ATTRACTOR DYNAMICS OF ENERGY TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy Transformer (ET) performs image completion by energy descent over image tokens, but its energy does not explicitly respect the translational structure of the image grid, limiting its ability to form continuous attractors. We introduce the Attractor Energy Transformer (AET), which incorporates a translation-invariant spatial kernel $\mathcal{K}$ on the toroidal token grid $\mathbb{T}^2$. We prove that the resulting spatial energy F_{sp} is invariant to cyclic translations, giving rise to a two-dimensional critical manifold of equivalent states. We further show that a rectangular boundary acts as a pinning perturbation that induces a restoring force along this manifold. On COCO2014 image completion, AET consistently outperforms ET, with the improvement persisting when the kernel topology is fixed. A sliding-window experiment provides further evidence for the theoretical analysis. The gap between the toroidal and rectangular topologies is smallest for an interior mask and larger at every offset that puts the masked region on the identified edge, consistent with the pinning predicted for the rectangle.

1 INTRODUCTION

A modern Hopfield network stores patterns as minima of an energy and retrieves them by descending that energy (Ramsauer et al., 2021). The Energy Transformer (ET) (Hoover et al., 2023) extends this principle to image completion by combining an attention energy with a Hopfield term and performing energy descent over image tokens. However, the tokens are arranged on a rectangular patch grid, while their interactions do not explicitly encode relative spatial displacement. As a result, the energy does not possess an explicit translational symmetry, and its attractors are not naturally organized into continuous families of translated states.

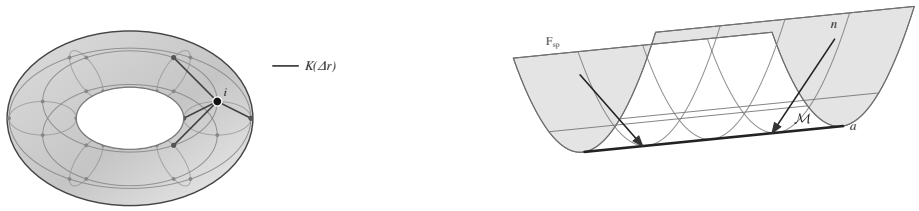
A continuous attractor network addresses this issue with a translation-invariant kernel $\mathcal{K}(\mathbf{r} - \mathbf{r}')$, so that shifted copies of a pattern have the same energy and a bump can stably occupy many positions (Wu et al., 2016). We introduce the Attractor Energy Transformer (AET), which adds this kernel on the torus $\mathbb{T}^2$. The implementation samples the torus on the image token grid. Each token therefore interacts with nearby locations according to their relative displacement, as sketched in Figure 1a. Translations along the two spatial directions are independent, giving a two-dimensional family of translated states.

On $\mathbb{T}^2$, the energy F_{sp} of that kernel is invariant under translations of the field, so the landscape is flat along each translation orbit. In the continuum analysis, any sufficiently regular critical field whose two spatial derivatives are linearly independent therefore generates a two-dimensional critical manifold. Figure 1b illustrates one translation coordinate of this manifold and shows nearby states descending toward it when the normal curvature is positive.

A rectangular boundary breaks this translation symmetry by removing interactions across the identified edges. We write the resulting spatial energy as $F_\varepsilon = F_{\mathrm{sp}} + \varepsilon P$, where P captures the boundary contribution. Let $(T_{\mathbf{a}}\mathbf{x})(\mathbf{r}) = \mathbf{x}(\mathbf{r} - \mathbf{a})$ denote translation on $\mathbb{T}^2$, with subtraction taken modulo the periods. For translations of a critical field $\mathbf{x}^*$, this perturbation induces the pinning potential $p(\mathbf{a}) = F_\varepsilon(T_{\mathbf{a}}\mathbf{x}^*) - F_\varepsilon(\mathbf{x}^*)$. Theorem 3 shows that its gradient governs the projected spatial motion along the former orbit. When the orbit is normally hyperbolic and the perturbation is sufficiently small, a nearby slow manifold persists and carries the corresponding reduced dynamics (Ságodi et al., 2024). The experiments below isolate this boundary effect.

A cyclic shift of the token field on trained weights traces a level set of F_{sp} on the torus, and sweeps out the pinning potential $p(\mathbf{a})$ on the rectangle. We evaluate image completion on COCO2014,

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.



(a) Kernel interactions on $\mathbb{T}^2$.

(b) Energy trough near the critical manifold.

Figure 1: Conceptual schematic of the spatial energy. (a) Selected token interactions under $\mathcal{K}(\Delta\mathbf{r})$. (b) In the normally attracting case, energy is flat along translations and rises in the normal direction. Arrows show descent. One translation coordinate is shown.

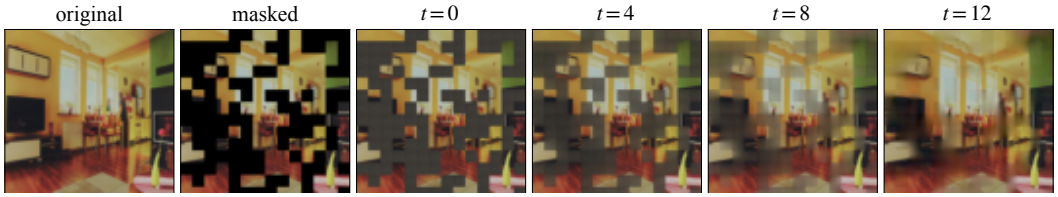


Figure 2: Example of iterative image completion by AET on $\mathbb{T}^2$. Starting from a randomly masked image, AET performs 12 update steps, restoring the observed patches after each step.

comparing AET with ET and using a rectangular kernel to examine boundary effects. Figure 2 illustrates iterative completion by AET, while a sliding-window probe compares the toroidal and rectangular kernels as the mask approaches the boundary.

In summary, our main contributions are as follows:

- We introduce AET, which adds a translation-invariant kernel on $\mathbb{T}^2$ and improves image completion over ET on COCO2014.
- We show that the spatial energy is invariant under translations. In the continuum, a sufficiently regular critical field whose two spatial derivatives are linearly independent generates a two-dimensional critical manifold, and rectangular boundaries induce a pinning potential along it.
- We measure this potential by cyclically translating token fields and show that the completion gap between the toroidal and rectangular kernels is smallest for an interior mask and larger whenever the mask reaches the edge.

