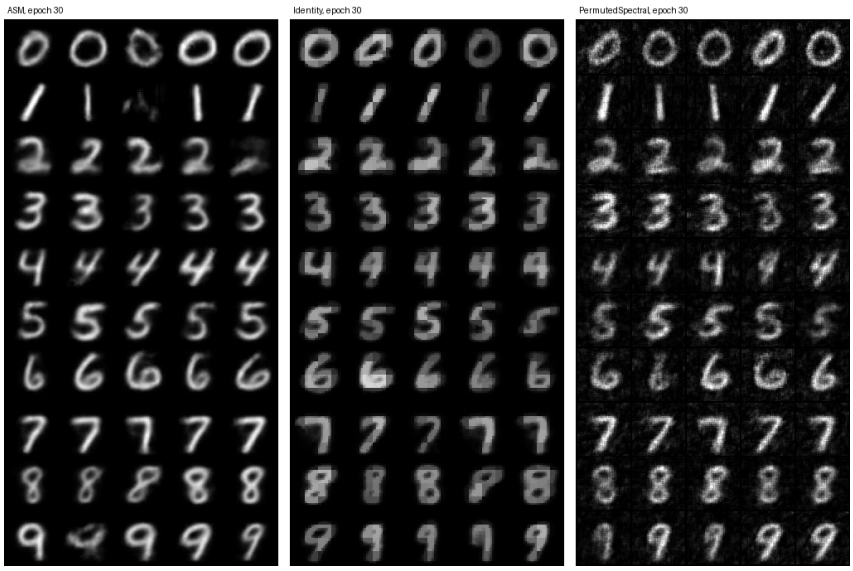

Figure 4: Wider models after 30 epochs, fresh source bank, seed 17. Rows are requested classes; columns are the first five draws, without selection or filtering.

This run changes width and training length together and has one seed. Both retrained controls close most of the class-accuracy gap, reaching 90.31% and 94.06% from fresh sources against 95.94% for ASM. ASM keeps the lowest fresh matched MSE, 0.03228 against 0.03826 and 0.03769. With more width and training, the learned path recovers much of what propagation gives at 6,000 updates. The classifier's 97.3% ceiling limits how far class accuracy can separate the three. With 320 samples per bank and one seed, the class-accuracy difference between ASM and PermutedSpectral is within sampling noise; the matched-MSE difference holds at every evaluation.

5.5 Aperture capacity controls whether propagation matters

The earlier capacity experiment studies a different, class-conditioned image-fitting model. A learned aperture at full 64×64 spatial resolution gives a retrained bypass almost the same validation error as ASM. At $m = 16$, ASM removes 30.7% of the bypass's excess error; at $m = 64$, the fraction is 2.1%. Here excess is error above an evaluation-only class-mean reference, and the reported fraction averages paired per-seed reductions (Appendix E). These are not 30.7% and 2.1% reductions in total MSE.

The dependence is nonmonotonic: $m = 8$ removes 25.4% of bypass excess, and strong restriction also worsens absolute error. Too much aperture freedom makes ASM unnecessary, and too little makes both ASM and the bypass worse. In the generator's own aperture sweep, $m = 32$ narrowly wins fixed-source class accuracy, while $m = 64$ is slightly better on fresh sources and matched MSE (Appendix D). We keep $m = 32$ rather than reselecting after the controls.

5.6 Other global operators also carry distant information

A separate synthetic task places instance-specific symbols in one region and requests their prediction in a distant region. Learned spatial operations have a bounded local path. With Identity evolution, the readout cannot depend on the source region and symbol accuracy remains at the 25% chance level. Global evolution restores a communication path.

In the exploratory distance-20 ASM comparison, accuracy is approximately 55%; random-unitary and permuted-spectral comparisons reach approximately 86% and 88%. The original shorter-distance ASM configuration reaches only about 34%. We use distance 20 here, which we tested after the original setting. These are exploratory runs on a different model, objective, and physical setup than the generator (Appendix F).

