LEARNING PHOTONIC DYNAMICS FOR HIGHER-ORDER ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Photonic hardware is usually used to accelerate fixed neural operators. We ask whether a photonic feedback loop can instead be trained as part of the neural model. We study associative memory—recovering a stored pattern from a corrupted cue—with recurrent photonic dynamics. Programmable optical projections and nonlinear interactions repeatedly update an internal state. Higher-order interactions reduce interference among stored memories, and a readout based on the stored memories restores information lost by normalization. At the highest tested load of 128 stored memories, a model trained for four feedback steps reconstructs 5.3 dB better at four steps than after running to convergence. Thus, more convergence can make the task result worse: runtime itself becomes part of what the model learns. Controlled retrieval shows when higher-order interactions help, while a real multi-hop retrieval test shows they can also hurt generalization. Finally, we map the model to a hybrid photonic/electronic circuit and simulate finite precision, loss, drift, and calibration. These hardware results are simulations, not measured device performance.

1 INTRODUCTION

Photonic neural hardware often accelerates familiar feed-forward operators: matrix multiplication, convolution, recurrent transforms, or diffractive propagation (Shen et al., 2017; Tait et al., 2017; Lin et al., 2018; Feldmann et al., 2021; Xu et al., 2021; Shastri et al., 2021). Feedback offers another role: repeated optical transforms can make the *trajectory itself* the computation of a neural-network layer. This suits energy-based memory and optimization, but leaves a key choice: train hardware for a useful finite-runtime output, or treat it as an approximate solver whose semantics are defined at equilibrium.

Higher-order associative memories make this question concrete. Increasing interaction order can suppress interference among competing memories and reshape storage behavior (Hopfield, 1982; Krotov & Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2021), while photonic Ising and nonlinear-optical systems can represent and relax nontrivial energy functions, including higher-order interactions (Inagaki et al., 2016; McMahon et al., 2016; Pierangeli et al., 2019; Bybee et al., 2023; Nikhar et al., 2024; Al-Kayed et al., 2025; Musa et al., 2026). Yet a useful neural layer requires more than a favorable energy minimum: its transient basin geometry must be trainable, its readout must preserve recovered information, its backward pass must differentiate the computation actually executed, and its circuit must tolerate finite precision and drift.

We call this recurrent computation a *photonic energy layer* (Fig. 1): programmable optical projections form low-dimensional interaction coordinates, pairwise or higher-order products shape an energy landscape, feedback updates the latent state, and the surrounding model reads either a finite trajectory or an explicitly intended fixed point. Figure 1 separates this algorithmic view (top) from the hybrid optical/electronic path used in our realism study (bottom). We therefore evaluate a specified converter-mediated loop rather than an unspecified all-optical implementation; Sec. 3 gives device-level details.

Associative memory provides the strongest positive result. At high load, higher-order dynamics resist interference that defeats the storage-matched pairwise model, while a memory-conditioned radial readout restores latent scale without changing identity. More importantly, task-optimal computation need not be converged: a short-runtime-trained model reconstructs substantially better near its training depth than at its fixed point, even as continued iteration makes the state and implicit gradient converge. This separates *learned finite-runtime dynamics* from *equilibrium semantics*. We then probe the

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.

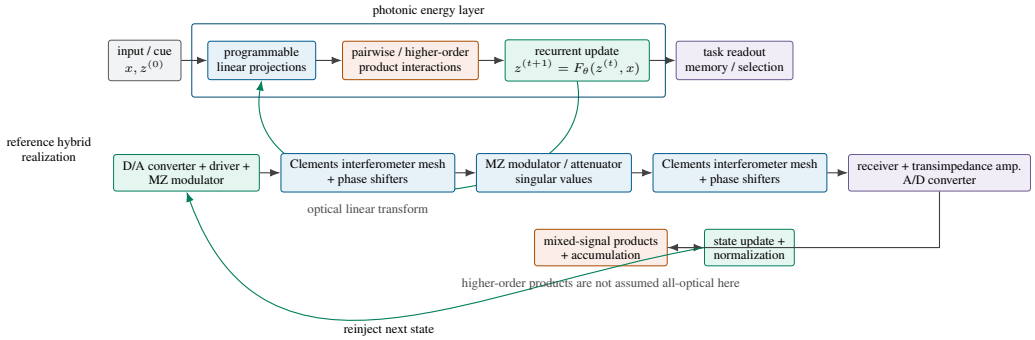

Figure 1: **Learned photonic dynamics and reference hybrid realization.** Top: programmable projections and higher-order interactions define a recurrent energy-layer computation whose finite trajectory or fixed point is read out by the surrounding model. Bottom: the physical-sensitivity study uses digital-to-analog conversion and optical modulation, Clements interferometer meshes, a modulator/attenuator singular-value bank, optical reception and analog-to-digital conversion, mixed-signal product accumulation, and recurrent state update. Hardware abbreviations are expanded in Sec. 3 and Appendix B. Fully optical product channels are not assumed.

boundary: a controlled selection task confirms irreducible higher-order expressivity, but a HotpotQA residual (Yang et al., 2018) overfits useful pretrained geometry. Finally, loop-level models expose sensitivity to repeated conversion precision, attenuation, drift, and fixed-point conditioning.

Our contributions are:

- **A differentiable formulation of learned photonic dynamics and explicit hybrid circuit mapping** that combines factorized higher-order interactions, recurrent state evolution, finite-runtime or fixed-point semantics, and continuous task readout.
- **A mechanism study of higher-order associative memory** that separates interference, trainable basin geometry, and reconstruction scale, including diagnostics in pretrained image latents and practical photos.
- **Task-facing evidence that physical runtime is part of the learned computation**: task quality under finite recurrence can peak before equilibrium even while forward states and gradients converge numerically to the fixed-point computation.
- **Boundaries on the claim**: matched training-depth controls, a negative real-data retrieval result, and modeled photonic nonidealities show where extra recurrence or higher-order structure helps, where it does not, and which conclusions still require hardware validation.

2 Related Work

Associative memory and energy models. Classical and dense/modern Hopfield memories use energy descent and nonlinear separation to increase capacity (Hopfield, 1982; Krotov & Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2021; Krotov & Hopfield, 2021; Millidge et al., 2022); recent work also studies useful transient retrieval outside equilibrium attractors (Clark, 2026). We study a trainable physical layer, jointly evaluating basin learning, continuous readout, finite-versus-equilibrium task/gradient semantics, and realization nonidealities.

Implicit and equilibrium layers. Deep equilibrium (DEQ) models and other optimization/fixed-point layers can be differentiated through optimality conditions rather than by storing every forward state (Amos & Kolter, 2017; Bai et al., 2019; Winston & Kolter, 2020; Blondel et al., 2022). Equilibrium propagation provides a physically motivated energy-based learning view (Scellier & Bengio, 2017), while Jacobian-free methods trade gradient exactness for lower backward cost (Fung et al., 2022). Our focus is the distinction between the derivative of the *finite computation actually*

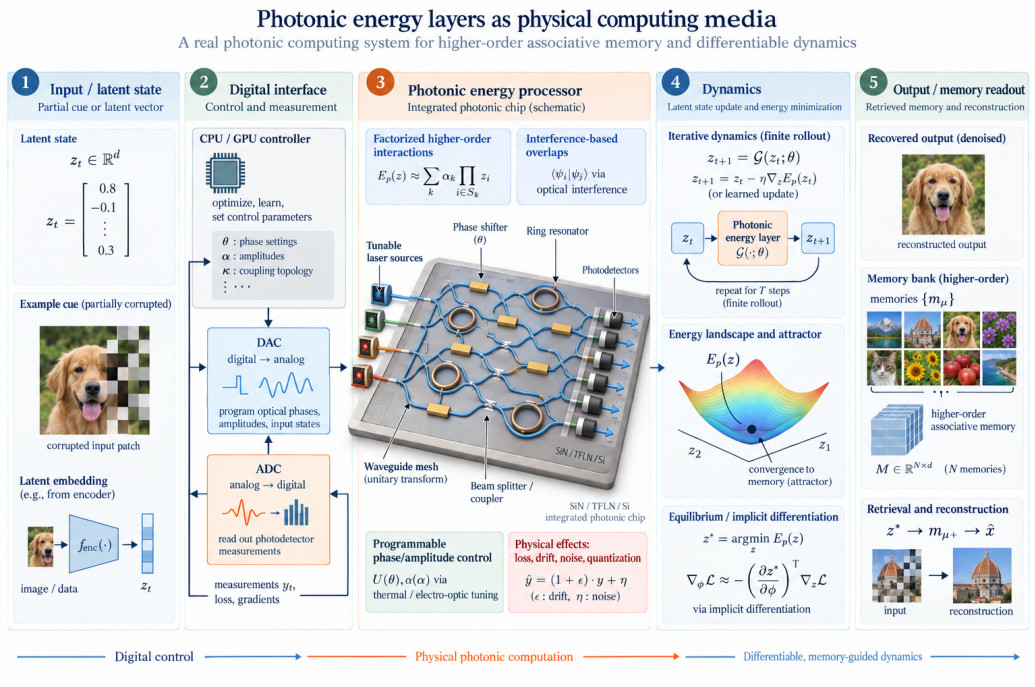

Figure 2: **System-level intuition for learned photonic dynamics.** A corrupted input becomes a latent state; a digital controller programs and observes the photonic processor; recurrent dynamics traverse a learned energy landscape; and memory-guided readout reconstructs the output. This conceptual visualization is not a fabricated-chip claim; Fig. 1 (bottom) gives the converter-mediated realization, and Sec. 3 expands its hardware abbreviations.

executed and the derivative of a *converged equilibrium*; the two can differ substantially under a strict physical iteration budget.

Photonic and physical neural computing. Photonic neural networks span interferometric, recurrent, diffractive, and tensor architectures, including systems trained in situ (Shen et al., 2017; Tait et al., 2017; Hughes et al., 2018; Lin et al., 2018; Wetzstein et al., 2020; Shastri et al., 2021; Wright et al., 2022; Pai et al., 2023). Relaxation-based systems include Ising/Hopfield machines and optical two-/four-body dense associative memory (Inagaki et al., 2016; McMahon et al., 2016; Pierangeli et al., 2019; Al-Kayed et al., 2025; Musa et al., 2026). We do not claim priority for higher-order photonic memory; we study it as a differentiable learned layer with explicit runtime, gradient, readout, and realization semantics.

Retrieval and multi-hop evidence selection. Dense retrieval and retrieval-augmented generation use semantic embeddings and external memories (Karpukhin et al., 2020; Lewis et al., 2020; Reimers & Gurevych, 2019). Multi-hop methods condition later retrieval on earlier evidence or reasoning state (Xiong et al., 2021; Trivedi et al., 2023), while Self-RAG learns adaptive retrieval and self-critique (Asai et al., 2024). We use evidence selection only to stress-test higher-order compatibility: the HotpotQA benchmark (Yang et al., 2018) shows that controlled higher-order expressivity need not transfer to pretrained semantic representations. Appendix A gives a detailed comparison.

3 Learning Photonic Dynamics

We model learned photonic dynamics as a differentiable stateful map, a *photonic energy layer*, whose recurrent relaxation admits photonic or hybrid realization. For state $z \in \mathbb{R}^N$, optional conditioning x,

and trainable parameters θ, we use factorized polynomial interactions,

$$E_\theta(z;x) = R(z;x) - \sum_{r=1}^{R_f} c_r \prod_{j=1}^{p_r} a_{rj}(z,x), \tag{1}$$

$$a_{rj}(z,x) = \langle w_{rj}, u_j(z,x)\rangle, \tag{2}$$

where R_f counts factors; c_r, p_r, and w_{rj} are the factor weight, order, and projection; $u_j(z,x)$ is its state/conditioning feature; and R stabilizes the state. A dense homogeneous order-p tensor can require $\mathcal{O}(N^p)$ coefficients, while the rank-R_f factorization uses $\mathcal{O}(R_f pN)$ projection parameters. For the common case $u_j = z$, the interaction field is evaluated directly from factor projections,

$$h_{\mathrm{int}}(z) = \sum_r c_r \sum_{k=1}^{p_r} w_{rk} \prod_{j\neq k} a_{rj},$$

without materializing a dense N^p tensor. This reduces storage and avoids dense-tensor evaluation, but not the physical arithmetic cost of increasing interaction order (Appendix C).

Reference hybrid photonic realization. Figure 1 (bottom) makes the physical mapping explicit. A learned projection $W = U\Sigma V^\dagger$ is compiled by singular-value decomposition (SVD): U and V are Clements meshes of Mach–Zehnder interferometers (MZIs) (Clements et al., 2016), with phase shifters programming their internal/external phases, while Σ is implemented by Mach–Zehnder modulators (MZMs) or variable optical attenuators (VOAs). Gain beyond passive transmission remains explicit and electronic. Each recurrent pass traverses an optical receiver with a transimpedance amplifier (TIA) and analog-to-digital converter (ADC); higher-order scalar products, accumulation, state update, and normalization are mixed-signal/electronic; the next state returns through a digital-to-analog converter (DAC), driver, and modulator. Thus our precision/drift study models a converter-mediated hybrid loop. Fully optical product channels are possible but not benchmarked, and we do not claim a foundry-qualified signed-field receiver.

One recurrent step is

$$z^{(t+1)} = F_\theta(z^{(t)},x) = \mathcal{N}\Big((1-\alpha)z^{(t)} + \alpha h_\theta(z^{(t)};x)\Big), \qquad h_\theta = -\nabla_z E_\theta, \tag{3}$$

with relaxation α and task-specific normalization $\mathcal{N}$. In pretrained-latent memory, $\mathcal{N}$ uses spherical normalization with a root-mean-square (RMS) radius convention, so recurrence chiefly determines latent direction and identity. The executed output is the finite state $z^{(T)}$; the fixed point $z^\star = F_\theta(z^\star,x)$ is a distinct computational limit.

For image reconstruction, stored memories—not the corrupted cue—restore latent radius,

$$\hat{z} = \left(\sum_\mu \pi_\mu(z^{(T)}) r_\mu\right) z^{(T)}, \qquad \pi_\mu \propto \exp\{\tau\,\mathrm{sim}(z^{(T)}, m_\mu)\}. \tag{4}$$

Here μ indexes memories; m_μ and r_μ are the stored direction and RMS radius; τ is the softmax temperature; and sim is cosine similarity, so π_μ forms a soft memory selection. This differentiable readout changes only the decoded magnitude, preserving terminal direction and retrieval identity. Finite unrolling differentiates the executed computation; fixed-point implicit differentiation applies only when equilibrium is the intended forward semantics (Sec. 5). Appendices B and C give the system mapping, factorization accounting, readout details, and quadratization control.

4 Higher-Order Associative Memory

For P stored bipolar memories $\{\xi_\mu\}_{\mu=1}^P$ with $\xi_\mu \in \{-1,+1\}^N$, the order-p energy and local field are

$$E_p(z) = -\frac{N}{p}\sum_{\mu=1}^{P}\left(\frac{\xi_\mu^\top z}{N}\right)^p, \qquad h_p(z) = \sum_{\mu=1}^{P}\left(\frac{\xi_\mu^\top z}{N}\right)^{p-1}\xi_\mu. \tag{5}$$

The bank stores PN factors per order, avoiding dense order-p storage. Storage is matched, arithmetic is not: higher order adds scalar powers/products (Appendix J). Unless noted, image memory uses

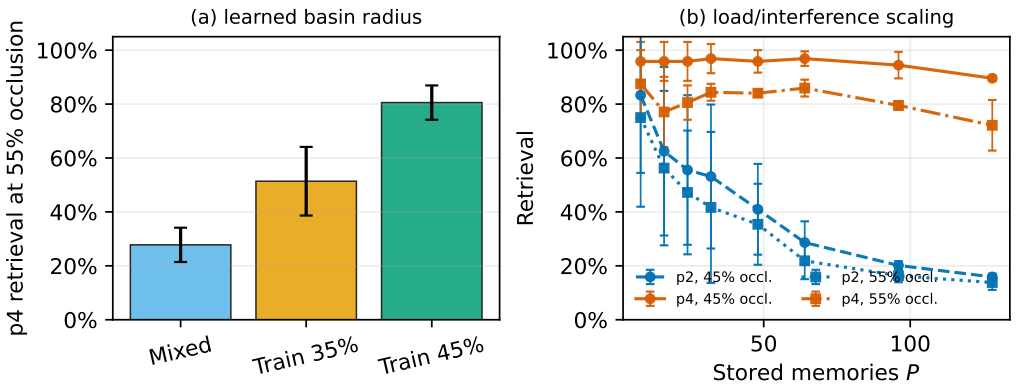

Figure 3: **Learnable basin radius and memory-load interference.** Harder-corruption training expands the fourth-order basin; increasing memory load perturbs terminal direction before identity fails. Pairwise dynamics degrade earlier under the same factorized-storage scaling. Error bars show one sample standard deviation over three seeds.