2 Related work

2.1 Energy-based transformers and associative memory

Hopfield networks retrieve stored patterns by descending an energy landscape, with patterns represented by local minima (Hopfield, 1982). Dense associative memories extend this framework through higher-order interactions (Krotov & Hopfield, 2016), while modern Hopfield networks connect associative retrieval to transformer attention (Ramsauer et al., 2021). More broadly, energy-based learning formulates inference through a scalar energy and its descent (LeCun et al., 2006). The Energy Transformer (ET) combines attention and Hopfield terms into an energy over tokens and applies energy descent to image completion and graph tasks (Hoover et al., 2023). For images, ET operates on a rectangular patch grid without explicitly encoding relative spatial displacement through a translation-invariant interaction.

2.2 CONTINUOUS ATTRACTOR NEURAL NETWORKS

Continuous attractor neural networks (CANNs) use translation-invariant interactions to support families of states related by spatial shifts, allowing an activity bump to move without changing its energy (Amari, 1977; Wu et al., 2016). Ring models have been used to represent orientation and head direction (Ben-Yishai et al., 1995; Zhang, 1996), while line attractors also arise in models of gaze holding (Seung, 1996). Such continuous attractors are structurally fragile: small perturbations can split the neutral directions along the attractor. Ságodi et al. (2024) analyze the slow manifolds that can persist after such perturbations and their role in finite-time memory. This motivates studying how rectangular image boundaries perturb translational symmetry.

2.3 CANN-INSPIRED TRANSFORMER MODELS

Recent work brings CANN-inspired dynamics into transformers. CDAT combines energy-based transformer dynamics with excitation–inhibition modulation for graph tasks (Zhang et al., 2026). For image modeling, AET instead defines interactions through relative displacement on a two-dimensional torus. This gives the spatial energy two independent translation directions, and our continuum analysis characterizes the resulting orbit and the pinning induced by rectangular boundaries.

3 METHOD

3.1 TORUS TOPOLOGY

We formulate the spatial model on the rectangular flat torus

$$\mathbb{T}^2 = (\mathbb{R}/L_u\mathbb{Z}) \times (\mathbb{R}/L_v\mathbb{Z}), \tag{1}$$

where L_u and L_v are the periods in the two spatial directions, and $\mathbb{Z}$ denotes the integers. The quotient identifies points that differ by an integer multiple of the corresponding period, so opposite sides of the domain represent the same location. A rectangular domain, where these sides remain distinct, is used below to model a boundary.

Write a spatial location as $\mathbf{r} = (u, v)$ with $u \in [0, L_u)$ and $v \in [0, L_v)$. Translations $\mathbf{r} \mapsto \mathbf{r} + \mathbf{a}$ are understood modulo L_u and L_v. For two locations $\mathbf{r} = (u, v)$ and $\mathbf{r}' = (u', v')$, let $\Delta u = u - u'$ and $\Delta v = v - v'$ denote their coordinate displacements. Their squared toroidal distance is

$$d_{\mathbb{T}^2}(\mathbf{r}, \mathbf{r}')^2 = \min_{m,n\in\mathbb{Z}} \left[(\Delta u + mL_u)^2 + (\Delta v + nL_v)^2\right]. \tag{2}$$

The terms mL_u and nL_v account for wrapping around the two periodic directions.

We define the spatial interaction using a kernel that depends only on the relative displacement:

$$\mathcal{K}(\mathbf{r} - \mathbf{r}') = J_0 \exp\left(-\frac{d_{\mathbb{T}^2}(\mathbf{r}, \mathbf{r}')^2}{2a^2}\right) + \mathcal{W}_{\text{train}}(\mathbf{r} - \mathbf{r}'). \tag{3}$$

Here, J_0 controls the strength of the fixed Gaussian term, and a sets its spatial range. The learned term $\mathcal{W}_{\text{train}}$ is periodic with periods L_u and L_v. We take the full kernel to be even,

$$\mathcal{K}(\Delta u, \Delta v) = \mathcal{K}(-\Delta u, -\Delta v). \tag{4}$$

which makes the corresponding spatial interaction symmetric with respect to the L^2 pairing.

Because the kernel depends only on relative displacement on the torus, its structure is unchanged by translations of the field. This translation invariance forms the basis for the continuous attractor analysis in Section 4.

3.2 TOKEN DYNAMICS

The field evolves as

$$\tau \frac{\partial \mathbf{x}(\mathbf{r}, t)}{\partial t} = -\mathbf{x}(\mathbf{r}, t) + \int_{\mathbb{T}^2} \mathcal{K}(\mathbf{r} - \mathbf{r}')\, \mathbf{x}(\mathbf{r}', t)\, d\mathbf{r}' - \frac{\partial E}{\partial \mathbf{g}(\mathbf{r}, t)}, \tag{5}$$

where $\tau > 0$ sets the time scale and $\mathbf{g}$ is obtained by applying standard channel-wise layer normalization at each spatial location of $\mathbf{x}$. The first two terms constitute a linear interaction on the torus, while the last term is the gradient of the attention and Hopfield energy,

$$E = E^{\mathrm{ATT}} + E^{\mathrm{HN}}. \tag{6}$$

Queries and keys are computed from the normalized tokens using learned maps W^Q and W^K:

$$Q_{\alpha hC} = \sum_j W^Q_{\alpha hj}\, g_{jC}, \qquad K_{\alpha hB} = \sum_j W^K_{\alpha hj}\, g_{jB}. \tag{7}$$

For key B and query C, define the score and its partition function by

$$A_{hBC} = \sum_\alpha K_{\alpha hB} Q_{\alpha hC}, \qquad Z_{hC} = \sum_{B \neq C} \exp(\beta A_{hBC}). \tag{8}$$

Following Hoover et al. (2023), the attention energy is

$$E^{\mathrm{ATT}} = -\frac{1}{\beta} \sum_{h=1}^{H} \sum_{C=1}^{N} \log Z_{hC}. \tag{9}$$

The Hopfield term is a learned prior over prototype memories ξ_μ (Ramsauer et al., 2021), with a rectified activation as in Hoover et al. (2023):

$$E^{\mathrm{HN}} = -\sum_{B=1}^{N} \sum_{\mu} G\left(\sum_j \xi_{\mu j} g_{jB} \right), \tag{10}$$

where $G(z) = \frac{1}{2}\mathrm{ReLU}(z)^2$. This term lowers the energy when a token aligns with a stored prototype memory.

3.3 Energy landscape

The attention term E^{ATT} is the Helmholtz free energy of the Boltzmann distribution over keys for each query. That identification applies only to E^{ATT}. The spatial energy is[1]

$$F_{\mathrm{sp}}(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|^2 - \tfrac{1}{2}\langle \mathbf{x}, \mathcal{K} * \mathbf{x} \rangle. \tag{11}$$

Here, $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$ with the pairing above, and F_{sp} is the variational potential of the even interaction $\mathcal{K}$. Its variation, computed via equation 4, is $\mathbf{x} - \mathcal{K} * \mathbf{x}$, which matches the linear part of equation 5. As analyzed in Section 4, F_{sp} is the quantity whose gradient governs the spatial update, and it also underpins Theorems 1–3 and the force calculation in Appendix A.5. Its spatial gradient is derived in Appendix A.1.

