GENERATIVE ASSOCIATIVE MEMORY VIA EQUILIBRIUM MATCHING

Adrian Rodriguez[1]* **Benjamin Hoover**[2] **Yunhui Guo**[1] **Yilun Du**[3]

[1]The University of Texas at Dallas [2]Georgia Institute of Technology [3]Harvard University

ABSTRACT

Modern generative approaches like Equilibrium Matching (EqM) train models to approximate energy gradients, yet they typically rely on unconstrained architectures that lack intrinsic energy guarantees. We address this by training the Energy Transformer (ET), a Modern Hopfield Network where the forward pass explicitly performs gradient descent on a global energy function, using the EqM objective. This combination yields a Generative Associative Memory where the architecture strictly enforces the conservative vector field required by the training objective. We evaluate this framework on CIFAR-10, systematically exploring the trade-offs between architectural depth (stacked blocks) and temporal recurrence (iterative refinement within blocks). While a baseline single-layer model demonstrates feasibility (FID 79.72), we find that scaling to multi-block configurations drastically improves generation quality (FID 28.56), suggesting that hierarchical energy landscapes are essential for capturing complex image distributions. We further ablate design choices such as 2D positional encodings, energy minimization timesteps, and guidance strategies, offering a comprehensive analysis of how explicit associative memories can be scaled to competitive generative modeling.

1 INTRODUCTION

Generative modeling has seen rapid advancements through diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020) and flow-based methods (Lipman et al., 2023; Liu et al., 2022; Albergo & Vanden-Eijnden, 2023), which map simple noise distributions to complex data manifolds. Recent innovations, such as Equilibrium Matching (EqM) (Wang & Du, 2025), propose learning a time-invariant gradient field compatible with an implicit energy landscape, rather than the non-equilibrium dynamics of traditional diffusion/flow methods. EqM offers a compelling optimization-based sampling procedure where generation equates to gradient descent on a learned energy landscape. However, while EqM trains models to behave like Energy-Based Models (EBMs) (LeCun et al., 2006; Hopfield, 1982; Du & Mordatch, 2019), it typically employs standard feed-forward transformer backbones (e.g., SiT, DiT) that are not architecturally constrained to parameterize conservative vector fields.

Parallel to these developments, Modern Hopfield Networks (or Dense Associative Memories) (Krotov & Hopfield, 2016; Krotov et al., 2025) have established a rigorous theoretical link between the attention mechanism (Vaswani et al., 2023) in transformers and the energy update rules of associative memory Ramsauer et al. (2021). The Energy Transformer (ET) (Hoover et al., 2023) explicitly leverages this connection, designing an attention layer that strictly minimize a specific global energy function. While ET has demonstrated strong capabilities in pattern completion and graph anomaly detection, its application as a pure generative model, creating coherent images from noise rather than completing partial inputs, remains underexplored.

In this work, we propose a unified framework for Generative Associative Memory by training the Energy Transformer using the Equilibrium Matching objective. By replacing standard backbones with ET, we ensure that the learned vector field is mathematically guaranteed to be the gradient of a scalar energy potential, aligning the generative process with the attractor dynamics of associative

*Correspondence to: axr190042@utdallas.edu

memory retrieval. We evaluate this framework on CIFAR-10 (Krizhevsky & Hinton, 2009), treating the generation process as an energy minimization problem.

Crucially, we systematically explore the structural trade-offs required to scale associative memories for generation. While the original ET focuses on recurrent processing for completion, we analyze whether generative fidelity is better served by temporal recurrence (iterating a single block) or architectural depth (stacking multiple blocks). Our experiments demonstrate that while a single-layer recurrent model offers a feasible proof-of-concept, scaling to multi-block hierarchies is essential for improved generation. Additionally, we investigate whether learnable or 2D positional encoding makes a difference, showcasing that the extra parameter count of learnable positional encoding could be used elsewhere. This work bridges the gap between explicit associative memory architectures and modern generative training, offering a theoretically grounded path toward interpretable, energy-based generation.

2 Background and Method

Our framework unifies the architectural constraints of the Energy Transformer (ET) with the training objective of Equilibrium Matching. Unlike standard approaches that force a non-conservative architecture to approximate an energy gradient, we employ a backbone where the forward pass is, by definition, the gradient descent step of a global scalar energy function. Equilibrium Matching trains an equilibrium energy landscape by following the negative gradient of the energy towards clean data samples, which makes it a perfect fit for training the equilibrium energy of ET.