The ASM advantage holds for this generator only. The locality task shows that distant regions need a global route, and nonphysical routes work better there.

6 Discussion and limitations

With a passive aperture, retrained models without propagation do not recover the full generator's results at 6,000 updates. An aperture that can amplify light, or more width and training, recovers most of them. In our tests, removing spatial variation in the aperture or persistence prevents the model from matching the full result, and so does removing ASM while the aperture stays passive.

ASM outperforms the three nonphysical propagation operators tested here, but the strict-locality task reverses that ordering. The local control separates reach from the rest of ASM, but the remaining edge could come from the shape of its impulse response rather than its physical validity. There is no GAN, diffusion, transformer, or learned global-mixer baseline among our comparisons.

Baseline seeds were reused from aperture selection. The classifier is imperfect, and the 30-epoch model's fixed-bank score (97.19%) nearly equals the classifier's accuracy on real digits (97.3%), so this metric cannot resolve further gains. Correct-only variation depends on which images it accepts. We do not report significance tests that treat generated images as independent training replicates. Controls may optimize at different rates, and longer training shrinks the ASM advantage (Section 5.4). We trained only two controls at that budget, on one seed. Equal parameter counts do not guarantee equal effective capacity.

Recognizable MNIST variation is a narrower result than natural-image generation. An exploratory one-seed CIFAR-10 transfer (a three-channel field over $4 \to 8 \to 16 \to 32$ with $m = 16$, evaluated after 30 million image presentations) produced blobs with a coarse lattice texture that only rarely resembled the requested class, and its later stages changed the rendered image less than earlier ones. One likely cause is that nearest-neighbor resizing carries the early coarse structure upward, and the per-scale loss rewards forming the image early, so later stages add little detail. That run changed resizing, persistence, shared supervision, the aperture grid, and direct intensity readout together, so it cannot show which of them caused the failure (samples in Appendix J).

7 Conclusion

cDFM develops a stochastic complex field through learned aperture modulation and fixed diffraction. In the persistent multiscale MNIST generator, retraining without ASM under a passive aperture, removing spatial aperture variation, replacing ASM with the tested nonphysical operators, or rebuilding the source at each scale all reduce recognizable generation. Propagation carries part of the computation when the learned path is constrained, most clearly when the aperture is passive, as a physical aperture is. An identity control whose aperture can amplify light comes within seed variation of ASM. A nonphysical operator with the same local reach recovers most of the gap, so much of what ASM contributes here is local energy transport. Trained for 30 epochs with wider stage networks, the same architecture reaches 97.19% class accuracy from fixed sources and 95.94% from fresh sources on one seed. At that budget the passive identity and permuted spectral controls close most of the class-accuracy gap but still fit held-out digits worse than ASM.

Reproducibility statement

Appendices A–J describe preprocessing, architecture, objective, controls, seed-level results, and the scope of the historical comparisons. The supplementary material contains code that trains the generator and every retrained control, together with the evaluation classifier. For each reported experiment, we preserve the run identifier, configuration, evaluation-bank hashes, checkpoints, samples, and logs.

AI use statement

AI tools assisted with writing and wording refinement, drafted some passages that the authors then revised, and assisted with related-work retrieval and discovery, and research ideation and execution. This included implementation assistance and experiment orchestration. The authors set the

research direction and made the decisions about architecture, experimental design, controls, interpretation, and scientific claims. The authors verified the implementation, evidence, citations, and final manuscript.

References

Shiqi Chen, Yuhang Li, Yuntian Wang, Hanlong Chen, and Aydogan Ozcan. Optical generative models. *Nature*, 644(8078):903–911, August 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09446-5. URL http://dx.doi.org/10.1038/s41586-025-09446-5.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 2018. URL https://proceedings.mlr.press/v84/genevay18a.html.