In equation 6, the Energy Transformer term is evaluated on the normalized token field $\mathbf{g}$. Its update follows Hoover et al. (2023), using the derivative $\partial E / \partial \mathbf{g}$.

4 Theoretical analysis

The following results concern the continuous spatial model defined in Section 3. We treat the field as an element of $\mathcal{X} = L^2(\mathbb{T}^2; \mathbb{R}^D)$, using the pairing defined in Section 3.3. The token grid used in the experiments is a discretization of this model. For differentiating translations, we also use $H^1(\mathbb{T}^2; \mathbb{R}^D)$, the Sobolev space of fields whose weak first derivatives are square-integrable.

Write the spatial coordinate as $\mathbf{r} = (u, v)$, and let $\mathbf{a} = (a_u, a_v) \in \mathbb{T}^2$ denote a translation on the torus. The unperturbed spatial energy is invariant under translations. A boundary can break this symmetry and induce a restoring force along the translation orbit.

[1] We use the L^2 pairing $\langle \mathbf{u}, \mathbf{v} \rangle = \int_{\mathbb{T}^2} \mathbf{u}(\mathbf{r}) \cdot \mathbf{v}(\mathbf{r})\, d\mathbf{r}$, which reduces to $\sum_i \mathbf{u}_i \cdot \mathbf{v}_i$ on the token grid. The notation $(\mathcal{K} * \mathbf{x})(\mathbf{r})$ denotes the spatial interaction $\int_{\mathbb{T}^2} \mathcal{K}(\mathbf{r} - \mathbf{r}')\, \mathbf{x}(\mathbf{r}')\, d\mathbf{r}'$.

4.1 UNPERTURBED INVARIANCE

Theorem 1. *Suppose the spatial kernel K depends only on $\mathbf{r} - \mathbf{r}'$ on $\mathbb{T}^2$. Then*

$$F_{\mathrm{sp}}(T_{\mathbf{a}}\mathbf{x}) = F_{\mathrm{sp}}(\mathbf{x}) \qquad \textit{for all } \mathbf{a} \in \mathbb{T}^2. \tag{12}$$

For H^1 fields, translation invariance makes the functional gradient of F_{sp} orthogonal to the infinitesimal translation directions, as derived in Appendix A.2.

4.2 ATTRACTOR MANIFOLD

Theorem 2. *Suppose the spatial kernel is translation invariant on $\mathbb{T}^2$, so that it depends only on the relative displacement $\mathbf{r} - \mathbf{r}'$. Let $\mathbf{x}^* \in H^1(\mathbb{T}^2; \mathbb{R}^D)$ be a critical point of F_{sp}, and suppose that $\frac{\partial \mathbf{x}^*}{\partial u}$ and $\frac{\partial \mathbf{x}^*}{\partial v}$ are linearly independent. Then the translation orbit*

$$\mathcal{S}(\mathbf{x}^*) = \left\{ T_{\mathbf{a}}\mathbf{x}^* : \mathbf{a} \in \mathbb{T}^2 \right\} \tag{13}$$

is a two-dimensional critical manifold of F_{sp}. Let $H_{\mathrm{sp}}(\mathbf{x}^*)$ denote the Hessian of F_{sp} at $\mathbf{x}^*$. Its translation directions satisfy

$$H_{\mathrm{sp}}(\mathbf{x}^*)\frac{\partial \mathbf{x}^*}{\partial u} = 0, \qquad H_{\mathrm{sp}}(\mathbf{x}^*)\frac{\partial \mathbf{x}^*}{\partial v} = 0, \tag{14}$$

and therefore

$$T_{\mathbf{x}^*}\mathcal{S}(\mathbf{x}^*) \subseteq \ker H_{\mathrm{sp}}(\mathbf{x}^*). \tag{15}$$

If $H_{\mathrm{sp}}(\mathbf{x}^*)$ is positive definite on a complement of this tangent space with spectral gap $\mu > 0$, then $\mathcal{S}(\mathbf{x}^*)$ is normally attracting under the spatial gradient flow.

The proof is given in Appendix A.3.

4.3 SYMMETRY BREAKING

Let P be a perturbing energy that breaks translation invariance, and let $\varepsilon \in \mathbb{R}$ denote its amplitude. Model the perturbed spatial energy as

$$F_\varepsilon(\mathbf{x}) = F_{\mathrm{sp}}(\mathbf{x}) + \varepsilon P(\mathbf{x}). \tag{16}$$

Define the pinning potential as the change in the perturbed energy under a translation:

$$p(\mathbf{a}) = F_\varepsilon(T_{\mathbf{a}}\mathbf{x}^*) - F_\varepsilon(\mathbf{x}^*). \tag{17}$$

Theorem 3 (Pinning)**.** *Let $\mathcal{X} = L^2(\mathbb{T}^2; \mathbb{R}^D)$, and let $\mathbf{x}^* \in \mathcal{X}$ be a critical point of F_{sp} as in Theorem 2. Suppose $F_\varepsilon = F_{\mathrm{sp}} + \varepsilon P$, where P is twice differentiable. Parameterize the unperturbed orbit by shifting the reference critical field,*

$$\varphi(\mathbf{a}) = T_{\mathbf{a}}\mathbf{x}^*. \tag{18}$$

For $i, j \in \{u, v\}$, let $\mathbf{M}(\mathbf{a})$ be the tangent Gram matrix with entries

$$M_{ij}(\mathbf{a}) = \left\langle \frac{\partial \varphi}{\partial a_i}, \frac{\partial \varphi}{\partial a_j} \right\rangle. \tag{19}$$

Assume that this matrix is invertible and that $H_{\mathrm{sp}}(\mathbf{x}^*)$ is positive definite on a complement of $\mathrm{span}\{\frac{\partial \mathbf{x}^*}{\partial u}, \frac{\partial \mathbf{x}^*}{\partial v}\}$, with spectral gap $\mu > 0$. For the spatial gradient flow $\dot{\mathbf{x}} = -\delta F_\varepsilon / \delta \mathbf{x}$, the orthogonal projection onto the tangent space of the unperturbed orbit is

$$\mathbf{M}(\mathbf{a})\dot{\mathbf{a}} = -\nabla_{\mathbf{a}} p(\mathbf{a}). \tag{20}$$

This projected identity on the orbit is exact. At a critical point of p, the Jacobian of the projected flow is

$$\mathbf{J}(\mathbf{a}) = -\mathbf{M}(\mathbf{a})^{-1} H_p(\mathbf{a}). \tag{21}$$

where H_p is the Hessian of p. If H_p is positive definite there, the equilibrium is linearly stable for the projected flow.