2.1 Energy Transformer Backbone

We adopt the Energy Transformer (ET) (Hoover et al., 2023), a Modern Hopfield Network architecture where the forward pass is mathematically defined as the explicit gradient descent of a global scalar energy function. Unlike standard transformers that process tokens through arbitrary feed-forward layers, ET treats the latent tokens as interacting particles relaxing towards a low-energy state.

Initialization and Positional Encoding. Following standard Vision Transformer practices (Dosovitskiy et al., 2021), we split the input image into non-overlapping patches and project them into a latent token space $x \in \mathbb{R}^{N \times D}$. To preserve spatial information essential for generation, we add positional information to these tokens. In our experiments, we explore both standard learnable parameters and fixed 2D sinusoidal encodings to determine if explicit spatial inductive bias aids the energy minimization process.

Energy-Constrained Layer Norm. To ensure the token dynamics strictly decrease a global energy, the standard Layer Normalization (Ba et al., 2016) is replaced with a specialized variant. This modification removes the standard element-wise affine parameters, retaining only a single learnable scale γ per tokens and bias δ_i for the token's elements. This constraint is crucial because it allows the normalization operation to be expressed as the partial derivative of a scalar Lagrangian function, a necessary condition for the system to represent a conservative vector field (Krotov et al., 2025).

Global Energy Function. The core of the architecture is the global energy function $E(x)$, which is the sum of two components, namely, energy-based attention and a Hopfield Network which acts as a multi-layer perceptron (MLP) from the standard transformer. Instead of the standard scaled dot-product, the attention mechanism is defined as the energy of a Modern Hopfield Network with softmax activation. Conceptually, this energy is the Log-Sum-Exp of the query-key correlations. Like standard transformers, this formulation uses multi-head attention, but unlike transformers, there is no head mixing.

Minimizing this energy corresponds to maximizing the alignment between tokens, dynamically routing information based on content similarity. The feed-forward component (MLP) is implemented as a Dense Associative Memory (Krotov & Hopfield, 2016) with a ReLU activation function. This creates a convex energy landscape (quadratic in the active regime) that encourages tokens to align with stored static memory patterns, effectively acting as a "clean-up" or "denoising" step within the energy descent (Hoover et al., 2023).

Forward Pass as Energy Minimization. The most distinct feature of ET is the energy computation in the forward pass. Instead of passing tokens through a sequence of layers, it computes the total energy of the energy-based attention and Hopfield Network, $E(x) = E^{\text{ATT}}(x) + E^{\text{HN}}(x)$, and uses automatic differentiation to calculate the gradient $\nabla_x E(x)$ with respect to the token states. The tokens are then updated via gradient descent:

$$x_{t+1} = x_t - \alpha \nabla_x E(x_t), \qquad t = 1, ..., T \tag{1}$$

where t is the current time step and $\alpha \in \mathbb{R}$ is the learning rate. This process is repeated for T time steps (recurrence). When there are multiple blocks (depth), each block computes T iterations of gradient descent, then passes x_T to the next block, and continues until the last block.

Conditioning and Projection. To enable class-conditional generation to guide the generation process to specific classes, we modulate the refined token output of the Energy Transformer using AdaLN-Zero (Peebles & Xie, 2022). The final energy-minimized token state is scaled by class-dependent embeddings before being linearly projected to the target dimension. This final projection predicts the equilibrium gradient field required by the Equilibrium Matching objective.

2.2 Equilibrium Matching

To train this architecture for image generation, we employ the Equilibrium Matching (EqM) objective (Wang & Du, 2025). EqM learns a time-invariant energy gradient field that vanishes on the data manifold and increases towards noise. During training, this objective shapes the energy landscape that ET will be parameterizing. After training, sampling is performed by traversing this energy landscape using gradient-based optimization on ET.