Pei Hu, Tengyu Cui, Yuanyuan Zhang, and Shuai Feng. Generative adversarial optical networks using diffractive layers for digit and action generation. *Photonics*, 13(1):94, January 2026. ISSN 2304-6732. doi: 10.3390/photonics13010094. URL http://dx.doi.org/10.3390/photonics13010094.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, March 1955. ISSN 1931-9193. doi: 10.1002/nav.3800020109. URL http://dx.doi.org/10.1002/nav.3800020109.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 0018-9219. doi: 10.1109/5.726791. URL http://dx.doi.org/10.1109/5.726791.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. FNet: Mixing tokens with Fourier transforms. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.319. URL https://aclanthology.org/2022.naacl-main.319/.

Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, September 2018. ISSN 1095-9203. doi: 10.1126/science.aat8084. URL http://dx.doi.org/10.1126/science.aat8084.

Kyoji Matsushima and Tomoyoshi Shimobaba. Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. *Optics Express*, 17(22):19662, October 2009. ISSN 1094-4087. doi: 10.1364/oe.17.019662. URL http://dx.doi.org/10.1364/OE.17.019662.

Ilker Oguz, Niyazi Ulas Dinc, Mustafa Yildirim, Junjie Ke, Innfarn Yoo, Qifei Wang, Feng Yang, Christophe Moser, and Demetri Psaltis. Optical diffusion models for image generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 59150–59173. Curran Associates, Inc., 2024. doi: 10.52202/079017-1887. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6cb81234ab47027e991728ed7dd76735-Paper-Conference.pdf.

Jumin Qiu, Ganqing Lu, Tingting Liu, Dejian Zhang, Shuyuan Xiao, and Tianbao Yu. Optoelectronic generative adversarial networks. *Communications Physics*, 8(1), April 2025. ISSN 2399-3650. doi: 10.1038/s42005-025-02081-6. URL http://dx.doi.org/10.1038/s42005-025-02081-6.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10819–10829, June 2022.

Ziyu Zhan, Hao Wang, Qiang Liu, and Xing Fu. Photonic diffractive generators through sampling noises from scattering media. *Nature Communications*, 15(1), December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-55058-4. URL http://dx.doi.org/10.1038/s41467-024-55058-4.

A Implementation and evaluation details

The source is a 4×4 real Gaussian draw mapped to complex unit phase by $\exp(i\pi \tanh z)$. All stages carry this field forward. Complex nearest-neighbor resizing applies independently to real and imaginary components. The class embedding has 80 parameters. Each stage has a $10 \to 96$ convolution with kernel three and padding one, a $96 \to 96$ convolution with the same kernel, and a $96 \to 2$ convolution with kernel one, all with biases. Each of the first two convolutions is followed by GELU. This gives $80 + 4(91{,}970) = 367{,}960$ trainable parameters. Stage parameters are not shared. ASM and resizing have no trainable parameters.

The MNIST split is generated by a random permutation with seed 20260907. Targets use bilinear $28 \to 64$ resizing with `align_corners=False`; stage supervision uses area interpolation from that target. The final output is not passed through a learned image nonlinearity. For classifier evaluation it is clipped to $[0, 1]$ and mapped to $[-1, 1]$. The stored intensity tensors remain available for loss and variation analysis. The classifier uses GroupNorm and remains frozen. Its weights and their SHA-256 hash are included in the supplementary material; the real-test check is classifier validation, not a held-out generator distribution evaluation.

Training uses float32 learned computation and complex64 fields. Adam uses learning rate 10^{-3} with the implementation defaults. The target-order, source, fixed-evaluation, and fresh-evaluation seeds are $1000 + s$, $2000 + s$, $3000 + s$, and $4000 + s$ for training seed s. Labels cycle by update. Within a class, 32 target IDs are sampled without replacement for that update. The detached full-batch forward determines a single assignment; the four microbatch forwards then accumulate its gradient before one optimizer update. Assignment is recomputed next update.

The final-scale matched MSE is an evaluation-time Hungarian match between 32 generated images and 32 fixed held-out targets of the requested class. Pairwise variation for a class with K generated images is

$$V = \frac{2}{K(K-1)} \sum_{i<j} \frac{\|I_i - I_j\|_2^2}{64^2}. \tag{6}$$

Class means are then averaged. V_{correct} applies the same expression to the classifier-correct subset when at least two examples remain. Eligible-class counts vary for weak controls. No source-swap causal intervention, distribution precision/recall, or memorization audit is implied by these metrics.