In addition, if $|\varepsilon| < \varepsilon_0$ for sufficiently small $\varepsilon_0 = \varepsilon_0(\mu, P) > 0$, normal hyperbolicity ensures that a nearby slow manifold persists (Ságodi et al., 2024). On this slow manifold, the reduced dynamics agree with equation 20 up to $O(\varepsilon^2)$. When P is invariant under translations, p is constant, so the projected pinning force vanishes.

Table 1: Image completion on COCO2014 (mask ratio 50%). Adding the kernel $\mathcal{K}$ raises PSNR relative to ET, and the two topologies of $\mathcal{K}$ stay close in the interior. Entries are means $\pm$ standard deviation over five evaluations of the same checkpoint with different random masks.

	AET on $\mathbb{T}^2$	**AET, rectangle**	**ET**	**ViT**
Params	13M	13M	13M	13M
PSNR	$\mathbf{22.186 \pm 0.007}$	22.101 ± 0.007	21.297 ± 0.011	21.989 ± 0.004
SSIM	0.7472 ± 0.0001	$\mathbf{0.7484 \pm 0.0001}$	0.7182 ± 0.0001	0.7405 ± 0.0001

The proof is given in Appendix A.4. On the original orbit, the projected identity equation 20 is exact: the gradient of F_{sp} vanishes there, and the tangential component of the perturbed flow has the coordinate representation in equation 20. The slow-manifold statement describes the nearby full dynamics for sufficiently small perturbations. For larger perturbations, the orbit need not approximate full trajectories.

For a rectangular boundary, let $\mathcal{K}_\partial$ denote the same spatial interaction evaluated without periodic boundary conditions. The difference between the rectangular and periodic spatial energies is

$$P(\mathbf{x}) = \frac{1}{2} \langle \mathbf{x}, (\mathcal{K} - \mathcal{K}_\partial) * \mathbf{x} \rangle . \tag{22}$$

Thus, the rectangular case is a symmetry-breaking perturbation with $\varepsilon = 1$. The pinning potential $p(\mathbf{a})$ measures the change in rectangular spatial energy as the field shifts.

5 Experiments

The experiments test two claims. Continuous attractor dynamics should improve completion, and the two topologies should stay close in the interior and split at the boundary. Image completion tests the first claim, comparing AET with ET while both topologies of $\mathcal{K}$ act on a complete grid, so that the interior gap between them separates the dynamics from the topology. A cyclic shift of the entire token field evaluates the pinning potential $p(\mathbf{a})$ induced by P, without holding any tokens fixed. A rectangle with the same kernel support serves as the control. Its open sides define the boundary term in equation 22. A sliding window then moves the masked region onto the identified edge, so that the periodic interaction is exercised and the two topologies split.

5.1 Image completion

Image completion is the setting of the Energy Transformer (Hoover et al., 2023). The model trains on the COCO2014 dataset (Lin et al., 2014). Images are 224×224 with patch size 16. Both train and test use random patch masking at ratio 0.5, and masked PSNR is scored with the known pixels pasted back. The models are AET on $\mathbb{T}^2$, AET on the rectangle, ET, and a matched-width ViT (Dosovitskiy et al., 2021) trained with identical settings. Hyperparameters for this setup are in Appendix A.6.

Table 1 reports image completion on the COCO2014 test set. Both AET models lie above ET, which has no spatial kernel $\mathcal{K}(\mathbf{r} - \mathbf{r}')$. Adding the spatial term improves completion. They also lie above the ViT reference. AET on $\mathbb{T}^2$ and AET on the rectangle stay within $0.1\,\mathrm{dB}$ of each other. The interactions across the identified sides are exercised only near the outer boundary of the frame, so the small gap between the two topologies is not a gain from the torus, and the lift relative to ET is the spatial term. Figure 3 shows the same random mask completed by the four models of Table 1.

5.2 Cyclic shift of the token field

The cyclic roll is the primary measurement of $p(\mathbf{a})$. We form token fields from unmasked COCO2014 images (Lin et al., 2014) at 224×224 with patch size 16, then apply every cyclic translation of the 14×14 token grid. This uniform grid is the experimental sampling of the continuous torus. The same trained weights are evaluated twice, and only the topology of $\mathcal{K}$ changes. The roll moves the whole field, and the positional encoding moves with it, so the measurement isolates the boundary term P.

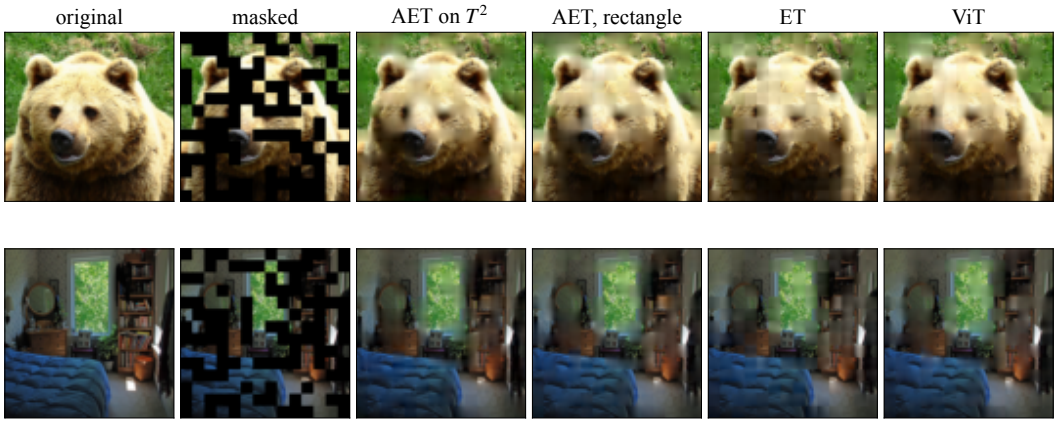


Figure 3: Image completion at mask ratio 50%, using the same images and the same random mask, with visible patches pasted back.