128×128 CIFAR-10 test images (Krizhevsky, 2009), a frozen Stable Diffusion variational autoencoder (SD-VAE) (Rombach et al., 2022), and stored-identity recall from corrupted cues; only the landscape is trained. Higher order improves finite-size load scaling, yet at severe corruption the fourth-order model retrieves 95.8% while the sixth-order model retrieves 0.0%; load tolerance need not imply wider basins. We report peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) (Wang et al., 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). Appendix E gives the image protocol; Appendix D gives the capacity/basin sweep.

Question	**Pretrained-latent result**
Learnable basin	$P = 24$, 55% occlusion: fourth-order retrieval 27.8% $\rightarrow$ **80.6%** when training moves from the mixed-corruption schedule to fixed-45% corruption.
High-load interference	$P = 128$, 45% occlusion: pairwise retrieval 15.9%; fourth-order retrieval 89.6%.
Radial reconstruction	Soft memory-radius calibration leaves identity unchanged and improves mean PSNR by **+2.34 dB** at $P = 48$ and **+1.75 dB** at $P = 128$.

Table 1: **Three separable controls in pretrained image-latent memory.** Order controls interference, corruption-aware training moves the basin, and radial readout restores scale. Main entries are three-seed means where repeated seeds exist; Appendix K reports uncertainty.

Figure 3 visualizes the first two effects. Harder-corruption training moves the fourth-order basin outward, while increasing memory load perturbs terminal direction before destroying identity. A simple independent-distractor analysis predicts less crosstalk at higher order and the classical $P = \mathcal{O}(N^{p-1})$ capacity trend under idealized assumptions. The same mechanism amplifies whichever memory already has the larger overlap, explaining why higher order can improve load tolerance yet sharpen failures after severe corruption; Appendix C gives the equations and assumptions. The third effect is separate: spherical dynamics can recover the correct memory direction but the wrong latent magnitude, so Eq. (4) restores radius without changing identity. At lighter-load Microsoft Common Objects in Context (COCO) (Lin et al., 2014) resolution scaling, matched 128/256/512-pixel images retain 100% identity for both orders at $T = 2$, while $p = 4$ gains 3.59–4.16 dB PSNR; Appendix F gives the protocol and latent dimensions.

5 FINITE-DEPTH AND FIXED-POINT SEMANTICS

For finite T, $z^{(T)} = F_\theta^{(T)}(z^{(0)}, x)$ and unrolled backpropagation is the exact derivative of the executed computation. If equilibrium $z^\star = F_\theta(z^\star, x)$ is the intended output, the implicit adjoint is

$$(I - J_z F_\theta)^\top \lambda = \nabla_z L, \qquad \frac{dL}{d\theta} = \partial_\theta L + \lambda^\top J_\theta F_\theta. \tag{6}$$

Here L is the task loss, I is the identity, $J_z F_\theta$ and $J_\theta F_\theta$ are Jacobians with respect to state and parameters, and λ is the adjoint variable. This covers normalized spherical memory directly; the Hessian form is only the gradient-step special case (Appendix H).

	$P = 48$		$P = 128$	
	$T = 4$	fixed	$T = 4$	fixed
Retrieval (%)	**95.8**	81.3	**89.6**	32.3
PSNR (dB)	**21.59**	20.27	**18.44**	13.13
SSIM ↑	**0.841**	0.787	**0.718**	0.524
LPIPS ↓	**0.144**	0.194	**0.293**	0.526

Table 2: **Finite trained depth versus fixed point for fourth-order image memory.** Three-seed means. Both landscapes were trained with four recurrent steps; fixed-point evaluation is outside the training-depth regime. Standard deviations are reported in Appendix K.

The trained landscape is therefore not merely an approximate deep-equilibrium (DEQ) solve (Bai et al., 2019). We evaluate $T = 1, \ldots, 8$ densely, then $T = 16, 32, 64, 128, 256, 512$, and a separately converged fixed point. At $P = 128$, 45% occlusion, the fourth-order task optimum is at the training depth $T = 4$ (Table 2); the dense $T = 1 \ldots 8$ neighborhood rules out the concern that $T = 4$ appeared optimal only because shorter budgets were absent. At $P = 48$, $T = 4$ and $T = 5$ form an essentially flat PSNR peak. Figure 4 shows the full compute-budget trajectory on a logarithmic iteration axis.

Near a contractive fixed point, finite-depth sensitivities approach the implicit fixed-point sensitivity geometrically with additional iterations. This explains an otherwise counterintuitive result: states and gradients can converge numerically while task quality worsens, because finite-depth training optimizes the executed state $z^{(T)}$, not necessarily the eventual fixed point $z^\star$. Appendix H gives the derivation, error bound, and conditioning controls.

On the $P = 128$, $p = 4$ gradient subset, mean finite-versus-implicit memory-gradient error falls from 1.20 at $T = 4$ to 5.36×10^{-3} at $T = 128$ and 9.69×10^{-7} at $T = 512$; state RMSE to the fixed point simultaneously falls from 0.465 to 2.97×10^{-4} and 1.30×10^{-7}. Thus numerical convergence to the fixed-point forward/backward computation does not imply task equivalence.

P	T_{train}	best T_{infer}	best PSNR	PSNR@T_{train}
48	2	3	21.08	21.06
48	4	4	**21.59**	**21.59**
48	8	7	19.03	19.03
128	2	2	**19.74**	**19.74**
128	4	4	18.44	18.44
128	8	6	14.47	14.28

Table 3: **Matched recurrent training-depth control** ($p = 4$). Three-seed mean PSNR. Increasing T_{train} shifts the preferred execution depth later (48: $3 \to 4 \to 7$; 128: $2 \to 4 \to 6$), but not by an exact one-to-one rule; deep recurrence also becomes difficult at high load. In five of six cells, T_{train} lies within 0.05 dB of the optimum. Standard deviations are in Appendix K.

The matched-depth control supports *depth-dependent learned recurrent semantics*, not an intrinsic disadvantage of equilibrium models. The $T_{\text{train}} = 8$, $P = 128$ condition refutes a simple "deeper is better" rule: it peaks at $T = 6$ and reconstructs substantially worse than shallower training. Appendix H gives full curves and retrieval metrics.

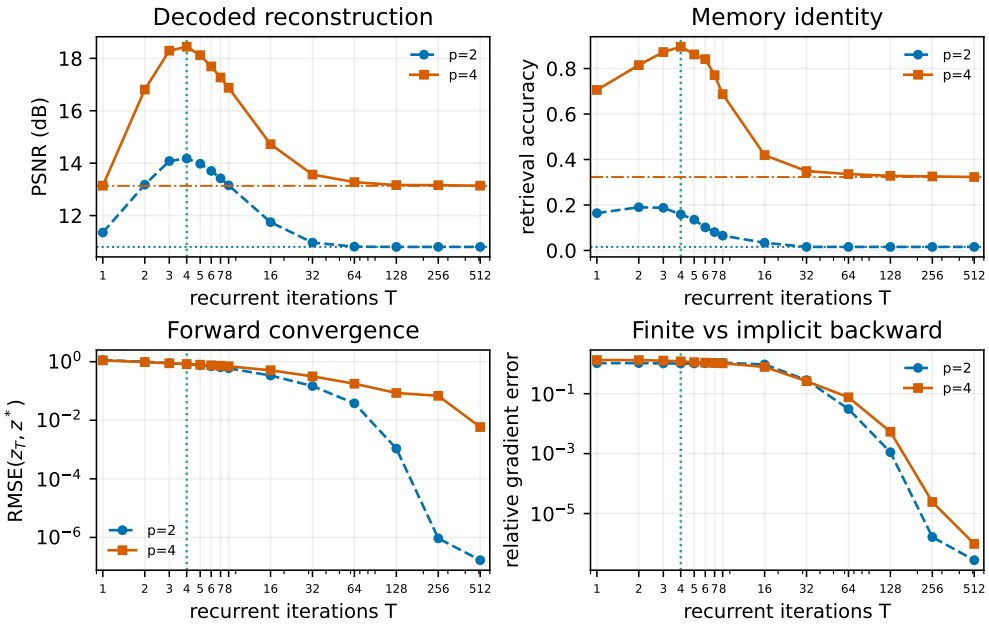

Figure 4: **Finite compute budget versus the fixed-point/DEQ limit on the trained $P = 128$ image-memory landscape.** The log-T sweep includes every measured early point $T = 1, \ldots, 8$ and all later powers of two through 512. Top: decoded reconstruction and retrieval; dashed horizontal lines are independently converged fixed-point values. Bottom: state and gradient convergence to the fixed-point computation. The $p = 4$ layer is trained at $T = 4$: task quality peaks near that finite depth even while forward and backward discrepancies continue to shrink toward the DEQ limit.

Further controls delimit the claim. At $P = 128$, cosine nearest-neighbor lookup reaches 22.9% identity retrieval and 16.13 dB PSNR, versus 98.2% and 19.74 dB for $T_{\text{train}} = 2$ recurrence. A separately equilibrium-trained model reaches 13.51 dB after the same 120 optimizer updates, though the budgets are not matched in floating-point operations (FLOPs) or wall-clock time (Appendix I).

Practical-photo check. The same readout decomposition appears on eight curated Creative Commons Zero (CC0)-licensed photo cases: the original fourth-order cue-RMS readout gives 7/8 correct identities and 19.19 dB mean PSNR, soft memory-conditioned RMS preserves 7/8 while improving to 24.13 dB, and trajectory regularization reaches 8/8 and 24.65 dB. Full SSIM/LPIPS and per-case results are in Appendix E.3; these remain qualitative diagnostics rather than benchmark estimates.

6 Structured Retrieval as a Boundary Test

Higher-order interactions help only when representations expose joint structure. In a controlled task, three- and four-way compatibility has uninformative lower-order marginals; factorized scoring matches explicit joint enumeration at about 99% success, while lower-order scores cannot represent the target. This benefit does not automatically transfer to pretrained language retrieval.

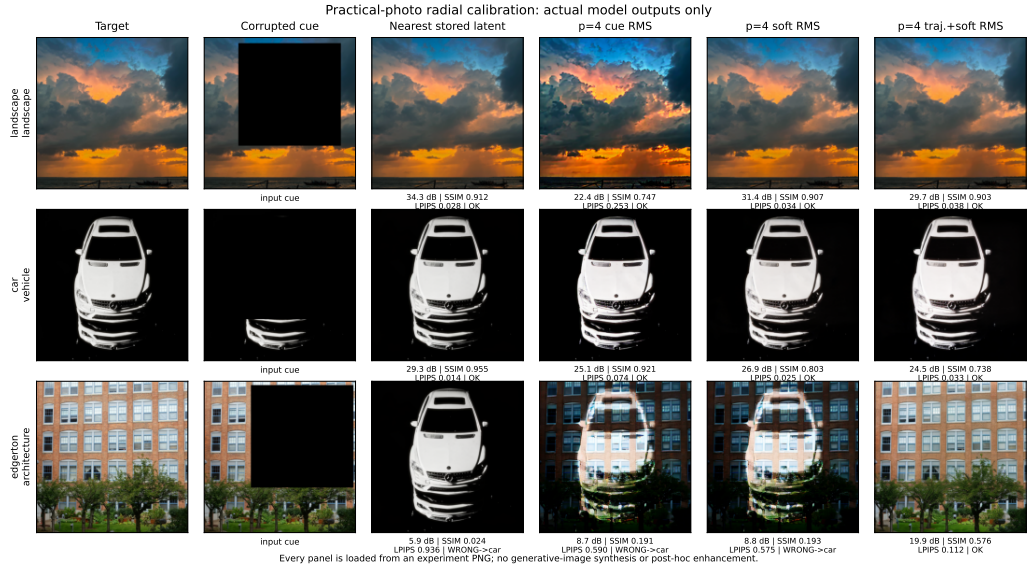

Figure 5: **Illustrative practical-photo reconstructions from retained model outputs.** Three cases show the target, 45% occluded cue, nearest stored latent, original fourth-order cue-RMS readout, soft memory-conditioned radial readout, and trajectory-regularized soft readout. Radial calibration changes continuous scale without changing the terminal direction; the Edgerton House photo case (CC0; Appendix E.1) also illustrates an identity error repaired by trajectory regularization. Every panel is loaded from a saved experiment output; no image synthesis or post hoc enhancement is used.

HotpotQA selector	Joint	Support recall	Oracle 2nd hop
Frozen cosine	32.42	61.91	58.01
Pairwise residual	**36.78**	**64.26**	**58.50**
Third-order residual	28.84	55.89	50.10

Table 4: **Real-data boundary on HotpotQA** (percent; three-seed means). "Joint" requires both gold support documents; support recall measures recovery of gold supports; "oracle 2nd hop" evaluates second-hop selection when the first gold support is supplied. A bounded pairwise residual improves the frozen retriever, whereas the parameter-matched third-order residual generalizes worse despite stronger training fit. Appendix G gives the protocol and Appendix K reports standard deviations.

We therefore use HotpotQA as a scope boundary rather than a target application: extra interaction order is not itself a source of generalization and can damage an already useful pretrained geometry. Appendix G gives the controlled construction, residual formulation, and failure diagnostics.

7 Photonic Realism and Realization

We stress executed feedback dynamics, not an ideal energy minimum. We simulate the converter-mediated loop defined in Fig. 1 and Sec. 3: optical projection is interleaved with conversion, mixed-signal nonlinear products, state update, and reinjection. Table 5 reports state root-mean-square error (RMSE); these are modeled sensitivities, not measured hardware performance.

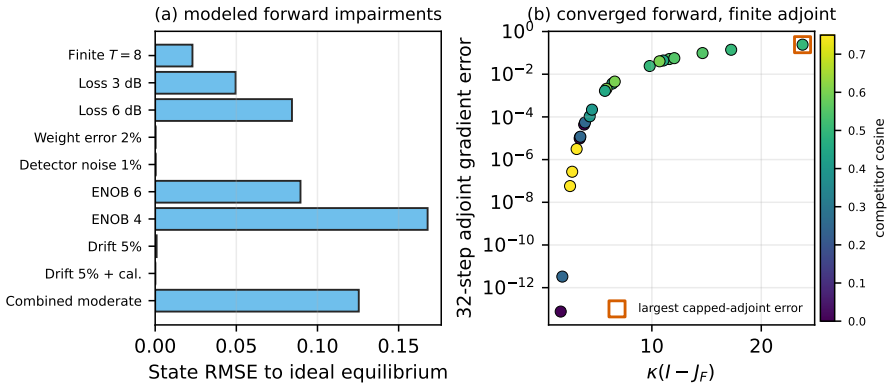

Figure 6: **Modeled photonic-realism and fixed-point conditioning stress.** Left: state RMSE under finite budget, loss/error/noise, repeated converter quantization (effective number of bits, ENOB), drift, and calibration. Right: forward residuals are $< 10^{-10}$, but a capped 32-step adjoint becomes inaccurate as the fixed-point system becomes ill-conditioned; the highlighted case reaches 24.2% gradient error. Simulation/model diagnostics only.

Stress	**Observed effect**
Finite budget	8 versus 64 updates: **0.0230** state RMSE before added device perturbations.
Repeated quantization	6-bit: 0.0895 RMSE; 4-bit: **0.1679** RMSE and a dead-zone/stall regime.
Slow drift + calibration	5% drift: 8.89×10^{-4} RMSE; 90% compensation reduces it to 6.76×10^{-5} (**92.4%** mean removal).

Table 5: **Modeled feedback-loop sensitivities.** Three-seed means; Appendix K reports standard deviations. Repeated precision error accumulates through recurrence, while slow drift is largely calibratable in this model.

Iteration count is part of the computation: repeated precision errors perturb each update, while slow drift is partly calibratable. Fixed-point gradients also require an adequate adjoint solve. Figure 6 shows these diagnostics; Appendix J gives the perturbation equations and realization inventory. We infer no chip area, latency, throughput, or energy from structural counts.

8 Discussion and Limitations

Interaction order controls interference, basin training controls corruption tolerance, radial calibration restores scale, and recurrent depth selects the computation. These axes are distinct: higher order can raise capacity without widening basins, while extra iterations can improve convergence but hurt the task. Runtime is also a resource choice because larger T requires more optical/conversion passes and accumulates latency and error. Factorization avoids dense N^p storage, not higher-order arithmetic or memory access. HotpotQA and severe-corruption failures show that extra order is not universally beneficial. Hardware evidence is simulated, and the image/equilibrium studies remain limited in backbone, seeds, task scope, and training recipe; Appendix N gives full boundaries.

9 Conclusion

We learn photonic dynamics for higher-order associative memory. Higher order reduces interference, basin training improves corruption tolerance, and memory-conditioned readout restores scale. Finite-depth models can approach the DEQ limit while task quality worsens, making runtime part of the model. Retrieval and realism bound the claim; hardware validation remains.

AI USE STATEMENT

In this work, we used generative AI tools (OpenAI ChatGPT, including code-execution and web-search capabilities) for conceptual brainstorming and hypothesis refinement; experimental methodology design; implementation, debugging, and test development; data reformatting and interpretation; literature search and synthesis; drafting and editing; reference-formatting assistance; and planning or generation of manuscript figures and tables. Generative image tools were used only for conceptual illustrations, i.e., Figs. 2 and 7; experimental image panels and quantitative figures were generated from retained data/model outputs. AI-assisted code was executed and tested, quantitative manuscript claims were checked against retained result artifacts, and bibliographic metadata was cross-checked against primary publisher or proceedings sources. We did not use generative AI to fabricate experimental measurements or to present simulated results as measured hardware evidence. The authors reviewed all AI-assisted work and take responsibility for the final manuscript, analyses, figures, and claims.

