Balancing Fidelity and Diversity in Diffusion Models via Symmetric Attention Decomposition: Hopfield Perspective

Hyunmin Cho [1] **Woo Kyoung Han** [1] **Kyong Hwan Jin** [1 ‡]

Abstract

We characterize the pre-softmax attention matrix $\mathbf{QK}^\top$ in transformers as an associative memory matrix encoding pairwise associations between input features. By decomposing this matrix into its symmetric and skew-symmetric parts, we interpret the symmetric component as governing the structure of the *energy landscape*, and the skew-symmetric component as driving *circulation* on that landscape. Leveraging the energy formulation induced by the symmetric component, we derive Hopfield-style stability measures that quantify the stability of retrieved features. We observe meaningful correlations between Hopfield-style stability measures and the fidelity–diversity trade-offs in generation. Finally, we propose a controllable knob to modulate this trade-off by modifying the circulation of the underlying dynamics. Code is available at our Project Page.

1. Introduction

Diffusion models (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Labs et al., 2025) have become a leading paradigm for image generation. Their success is largely driven by the attention mechanism (Vaswani et al., 2017), which enables the integration of global context and long-range dependency modeling throughout the denoising process (Nichol & Dhariwal, 2021). While this global connectivity facilitates richer compositional associations that enhance novelty and variety (Zhang et al., 2019), it is simultaneously prone to causing spurious mixing of incompatible features, such as the blending of materials between two distinct objects (Oriyad et al., 2025). Crucially, distinguishing between such beneficial context integration and harmful semantic leakage remains non-trivial, as they share the same underlying mechanism. To address this ambiguity, our goal is to (i) *identify* when attention settles into spurious mixtures, and (ii) *control* this behavior to navigate the trade-off between coherent structure and diversity.

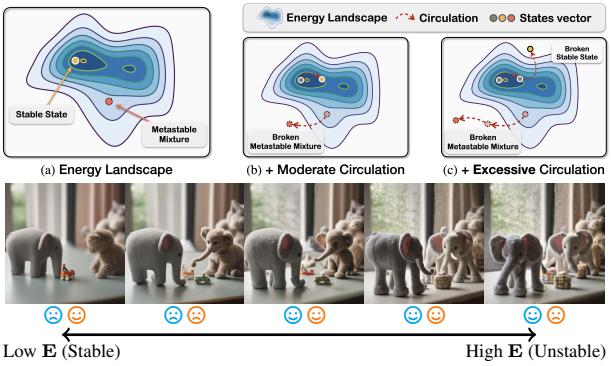

Figure 1. **Skew perturbation and the fidelity–diversity trade-off. Top**: We decompose $\mathbf{QK}^\top$ into symmetric (energy) and skew (circulation) parts. (a) The symmetric part gives stable but low-diversity retrieval. (b) Moderate skew perturbation breaks metastable mixtures while preserving stable states. (c) Excessive perturbation destabilizes even well-formed retrievals, producing artifacts. **Bottom**: Moderate skew perturbation improves diversity, but excessive perturbation causes hallucinations. ☺/☹ denote positive/negative *diversity*; ☺/☹ denote positive/negative *fidelity*.

The perspective of associative memory provides a principled lens on these challenges (Amari, 1972; Nakano, 1972; Little, 1974; Hopfield, 1982). Recent dense associative memory work further suggests that the choice of energy function can substantially reshape the landscape of local minima, even giving rise to additional emergent memories beyond stored patterns (Hoover et al., 2026). Building on the insight that transformer self-attention approximates the update rule of a modern Hopfield network (Ramsauer et al., 2021), we re-frame spurious mixing as entrapment in metastable states (local energy minima where the model settles on an incoherent combination of distinct patterns). However, standard analyses typically operate at a *token-wise* level, treating attention merely as a retrieval mechanism. This token-centric view restricts the capture of the rich interaction dynamics encoded in the attention matrix itself.

Furthermore, these interpretations often overlook the dynamical consequences of asymmetric association matrices. In recurrent associative memories, such asymmetry is known

[1]Department of Electrical Engineering, Korea University, Seoul, South Korea. Correspondence to: Kyong Hwan Jin <kyong_jin@korea.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