Training. Given a clean image $x \in \mathbb{R}^D$, Gaussian noise $\epsilon \in \mathbb{R}^D$, and an interpolation factor $\gamma \in [0, 1]$, we construct a perturbed sample $x_\gamma = \gamma x + (1 - \gamma)\epsilon$. The Energy Transformer f_θ is trained to predict a target energy gradient that points from data to noise. The objective function is defined as:

$$\mathcal{L}_{\text{EqM}} = (f_\theta(x_\gamma) - (\epsilon - x)c(\gamma))^2 \tag{2}$$

where $c(\gamma)$ is a weighting function that controls the gradient magnitude, ensuring the energy landscape has vanishing gradients at real samples ($c(1) = 0$). In this work, we use the truncated decay schedule:

$$c_{\text{trunc}}(\gamma) = \begin{cases} 1, & \gamma \le a \\ \frac{1-\gamma}{1-a}, & \gamma > a \end{cases}, \qquad a \in [0, 1) \tag{3}$$

This weighting function maintains a constant gradient magnitude far from the data distribution to encourage smooth transport, then decays linearly to 0 as it approaches the manifold. To control the overall scale, we apply a multiplier $\lambda \in \mathbb{R}^+$ such that $c(\gamma) = \lambda c_{\text{trunc}}(\gamma)$.

Sampling. Due to the equilibrium nature of EqM, where valid samples correspond to local minima of the energy landscape, we can employ various optimization-based samplers [1]. To keep it simple, we use standard gradient descent:

$$x_{k+1} = x_k - \eta f_\theta(x_k), \qquad k = 1, ..., K \tag{4}$$

where k denotes the current optimization step, $\eta \in \mathbb{R}^+$ is the step size (learning rate), and the gradient of the implicit energy landscape is given by the model output. This formulation allows for flexible inference compute since we can scale the number of sampling steps to refine sample quality arbitrarily. Moreover, we can keep track of the gradient magnitude and cut off computation when it reaches below a certain threshold.

3 Experiments

We evaluate our Generative Associative Memory framework on CIFAR-10 (32×32). Our experiments investigate the scalability of Energy Transformers (ET) trained with Equilibrium Matching (EqM) and analyze the trade-offs between parameter efficiency, architectural depth, and temporal recurrence.

[1] See Wang & Du (2025) for different $c(\gamma)$ choices and sampling strategies.

Table 1: Main results on CIFAR-10. **Depth** (L) is the primary driver of performance. While learned PE is slightly better for shallow models, **Fixed 2D PE** scales better at depth, achieving the best results with fewer parameters.

Model	**Blocks** (L)	**PE Type**	**Params**	**IS** ↑	**FID** ↓
Baselines					
EqM SiT-S/2 (Wang & Du, 2025)	12	Fixed 2D	32.00M	8.80	8.07
Ours (Energy Transformer)					
Shallow Recurrent	1	Learned	4.94M	4.77	74.65
Shallow Recurrent	1	Fixed 2D	4.74M	4.76	75.83
Medium Hierarchical	4	Learned	15.56M	6.83	32.04
Medium Hierarchical	4	Fixed 2D	15.36M	**6.95**	30.49
Deep Hierarchical	8	Learned	29.72M	6.62	31.39
Deep Hierarchical	8	Fixed 2D	29.52M	6.78	**28.56**

3.1 Experimental Setup

Implementation Details. We tokenize images into non-overlapping patches of size 2×2, resulting in sequence lengths of $N = 256$. Our backbone configurations use an embedding dimension of $D = 768$ and $H = 12$ attention heads. We compare configurations across varying depths ($L \in \{1, 2, 4, 8\}$) and recurrence steps (T). Models are trained for 1000 epochs using the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 0.0, and a batch size of 256. For the EqM objective, we use the truncated decay schedule with $a = 0.8$ and gradient multiplier $\lambda = 4.0$. During inference, we use Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) with a scale of 1.5 and perform 500 sampling steps.

Baselines. We compare our method against the original unconstrained Equilibrium Matching model (EqM) (Wang & Du, 2025). We train this EqM SiT-S/2 backbone (32M parameters) on CIFAR-10 with as similar settings as possible, which is comparable in scale to our largest models.

3.2 Main Results

Our results confirm that Energy Transformers can successfully function as generative models, with performance scaling dramatically with depth. Table 1 presents the quantitative results.

The Dominance of Depth. The single-block ET ($L = 1$) achieves an FID of 74.65. While feasible as a proof-of-concept, the model lacks capacity. Scaling depth yields massive improvements. Moving from 1 to 4 blocks improves FID from 74.65 to 30.49. Further scaling to 8 blocks pushes performance to 28.56. This suggests that the "energy capacity" of the model is fundamentally limited by the number of distinct energy functions (blocks) stacked hierarchically. Notably, the gain from 4 to 8 blocks is diminishing despite the parameter count nearly doubling. This implies that simply adding depth yields diminishing returns and further refinement of the energy function itself (e.g., the Hopfield Network or energy-based attention mechanism) may be required for future improvements.