B Exact control definitions

Evolution replacements. Identity sets the transfer array to one, preserving the FFT round trip in the implementation. PermutedSpectral replaces H_s with a fixed permutation of its flattened entries reshaped to the same grid. The permutation is generated with seed $o + 104729s$, where o is the recorded operator seed and s is the code's stage index. This preserves transfer values, not spatial geometry. Structured RandomUnitary is

$$\mathcal{E}_{\text{RU}}(U) = \mathcal{F}^{-1} D_f \mathcal{F} D_x U, \tag{7}$$

where D_x and D_f are diagonal unit-modulus masks with independently sampled uniform phases. The stage RNG seed is $o + 130363s + 1009n_s$, and o equals the training seed. Reported seed variation therefore includes the draw of the replacement operator. It is a structured unitary transform, not a Haar-distributed dense random unitary. Local random spectral replaces H_s with $\exp(i\alpha_s \phi_s)$, where ϕ_s is standard-normal noise on the frequency grid, circularly smoothed with a Gaussian of width $\max(n_s/8, 0.5)$ frequency bins and standardized, drawn with seed $o + 7919s + n_s$. The scale α_s is chosen from a grid on $[0, 20]$ so that the RMS radius of the intensity impulse response matches ASM at the same grid size; the matched radii are 1.77, 1.77, 1.73, and 1.71 cells. The operator is unitary and translation-equivariant but follows no dispersion relation. Figure 5 shows the impulse responses of all four operators. These controls are fixed after initialization and add no learned weights.

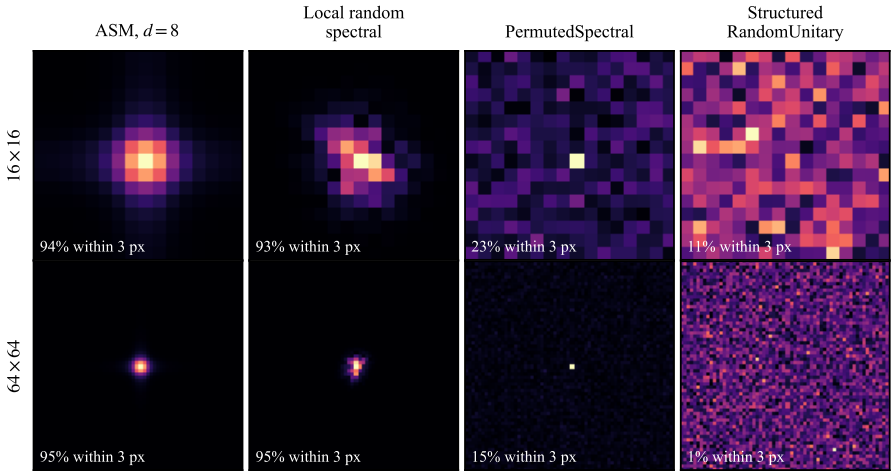

Figure 5: Intensity impulse responses of the fixed operators at the 16×16 and 64×64 stages (seed 17), shown on a square-root scale. Labels give the share of energy within 3 pixels of the source.

Aperture neutralization. After the usual aperture pooling, the two logits are averaged over spatial positions and broadcast before sigmoid and phase tanh. The model retains conditioning and state dependence through those scalars. This control tests spatial mask freedom rather than removing every aperture function. Nominal parameter count is unchanged, but the effective output degrees of freedom are much smaller. No unused learned image head compensates for them.

Identity with gain. This control uses the Identity evolution and replaces $\sigma(a)$ with $\mathrm{softplus}(a - c)$, where $c = 4 - \log(e^{\sigma(4)} - 1) \approx 3.49$. The amplitude logit is initialized with bias 4, so both versions start from the same mask amplitude, $\sigma(4) \approx 0.982$. The softplus amplitude has no upper bound, so the mask can add light. It also has a larger gradient than the sigmoid at initialization, so this control changes optimization as well as the bound.