Table 2: Mean $|\Delta F_{\mathrm{sp}}|/|F_{\mathrm{sp}}|$ under cyclic rolls, i.e. the size of $p(\mathbf{a})$ relative to F_{sp}. Same trained weights within each row, and only the topology of $\mathcal{K}$ at evaluation changes. COCO2014, 224×224, patch size 16.

	eval torus	**eval rectangle**
trained on torus	4.1×10^{-8}	5.5×10^{-3}
trained on rectangle	2.6×10^{-8}	7.4×10^{-3}

Table 2 reports the mean of $|\Delta F_{\mathrm{sp}}|/|F_{\mathrm{sp}}|$ over 256 images and all 196 shifts, with $\Delta F_{\mathrm{sp}} = F_{\mathrm{sp}}(T_{\mathbf{a}}\mathbf{x}) - F_{\mathrm{sp}}(\mathbf{x})$ the change under a shift $\mathbf{a}$. On $\mathbb{T}^2$ the relative change sits at floating-point noise for weights trained on either topology, so $\nabla p = \mathbf{0}$, as Theorem 3 requires when P is invariant. On the rectangle the same weights move by about 10^{-3}, and the roll sweeps out the pinning potential $p(\mathbf{a})$ of the boundary term equation 22. The roll permutes the tokens, so the attention and Hopfield energies are exactly unchanged for either topology and the entire gap is the spatial term.

Appendix A.5 plots the restoring force along both axes, and the force on the rectangle is almost five orders of magnitude larger than on $\mathbb{T}^2$ and nearly isotropic in strength.

5.3 Sliding-window probe

A sliding window moves a fixed canvas mask toward the crop seam, where the boundary contribution P is expected to be larger. This probe measures how boundary topology affects completion as the masked region approaches an edge. At each offset, all models receive the same 128×128 crop per image. The canvas is 256×256, the patch size is 8, and the test split has 5000 images. Shifts are $(0, 0)$, the axis-aligned cuts $(0, 64)$ and $(64, 0)$, and the oblique cuts $(32, 48)$, $(40, 40)$, and $(48, 32)$.

We draw one block mask of ratio 0.4 in canvas coordinates, centered in the anchor crop W_0. The window moves to W_s, the crop shifted by offset s, so the masked region slides from the center toward the seam. On $\mathbb{T}^2$ the seam is joined to the opposite side, so a masked region at the edge still interacts with tokens by relative position, exactly as it did at the center. On the rectangle the seam is a boundary, and the masked region meets the edge.

Both windows are cut from the same canvas, so the overlapping pixels are the same content, and each is completed independently. We score PSNR against the ground truth on the overlap of W_0 and W_s, restricted to the mask, and Appendix A.7 sketches this geometry.

Table 3 reports those scores. The gap Δ is smallest at $(0, 0)$, where the masked region is interior, and larger at every offset that brings it to a window seam, from 0.56 dB at $(48, 32)$ to 2.06 dB at $(0, 64)$. For each offset, the toroidal and rectangular models receive the same token field, and only

Table 3: A constrained reading of P, scored as PSNR (dB) on the overlap restricted to the mask (COCO2014, block mask of ratio 0.4). Δ is AET on $\mathbb{T}^2$ minus AET on the rectangle. At $(0, 0)$ the masked region is interior, and the later shifts put it on the identified edge.

Model	$(0, 0)$	$(0, 64)$	$(64, 0)$	$(32, 48)$	$(40, 40)$	$(48, 32)$
AET on $\mathbb{T}^2$	17.08	15.13	15.39	14.77	14.73	14.63
AET, rectangle	16.75	13.07	14.52	13.65	13.89	14.07
Δ	0.33	2.06	0.87	1.12	0.84	0.56

the domain on which $\mathcal{K}$ is evaluated changes. Across offsets, the crop context also changes. The two axes differ in the size of that gap, and the kernel does not explain it, since $|\mathcal{K}|$ on the crop weights is isotropic to a few percent, and so is the unconstrained force in Appendix A.5.

6 Discussion

6.1 Summary

AET defines the spatial interaction of an Energy Transformer through a kernel $\mathcal{K}(\mathbf{r} - \mathbf{r}')$ on a torus, so that it depends only on relative displacement. The analysis in Section 4 gives the construction three properties. The spatial energy is unchanged by a cyclic translation of the token field. The orbit of a critical field whose two spatial derivatives are linearly independent is a two-dimensional critical manifold, and a rectangular boundary perturbs that symmetry so that the flow projected along the orbit follows the negative gradient of a pinning potential p. The COCO2014 experiments support both claims made for the construction. Adding $\mathcal{K}$ raises PSNR over ET and over the matched-width ViT reference, and the two topologies of $\mathcal{K}$ stay within 0.09 dB of each other under random masking, which is not a gain from the torus, since the interactions cross the identified sides only near the outer boundary of the frame. The cyclic roll separates the topologies cleanly: the relative change in F_{sp} sits at floating-point noise on $\mathbb{T}^2$ and at about 10^{-3} on the rectangle (Table 2). Under the constrained read of the sliding window, the completion gap between the two topologies is 0.33 dB for an interior mask and between 0.56 and 2.06 dB once the mask reaches a window seam, though it does not grow monotonically across offsets, and the two axes differ by more than the kernel accounts for.

6.2 Positioning

The closest work is CDAT, which applies energy-based transformer dynamics to graph tasks (Zhang et al., 2026). AET differs in that its coupling depends only on relative displacement on a two-dimensional torus, where the neighbourhood structure is uniform and translation symmetry is preserved. The analysis concerns the resulting critical manifold and how a rectangular boundary pins it.

6.3 Broader implications

The pinning potential describes how a translation-invariant spatial interaction behaves on a bounded domain, so boundary effects follow from the interaction rather than from a mechanism designed to break translation invariance.

On a normally attracting manifold, motion along the manifold costs no energy while perturbations off it decay. This is one way to state the smoothness commonly assumed of learned representations, though the analysis does not show that a trained kernel realizes such a manifold.

The torus construction also shows that a topological choice in the spatial part of the model has measurable consequences for the energy landscape. Whether the same effect appears for longer sequences or for video is not known.

6.4 Future work

The immediate next step is to connect the continuum conditions to the learned model by establishing when its spatial energy has nontrivial critical fields with a normal spectral gap and analyzing how the finite-step AET update behaves near those fields. Beyond that, we plan to derive discrete-to-continuum approximation bounds to test whether the account carries over to language modeling and to examine whether proximity to the critical manifold is useful as a training signal.

AI use statement

The authors used generative AI tools only for language polishing (grammar, clarity, and readability) and LaTeX formatting. All technical content, including mathematical claims, proofs, model implementation, analysis, figures, and the original writing, was conceived and produced by the authors. The authors reviewed and edited all AI-assisted text and take full responsibility for the accuracy and integrity of all content.

Reproducibility statement

Section 3 defines our notation and the spatial kernel $\mathcal{K}$, and Section 3.3 states the energy landscape. The experiments are described in Section 5, with hyperparameters in Appendix A.6 and the sliding-window geometry in Appendix A.7. Complete proofs of Theorems 1, 2, and 3, including the conditions on the critical field and the normal spectral gap, appear in Appendices A.1–A.4, and Appendix A.5 details the force computation along the roll. Code is included in the anonymous supplementary material.