ETHICS STATEMENT

This work uses public benchmark data and simulation/modeling; it does not involve human-subject experimentation or personal-data collection. The primary research-integrity risk is over-interpreting modeled photonic resource or nonideality results as measured hardware performance. We therefore distinguish modeled/simulated evidence from measured hardware evidence, retain negative results, and make no measured speed, energy, area, or throughput claim.

REFERENCES

Nayem Al-Kayed, Charles St-Arnault, Hugh Morison, A. Aadhi, Chaoran Huang, Alexander N. Tait, David V. Plant, and Bhavin J. Shastri. Programmable 200 GOPS hopfield-inspired photonic Ising machine. *Nature*, 648:576–584, 2025. doi: 10.1038/s41586-025-09838-7.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 2017.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0378.

Connor Bybee, Denis Kleyko, Dmitri E. Nikonov, Amir Khosrowshahi, Bruno A. Olshausen, and Friedrich T. Sommer. Efficient optimization with higher-order Ising machines. *Nature Communications*, 14:6033, 2023. doi: 10.1038/s41467-023-41214-9.

David G. Clark. Transient dynamics of associative memory models. *Physical Review E*, 113:054301, 2026. doi: 10.1103/42y2-bsh1.

William R. Clements, Peter C. Humphreys, Benjamin J. Metcalf, W. Steven Kolthammer, and Ian A. Walmsley. Optimal design for universal multiport interferometers. *Optica*, 3(12):1460–1465, 2016. doi: 10.1364/OPTICA.3.001460.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017. doi: 10.1007/s10955-017-1806-y.

Johannes Feldmann, Nathan Youngblood, Maxim Karpov, Hannes Gehring, Xuan Li, Marcel Stappers, Manuel Le Gallo, Xue Fu, Anton Lukashchuk, Arslan S. Raja, Junqiu Liu, C. David Wright,

Abu Sebastian, Tobias J. Kippenberg, Wolfram H. P. Pernice, and Harish Bhaskaran. Parallel convolutional processing using an integrated photonic tensor core. *Nature*, 589:52–58, 2021. doi: 10.1038/s41586-020-03070-1.

Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel Mckenzie, Stanley Osher, and Wotao Yin. JFB: Jacobian-free backpropagation for implicit networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6648–6656, 2022. doi: 10.1609/aaai.v36i6.20619.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27532–27559, 2023. doi: 10.52202/075280-1197.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

Tyler W. Hughes, Momchil Minkov, Yu Shi, and Shanhui Fan. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica*, 5(7):864–871, 2018. doi: 10.1364/OPTICA.5.000864.

Takahiro Inagaki, Yoshitaka Haribara, Koji Igarashi, Tomohiro Sonobe, Shuhei Tamate, Toshimori Honjo, Alireza Marandi, Peter L. McMahon, Takeshi Umeki, Koji Enbutsu, Osamu Tadanaga, Hirokazu Takenouchi, Kazuyuki Aihara, Ken-ichi Kawarabayashi, Kyo Inoue, Shoko Utsunomiya, and Hiroki Takesue. A coherent Ising machine for 2000-node optimization problems. *Science*, 354 (6312):603–606, 2016. doi: 10.1126/science.aah4243.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1172–1180, 2016.

Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

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, 2018. doi: 10.1126/science.aat8084.

Peter L. McMahon, Alireza Marandi, Yoshitaka Haribara, Ryan Hamerly, Carsten Langrock, Shuhei Tamate, Takahiro Inagaki, Hiroki Takesue, Shoko Utsunomiya, Kazuyuki Aihara, Robert L. Byer, M. M. Fejer, Hideo Mabuchi, and Yoshihisa Yamamoto. A fully programmable 100-spin coherent Ising machine with all-to-all connections. *Science*, 354(6312):614–617, 2016. doi: 10.1126/science.aah5178.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15561–15583. PMLR, 2022.

Khalid Musa, Santosh Kumar, Michael Katidis, and Yu-Ping Huang. Dense associative memory in a nonlinear-optical hopfield neural network. *Physical Review Applied*, 25(1):014011, 2026. doi: 10.1103/7jh8-2vny.

Srijan Nikhar, Sidharth Kannan, Navid Anjum Aadit, Shuvro Chowdhury, and Kerem Y. Camsari. All-to-all reconfigurability with sparse and higher-order Ising machines. *Nature Communications*, 15:8977, 2024. doi: 10.1038/s41467-024-53270-w.

Sunil Pai, Zhanghao Sun, Tyler W. Hughes, Taewon Park, Ben Bartlett, Ian A. D. Williamson, Momchil Minkov, Maziyar Milanizadeh, Nathnael Abebe, Francesco Morichetti, Andrea Melloni, Shanhui Fan, Olav Solgaard, and David A. B. Miller. Experimentally realized in situ backpropagation for deep learning in photonic neural networks. *Science*, 380(6643):398–404, 2023. doi: 10.1126/science.ade8450.

D. Pierangeli, G. Marcucci, and C. Conti. Large-scale photonic Ising machine by spatial light modulation. *Physical Review Letters*, 122(21):213902, 2019. doi: 10.1103/PhysRevLett.122.213902.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24, 2017. doi: 10.3389/fncom.2017.00024.

Bhavin J. Shastri, Alexander N. Tait, Thomas Ferreira de Lima, Wolfram H. P. Pernice, Harish Bhaskaran, C. David Wright, and Paul R. Prucnal. Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics*, 15:102–114, 2021. doi: 10.1038/s41566-020-00754-y.

Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, and Marin Soljačić. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11:441–446, 2017. doi: 10.1038/nphoton.2017.93.

Alexander N. Tait, Thomas Ferreira de Lima, Ellen Zhou, Allie X. Wu, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal. Neuromorphic photonic networks using silicon photonic weight banks. *Scientific Reports*, 7:7430, 2017. doi: 10.1038/s41598-017-07754-z.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David A. B. Miller, and Demetri Psaltis. Inference in artificial intelligence with deep optics and photonics. *Nature*, 588:39–47, 2020. doi: 10.1038/s41586-020-2973-6.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Logan G. Wright, Tatsuhiro Onodera, Martin M. Stein, Tianyu Wang, Darren T. Schachter, Zoey Hu, and Peter L. McMahon. Deep physical neural networks trained with backpropagation. *Nature*, 601:549–555, 2022. doi: 10.1038/s41586-021-04223-6.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*, 2021.

Xingyuan Xu, Mengxi Tan, Bill Corcoran, Jiayang Wu, Andreas Boes, Thach G. Nguyen, Sai T. Chu, Brent E. Little, Damien G. Hicks, Roberto Morandotti, Arnan Mitchell, and David J. Moss. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature*, 589:44–51, 2021. doi: 10.1038/s41586-020-03063-0.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A Detailed Related Work

A.1 Associative memory, higher-order interactions, and energy-based computation

Hopfield's classical model connects recurrent neural dynamics, Lyapunov energy, and content-addressable memory (Hopfield, 1982). Its pairwise form is capacity-limited by interference among stored patterns. Dense associative memory replaces the quadratic interaction with nonlinear polynomial separation functions, increasing capacity and connecting energy-based memories to ordinary neural networks (Krotov & Hopfield, 2016). Demircigil et al. formalized the storage advantage for polynomial and exponential interactions and showed that very large capacity need not eliminate useful basins (Demircigil et al., 2017). Modern Hopfield networks use continuous states, recover an attention-like update, and obtain exponentially large capacity under suitable assumptions (Ramsauer et al., 2021). Krotov and Hopfield further showed how apparently many-body associative interactions can arise from hidden-variable systems with only two-body microscopic couplings (Krotov & Hopfield, 2021). Universal Hopfield Networks organize this family into similarity, separation, and projection operations (Millidge et al., 2022), while Energy Transformer uses explicitly constructed energies to organize token interactions (Hoover et al., 2023).

Recent work makes the transient/equilibrium distinction especially relevant: Clark analyzes associative-memory models whose useful retrieval can be transient even when the corresponding equilibrium attractor is absent (Clark, 2026). Our use of higher order is narrower than a universal memory or transient-retrieval priority claim. We use a factorized polynomial energy as a differentiable physical-computation primitive and separately measure (i) representation/storage, (ii) attraction-basin learning, (iii) continuous readout, (iv) finite-runtime task behavior, and (v) finite-versus-implicit differentiation. This separation matters because higher-order expressivity does not guarantee favorable basins or solver dynamics: our controlled $p = 6$ memory and real-data structured-retrieval study provide concrete counterexamples.

A.2 Differentiable optimization, equilibrium models, and physical learning

Optimization layers make the solution of a parameterized optimization problem part of a neural network and differentiate it through Karush–Kuhn–Tucker (KKT) or other optimality conditions

(Amos & Kolter, 2017). Deep equilibrium (DEQ) models directly represent the fixed point of a weight-tied transformation and use implicit differentiation to avoid storing an arbitrarily deep computation graph (Bai et al., 2019); monotone operator equilibrium networks impose structure to improve existence and convergence guarantees (Winston & Kolter, 2020). Modular implicit-differentiation frameworks generalize the approach by differentiating user-specified optimality maps (Blondel et al., 2022). Jacobian-free backpropagation trades exactness for a simpler fixed-memory backward approximation (Fung et al., 2022). Equilibrium propagation provides a complementary energy-based physical view in which learning is expressed through nearby equilibria (Scellier & Bengio, 2017).

Our emphasis is the distinction between the *gradient of the executed finite trajectory* and the *gradient of a converged equilibrium*. The two coincide only as the forward solve converges under suitable regularity. This issue is particularly relevant for physical accelerators, where wall-clock or round-trip budgets can make a truncated trajectory the actual computation rather than an implementation detail.

A.3 Photonic neural computing and physical optimization

Programmable photonics has demonstrated coherent interferometric neural inference (Shen et al., 2017), recurrent neuromorphic photonic networks (Tait et al., 2017), diffractive optical inference (Lin et al., 2018), photonic convolution/tensor acceleration (Feldmann et al., 2021; Xu et al., 2021), and broad architectures for optical AI (Wetzstein et al., 2020; Shastri et al., 2021). Training-oriented work includes in situ photonic gradient measurement (Hughes et al., 2018), general physical neural networks trained with backpropagation (Wright et al., 2022), and experimentally realized in situ backpropagation in photonic neural networks (Pai et al., 2023). These systems demonstrate that photonic computation can be programmable and trainable, but they largely target feed-forward transformations or device-specific physical layers.

Optical Ising machines instead use dynamical relaxation to search an energy landscape. Coherent and spatial-light-modulator systems realize programmable pairwise interactions (Inagaki et al., 2016; McMahon et al., 2016; Pierangeli et al., 2019), while recent systems extend programmable Hopfield/Ising operation and explicitly higher-order interactions (Bybee et al., 2023; Nikhar et al., 2024; Al-Kayed et al., 2025). Most directly, Musa et al. experimentally demonstrate a nonlinear-optical dense associative memory with two- and four-body interactions (Musa et al., 2026). We therefore make no priority claim for higher-order photonic associative memory itself. Our contribution is the learned-layer formulation and the coupled study of trainable basins, continuous pretrained-latent readout, finite-runtime versus equilibrium semantics and gradients, and converter-mediated realization constraints.

A.4 Dense retrieval, retrieval-augmented generation, and multi-hop evidence selection

Dense Passage Retrieval learns semantic embeddings for open-domain question answering (Karpukhin et al., 2020); Sentence-BERT provides compact sentence embeddings useful for cosine retrieval (Reimers & Gurevych, 2019). Retrieval-Augmented Generation (RAG) couples a neural retriever to a generator with explicit non-parametric memory (Lewis et al., 2020). For multi-hop questions, HotpotQA provides annotated supporting documents and facts (Yang et al., 2018); multi-hop dense retrieval conditions later retrieval on earlier hops (Xiong et al., 2021), while Interleaving Retrieval with Chain-of-Thought (IRCoT) alternates retrieval and explicit reasoning (Trivedi et al., 2023). Self-RAG learns to retrieve and critique adaptively (Asai et al., 2024).

Our structured-retrieval experiments are not proposed as a competitive RAG system. Rather, they test whether the third-order interaction $q \times e_i \times e_j$ adds useful conditional compatibility beyond strong semantic embeddings. It does on the controlled parity-like construction but not in our HotpotQA residual study. This negative result motivates a possible future revisit with a frozen first hop, bridge-focused hard negatives, or stronger compositional document representations, rather than further tuning the same residual architecture.

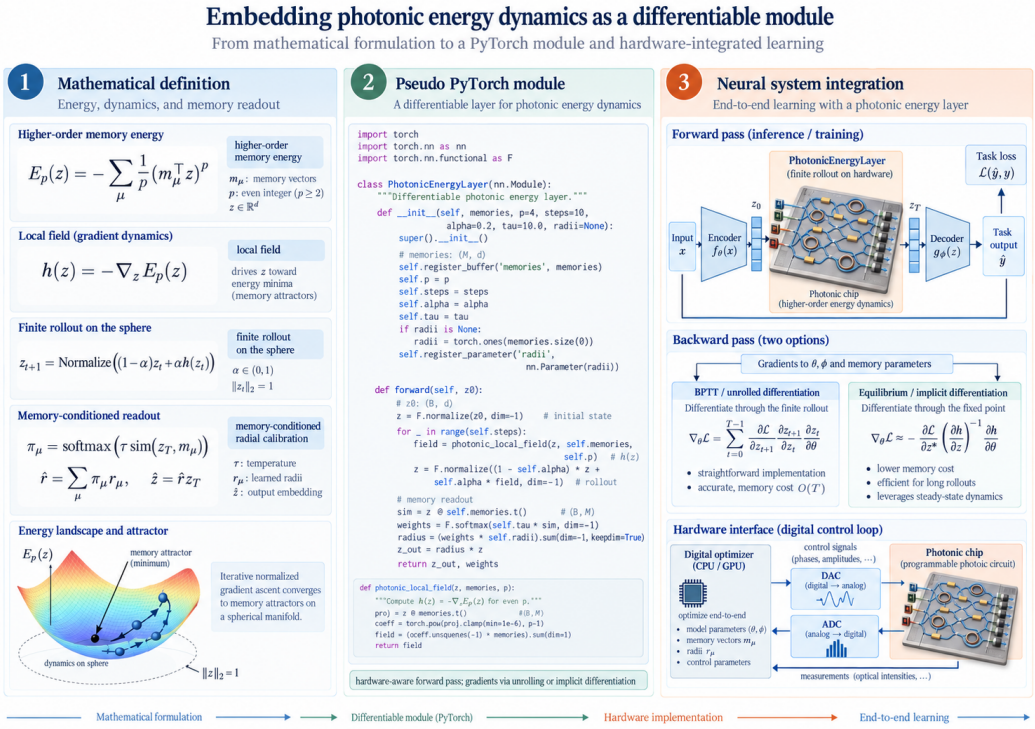

Figure 7: **Embedding photonic energy dynamics as a differentiable module.** The figure summarizes the mathematical energy definition, local-field dynamics, finite rollout, and memory-conditioned radial readout together with PyTorch-style pseudocode and two training interfaces: finite-step unrolling and equilibrium/implicit differentiation. It is intended to make clear that the photonic block can be treated as a neural-network module whose forward pass is a dynamical solver and whose backward pass may follow either the executed rollout or the equilibrium equations.

B Module and System View

The main paper's system visualization (Fig. 2) gives the end-to-end conceptual view. Here we complement it with a module-level bridge from equations to implementation: Fig. 7 summarizes the higher-order energy, finite rollout, memory-conditioned readout, and a PyTorch-like module skeleton. Both views are explanatory schematics rather than measured chip photographs and do not imply a fabricated layout, throughput, or energy claim.

C Energy-Layer Formulation and Factorization

The generic energy in Eq. (2) contains terms with interaction orders p_r. For a homogeneous order-p model,

$$E(z;x) = R(z;x) - \sum_{r=1}^{R_f} c_r \prod_{j=1}^{p} \langle w_{rj}, u_j(z,x) \rangle. \tag{7}$$

An explicit dense polynomial over $z \in \mathbb{R}^N$ can require $\mathcal{O}(N^p)$ coefficients. A rank-R_f factorization requires $\mathcal{O}(R_f p N)$ projection coefficients and $\mathcal{O}(R_f p)$ product/mixing operations before task-specific sharing. Figure 8 illustrates the distinction.

Local field without a dense order-p tensor. For the common case $u_j(z,x) = z$, define $a_{rj} = w_{rj}^\top z$. The interaction contribution to the negative energy gradient is

$$h_{\text{int}}(z) = -\nabla_z (E - R) = \sum_{r=1}^{R_f} c_r \sum_{k=1}^{p} w_{rk} \prod_{j \neq k} a_{rj}. \tag{8}$$

Thus the local field can be evaluated from factor projections and products without ever materializing an N^p coefficient tensor. With untied factors the logical parameter count is $\mathcal{O}(R_f pN)$; with the tied associative-memory factors below, one stores only the $P \times N$ memory matrix, independent of p. This is a representation and operation-count statement, not a claim that physical area or energy is independent of interaction order.