Persistence and scale. For same-source reinjection, each stage after the first receives $R_s(U_0)$ instead of $R_s(U_{s-1})$. It is the same source, not a new draw and not a corrupted target. Intermediate losses and one shared assignment remain. Consequently, later output losses cannot train earlier stage fields through a carried-state path. The single-scale control uses four 64×64 stages, the original source expanded to 64 before the first, and all four supervision targets at 64. It retains persistence but changes both the scale schedule and its loss resolutions. Neither comparison isolates every possible optimization confound.

Budgets and recovery. All final models receive 6,000 updates and retain 367,960 nominal parameters. Source snapshots, initial weights, gradients, checkpoints, and intermediate field records are preserved. Two jobs failed at step zero from an infrastructure error, before any training, and were rerun with identical configurations. Runs that never trained are excluded from the analysis. No wall-clock or FLOP equivalence is claimed, especially for the single-scale control.

C Seed-level generator results

Configuration	Fixed class acc. (%) 17	29	43	Fresh class acc. (%) 17	29	43
Full ASM	85.94	87.50	88.44	78.75	83.44	87.19
Identity	10.00	25.62	26.25	10.00	26.56	24.38
Identity with gain	70.63	87.19	80.00	72.19	86.56	76.88
Uniform aperture	11.88	10.00	11.25	10.31	10.00	10.94
Structured random unitary	17.81	18.75	20.31	19.69	19.69	20.00
Local random spectral	70.31	80.63	79.38	73.44	79.06	78.44
Permuted spectral	74.69	53.75	50.00	67.19	51.25	48.75
Same-source reinjection	13.75	18.44	18.44	11.88	19.38	17.81
Single-scale	18.75	13.75	16.88	20.00	14.69	19.06

Table 3: Class accuracy at update 6,000. Fixed and fresh sources are paired across models within a seed. The full-model rows were reused from aperture selection.

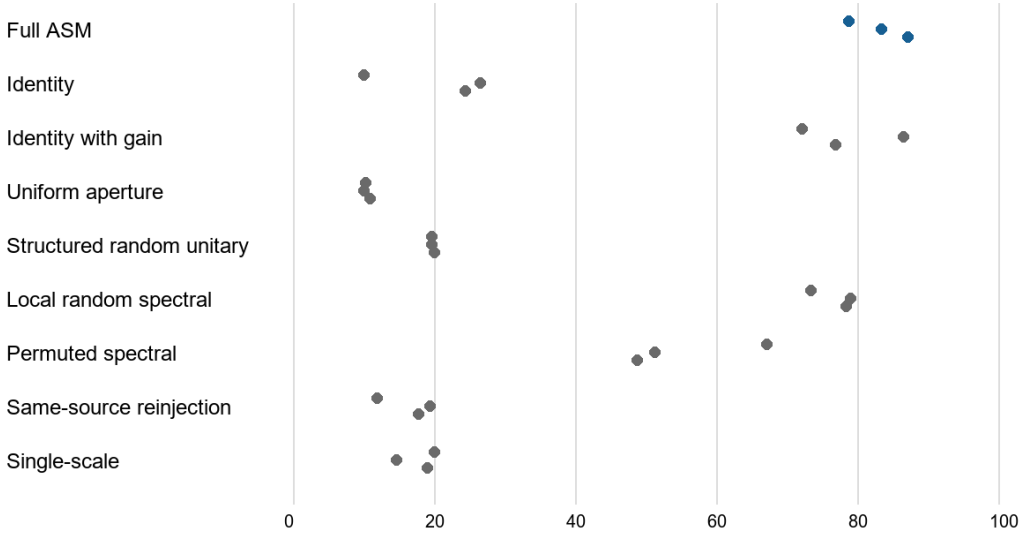

Figure 6: Fresh class accuracy (%). Each dot is one training seed, vertically offset to show overlap. These are descriptive replications, not confidence intervals.