References

Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

Rani Ben-Yishai, R. L. Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

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. In *International Conference on Learning Representations*, 2021.

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. doi: 10.52202/075280-1197. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/57a9b97477b67936298489e3c1417b0a-Paper-Conference.pdf.

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.

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

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.

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

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 K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter.

Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Ábel Ságodi, Guillermo Martín-Sánchez, Piotr Sokół, and Il Memming Park. Back to the continuous attractor. In *Advances in Neural Information Processing Systems*, 2024.

H. Sebastian Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344, 1996.

Si Wu, K. Y. Michael Wong, C. C. Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: Candidate of a canonical model for neural information representation. *F1000Research*, 5(F1000 Faculty Rev):156, 2016. doi: 10.12688/f1000research.7387.1. URL https://doi.org/10.12688/f1000research.7387.1.

Cheng Zhang, Minnan Luo, Zesheng Yang, Ming Li, Yong-Jin Liu, and Qinghua Zheng. Controlled dynamics attractor transformer. In *Forty-third International Conference on Machine Learning (ICML)*, 2026.

Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

A Appendix

A.1 From the ODE to the energy landscape

Start from equation 5,

$$\tau \frac{\partial \mathbf{x}}{\partial t} = -\mathbf{x} + \mathcal{K} * \mathbf{x} - \frac{\partial E}{\partial \mathbf{g}}, \tag{23}$$

with $E = E^{\mathrm{ATT}} + E^{\mathrm{HN}}$ and $\mathcal{K} * \mathbf{x}$ the spatial interaction on $\mathbb{T}^2$. The first two terms are the negative functional gradient of F_{sp}.

Quadratic interaction. By equation 4, the spatial interaction is symmetric under the pairing, $\langle \mathbf{x}, \mathcal{K} * \mathbf{v} \rangle = \langle \mathcal{K} * \mathbf{x}, \mathbf{v} \rangle$, so only the term linear in the test field survives in the variation of F_{sp}:

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} F_{\mathrm{sp}}(\mathbf{x} + \varepsilon \mathbf{v}) = \langle \mathbf{x} - \mathcal{K} * \mathbf{x},\ \mathbf{v} \rangle. \tag{24}$$

Hence $-\delta F_{\mathrm{sp}}/\delta \mathbf{x}$ is exactly the first two terms of equation 5.

A.2 Proof of Theorem 1

Fix $\mathbf{a} \in \mathbb{T}^2$ and set $(T_{\mathbf{a}}\mathbf{x})(\mathbf{r}) = \mathbf{x}(\mathbf{r} - \mathbf{a})$. Translation preserves the measure on $\mathbb{T}^2$, so $\|T_{\mathbf{a}}\mathbf{x}\|^2 = \|\mathbf{x}\|^2$. Substituting $\mathbf{s} = \mathbf{r}' - \mathbf{a}$ in the interaction gives $\mathcal{K} * T_{\mathbf{a}}\mathbf{x} = T_{\mathbf{a}}(\mathcal{K} * \mathbf{x})$, and the same substitution in the pairing gives $\langle T_{\mathbf{a}}\mathbf{x},\ \mathcal{K} * T_{\mathbf{a}}\mathbf{x} \rangle = \langle \mathbf{x},\ \mathcal{K} * \mathbf{x} \rangle$. Hence $F_{\mathrm{sp}}(T_{\mathbf{a}}\mathbf{x}) = F_{\mathrm{sp}}(\mathbf{x})$.

For $\mathbf{x} \in H^1(\mathbb{T}^2; \mathbb{R}^D)$ and a translation direction $\mathbf{e} = (e_u, e_v) \in \mathbb{R}^2$, differentiating equation 12 along $t\mathbf{e}$ at $t = 0$ gives

$$\left\langle \frac{\delta F_{\mathrm{sp}}}{\delta \mathbf{x}}, -e_u \frac{\partial \mathbf{x}}{\partial u} - e_v \frac{\partial \mathbf{x}}{\partial v} \right\rangle = 0. \tag{25}$$

A.3 Proof of Theorem 2

Let $\mathbf{x}^*$ be a critical point of F_{sp}. For any $\mathbf{a} \in \mathbb{T}^2$, Theorem 1 gives $F_{\mathrm{sp}}(T_{\mathbf{a}}(\mathbf{x}^* + \varepsilon \mathbf{w})) = F_{\mathrm{sp}}(\mathbf{x}^* + \varepsilon \mathbf{w})$ for every test field $\mathbf{w}$, so differentiating in ε at $\varepsilon = 0$ yields

$$\left\langle \delta F_{\mathrm{sp}}/\delta \mathbf{x}\big|_{T_{\mathbf{a}}\mathbf{x}^*},\ T_{\mathbf{a}}\mathbf{w} \right\rangle = 0.$$

Since this holds for every $\mathbf{w}$, every point of the orbit equation 13 is critical for F_{sp}.

Write $\varphi(\mathbf{a}) = T_{\mathbf{a}}\mathbf{x}^*$ for the orbit map $\mathbb{T}^2 \to \mathcal{X}$, and let $\mathbf{e} = (e_u, e_v) \in \mathbb{R}^2$ be an arbitrary translation direction. Then

$$\frac{d}{dt}\Big|_{t=0} \varphi(t\mathbf{e}) = \frac{d}{dt}\Big|_{t=0} \mathbf{x}^*(\mathbf{r} - t\mathbf{e}) = -e_u \frac{\partial \mathbf{x}^*}{\partial u} - e_v \frac{\partial \mathbf{x}^*}{\partial v}. \tag{26}$$

If the two spatial derivatives are linearly independent, $d\varphi_{\mathbf{0}}$ has rank two and $\mathcal{S}(\mathbf{x}^*)$ is a two-dimensional immersed submanifold of $\mathcal{X}$; if $T_{\mathbf{a}}\mathbf{x}^* = \mathbf{x}^*$ only at $\mathbf{a} = \mathbf{0}$, the stabilizer is trivial and $\mathcal{S}(\mathbf{x}^*)$ is an embedded copy of $\mathbb{T}^2$.