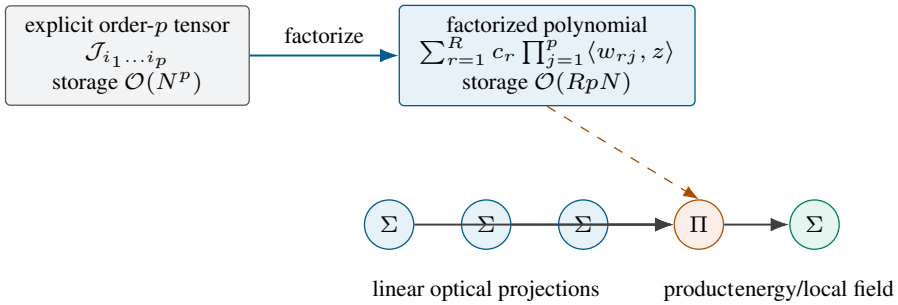

Figure 8: A dense order-p tensor versus a factorized polynomial energy. The factorized form maps naturally to programmable linear projections followed by product interactions and accumulation. These are logical operation counts, not measured photonic area or energy.

For the controlled associative memory, define overlaps

$$m_\mu(z) = \frac{1}{N} \xi_\mu^\top z, \tag{9}$$

and the order-p energy

$$E_p(z) = -\frac{N}{p} \sum_{\mu=1}^{P} m_\mu(z)^p. \tag{10}$$

A relaxed binary-state update is

$$z^{(t+1)} = (1-\alpha) z^{(t)} + \alpha \tanh\left(\beta \sum_\mu m_\mu(z^{(t)})^{p-1} \xi_\mu \right). \tag{11}$$

The stored factor matrix contains PN values for every order p in this construction; changing p changes the local field but not the factor count. For the image-latent study, the state is instead root-mean-square (RMS) normalized after each relaxed update so that the learned energy changes direction/identity rather than merely vector norm.

Why higher order suppresses interference but can narrow basins. Let ν be the intended memory and write its overlap as $a = m_\nu(z)$. The local field decomposes as

$$h_i(z) = a^{p-1} \xi_{\nu i} + \underbrace{\sum_{\mu \neq \nu} m_\mu(z)^{p-1} \xi_{\mu i}}_{\eta_i \text{ (crosstalk)}}. \tag{12}$$

For independent Rademacher distractor memories and a unit-RMS state, the usual leave-one-coordinate-out/Gaussian crosstalk approximation gives $m_\mu \approx \mathcal{N}(0, 1/N)$ for $\mu \neq \nu$. Since $\mathbb{E}[m_\mu^{2(p-1)}] \approx (2p-3)!!\, N^{-(p-1)}$, the distractor RMS on one coordinate is approximately

$$\sigma_{\eta,p} \approx \sqrt{(P-1)(2p-3)!!}\; N^{-(p-1)/2}. \tag{13}$$

Consequently, at a fixed target overlap $|a|$, the idealized signal-to-crosstalk ratio (SIR) scales as

$$\mathrm{SIR}_p \approx \frac{|a|^{p-1} N^{(p-1)/2}}{\sqrt{(P-1)(2p-3)!!}}. \tag{14}$$

Holding this proxy constant suggests the classical polynomial-memory trend $P = \mathcal{O}(N^{p-1})$ under independent-pattern assumptions (Krotov & Hopfield, 2016; Demircigil et al., 2017). We use this only

as a directional explanation: our measured exponents are finite-size empirical fits under nonlinear relaxed dynamics, not estimates of the asymptotic law.

The same exponent also explains why capacity and basin width need not improve together. If the target and a competing memory have overlaps $|a|$ and $|r|$, their local-field weights obey

$$\frac{|a|^{p-1}}{|r|^{p-1}} = \left(\frac{|a|}{|r|}\right)^{p-1}. \tag{15}$$

Higher order therefore amplifies an existing overlap margin: it suppresses many smaller distractors when $|a| > |r|$, but it also amplifies a wrong competitor once severe corruption makes $|r| \gtrsim |a|$. In the spherical layer, subsequent field normalization removes a common positive scale but preserves these relative weights, so the margin-amplification argument still applies. This gives a simple mechanism for the empirical combination of improved load tolerance and a sharper severe-corruption boundary.

Quadratization control. Higher-order polynomials can be represented using pairwise objectives by introducing auxiliary variables and constraints. Our quadratization control verifies equality of native and lifted energies on valid auxiliary assignments. However, free-auxiliary pairwise descent can follow a different relaxation geometry and performs worse than native/constraint-preserving retrieval. Thus representational equivalence does not imply dynamical equivalence at finite solver budget.

D Controlled Interaction-Order Capacity and Basin Geometry

Figure 9 summarizes two complementary controlled observations. First, empirical finite-size fits of the 50%-retrieval load threshold give exponents of approximately 0.72, 2.04, and 3.68 for $p = 2, 4, 6$, respectively. These are empirical finite-size fits, not asymptotic capacity laws. Second, at a low-load 30%-flip checkpoint, exact retrieval is 50.0% for $p = 2$, 95.8% for $p = 4$, and 0.0% for $p = 6$. Increasing interaction order therefore increases storage/interference tolerance but can simultaneously sharpen or narrow basins.

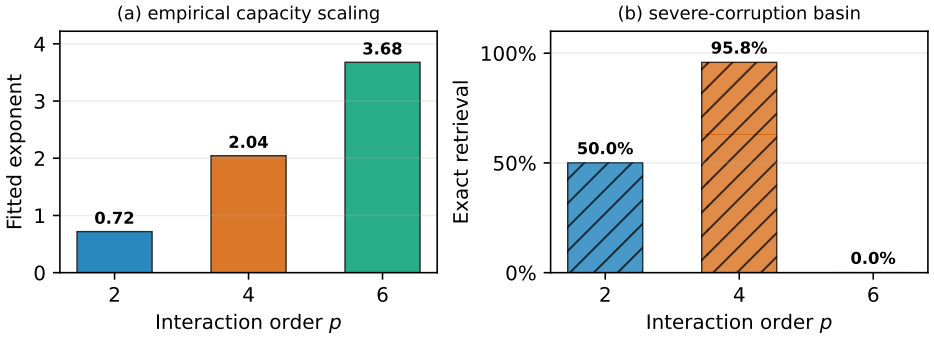

Figure 9: **Higher order increases load scaling but not basin width.** All bar values are shown explicitly. The severe-corruption checkpoint confirms that the sixth-order retrieval result is exactly 0.0%, rather than a missing bar.

E Learned Associative Basins in Pretrained Image Latents

The flagship pretrained-latent study uses CIFAR-10 test images (Krizhevsky, 2009) resized to 128×128 and encoded by the frozen `stabilityai/sd-vae-ft-mse` variational autoencoder associated with latent diffusion (Rombach et al., 2022); only the associative landscape is trained. The task is content-addressable recall of identities already stored in each memory bank from corrupted cues, not generalization to unseen identities. For clean target latents m_μ and corrupted cues q_μ, a

spherical higher-order layer iterates

$$f(z) = \nabla_z[-E_p(z)], \tag{16}$$

$$\hat{f}(z) = f(z)/\,\mathrm{RMS}(f(z)), \tag{17}$$

$$z^{(t+1)} = \mathrm{RMSNorm}\left((1-\alpha)z^{(t)} + \alpha\hat{f}(z^{(t)})\right). \tag{18}$$

The default objective is a target cosine retrieval loss plus a memory-anchor penalty. To isolate dynamics from training-depth confounds, we train each landscape at a fixed rollout depth, freeze its parameters, and then evaluate the same learned landscape across many rollout depths.

Fixed landscape and learned basin radius. At $P = 24$, fourth-order retrieval on a single frozen landscape remains high over multiple iterations while pairwise dynamics increasingly enter mixed/incorrect attractors. Clean stored memories are naturally closer to fourth-order fixed points. Training-distribution controls then show that the basin boundary is learned rather than fixed. At 55% test occlusion, fourth-order retrieval at $T = 4$ is 27.8% under the mixed-corruption schedule, 51.4% after fixed-35% training, and 80.6% after fixed-45% training.

Load/interference scaling. With fixed-45% corruption training, fourth-order retrieval remains around 95–97% at 45% test occlusion through $P = 64$ and then declines gradually to 94.4% at $P = 96$ and 89.6% at $P = 128$. Pairwise retrieval falls to 15.9% at $P = 128$. At 55% corruption, fourth-order retrieval is 72.1% at $P = 128$ versus 13.8% pairwise. This is *matched factorized storage* scaling—the number of trainable factor parameters grows with P for both models—not a fixed-total-parameter capacity claim.

A useful distinction appears before identity retrieval fails: continuous decoded PSNR degrades with increasing P even when the final state remains in the correct identity basin. Reading out the nearest stored latent from the final identity recovers substantially higher PSNR. We interpret this only as a diagnostic separating *selection* from *continuous reconstruction*; exact stored-latent readout is not a fair replacement for a continuous generative state when comparing reconstruction methods.

Memory-conditioned radial calibration at larger load. The practical-photo diagnostic isolates a radial readout mismatch that is orthogonal to memory identity: spherical dynamics recover a latent direction, whereas the original continuous decoder restores its magnitude using the corrupted cue RMS. We therefore repeat the high-load pretrained-latent study while holding each terminal direction fixed and replacing only this radial restoration with hard or soft memory-conditioned RMS. For terminal direction z_T, stored directions m_μ, and stored latent RMS values r_μ, the soft inference-valid scale is

$$\pi_\mu(z_T) = \frac{\exp(\tau z_T^\top m_\mu/d)}{\sum_\nu \exp(\tau z_T^\top m_\nu/d)}, \qquad \hat{r}(z_T) = \sum_\mu \pi_\mu(z_T) r_\mu, \tag{19}$$

and the frozen decoder receives $\hat{z}_T = z_T\hat{r}(z_T)$. Because $\hat{r}(z_T) > 0$, radial rescaling leaves every cosine similarity invariant:

$$\cos(\hat{r}z_T, m_\mu) = \cos(z_T, m_\mu), \qquad \hat{r} > 0. \tag{20}$$

The calibration weights are computed from the unscaled terminal state and the positive scale is applied only afterward. Hence radial calibration can change decoded continuous magnitude and PSNR/SSIM while being mathematically unable to change the cosine-selected memory identity. The soft readout does require a terminal similarity sweep over the stored memory bank plus stored radii; this inference-time access is separate from the recurrent energy-layer cost. The study verifies this invariance numerically across cue-, hard-memory-, and soft-memory-RMS readouts and additionally records soft-weight entropy/effective-memory count to detect whether a fixed temperature becomes too diffuse as memory load grows.

On the external three-seed run, identity retrieval is exactly invariant under calibration at every load and corruption level, reproducing the fourth-order retrieval values from 95.8%/84.0% at $P = 48$ to 89.6%/72.1% at $P = 128$ for 45%/55% occlusion. Soft memory-RMS calibration consistently improves continuous reconstruction: relative to cue-RMS decoding, PSNR gains are about $+2.34/+2.26$ dB at $P = 48$ and remain $+1.75/+1.72$ dB at $P = 128$ for 45%/55% occlusion, with corresponding SSIM gains of $+0.083/+0.091$ and $+0.059/+0.077$. The radial-scale error drops

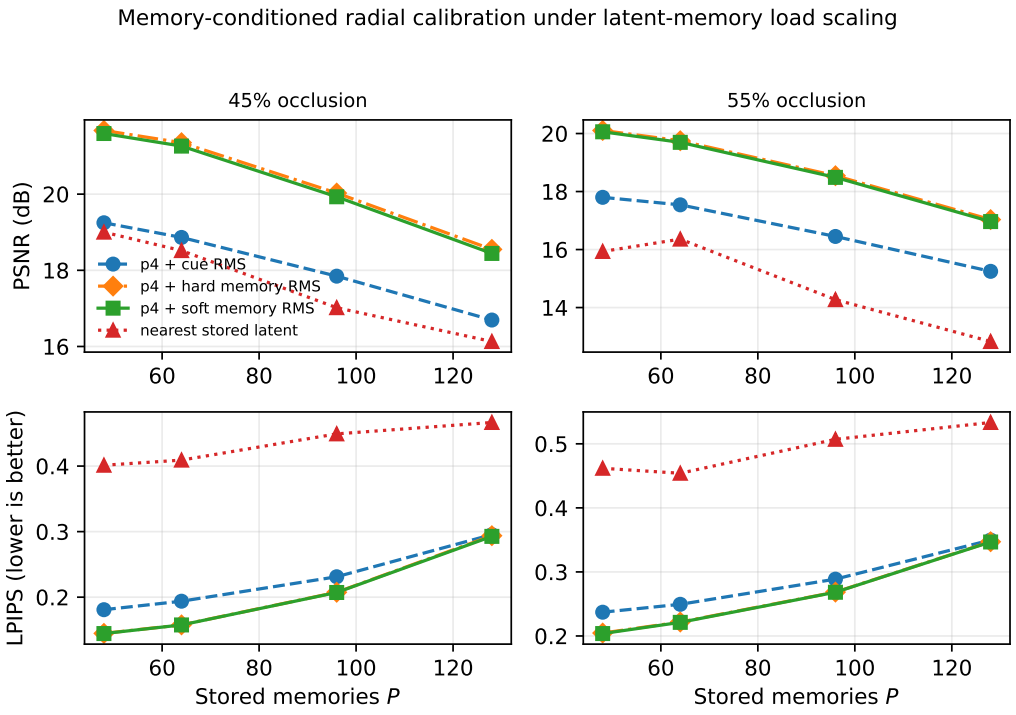

Figure 10: **Memory-conditioned radial calibration under pretrained-latent load scaling.** Cue-RMS decoding is compared with inference-valid soft memory-RMS calibration and exact nearest stored-latent readout at matched terminal directions. Calibration changes continuous reconstruction only; the experiment verifies that the energy-layer identity decisions are unchanged. PSNR is higher-is-better and LPIPS is lower-is-better.

by roughly 0.21–0.26 in mean absolute log-ratio across all operating points. LPIPS improves substantially at lower loads, but the perceptual gain narrows at $P = 128$ even though PSNR/SSIM and radial fidelity still improve. Hard and soft memory-RMS readouts are very close; hard memory RMS gives slightly higher PSNR, whereas soft RMS gives comparable or slightly better LPIPS while remaining differentiable.

The soft weighting becomes more diffuse with load: its effective number of contributing memories rises from about 4.8/6.6 at $P = 48$ to 25.7/32.0 at $P = 128$ for 45%/55% occlusion. We therefore keep temperature $\tau = 8$ as the prespecified primary setting but treat temperature scaling with P as a secondary diagnostic rather than retuning the main result post hoc.

Temperature sensitivity of soft radial calibration. We next keep the pretrained-latent memory dynamics fixed and vary only the soft-memory radial temperature $\tau \in \{4, 8, 16, 32, 64\}$ at the two largest memory loads, $P \in \{96, 128\}$. The pre-existing default $\tau = 8$ is retained as the primary setting; the remaining temperatures form a sensitivity analysis rather than a retuning of the main result. Increasing τ sharply concentrates the radial weights: at $P = 128$, the effective number of contributing memories falls from roughly 93–95 at $\tau = 4$ to 26–32 at $\tau = 8$, about 5–6 at $\tau = 16$, and nearly one by $\tau = 64$. Reconstruction quality improves mainly from $\tau = 4$ to 8–16 and then saturates. For example, at $P = 128$ and 55% occlusion, PSNR changes from 16.90 dB at $\tau = 4$ to 16.96 at 8, 17.02 at 16, and 17.03 at 32–64; SSIM and LPIPS change only marginally beyond $\tau = 8$–16. Thus the default $\tau = 8$ is somewhat diffuse at high load but already lies close to the fidelity plateau, while larger τ primarily sharpens the memory weighting without materially changing reconstruction quality.