D Aperture selection and objective comparison

m	Fixed class acc. (%)	Fresh class acc. (%)	Fixed matched MSE
4	30.31 ± 3.01	30.10	0.05454
8	69.06 ± 9.63	67.71	0.04301
16	83.85 ± 3.61	82.08	0.03869
32	87.29 ± 1.26	83.12	0.03772
64	86.77 ± 1.18	83.65	0.03733

Table 4: Exploratory aperture sweep on the stochastic multiscale family, three seeds, 6,000 updates. The fixed-bank criterion selected $m = 32$; the fresh bank does not show the same ranking of 32 and 64.

The fixed-bank class-accuracy difference between $m = 32$ and 64 is only about 0.52 percentage points. The selected cap is an operating choice, not an established optimum. The full-generator comparison does not repeat the complete capacity-by-operator sweep; it cannot show that $m = 32$ itself causes the final ASM advantage.

A preceding one-seed comparison on the unrestricted persistent multiscale family changed only the assignment objective. Final-only Hungarian reached 78.44% fixed class accuracy; shared multiscale Hungarian reached 88.13%. Matched MSE changed from 0.03884 to 0.03666. Pairwise MSE decreased from 0.01699 to 0.01546, and correct-only pairwise MSE from 0.01472 to 0.01337. This is evidence of improved recognition under that objective, not increased measured diversity. Both runs already used the multiscale topology. Since each scale is supervised toward its own target, the change could favor early target formation; no stagewise specialization metric was collected.

E Earlier aperture-capacity evidence

These experiments use an earlier class-conditioned image-fitting implementation and its two-epoch budget. They do not use the final stochastic multiscale generator. Let $M_A^{(s)}$ and $M_B^{(s)}$ be ASM and retrained-bypass MSE at seed s, and M_μ the evaluation-only validation class-mean reference. The reported reduction is

$$\frac{1}{3}\sum_s \frac{M_B^{(s)} - M_A^{(s)}}{M_B^{(s)} - M_\mu}. \tag{8}$$

It differs from computing a ratio after averaging errors. For this table $M_\mu \simeq 0.043238$; a training-fit class-mean reference evaluated on validation has MSE 0.043317. The validation-fit reference is used only for error decomposition, not as a training target or a deployable oracle.

m	ASM total MSE	Retrained bypass total MSE
4	0.059284 ± 0.000030	0.060459 ± 0.000065
8	0.048798 ± 0.000117	0.050693 ± 0.000083
16	0.045350 ± 0.000063	0.046295 ± 0.000189
32	0.044667 ± 0.000084	0.044898 ± 0.000097
64	0.044454 ± 0.000111	0.044485 ± 0.000065

Table 5: Older capacity study, mean $\pm$ sample SD across three seeds. Smaller apertures increase some relative propagation benefits while worsening absolute image-fitting error.

A separate matched $m = 8$ rerun obtains MSE 0.048808 for ASM, 0.050758 for bypass, and 0.046955 for PermutedSpectral. Thus the earlier image-fitting task also contains a nonphysical operator stronger than ASM. Neither these errors nor the capacity percentages should be merged numerically with the stochastic-generator Hungarian errors.

F Earlier strict-locality communication experiment

The source contains a 4×4 array of four possible complex phase symbols at rows 2–5 and columns 2–5. Their RGB-coded predictions are requested at rows 18–21 and columns 18–21. A separate strict-locality model has 2,327 trainable parameters and four repeated evolution steps. Its Identity path has local radius four, insufficient to connect the two regions. The task uses 4,096 training patterns and 1,024 test patterns, a batch of 64, and 5,000 updates for three seeds. These are synthetic symbols, not class-conditioned generated images.