It remains to place the tangent space in $\ker H_{\mathrm{sp}}(\mathbf{x}^*)$. Identity equation 25 holds for every H^1 field, so substituting $\mathbf{x} = \mathbf{x}^* + \varepsilon\mathbf{w}$ and differentiating in ε at $\varepsilon = 0$ gives, by the product rule,

$$\big\langle H_{\mathrm{sp}}(\mathbf{x}^*)\,\mathbf{w},\, -e_u \frac{\partial \mathbf{x}^*}{\partial u} - e_v \frac{\partial \mathbf{x}^*}{\partial v} \big\rangle = 0 \qquad \text{for all } \mathbf{w}, \tag{27}$$

where $H_{\mathrm{sp}} = \delta^2 F_{\mathrm{sp}}/\delta\mathbf{x}^2$ and the second term of the product rule vanishes because $\mathbf{x}^*$ is critical. The Hessian of a scalar functional is symmetric under the same pairing, and since $\mathbf{e}$ is arbitrary, both spatial derivatives lie in the Hessian kernel, as stated in equation 14. The tangent space is $\mathrm{span}\{\frac{\partial \mathbf{x}^*}{\partial u}, \frac{\partial \mathbf{x}^*}{\partial v}\}$, and therefore $T_{\mathbf{x}^*}\mathcal{S}(\mathbf{x}^*) \subseteq \ker H_{\mathrm{sp}}(\mathbf{x}^*)$. If $H_{\mathrm{sp}}(\mathbf{x}^*)$ is positive definite on a complement of this plane, $\mathcal{S}(\mathbf{x}^*)$ is normally attracting for the spatial flow.

A.4 Proof of Theorem 3

Write $\mathbf{a} = (a_u, a_v) \in \mathbb{T}^2$, $\varphi(\mathbf{a}) = T_{\mathbf{a}}\mathbf{x}^*$, and $F_\varepsilon = F_{\mathrm{sp}} + \varepsilon P$. On the orbit, Theorem 1 gives

$$F_\varepsilon(\varphi(\mathbf{a})) = F_{\mathrm{sp}}(\mathbf{x}^*) + \varepsilon P(\mathbf{x}^*) + p(\mathbf{a}),$$

with p as in equation 17, so differentiating with respect to a_i for $i \in \{u, v\}$ yields

$$\frac{\partial p(\mathbf{a})}{\partial a_i} = \varepsilon \Big\langle \frac{\delta P}{\delta \mathbf{x}}\Big|_{\varphi(\mathbf{a})}, \frac{\partial \varphi}{\partial a_i} \Big\rangle. \tag{28}$$

Theorem 2 makes $\varphi(\mathbf{a})$ critical for F_{sp}, so $\delta F_\varepsilon/\delta\mathbf{x}|_{\varphi(\mathbf{a})} = \varepsilon\, \delta P/\delta\mathbf{x}|_{\varphi(\mathbf{a})}$. Pairing with $\frac{\partial \varphi}{\partial a_i}$ and using equation 28 gives

$$\big\langle \delta F_\varepsilon/\delta\mathbf{x}|_{\varphi(\mathbf{a})}, \frac{\partial \varphi}{\partial a_i} \big\rangle = \frac{\partial p(\mathbf{a})}{\partial a_i}.$$

The spatial gradient flow is $\dot{\mathbf{x}} = -\delta F_\varepsilon/\delta\mathbf{x}$. If the state remains on the orbit, $\dot{\mathbf{x}} = D\varphi(\mathbf{a})\,\dot{\mathbf{a}}$, and pairing against $\frac{\partial \varphi}{\partial a_i}$ produces equation 20. When $\varepsilon = 0$, or when P is invariant so that p is constant, this is equation 25.

The identity is exact on $\mathcal{S}(\mathbf{x}^*)$. If that manifold is normally hyperbolic, a compact slow manifold persists for $|\varepsilon| < \varepsilon_0(\mu, P)$ in a neighbourhood of the orbit (Ságodi et al., 2024). The state on that slow manifold is $O(\varepsilon)$ from $\varphi(\mathbf{a})$, so the reduced vector field is equation 20 plus $O(\varepsilon^2)$.

At a critical point of p, $\nabla p = \mathbf{0}$, so linearizing $\dot{\mathbf{a}} = -\mathbf{M}^{-1}\nabla p$ in the orbit coordinates produces the Jacobian equation 21. The derivative of $\mathbf{M}^{-1}$ drops out because it multiplies ∇p. The perturbation therefore lifts the degeneracy along the translation orbit, with leading-order dynamics given by this matrix. If H_p is positive definite at an isolated minimum of p, both eigenvalues of $\mathbf{J}$ are negative and that location is linearly stable for equation 20.

For the rectangle, F_{sp} with the periodic kernel $\mathcal{K}$ differs from the same quadratic with $\mathcal{K}_\partial$, the kernel taken on the rectangle, by equation 22, which is the pairing of the interactions that the rectangle drops. Setting $\varepsilon = 1$ puts the rectangle in the form equation 16. If $\mathcal{K}$ is local, those interactions are small and the perturbation is weak except near the identified edge.

A.5 Force along the roll

Figure 4 plots $p(\mathbf{a})$ and the restoring force $-\nabla_{\mathbf{a}} p$ along horizontal and vertical rolls. On $\mathbb{T}^2$ both axes stay on a level set up to numerics. The RMS of $-\nabla_{\mathbf{a}} p$ is $5.5{\times}10^{-4}$ and $5.3{\times}10^{-4}$ for weights trained on $\mathbb{T}^2$, and $2.6{\times}10^{-3}$ on both axes for weights trained on the rectangle. On a rectangle, the same weights give a pinning potential on both axes, and the force grows to RMS 36 (horizontal) and 40 (vertical) after $\mathbb{T}^2$ training, and 172 and 184 after rectangle training. That is almost five orders of magnitude above the values on $\mathbb{T}^2$, and nearly isotropic in strength. The shape of $p(\mathbf{a})$ still