Illustrative image cases. To make the latent-memory behavior more concrete, Figure 12 shows three individual reconstruction cases from the frozen qualitative grid. For each case we show the

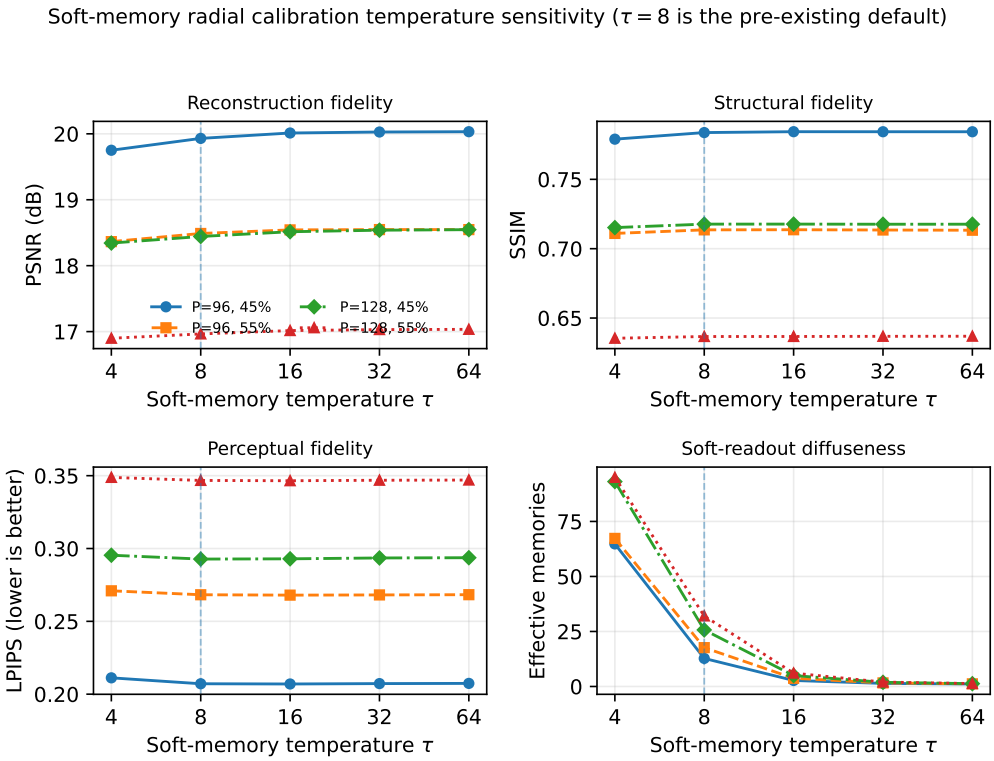

Figure 11: **Soft-memory radial-calibration temperature sensitivity.** The horizontal axis is the soft-memory temperature τ (not memory load). Curves show $P = 96$ and 128 at 45% and 55% occlusion. PSNR, SSIM, and LPIPS rapidly approach a plateau by $\tau \approx 8$–16, whereas the effective number of contributing memories continues to collapse toward a hard selection as τ increases. The vertical dashed line marks the pre-existing default $\tau = 8$.

target, the corrupted cue, the nearest retrieved stored identity, and the decoded outputs from the learned pairwise and fourth-order energy layers. The reported PSNR and SSIM values are computed per displayed case against the target image and are included only as case records, not as aggregate statistics. The ensemble-level claims in the paper continue to rely on the frozen quantitative summaries rather than on these examples alone.

E.1 Practical high-resolution photo case studies

We additionally provide a small practical-photo case study using natural photographs whose Wikimedia Commons file pages declare Creative Commons Zero (CC0) 1.0. The "Edgerton" label used in the figures refers to the Edgerton House photograph in this set, not to a method or model; its source is a CC0 Wikimedia Commons image by Daderot. The authors' internal provenance records preserve each source page, verified license metadata, source SHA-256, and per-case metrics. The acquisition pipeline keeps the raw originals separate from the analysis workspace, center-crops each source to 512×512, and evaluates the same frozen Stable Diffusion variational autoencoder (SD-VAE) latent-memory pipeline under 45% occlusion. This is intentionally a qualitative appendix study rather than an ensemble benchmark: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) are reported for each displayed image individually, with no aggregate mean used as a primary claim. To avoid a redundant image grid, Fig. 14 shows the more complete comparison across all eight cases and calibrated readouts.

The hard nearest-stored-latent readout is strongest when it selects the correct stored identity, because it directly decodes a memorized latent rather than the continuous terminal energy state. In this eight-case diagnostic set, nearest-neighbor identity is correct in five cases, the pairwise energy state in eight, and the fourth-order energy state in seven; these are descriptive counts for the illustrative

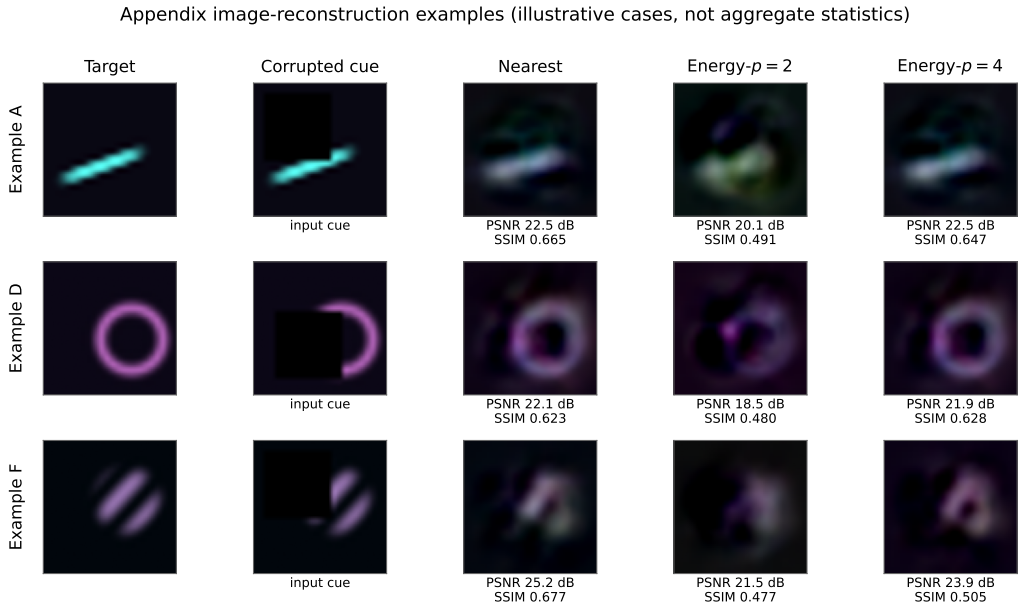

Figure 12: **Illustrative image-reconstruction cases from the latent-memory study.** Three representative cases from the frozen qualitative grid are shown at larger scale. Under each decoded reconstruction we report its individual PSNR and SSIM relative to the target image. These values are intended as qualitative case documentation rather than as a replacement for the aggregate quantitative results. In these selected examples, the fourth-order energy layer improves over the pairwise energy layer and, in some cases, approaches the nearest-memory baseline. The per-case metrics for all six qualitative-grid examples were checked against the retained internal experiment records; the aggregate claims continue to rely on the frozen quantitative summaries.

set rather than benchmark rates. The dog and apple examples are particularly informative: nearest neighbor selects the building memory, whereas both learned energy dynamics preserve the intended identity and substantially improve all three image metrics. Conversely, the Edgerton House photo case is retained as a fourth-order failure. Several continuous energy-state decodes also exhibit visibly stronger contrast or saturation than the target, motivating the controlled diagnostic below.

E.2 Contrast, latent drift, and radial readout diagnostics

The spherical energy layer evolves latent direction at fixed RMS. In the original continuous readout, the terminal direction is rescaled by the corrupted cue's latent RMS before frozen-VAE decoding. We therefore separate three possible causes of the visual contrast shift: rollout depth, learned-landscape regularization/dynamics, and latent radial-scale restoration. The follow-up evaluates the same baseline landscape across multiple depths; retrains controlled variants with stronger memory anchoring, clean-memory fixed-point loss, intermediate-trajectory loss, or lower relaxation; and finally holds the learned direction fixed while changing only the radial readout. The memory-derived radial readouts are inference-valid because they use only the stored memory bank and terminal state; an oracle target-RMS readout is included solely as a diagnostic and is not a valid inference method.

E.3 Inference-valid radial calibration and auditable sample rendering

The contrast diagnostic motivates an inference-valid correction: rather than restoring the terminal spherical state with the corrupted cue RMS, we estimate a radial scale from the stored memory bank. For terminal state z_T, normalized stored memories m_μ, and stored latent RMS values r_μ, the soft

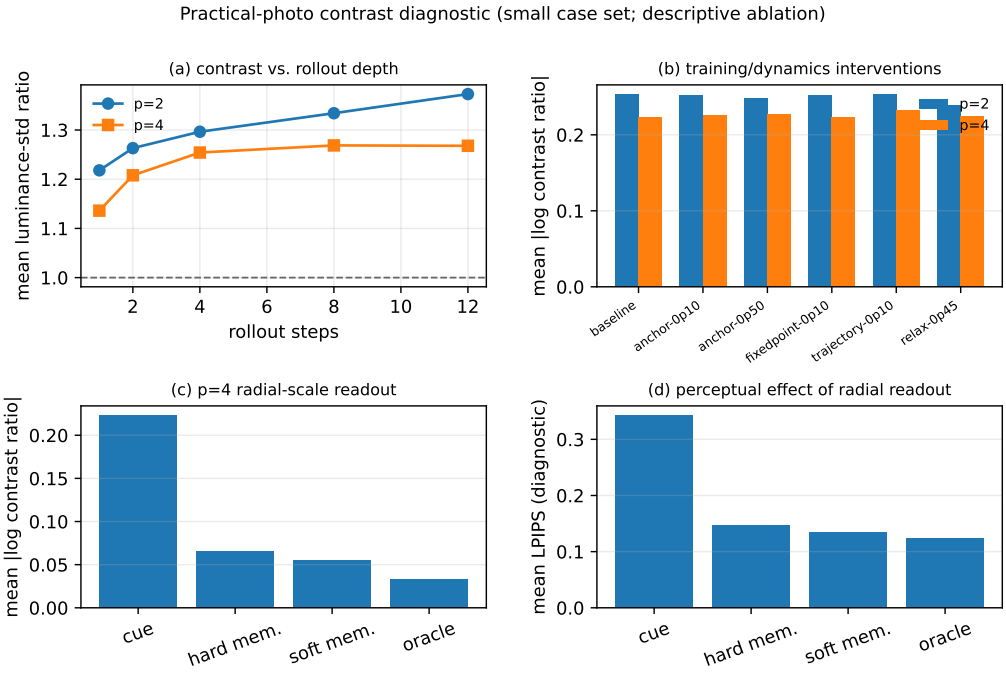

Figure 13: **Diagnostic decomposition of contrast inflation in practical-photo latent memory.** The follow-up separates finite-rollout depth, training regularization/relaxation, and latent radial readout. Luminance-contrast error is measured as the absolute log ratio of reconstructed to target luminance standard deviation. LPIPS summaries are descriptive means over this same small case set, not benchmark results. The oracle radial-scale condition is diagnostic only.

calibration uses

$$\pi_\mu(z_T) = \frac{\exp\{\tau\, z_T^\top m_\mu / d\}}{\sum_\nu \exp\{\tau\, z_T^\top m_\nu / d\}}, \tag{21}$$

$$\hat{r}(z_T) = \sum_\mu \pi_\mu(z_T) r_\mu, \tag{22}$$

$$\hat{z}_T = z_T \hat{r}(z_T). \tag{23}$$

This preserves a continuous terminal state while restoring a latent radius from inference-available stored-memory statistics. We also retain hard selected-memory RMS as an ablation and combine the soft readout with the trajectory-regularized fourth-order landscape as a basin-robust variant.

Observed calibration effect. On the same eight curated practical-photo cases, the original cue-RMS fourth-order readout selects the correct identity in 7/8 cases and has descriptive means of 19.19 dB PSNR, 0.572 SSIM, and 0.342 LPIPS. Holding the learned fourth-order direction fixed while replacing cue RMS with soft memory-derived RMS keeps the same 7/8 identity decisions but changes the descriptive means to 24.13 dB, 0.664, and 0.134, respectively. Combining trajectory regularization with soft-memory radial calibration corrects the remaining Edgerton House identity error (8/8 on this small set) and gives descriptive means of 24.65 dB PSNR, 0.702 SSIM, and 0.079 LPIPS. These are diagnostic summaries of a curated case set rather than benchmark estimates. The decomposition supports two distinct roles: memory-conditioned radial calibration repairs the spherical readout scale, whereas trajectory regularization improves basin/identity robustness.

The qualitative comparison in Figure 14 is assembled only after the experiment has written the per-case PNG files and corresponding metric tables. The rendering stage only arranges saved artifacts; it cannot synthesize, replace, or enhance image content. In particular, a failed nearest-memory lookup displays the actual wrongly selected image from the same memory bank and is annotated with the selected case identifier.

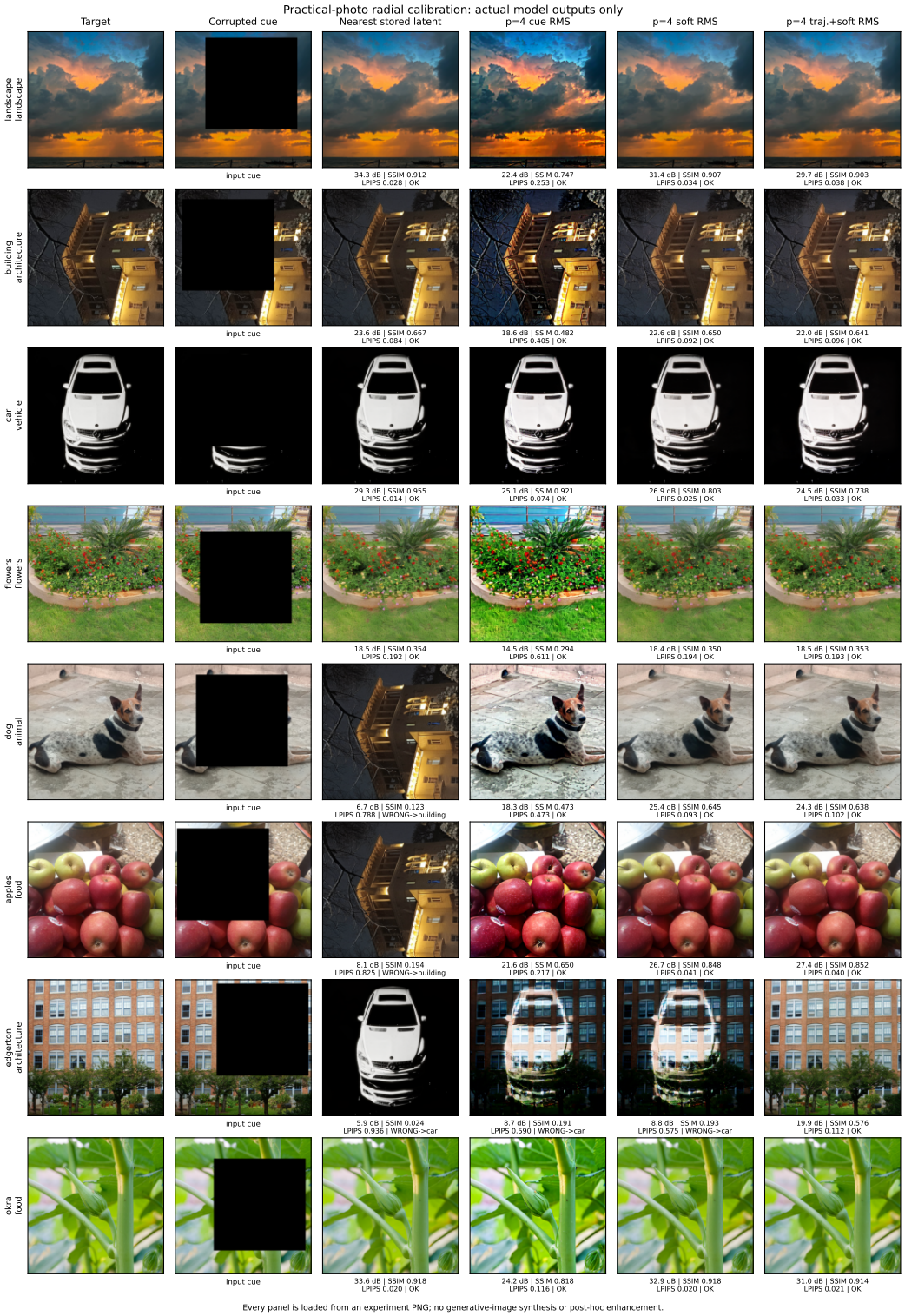

Figure 14: **Auditable practical-photo comparison after radial calibration.** Every panel is a saved model/data output from the frozen experiment pipeline and the figure is assembled only from those saved artifacts; no generative-image synthesis or post hoc enhancement is used. Columns compare the corrupted-cue radial baseline, soft-memory radial calibration, and the trajectory-regularized soft-calibrated variant. Wrong memory retrievals are labeled with the actually selected stored case.