The original ASM setting reaches approximately 34.4% symbol accuracy; the later distance-20 setting reaches 54.94%, 55.44%, and 55.03% (mean 55.14%). Distance 22 is slightly worse in all three paired seeds. PermutedSpectral reaches 87.8–88.0%, and the spatial random-unitary comparison reaches 85.0–86.6%. The latter uses $\mathcal{F}^{-1}D_f\mathcal{F}$ with random unit spectral phases (seed 2026091605); it is not the final generator's $\mathcal{F}^{-1}D_f\mathcal{F}D_x$ replacement. The abstract's rounded values refer to these settings. We treat these results as exploratory evidence about a mechanism, and they establish no general physical law.

G Stage-by-stage outputs

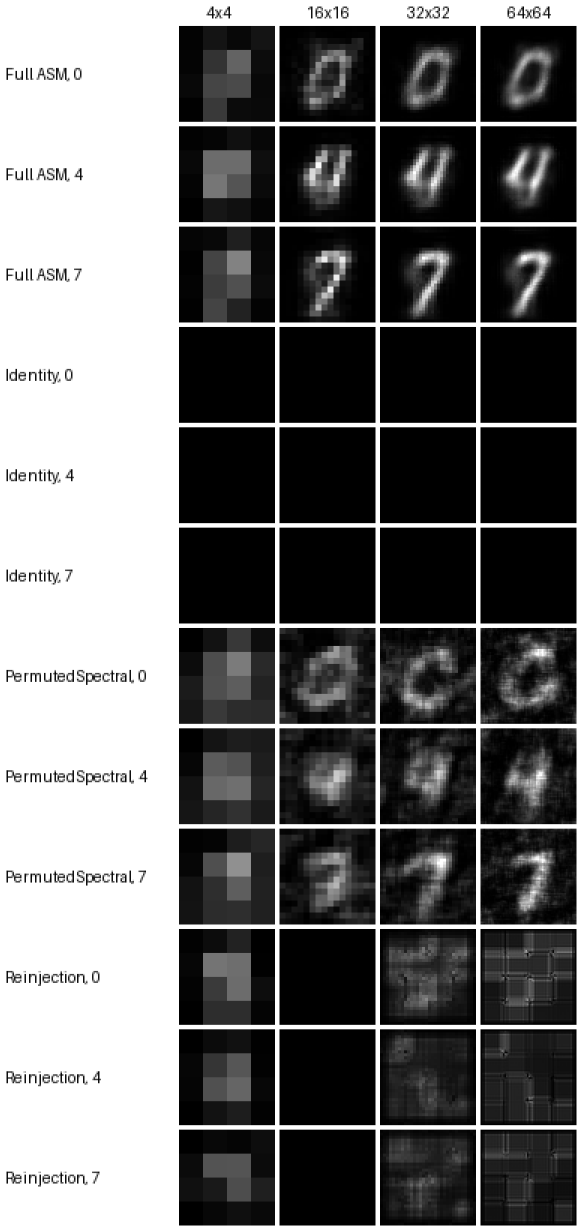

Figure 7: Intensity after each stage for the full model and three retrained controls, training seed 17, fresh sources. Each row is the first draw of the requested class; nothing is selected. Stages are enlarged to 64×64 by nearest-neighbor resizing for display, with a fixed $[0, 1]$ range.

Identity is black from the first stage onward, matching the amplitude collapse described in Section 5.1; Identity with gain does not collapse (Table 1). PermutedSpectral forms digit-like layouts with speckle at every stage. Reinjection is black at 16×16 and ends in a lattice-like pattern. These are three samples per model from one seed; they illustrate the quantitative results and do not measure refinement.

H Longer MNIST training

The supplementary README lists the command-line settings for these runs. One epoch is 1,719 class-balanced updates of batch 32. Evaluation seeds are 3017 (fixed) and 4017 (fresh). The official MNIST test split was not read.

Epochs	Fixed (%)	Fresh (%)	Fixed MSE	Fresh MSE
5	89.06	84.38	0.03671	0.03813
10	96.25	95.31	0.03335	0.03487
15	95.94	94.06	0.03247	0.03345
20	96.88	95.63	0.03244	0.03384
25	97.19	96.88	0.03239	0.03303
30	97.19	95.94	0.03187	0.03228

Table 6: All evaluations of the wider 30-epoch ASM run.