Positional Encoding (PE). We analyze the impact of spatial inductive bias by comparing standard learned PE against fixed 2D Sinusoidal PE. For shallow models ($L = 1$), learned PE performs slightly better (FID 74.65 vs. 75.83). With limited capacity, the model benefits from the flexibility of adapting the spatial parameters. However, in deeper regimes ($L \geq 4$), the trend reverses. At 4 blocks, 2D PE overtakes learned PE (FID 30.49 vs. 32.04). At 8 blocks, the gap widens further (FID 28.56 vs. 31.39). This indicates that as the hierarchy deepens, a rigid spatial grid provides a consistent coordinate system that aids the optimization of complex energy landscapes. Furthermore, fixed PE reduces the parameter count by $\sim$0.2M per model, allowing resources to be allocated closer to the other components of the computational backbone.

Comparison to Generative Transformers. Despite comparable parameter counts, our Energy Transformer lags behind the unconstrained SiT backbone, suggesting that the structural constraints required for energy guarantees limit immediate expressivity. Specifically, the Hopfield Network utilizes a single projection, offering significantly less capacity than the two-layer MLPs found in standard Transformers. Furthermore, our current design lacks inter-head communication, a key driver of generalization in standard attention mechanisms. Closing this performance gap requires architectural innovations that respect energy minimization principles by implementing energy-preserving head mixing, integrating class-conditional energy terms at every block instead of the output projection of ET, and developing methods to impose hierarchy without fracturing the global energy landscape.

3.3 Ablation Studies

Temporal Recurrence (T) vs. Depth (L). We investigate whether strictly minimizing the energy within a layer (increasing recurrence steps T) is as effective as adding new layers. Using a 2-block model ($L = 2$), we varied $T \in \{1, 2, 3\}$. As shown in Table 2, the results are non-monotonic and show high variance (FID 45.32 $\rightarrow$ 56.61 $\rightarrow$ 42.81). Compared to the consistent gains from adding blocks, temporal recurrence appears to be a less reliable scaling lever for generation. Moreover, the computational cost of backpropagating through these recurrent gradient steps increases memory usage and training time significantly, especially when multiple blocks are introduced.

Table 2: Ablation on temporal recurrence (T) for the 2-block architecture. Increasing internal steps does not yield consistent improvements compared to depth.

Time Steps (T)	**Blocks (L)**	**IS** $\uparrow$	**FID** $\downarrow$
1	2	6.15	45.32
2	2	5.24	56.61
3	2	6.00	42.81

Guidance Strategy. We evaluate the impact of Classifier-Free Guidance (CFG). As shown in Table 3, removing guidance ($w = 0$) results in a performance drop (FID $74.65 \rightarrow 79.72$), confirming that guidance effectively sharpens the energy minima around class-conditional modes, as shown in Diffusion and Flow Matching models.

Table 3: Impact of classifier-free guidance (CFG) on the shallow recurrent model ($L = 1$).

Configuration	**IS** $\uparrow$	**FID** $\downarrow$
No Guidance ($w = 0.0$)	4.27	79.72
With Guidance ($w = 1.5$)	4.77	74.65

4 Conclusion

We have presented a unified framework for Generative Associative Memory by training Energy Transformers with Equilibrium Matching. Our extensive empirical analysis on CIFAR-10 reveals that architectural depth is the critical factor for scaling explicit energy models, vastly outweighing the benefits of temporal recurrence. While we establish that stable, deep energy-based generation is feasible, the remaining gap to unconstrained baselines highlights the trade-off between strict mathematical guarantees and architectural flexibility. Future work lies in bridging this gap by developing energy-compatible head mixing strategies, integrating hierarchical conditioning into the energy function itself, and exploring architectures that can capture multiscale hierarchy while preserving a single global energy potential.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27532–27559. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/57a9b97477b67936298489e3c1417b0a-Paper-Conference.pdf.

J 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. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.

Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. Modern methods in associative memory, 2025. URL https://arxiv.org/abs/2507.06211.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu-Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006. URL http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2021. URL https://arxiv.org/abs/2008.02217.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models, 2025. URL https://arxiv.org/abs/2510.02300.