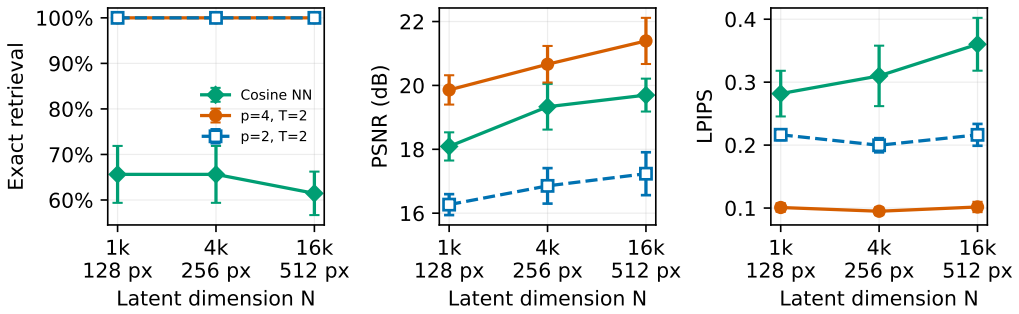

Figure 15: **Natural-image resolution scaling at fixed memory load $P = 32$.** Error bars are one sample standard deviation over three seeds. Image identities, native center crops, and relative occlusion placement are matched across 128, 256, and 512 pixels. Both learned models retain 100% exact identity retrieval at the prespecified $T = 2$ execution depth, while the fourth-order terminal state consistently improves PSNR and LPIPS over the pairwise state. The cosine control remains near 61–66% retrieval.

F Natural-Image Resolution Scaling

To test whether the pretrained-latent memory behavior is specific to 128×128 inputs, we run a small controlled study on Microsoft Common Objects in Context (COCO) 2017 validation images (Lin et al., 2014). For each of three seeds, the same 32 source-image identities are center-cropped once and evaluated at 128, 256, and 512 pixels, corresponding to frozen SD-VAE latent dimensions $N = 1{,}024$, $4{,}096$, and $16{,}384$. The 45% square occlusion uses matched normalized coordinates across resolutions. Both learned memories use the same 120-update recipe with $T_{\text{train}} = 2$, and the primary evaluation depth $T = 2$ is fixed before observing the sweep. Cosine nearest-neighbor lookup of the stored latent bank provides a training-free control.

At this relatively light memory load, interaction order no longer separates *identity* success: both $p = 2$ and $p = 4$ retrieve the correct stored item in every measured seed/resolution cell. It does, however, separate continuous terminal-state fidelity. The fourth-order PSNR is 19.86, 20.66, and 21.39 dB at 128, 256, and 512 pixels, respectively, compared with 16.27, 16.86, and 17.24 dB for the pairwise model, a gap of 3.59–4.16 dB. Fourth order also maintains LPIPS near 0.10 and latent cosine near 0.975 across the $16\times$ latent-dimensionality increase. Cosine nearest-neighbor retrieves only 65.6%, 65.6%, and 61.5% of identities, respectively. The plotted error bars give the sample standard deviations over the three seeds.

The secondary $T = 1 \ldots 8$ trajectory is retained only as a descriptive check. It places the fourth-order PSNR maximum at $T = 5$ for all three resolutions, but none of the primary claims use this post-hoc depth selection; Fig. 15 uses the prespecified $T = 2$ result. This experiment tests numerical resolution/latent-dimensionality robustness with a frozen pretrained codec. It is not a measured photonic throughput, area, energy, or device-scaling experiment, and cross-resolution changes in image metrics should not be interpreted as hardware scaling laws.

G Structured Evidence Selection Beyond Associative Memory

G.1 Controlled test of irreducible higher-order compatibility

The controlled task contains S evidence slots with C candidates per slot. Each candidate carries balanced $\{\pm 1\}$ code channels, and the query syndrome is formed from the product of the codes associated with the correct candidate in every slot. By construction, compatibility marginals below order S carry essentially no information about the target combination. This gives a clean test of whether the score can represent an interaction that is genuinely joint rather than merely correlated through pairwise statistics.

A factorized scorer forms slot-specific projections followed by channelwise products, avoiding explicit enumeration of all C^S tuples. With $C = 8$, exact factorized scoring reaches 99.0% joint success for the three-way task and 98.4% for the four-way task, matching explicit digital joint scoring. Independent and best pairwise-projected scorers achieve only 30.7% and 15.1%, respectively. A finite coordinate-relaxation solver improves to 51.0%/32.8% but remains far below the exact factorized score. The controlled result therefore separates two claims: the higher-order representation contains the required compatibility, while a finite dynamical solver may still fail to locate its optimum.

G.2 HotpotQA as a real-data boundary

For real multi-hop retrieval, a strong pretrained embedding already supplies useful semantic geometry. In the initial from-scratch selector study, frozen cosine retrieval obtained 29.7% joint support and retrieved at least one gold support for 118/128 examples, while learned pairwise, third-order, and set-attention selectors all degraded this starting point. We therefore preserve cosine and restrict learning to a residual correction,

$$S(i, j) = S_{\cos}(i, j) + \alpha R(q, e_i, e_j), \tag{24}$$

where the bounded gate α is initialized at zero. Pairwise and third-order variants use the same projection weights and matched parameter counts; the interaction order is the intended difference.

On 4096 training and 512 validation questions, the pairwise residual increases joint support from 32.42% to 36.78%, whereas the matched third-order residual falls to 28.84%. Oracle second-hop accuracy is 58.01% for cosine, 58.50% for the pairwise residual, and 50.10% for the third-order residual. The third-order model also changes roughly 72% of selections and damages about 52% of cases that cosine originally solved correctly. Importantly, the third-order model attains lower training loss while generalizing worse, indicating that the failure is not simply insufficient optimization: the extra interaction capacity is misaligned with the frozen representation and available supervision.

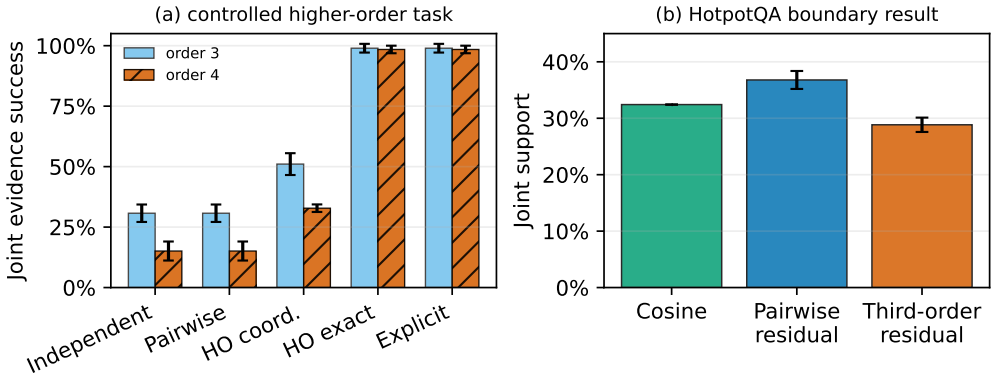

Figure 16: **Controlled expressivity versus a real-data boundary.** Bars are means and error bars are one sample standard deviation over three seeds. (a) The controlled construction makes the target compatibility irreducibly higher-order, so exact factorized scoring matches explicit joint enumeration while pairwise scores cannot. Finite coordinate relaxation leaves a separate optimization gap. (b) On HotpotQA, the same representational advantage does not transfer: a cosine-anchored pairwise residual improves joint retrieval, whereas a parameter-matched third-order residual generalizes worse.

The scientific boundary is therefore sharper than "higher order helps retrieval." Higher-order interactions help when the representation exposes a joint compatibility that lower-order statistics cannot express; they need not help when a pretrained embedding has already organized the task around a strong lower-order semantic geometry. A future revisit should change that representation hypothesis—for example by freezing the first hop and modeling conditional second-hop compatibility, using bridge-focused hard negatives, or employing a stronger compositional encoder—rather than merely tuning the failed residual more aggressively.

H FINITE-DEPTH VERSUS FIXED-POINT DIFFERENTIATION

For a general recurrent map $z^+ = F_\theta(z, x)$, a fixed point satisfies

$$G(z^\star, \theta) = z^\star - F_\theta(z^\star, x) = 0. \tag{25}$$

Differentiating gives

$$\frac{dz^\star}{d\theta} = (I - J_z F_\theta)^{-1} J_\theta F_\theta. \tag{26}$$

Rather than forming the inverse, solve

$$(I - J_z F_\theta)^\top \lambda = \nabla_z L, \tag{27}$$

then

$$\frac{dL}{d\theta} = \partial_\theta L + \lambda^\top J_\theta F_\theta. \tag{28}$$

For the special gradient step $F(z) = z - \eta \nabla E(z)$, $I - J_F = \eta \nabla^2 E$, recovering the familiar Hessian form. The fixed-point form is required for the RMS-normalized spherical associative-memory update because its equilibrium is defined by the full normalized map, not by unconstrained stationarity $\nabla E = 0$.

Finite unrolling as a truncated implicit series. The relation between finite-depth backpropagation and the implicit gradient is explicit in the local linearization around a fixed point. Let

$$J = J_z F_\theta(z^\star, x), \qquad A = J_\theta F_\theta(z^\star, x), \tag{29}$$

and suppose the initial state is parameter independent. The tangent recursion $U_{t+1} = JU_t + A$, $U_0 = 0$, gives

$$U_T = \sum_{k=0}^{T-1} J^k A, \qquad U_\star = (I - J)^{-1} A = \sum_{k=0}^{\infty} J^k A, \tag{30}$$

whenever $\|J\| < 1$. If $\|J\| \le q < 1$, then

$$\|U_\star - U_T\| \le \frac{q^T}{1 - q} \|A\|. \tag{31}$$

For the nonlinear layer, the Jacobians vary along the trajectory, so Eq. (31) is a local explanation rather than a global guarantee. It nevertheless clarifies the experiment: as the finite trajectory enters the fixed-point neighborhood, unrolled and implicit gradients can agree arbitrarily well even though the task loss of the converged state is worse than that of the trained finite-depth state.

Why state convergence does not imply task improvement. A contraction controls distance to $z^\star$, not the downstream task loss. For $e_T = z^{(T)} - z^\star$,

$$L(z^{(T)}) - L(z^\star) = \nabla L(z^\star)^\top e_T + \mathcal{O}(\|e_T\|^2). \tag{32}$$

Finite-depth training minimizes the loss at the executed depth; it does not in general enforce $\nabla L(z^\star) = 0$ for the out-of-training-depth equilibrium. Therefore $\|e_T\| \to 0$ is fully compatible with non-monotone task quality or with a finite-depth optimum. A monotone-energy or contraction argument alone cannot justify "more iterations is better" for the learned task.

Contraction and adjoint conditioning. The same Neumann structure governs the implicit adjoint. If $\|J\|_2 \le q < 1$, then

$$\|(I - J)^{-1}\|_2 \le \frac{1}{1 - q}, \qquad \kappa_2(I - J) \le \frac{1 + q}{1 - q}. \tag{33}$$

Thus a contraction factor near one permits slow and sensitive backward solves even after the forward residual is small. The bound requires an operator-norm contraction; spectral radius $\rho(J) < 1$ alone is weaker, especially for non-normal Jacobians. This is why we report both forward residual and direct conditioning/adjoint-error diagnostics rather than treating forward convergence as a certificate of gradient accuracy.

Task-facing finite-depth versus fixed-point study. The main experiment evaluates the same finite-depth-trained image-memory landscape over $T \in \{1, 2, 3, 4, 5, 6, 7, 8, 16, 32, 64, 128, 256, 512\}$ and at a separately converged fixed point. All budgets are reported. Reconstruction uses the same terminal direction with the inference-valid soft memory-radius readout; radial calibration therefore cannot alter the retrieval identity. The main-text Table 2 gives the two trained fourth-order loads.

In the completed three-seed $T \geq 4$ sweep at $P = 128$, the $T = 4$ PSNR advantage over the fixed point is +4.40, +5.42, and +6.11 dB for the three individual seeds. On the 12-case gradient-diagnostic subset, the mean finite-versus-implicit memory-gradient error drops from 1.20 at $T = 4$ to 5.36×10^{-3} at $T = 128$ and 9.69×10^{-7} at $T = 512$. Thus task quality and equilibrium fidelity move in opposite directions because the learned finite-depth computation itself changes when it is iterated further.

Matched recurrent training-depth control. The matched-depth control changes only the recurrent depth used during otherwise matched fourth-order training, using $T_{\text{train}} \in \{2, 4, 8\}$ with the same seeds, image indices, fixed-45% corruption schedule, 120 Adam updates, learning rate, memory anchor, and relaxation. The reconstruction optimum shifts monotonically later with training depth, but not in exact lockstep: at $P = 48$, the best PSNR depths are $3, 4, 7$, while at $P = 128$ they are $2, 4, 6$ for $T_{\text{train}} = 2, 4, 8$, respectively. Five of the six load/depth cells place the training depth inside the prespecified 0.05 dB near-optimal band. The exception is $P = 128, T_{\text{train}} = 8$, where $T = 6$ is best and exceeds $T = 8$ by about 0.188 dB. Moreover, this deeper high-load model is much weaker in absolute quality (best PSNR 14.47 dB; retrieval at $T = 8$ is 12.8%). Thus this control rejects a universal fixed early-stopping depth and supports depth-dependent learned dynamics, but it does not establish an exact $T^*_{\text{infer}} = T_{\text{train}}$ law or a general advantage for deeper recurrence.

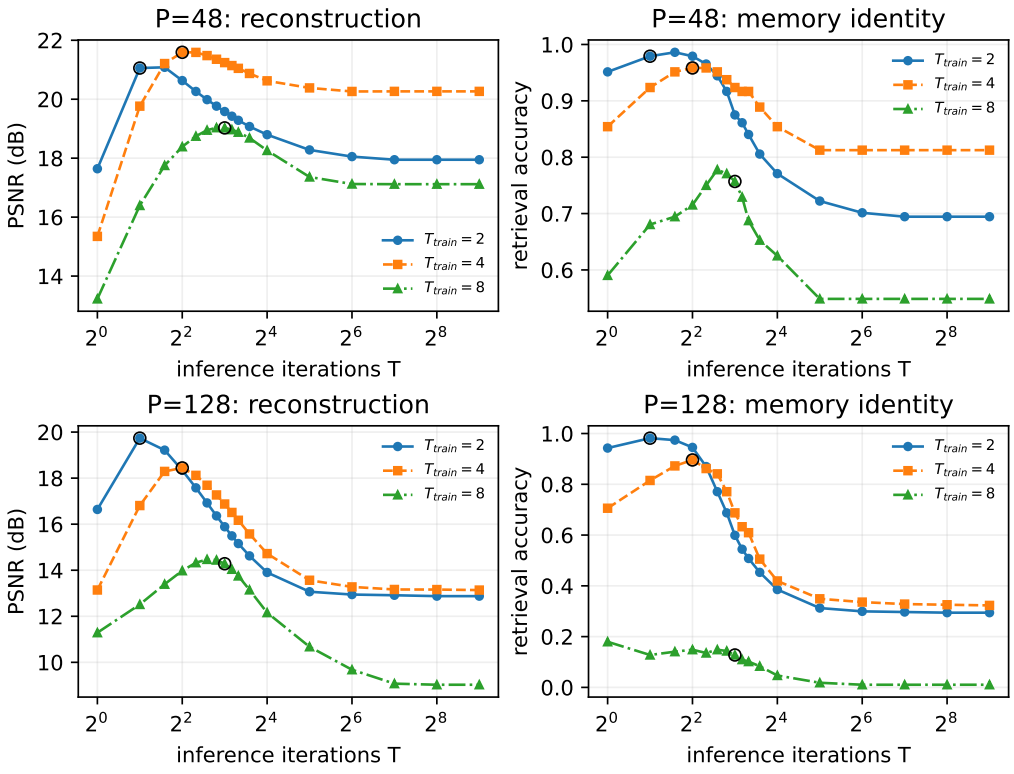

Figure 17: **Matched recurrent training depth changes the preferred finite execution depth.** Reconstruction PSNR and memory identity are shown versus inference iterations for fourth-order models trained with $T_{\text{train}} = 2, 4, 8$. The optimum moves later as training depth increases at both $P = 48$ and $P = 128$, but the high-load $T_{\text{train}} = 8$ model peaks before its training depth and is substantially weaker overall. This supports depth-dependent recurrent semantics rather than exact train/inference-depth equality.

Pretrained-latent fixed-point diagnostic. A separate untrained-factor diagnostic isolates forward/backward convergence without using retrieval accuracy as a benchmark. On frozen 1024-dimensional SD-VAE latents, all 12 pairwise cases are within 1% of the fixed-point implicit memory gradient by $T = 32$. For fourth order, 11/12 cases are within 1% by $T = 128$, but one slow tail still has 18.6% error after 128 steps and requires 524 forward iterations and 746 adjoint iterations to meet the fixed-point/adjoint tolerances. This tail motivates checking convergence per case rather than using a single nominal rollout depth.

Adjoint conditioning after forward convergence. Forward convergence alone does not certify an implicit gradient. In a controlled spherical-memory stress test, the forward residual is below 10^{-10} while the local fixed-point linearization reaches $\rho(J_F) \approx 0.958$ and $\kappa(I - J_F) \approx 23.8$. Limiting the adjoint to 32 iterations leaves 24.2% memory-gradient error. When the matrix-free adjoint is converged, it agrees with a dense exact implicit solve to within 2.7×10^{-10} relative error. This isolates backward conditioning from forward nonconvergence.