	Identity			PermutedSpectral		
Epochs	Fixed (%)	Fresh (%)	Fresh MSE	Fixed (%)	Fresh (%)	Fresh MSE
5	59.69	63.75	0.04247	63.75	67.50	0.04698
10	85.31	86.88	0.03975	87.81	87.50	0.04193
15	88.13	85.31	0.04027	93.75	92.81	0.04087
20	84.06	86.56	0.04150	90.94	93.44	0.03975
25	89.06	90.94	0.03883	95.63	95.00	0.03927
30	92.19	90.31	0.03826	93.13	94.06	0.03769

Table 7: Identity and PermutedSpectral retrained at the wider 30-epoch budget, seed 17.

I Propagation-distance sweep

The full model was retrained with only the ASM distance d changed, at width 96, seed 17, and 6,000 updates. Other settings match Table 1.

d	Fixed class acc. (%)	Fresh class acc. (%)	Fresh matched MSE
0	10.00	10.00	0.10115
1	55.94	60.63	0.04346
2	66.88	66.56	0.04135
4	80.31	81.25	0.03913
8	85.94	78.75	0.03913
16	84.69	86.56	0.03782
32	77.81	79.69	0.03852

Table 8: Propagation-distance sweep, one seed. $d = 8$ is the paper setting.

J CIFAR-10 TRANSFER SAMPLES

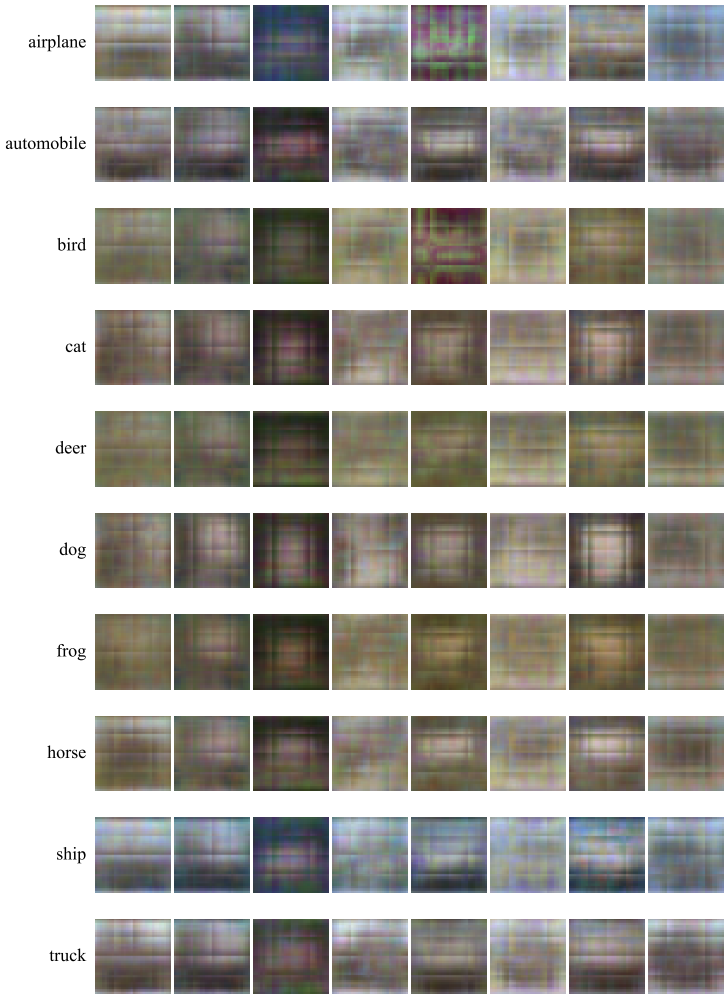

Figure 8: CIFAR-10 transfer after 30 million image presentations, fresh sources, one seed. Each row is a requested class; columns are the first eight saved samples, without selection or filtering, at the native 32×32 resolution.