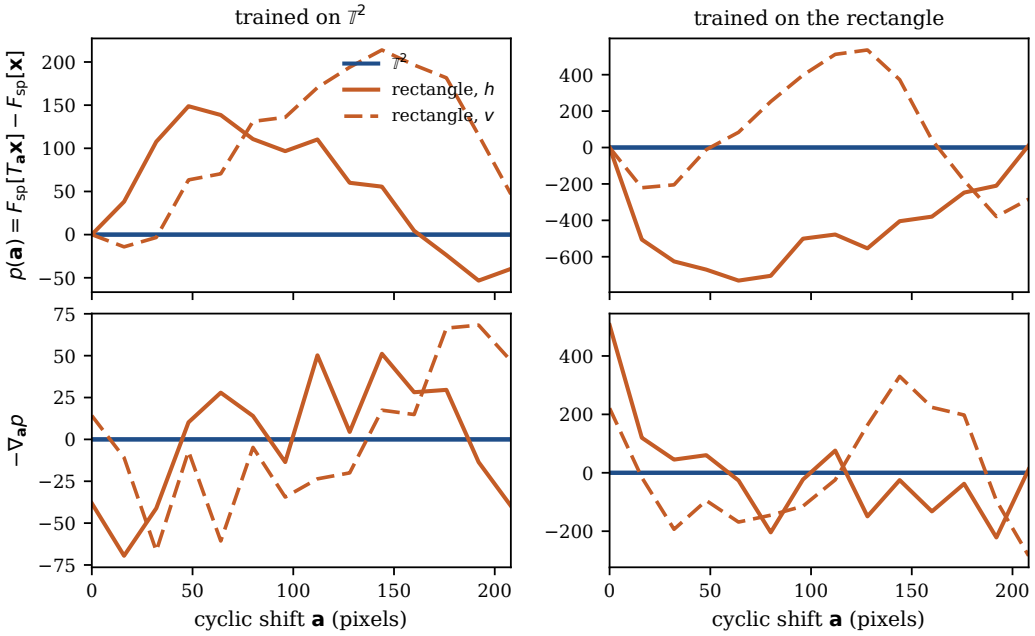


Figure 4: The top row plots the pinning potential $p(\mathbf{a})$ along cyclic shifts, and the bottom row the restoring force $-\nabla_{\mathbf{a}} p$ over one token step. The curves on $\mathbb{T}^2$ are plotted on the horizontal axis (the vertical $\mathbb{T}^2$ curve lies on top of it), and the rectangle curves are horizontal (solid) and vertical (dashed). Each column uses the same trained AET weights (left trained on $\mathbb{T}^2$, right trained on the rectangle), and within a column only the topology of $\mathcal{K}$ changes. Vertical scales are not shared. The forces on $\mathbb{T}^2$ sit at the numerical floor on both axes, and the rectangle forces have RMS 36–40 (left) and 172–184 (right). Mean over 256 COCO2014 images.

differs between axes. The difference between the axes is unconstrained and does not account for the sliding-window gap of Section 5.3, where the seams are read under constraints, not along the free $\nabla_{\mathbf{a}} p$. Each plotted value is a one-token-step difference of p, which estimates the projected restoring force of Theorem 3.

A.6 Image completion hyperparameters

AET for Table 1 uses the settings in Table 4. The two topologies share the same width and the same optimizer, and they differ only in how $\mathcal{K}$ is evaluated. Of the tokens that enter the loss, 10% stay visible so the Hopfield term sees real patches (Hoover et al., 2023).

Table 4: AET settings for image completion (Table 1).

Setting	**Value**	**Setting**	**Value**
Patch size	16	(β_1, β_2)	$(0.9, 0.99)$
Token dimension	770	Weight decay	0.05
Attention heads	12	Peak learning rate	5×10^{-4}
Key/Query dimension	64	Warmup	2 epochs
Update steps	12	Schedule	cosine to 5×10^{-7}
Step size	0.1	Gradient clip	1.0
Attention temperature	$\beta = 1/\sqrt{64}$	Epochs	100
Hopfield multiplier	4	Batch size	768
Optimizer	AdamW		

A.7 SLIDING-WINDOW GEOMETRY

Figure 5 is the layout of Section 5.3. The masked region is drawn once on the canvas, centered in W_0, and keeps those canvas coordinates in W_s: the window moves, the masked region does not. Both windows are 128×128 crops of a 256×256 canvas (patch size 8, block mask 0.4). The score is PSNR on the overlap restricted to the mask.

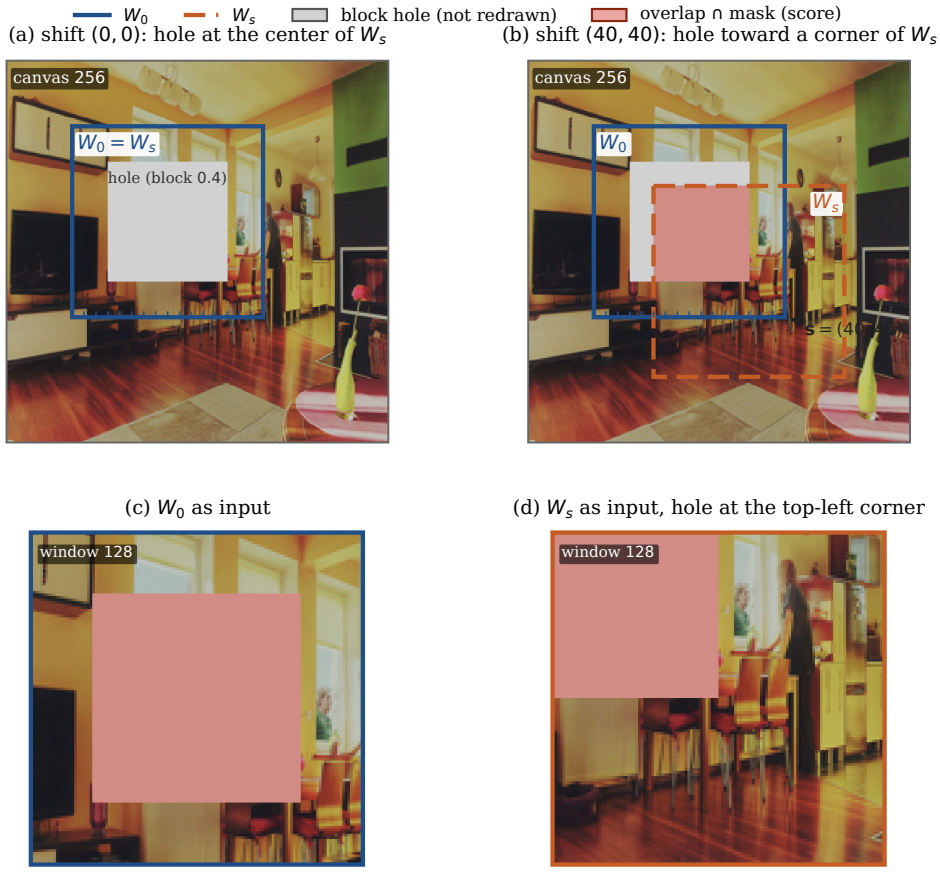


Figure 5: Sliding-window probe on a COCO2014 canvas. The masked region is drawn once on the canvas, centered in W_0, and keeps those canvas coordinates in W_s: the window moves, the masked region does not. Both windows are 128×128 crops of a 256×256 canvas (patch size 8, block mask of ratio 0.4). The score is PSNR on the overlap restricted to the mask (red). At $(0, 0)$ the masked region sits at the center of W_s, where P is weak, while at the oblique shift $(40, 40)$ it sits toward a corner, where P is large. Panels (c)–(d) show the 128×128 inputs seen by the model.