Controlled convergence and storage reference. A low-dimensional gradient-descent reference provides a software accounting control. It compares same-budget unrolling, implicit differentiation at the terminal state, a straight-through/Jacobian-free approximation, and a long-rollout equilibrium reference. Figure 18 shows that implicit gradient error collapses as the forward dynamics approach equilibrium; at $T = 32$ relative error is about 8.3×10^{-3} and at $T = 64$ about 4.4×10^{-6}. The PyTorch saved-tensor proxy at $T = 64$ is about $113\times$ the implicit path. This is an autodiff-storage proxy, not measured photonic memory.

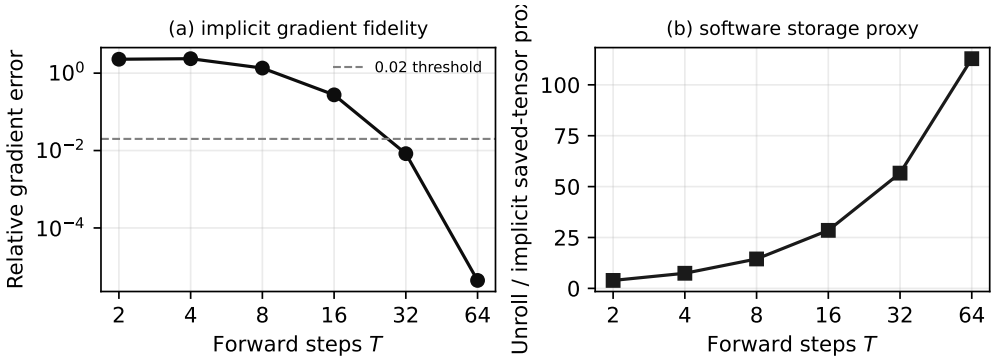

Figure 18: **Controlled fixed-point differentiation reference.** (a) Relative gradient error collapses only once the forward dynamics approach the equilibrium computation. (b) The deterministic PyTorch saved-tensor proxy grows approximately linearly for unrolling relative to implicit differentiation. This controlled gradient-descent reference is supporting evidence; the main text uses the task-facing trained associative-memory result.

A teacher/student experiment in the same controlled reference gives the complementary lesson: at short depth, estimators optimize the truncated computation; once the forward dynamics converge, unrolled, implicit, and straight-through estimators all learn successfully in the smooth regime. The estimator should therefore be chosen only after the intended forward semantics are fixed.

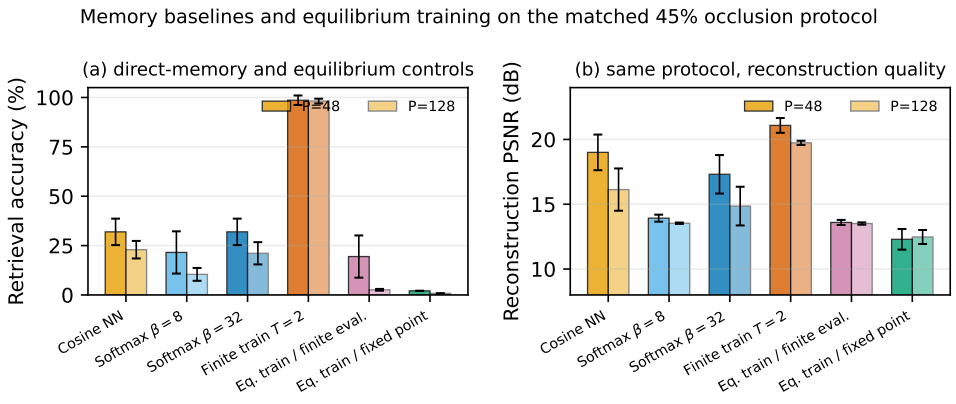

Figure 19: **Direct-memory and equilibrium-training controls under the matched frozen-latent protocol.** Retrieval accuracy (left) and reconstruction PSNR (right) are means with one sample standard deviation over three seeds. Cosine nearest neighbor and one-step softmax memory remain below the shallow finite-trained recurrent model at both loads. The $\beta = 32$ softmax point is the strongest member of the fixed temperature sweep, shown only as a post hoc upper-envelope control; $\beta = 8$ remains the prespecified primary softmax baseline. The equilibrium-trained rows are negative controls for the current optimization recipe, not evidence that equilibrium learning is intrinsically inferior.

I Additional Memory Baselines and Equilibrium Training

The main finite-depth result admits three straightforward alternative explanations: the recurrent layer might merely approximate direct lookup in the stored memory bank, the one-step attention baseline might be under-tuned, or a sufficiently deep/equilibrium-trained recurrent model might recover the same quality. We therefore ran matched controls on the same frozen SD-VAE memories, image indices, three seeds, and fixed 45% occlusion protocol used by the high-load image-memory study. These controls are deliberately reported as boundary checks rather than as a new primary benchmark.

I.1 Direct stored-memory lookup

Cosine nearest-neighbor retrieval provides the strongest simple hard-lookup reference among the prespecified direct baselines. At $P = 48$, it reaches $31.9 \pm 6.7\%$ retrieval and 19.00 ± 1.37 dB PSNR; at $P = 128$, it reaches $22.9 \pm 4.4\%$ retrieval and 16.13 ± 1.63 dB PSNR. The matched fourth-order model trained with two recurrent steps reaches $98.6 \pm 2.4\%$ retrieval and 21.08 ± 0.57 dB PSNR at $P = 48$, and $98.2 \pm 1.2\%$ and 19.74 ± 0.16 dB at $P = 128$. Thus the high-load gain is not reproduced by simply selecting the nearest stored latent.

I.2 Fixed softmax-temperature sweep

To test whether the primary $\beta = 8$ softmax memory is simply too diffuse, we report the complete fixed sweep $\beta \in \{1, 2, 4, 8, 16, 32\}$ rather than selecting a hidden test-set temperature. Increasing β moves the one-step softmax memory toward hard lookup and improves both retrieval and PSNR, but even the strongest tested point remains below the trained recurrent memory at both loads. We retain $\beta = 8$ as the prespecified primary comparison and use $\beta = 32$ only to show the upper envelope of this fixed sweep.

I.3 Deep finite training and true equilibrium training

The matched-depth experiment used $T_{\text{train}} \in \{2, 4, 8\}$. We extended the $P = 48$ bridge to $T_{\text{train}} \in \{16, 32, 64\}$ under the same 120-update optimization budget. Performance does not smoothly approach a useful equilibrium regime: the shallow models remain strongest, $T_{\text{train}} = 8$ already degrades, and the $T_{\text{train}} = 16, 32, 64$ models fall to roughly 13.6–14.3 dB with retrieval below 22%.

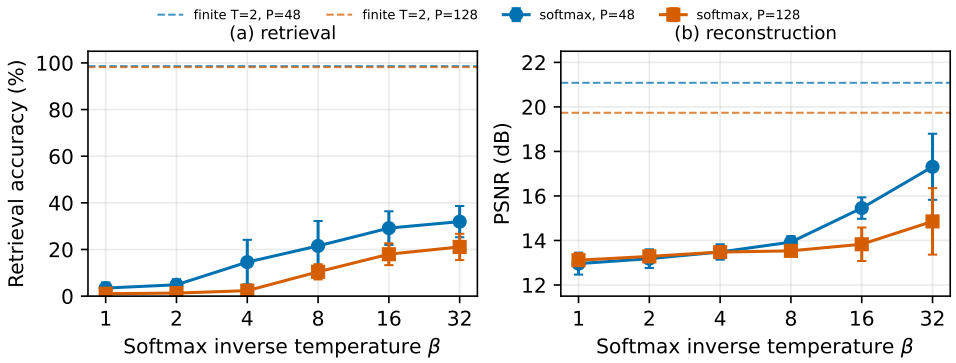

Figure 20: **Complete fixed softmax-memory temperature sweep.** Every tested β is shown for $P = 48$ and $P = 128$; dashed horizontal references mark the matched fourth-order model trained with two recurrent steps. Higher temperature sharpens the one-step memory readout but does not close the recurrent-model gap. No temperature is selected by test-set optimization.

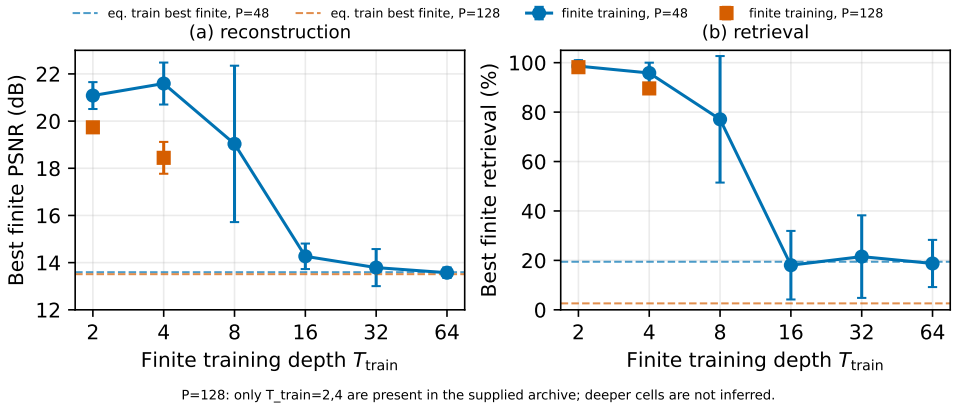

Figure 21: **Deep finite-training bridge and equilibrium-trained reference.** The complete $P = 48$ sweep shows that simply increasing recurrent training depth under the fixed 120-update recipe does not approach a high-quality equilibrium computation. For $P = 128$, only the available $T_{\text{train}} = 2, 4$ cells are plotted and are intentionally not connected across missing depths. Horizontal equilibrium-trained references report the separately trained fixed-point model. These results diagnose the present optimization regime rather than establish a universal depth law.

The $P = 128$ bridge contains only $T_{\text{train}} = 2, 4$; deeper cells are intentionally left missing rather than inferred from the lower-load sweep or other experiments.

We also trained a separate true fixed-point model: every training update solves $z^\star = F_\theta(z^\star, x)$ and differentiates the converged fixed point with the matrix-free implicit adjoint. Its evaluation fixed points converge for all recorded seeds, but the learned solution is weak under the present 120-update recipe. The best finite evaluation reaches 13.60±0.19 dB at $P = 48$ and 13.51±0.09 dB at $P = 128$, while the actual fixed points reach 12.29±0.80 and 12.47±0.54 dB. We interpret this as an optimization boundary of the current equilibrium-training recipe, not as a general statement that fixed-point learning is inferior.

Exact three-seed values. Tables 6 and 7 give the numerical values underlying Figs. 19–21. They report mean ± sample standard deviation; boldface marks the task-relevant best finite values among the available matched cells and does not imply statistical significance.

Table 6: Memory baselines and equilibrium-training controls on the matched frozen-latent 45% occlusion protocol. Values are mean ± sample standard deviation over three seeds. Softmax $\beta = 8$ is the prespecified primary attention baseline; $\beta = 32$ is the strongest point in the fixed $\{1, 2, 4, 8, 16, 32\}$ sweep and is shown as a post-hoc upper-envelope control. The equilibrium-trained rows report the best finite evaluation point and the actual fixed point of the same trained model.

Load	Method	Retrieval (%)	PSNR (dB)
48	Cosine nearest neighbor	31.9 ± 6.7	19.00 ± 1.37
	Softmax memory ($\beta = 8$)	21.5 ± 10.7	13.93 ± 0.27
	Softmax memory ($\beta = 32$, sweep best)	31.9 ± 6.7	17.31 ± 1.48
	Finite-trained $T_{\text{train}} = 2$ (best finite)	$\mathbf{98.6 \pm 2.4}$	$\mathbf{21.08 \pm 0.57}$
	Equilibrium-trained (best finite)	19.4 ± 10.7	13.60 ± 0.19
	Equilibrium-trained (fixed point)	2.1 ± 0.0	12.29 ± 0.80
128	Cosine nearest neighbor	22.9 ± 4.4	16.13 ± 1.63
	Softmax memory ($\beta = 8$)	10.4 ± 3.3	13.53 ± 0.05
	Softmax memory ($\beta = 32$, sweep best)	21.1 ± 5.6	14.86 ± 1.49
	Finite-trained $T_{\text{train}} = 2$ (best finite)	$\mathbf{98.2 \pm 1.2}$	$\mathbf{19.74 \pm 0.16}$
	Equilibrium-trained (best finite)	2.6 ± 0.5	13.51 ± 0.09
	Equilibrium-trained (fixed point)	0.8 ± 0.0	12.47 ± 0.54

Table 7: Training-depth bridge. For finite training, the best finite evaluation depth is selected by the study's declared soft-memory-RMS PSNR criterion. The equilibrium row uses true fixed-point forward training with implicit differentiation; its "best finite" evaluation is reported separately from its fixed point. Values are mean ± sample standard deviation over three seeds. The supplied $P = 128$ bridge archive contains only $T_{\text{train}} = 2, 4$; deeper cells are intentionally left absent rather than interpolated.

Load	Training	Evaluation	Retrieval (%)	PSNR (dB)	Final loss
48	$T = 2$	$T = 3$	$\mathbf{98.6 \pm 2.4}$	21.08 ± 0.57	0.064 ± 0.016
	$T = 4$	$T = 4$	95.8 ± 4.2	$\mathbf{21.59 \pm 0.89}$	0.064 ± 0.015
	$T = 8$	$T = 7$	77.1 ± 25.6	19.03 ± 3.31	0.152 ± 0.100
	$T = 16$	$T = 10$	18.1 ± 13.9	14.27 ± 0.54	0.363 ± 0.022
	$T = 32$	$T = 3$	21.5 ± 16.7	13.79 ± 0.79	0.449 ± 0.058
	$T = 64$	$T = 2$	18.8 ± 9.5	13.57 ± 0.22	0.509 ± 0.028
	equilibrium	$T = 2$	19.4 ± 10.7	13.60 ± 0.19	0.509 ± 0.028
	equilibrium	fixed point	2.1 ± 0.0	12.29 ± 0.80	–
128	$T = 2$	$T = 2$	$\mathbf{98.2 \pm 1.2}$	$\mathbf{19.74 \pm 0.16}$	0.082 ± 0.001
	$T = 4$	$T = 4$	89.6 ± 0.9	18.44 ± 0.67	0.136 ± 0.023
	$T = 8, 16, 32, 64$	not supplied	–	–	–
	equilibrium	$T = 3$	2.6 ± 0.5	13.51 ± 0.09	0.511 ± 0.012
	equilibrium	fixed point	0.8 ± 0.0	12.47 ± 0.54	–

Interpretation boundary. Together these controls rule out two narrow explanations of the finite-depth result: neither direct nearest-memory lookup nor the reported fixed softmax-memory sweep reproduces the shallow recurrent gain, and simply making finite training deeper does not recover a strong equilibrium-like solution under the current optimization budget. They do *not* prove that finite-depth learning is intrinsically superior to equilibrium learning. The missing high-load deep-finite cells, stronger equilibrium optimizers, alternative parameterizations, and larger training budgets remain valid follow-ups.

J PHYSICAL NONIDEALITIES, CALIBRATION, AND CONDITIONING

The realism study perturbs the *executed feedback dynamics* rather than modifying the target energy after the fact. We group the model into four effects: finite iteration budget, deterministic attenuation/parameter error, repeated converter precision and additive noise, and slow drift with partial calibration.

Finite-time execution. The ideal reference uses 64 feedback steps. A shorter physical budget is represented by truncating the same update rule after a prescribed number of steps. This control separates solver-budget error from device nonideality: an eight-step solve has about 0.023 root-mean-square error (RMSE) to the 64-step reference before any additional hardware perturbation is applied.

Attenuation, coefficient error, and additive noise. Optical attenuation is modeled as

$$a = 10^{-L_{\mathrm{dB}}/20}, \tag{34}$$

so that the interaction/drive path is attenuated while the local state regularization remains in its reference normalization. Static coefficient error uses $\tilde{\theta} = \theta(1+\epsilon)$, and additive detector/converter noise is injected in normalized state units. In this small reference, 3 dB and 6 dB attenuation produce about 0.050 and 0.084 state RMSE, respectively, whereas 2% static coefficient error and 1% detector noise remain much smaller. The point is not that loss is universally benign, but that different physical errors act through different mechanisms: attenuation shifts the equilibrium continuously, while small static coefficient/noise perturbations leave this particular attractor largely intact.

Repeated converter precision. Feedback precision is represented by the uniform quantizer

$$Q_b(x) = \Delta\ \mathrm{round}(x/\Delta). \tag{35}$$

Because quantization is applied on every state/drive exchange, its error compounds through the feedback loop. Six-bit quantization produces about 0.090 RMSE, while four-bit quantization produces about 0.168 RMSE and enters a dead-zone/stall regime. These are sensitivity values for the modeled converter-mediated implementation, not universal effective-number-of-bits (ENOB) requirements and not claims about a fully analog or all-optical loop.

Drift and calibration. Slow coefficient drift with a compensated fraction ρ is modeled as

$$\tilde{\theta}_t = \theta\left[1 + (1-\rho)\delta_t\right]. \tag{36}$$

A modeled 5% drift produces about 8.9×10^{-4} state RMSE; compensating 90% of that drift reduces the error to about 6.8×10^{-5}, a 92.4% reduction. This separation is useful architecturally: fast precision loss must be addressed within the feedback path, whereas slow drift can be tracked by a calibration loop.

Conditioning of the backward problem. Forward physical stability does not by itself certify a fixed-point gradient. The convergence-matched spherical-memory stress in Appendix H holds the forward residual below 10^{-10} while varying the local fixed-point linearization: at $\rho(J_F) \approx 0.958$ and $\kappa(I - J_F) \approx 23.8$, a 32-iteration adjoint still has about 24.2% memory-gradient error. This supports the two-part requirement stated in Sec. 5: fixed-point differentiation needs both a converged forward state and an adequately solved, sufficiently regular adjoint system.

A separate conditioning sweep is retained as a provenance diagnostic, not as the convergence-matched isolation used in the main text. Its ill-conditioned cases also have non-negligible forward stationarity residuals, so forward nonconvergence and conditioning are confounded. To make that provenance visible, Fig. 22 now shows every retained higher-order seed rather than collapsing the right panel to five aggregate points.

Logical realization inventory. The higher-order reference uses five state channels, seven sparse programmable signed weights, 17 variable-fanout paths, ten energy-product junctions, 11 higher-order local-field product junctions, and ten analog-to-digital/digital-to-analog (ADC/DAC) samples per feedback iteration. The seven sparse terms are far fewer than the 625 entries in a dense fourth-order N^4 tensor for $N = 5$. These are structural/model-derived counts only. They do not establish fabricated area, speed, energy, thermal behavior, bandwidth, or throughput.

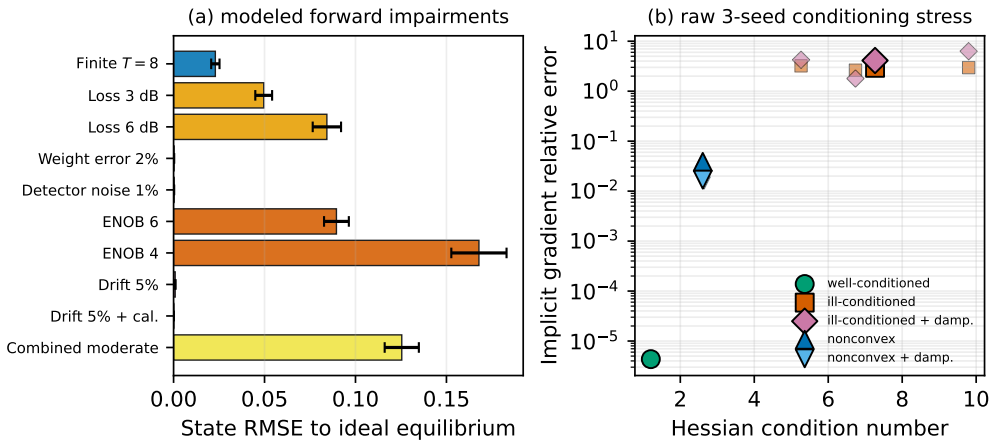

Figure 22: **Modeled photonic-realism and conditioning stress.** (a) Bars show mean ± sample standard deviation over three higher-order seeds. Repeated converter quantization is the dominant selected impairment in the small converter-mediated reference, whereas slow drift is strongly reduced by calibration. (b) All 15 retained higher-order seed points are shown; larger markers are scenario means. This diagnostic is *not* convergence-matched, so it should not be read as conditioning-only evidence. All quantities are simulation/model diagnostics rather than measured hardware performance.

P	factors	linear MAC-eq.	pow.	DAC	ADC	MZI cells	phase ctrl.
48	49,152	98,304	48	1,072	1,072	2,095,104	4,194,304
128	131,072	262,144	128	1,152	1,152	2,095,104	4,194,304

Table 8: **Structural scaling at the flagship latent dimension $N = 1024$ (per recurrent step, $p = 4$).** "Factors" is the learned PN memory-factor storage; "linear MAC-eq." is the logical digital equivalent of overlap plus backprojection ($2PN$); "pow." counts scalar $(p-1)$-power evaluations. DAC/ADC columns count the converter channels/samples for a converter-mediated overlap pass ($N \to P$) followed by a backprojection pass ($P \to N$). The MZI/phase columns are a conservative structural count for *two independent generic* SVD/Clements projection cores under the current zero-padded `MZILinear` lowering; one core is two N-mode Clements meshes plus an N-element singular-value bank. Time-reusing a projection core can halve these optical structural counts but introduces reconfiguration. These are not area, energy, latency, or throughput measurements, and they expose that the current generic dense lowering is not yet rank-aware.

Flagship-scale structural scaling. Table 8 separates the logical factorized workload from the current generic optical lowering at the 1024-dimensional latent scale used in the main image-memory experiment. Counts are per recurrent step and exclude the optional terminal soft radial readout, which performs a separate memory-bank similarity sweep. The logical factorization grows as $O(PN)$ rather than materializing an N^p interaction tensor. The current generic rectangular-matrix lowering, however, pads each projection to the larger mode count before singular-value-decomposition (SVD)/Clements synthesis, so its structural Mach–Zehnder-interferometer (MZI) count is an intentionally conservative upper bound rather than a claim of an optimized rank-aware photonic implementation.

K Uncertainty for Main-Text Result Tables

The main paper reports compact mean-only tables for readability. Tables 9–13 give the corresponding sample standard deviations across the three executed seeds wherever repeated-seed evidence is available. Boldface matches the task-relevant highlights used in the main paper; it does not indicate a statistical significance test.

Question	Pretrained-latent result with uncertainty
Learnable basin	$P = 24$, 55% occlusion: fourth-order retrieval $27.8 \pm 6.4\% \rightarrow 80.6 \pm 6.4\%$.
High-load interference	$P = 128$, 45% occlusion: pairwise retrieval $15.9 \pm 1.2\%$; fourth-order retrieval $89.6 \pm 0.9\%$.
Radial reconstruction	Soft memory-radius calibration improves mean PSNR by **+2.34 dB** at $P = 48$ and **+1.75 dB** at $P = 128$; only the retained aggregate gain is available for this row.

Table 9: **Uncertainty companion to Table 1.** Retrieval entries are mean $\pm$ sample standard deviation over three seeds.

P	Metric	$T = 4$	fixed
48	Retrieval (%)	95.8 ± 4.2	81.3 ± 17.8
48	PSNR (dB)	21.59 ± 0.89	20.27 ± 2.14
48	SSIM ↑	0.841 ± 0.025	0.787 ± 0.074
48	LPIPS ↓	0.144 ± 0.036	0.194 ± 0.101
128	Retrieval (%)	89.6 ± 0.9	32.3 ± 13.3
128	PSNR (dB)	18.44 ± 0.67	13.13 ± 1.54
128	SSIM ↑	0.718 ± 0.031	0.524 ± 0.059
128	LPIPS ↓	0.293 ± 0.040	0.526 ± 0.108

Table 10: **Uncertainty companion to Table 2.** Mean $\pm$ sample standard deviation over three seeds.

P	T_{tr}	best T	best PSNR	PSNR@T_{tr}
48	2	3	21.08 ± 0.57	21.06 ± 0.62
48	4	4	21.59 ± 0.89	21.59 ± 0.89
48	8	7	19.03 ± 3.31	19.03 ± 3.51
128	2	2	19.74 ± 0.16	19.74 ± 0.16
128	4	4	18.44 ± 0.67	18.44 ± 0.67
128	8	6	14.47 ± 0.07	14.28 ± 0.18

Table 11: **Uncertainty companion to Table 3.** PSNR is mean $\pm$ sample standard deviation over three seeds; T_{tr} abbreviates training depth.

Selector	Joint	Recall	Oracle 2nd
Frozen cosine	32.42 ± 0.00	61.91 ± 0.00	58.01 ± 0.00
Pairwise residual	36.78 ± 1.59	64.26 ± 1.36	58.50 ± 0.20
Third-order residual	28.84 ± 1.27	55.89 ± 1.05	50.10 ± 1.13

Table 12: **Uncertainty companion to Table 4.** Percent, mean $\pm$ sample standard deviation over three seeds.

Stress	Observed effect with uncertainty
Finite budget	8 versus 64 updates: 0.0230 ± 0.0023 state RMSE.
Repeated quantization	6-bit: 0.0895 ± 0.0068 RMSE; 4-bit: 0.1679 ± 0.0152 RMSE and a dead-zone/stall regime.
Slow drift + calibration	5% drift: $(8.89 \pm 2.29) \times 10^{-4}$ RMSE; 90% compensation reduces it to $(6.76 \pm 2.57) \times 10^{-5}$ (**92.4%** mean removal).

Table 13: **Uncertainty companion to Table 5.** Mean $\pm$ sample standard deviation over three higher-order seeds.

L Evidence Provenance and Statistical Notes

Quantitative manuscript claims were checked against retained internal experiment records, generated summaries, and source hashes rather than reconstructed from manuscript prose. The records distinguish deterministic reference simulations from experiments executed on the target software/model pipelines, and the compute-budget result remains linked internally to the aggregate summaries used for the paper-facing figures.

For six sign-sensitive effects in the frozen core evidence snapshot, the current three-seed paired 95% confidence intervals lie entirely on one side of zero: (i) fixed-45% versus mixed basin training at 55% corruption; (ii) fourth- versus second-order retrieval at $P = 128$, 45% corruption; (iii) Hotpot pairwise residual versus cosine; (iv) Hotpot third-order residual versus cosine; (v) the controlled $T = 32$ implicit-gradient margin to a 0.02 relative-error threshold; and (vi) drift-calibration improvement. The additional $P = 128$ finite-depth PSNR advantage is separately positive in every executed seed (+4.40/+5.42/+6.11 dB). These checks support the current qualitative conclusions but do not substitute for broader task/dataset replication. Most primary repeated-seed experiments use three seeds; their intervals and sample standard deviations should therefore be read as descriptive robustness checks rather than well-powered population-level significance tests.

Core figures remain tied to the frozen internal evidence snapshot. The finite-depth/fixed-point figure is generated from a separately retained internal compute-budget record with raw and aggregate results plus source hashes; this separation preserves the original evidence snapshot while supporting the authors' manuscript checks.

M REFERENCE GUIDE: ACRONYMS, DATASETS, AND MODELS

This section collects manuscript abbreviations and the external data/model assets used in the reported experiments. License labels follow the linked source or model-card statements used for this study; they are provenance notes rather than legal advice. The manuscript does not distribute raw third-party benchmark data or pretrained model weights.

M.1 ACRONYMS AND ABBREVIATIONS

Term	Meaning in this paper
ADC	analog-to-digital converter
BERT	Bidirectional Encoder Representations from Transformers
CC0	Creative Commons Zero public-domain dedication
CIFAR-10	Canadian Institute for Advanced Research 10-class image dataset
COCO	Microsoft Common Objects in Context
DAC	digital-to-analog converter
DEQ	deep equilibrium model
ENOB	effective number of bits
FLOP	floating-point operation; plural FLOPs
HIL	hardware-in-the-loop
IRCoT	Interleaving Retrieval with Chain-of-Thought
KKT	Karush–Kuhn–Tucker conditions
LPIPS	learned perceptual image patch similarity
MZI	Mach–Zehnder interferometer
MZM	Mach–Zehnder modulator
PNG	Portable Network Graphics
PSNR	peak signal-to-noise ratio
RAG	retrieval-augmented generation
RMS	root-mean-square
RMSNorm	root-mean-square normalization
RMSE	root-mean-square error
SD-VAE	Stable Diffusion variational autoencoder
SHA-256	256-bit Secure Hash Algorithm
SIR	signal-to-interference/crosstalk ratio
SSIM	structural similarity index
SVD	singular-value decomposition
TIA	transimpedance amplifier
VAE	variational autoencoder
VOA	variable optical attenuator

Table 14: **Acronyms and abbreviations used in the manuscript and Appendix.** Common mathematical symbols are defined locally at first use rather than repeated here.

M.2 DATASETS AND IMAGE SOURCES

Dataset/source	Use and license/provenance note
CIFAR-10 (Krizhevsky, 2009)	**Use:** primary image-memory benchmark; test images are resized to 128×128 and encoded by the frozen SD-VAE. **License:** the original distribution page provides the data and required citation but states no formal dataset license. Raw CIFAR-10 files are not redistributed.
COCO 2017 validation (Lin et al., 2014)	**Use:** natural-image resolution scaling at matched source identities and 128/256/512-pixel resolution. **License:** COCO annotations/site content are CC BY 4.0; image copyrights remain with the original image owners and individual image licenses are recorded in COCO metadata/terms. Raw images are not redistributed.
HotpotQA (distractor) (Yang et al., 2018)	**Use:** real multi-hop retrieval boundary test with train/validation examples and ten candidate documents per question. **License:** CC BY-SA 4.0, as stated by the HotpotQA project and dataset card.
Practical-photo set	**Use:** eight curated natural photographs for qualitative/calibration diagnostics. **License:** every selected Wikimedia Commons file page declares CC0 1.0. The experiment pipeline checks license metadata and records each source page, license string, dimensions, and SHA-256 in the authors' internal provenance records.
Synthetic memory/compatibility tasks	**Use:** controlled associative-memory and higher-order compatibility tests. **License/provenance:** no external dataset is used; instances are generated procedurally for the experiments.

Table 15: **Datasets and image sources.** Benchmark-level licenses are distinguished from per-item source licensing.

M.3 PRETRAINED MODELS AND LEARNED METRICS

Model identifier	Use and license/source note
`stabilityai/sd-vae-ft-mse`	**Use:** frozen AutoencoderKL codec for CIFAR-10, COCO, and practical-photo latent-memory experiments. **License:** MIT, per the model card.
`sentence-transformers/all-MiniLM-L6-v2`	**Use:** frozen semantic sentence encoder for the HotpotQA cosine-retrieval geometry. **License:** Apache-2.0, per the model card.
`google/flan-t5-small`	**Use:** frozen sequence-to-sequence answer model in the HotpotQA boundary experiment; the main residual-selector table does not depend on generated-answer metrics. **License:** Apache-2.0, per the model card.
LPIPS (AlexNet backbone) (Zhang et al., 2018)	**Use:** evaluation-only learned perceptual metric; the protocol uses the default AlexNet backbone. **License:** the LPIPS/PerceptualSimilarity package and calibration weights are distributed under the BSD 2-Clause license by the upstream project; upstream weights are not redistributed here.

Table 16: **Externally pretrained models and learned evaluation assets referenced by the manuscript evidence.** The manuscript reports identifiers and upstream license/source notes; pretrained weights are not distributed with the paper.

N Additional Limitations and Deferred Revisit Points

Scope boundary	What the current evidence does and does not establish
Measured hardware	Device, converter, calibration, and realization results are simulation/model diagnostics; no measured speed, energy, area, or throughput advantage is claimed.
Finite depth vs. equilibrium	The main fixed-point comparison extends a $T = 4$-trained model beyond its training depth. A true equilibrium-trained control is weak under the present 120 optimizer-update recipe, but this does not establish general finite-depth superiority; the budgets are not matched in wall-clock time or FLOPs.
Image/model scope	The strongest quantitative image results use one frozen SD-VAE family, modest memory banks, and mostly three seeds. The task recalls stored identities from corrupted cues rather than testing unseen-identity generalization; practical-photo results are descriptive diagnostics.
Finite-size scaling	Interaction-order capacity trends are empirical over the studied sizes and should not be read as asymptotic capacity theorems.
Retrieval generality	The controlled higher-order construction proves representational expressivity; the negative HotpotQA result shows that this does not imply a universal gain in pretrained semantic spaces.
Physical realization	Factorized products admit optical, optoelectronic, and mixed-signal implementations whose device-level tradeoffs require calibrated hardware-in-the-loop (HIL) or fabricated-device validation.

Table 17: **Detailed scope boundaries.** These caveats separate demonstrated properties from deferred hardware, scaling, and generality questions.

No measured hardware performance. The photonic experiments are software simulation and logical realization studies. We do not infer a measured speedup, energy efficiency, chip area, or throughput advantage until a qualified device/runtime experiment exists.

Finite-size memory scaling. The reported capacity exponents are empirical finite-size fits. They establish interaction-order trends in the studied regime, not asymptotic theorems.

Image-memory scope. The image experiments use one frozen pretrained SD-VAE family and comparatively small memory banks, with mostly three seeds. They test content-addressable recovery of stored identities from corrupted cues. They support mechanism claims about learned basins, interference, and continuous readout, not unseen-identity generalization or a competitive image-generation benchmark claim.

Higher-order retrieval revisit. The real HotpotQA third-order residual is negative and remains part of the paper. A future revisit should change the hypothesis materially (for example conditional second-hop factorization with a frozen first hop), not merely tune the failed residual harder.

Fixed-point differentiation. The constant-state-storage benefit assumes a fixed-point computation worth differentiating. It is not a drop-in replacement for finite-depth training: the forward fixed point must be the intended semantics, and both the forward and adjoint solves must converge. We do include a separately equilibrium-trained control; it is weak under the present 120 optimizer-update recipe, so it serves as an optimization boundary rather than evidence that equilibrium training is intrinsically inferior.