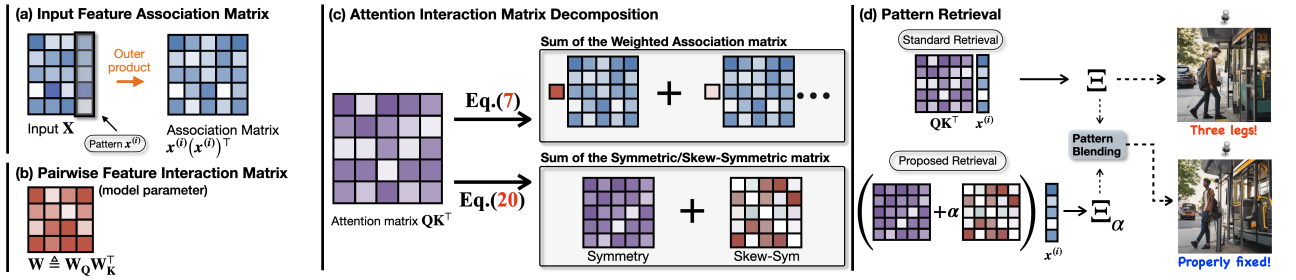

Figure 2. **Associative memory framework encoding pairwise feature interactions and its decomposition. (a)–(b)** We characterize the attention mechanism as an associative memory encoding *pairwise feature interactions*. (a) Viewing input features $\mathbf{X} \in \mathbb{R}^{L \times d_{\text{in}}}$ as a set of features $\boldsymbol{x}^{(i)} \in \mathbb{R}^L$, (b) the learned interaction matrix $\mathbf{W}$ encodes the association strength between these feature pairs. **(c)** The resulting attention matrix $\mathbf{QK}^\top$ can be decomposed into a *symmetric component* and a *skew-symmetric component*. The symmetric term defines a static *energy landscape* governing the stability of retrieved features, while the skew-symmetric term drives *circulation*, acting as a directional force. (d) Control via Circulation. Standard retrieval often settles into *metastable mixtures* (e.g., the incoherent 'Three legs' structure). We propose using the skew-symmetric component as a controllable knob. Amplifying this component injects circulation-driven drift to *perturb* the metastable state, *restoring structural coherence*.

to reshape the attractor structure, allowing non-fixed-point attractors such as limit cycles (Hwang et al., 2019). This structural property is crucial, as it induces circulation that helps perturb and *destabilize* metastable mixtures (Singh et al., 1995; Chengxiang et al., 2000).

In this work, we characterize the attention matrix $\mathbf{QK}^\top$ as a dynamic associative memory that encodes pairwise feature associations (Figure 2). Unlike prior token-level analyses, our view exposes the association structure that governs the mixing dynamics. Concretely, we decompose $\mathbf{QK}^\top$ into a symmetric and a skew component: the symmetric component defines a Hopfield-style *energy landscape*. In contrast, the skew-symmetric component drives *circulation*, acting as a directional force to perturb metastable states (Figure 1). This decomposition reveals that generation quality hinges on the balance between energy-based stability and circulation-driven dynamics. Leveraging this insight, we derive Hopfield-style stability measures, enabling us to *identify* metastable mixtures (Goal (i)). Finally, we exploit the skew-symmetric circulation as a tunable knob to *control* the retrieval process, facilitating the perturbation of metastable mixtures (Goal (ii)). To summarize our contributions:

- We establish an associative memory framework that encodes pairwise feature associations for the attention matrix and introduce a symmetric/skew-symmetric decomposition that disentangles energy-based stability from circulation-driven drift.
- Leveraging the symmetric component, we derive Hopfield-style stability measures that quantify the stability of retrieved features, demonstrating their correlation with the fidelity–diversity trade-off (Table 1).
- We propose the skew-symmetric component as a controllable 'circulation knob' for test-time intervention, which injects directional drift to perturb metastable mixtures and restore structural coherence (Table 3).

2. Related Work

Denoising diffusion models generate samples by learning to invert a progressive noising process, initially introduced in Sohl-Dickstein et al. (2015) and popularized as DDPMs in Ho et al. (2020). Subsequent formulations unify diffusion with score matching and continuous-time SDE/ODE views (Song & Ermon, 2019; Song et al., 2021), and related continuous-time objectives such as flow matching regress vector fields that transport noise to data (Lipman et al., 2023; Liu et al., 2023). Complementing these algorithmic formulations, recent theoretical works have reinterpreted these generative dynamics through the lens of associative memory, analyzing how diffusion trajectories disperse information and balance memorization with generalization (Ambrogioni, 2023; Hoover et al., 2023; Pham et al., 2025).

Associative memory networks are grounded in the classical Hopfield network, which defines an energy landscape over binary states. In these models, the system evolves to minimize energy based on the local field inputs (Amari, 1972; Nakano, 1972; Little, 1974; Hopfield, 1982). To overcome the storage limitations inherent to these classical pairwise-interaction models, Krotov & Hopfield (2016) introduced Dense Associative Memories (DAMs), which generalize the energy function by replacing the quadratic interaction term with a rapidly growing nonlinear function (e.g., polynomial or exponential) defined over the stored patterns. The gradient of this energy governs the update dynamics, resulting in sharper basins of attraction and a significantly higher storage capacity.

Asymmetric associative memories extend classical associative memory models beyond symmetric couplings by allowing directed interactions between stored states. Whereas symmetric Hopfield-type memories admit an energy-based interpretation with detailed balance, asymmetric interactions break this reversibility and can substantially alter retrieval dynamics (Peretto, 1984; Derrida et al., 1987; Chengxiang

et al., 2000). In the Hopfield model with random asymmetric interactions, the synaptic matrix $\mathbf{J}$ is decomposed into symmetric and asymmetric components:

$$\mathbf{J}_{ij} = \mathbf{J}^{\mathrm{s}}_{ij} + k\,\mathbf{J}^{\mathrm{as}}_{ij} \quad (i \neq j), \qquad \mathbf{J}^{\mathrm{as}}_{ij} = -\mathbf{J}^{\mathrm{as}}_{ji}, \tag{1}$$

where the symmetric part $\mathbf{J}^{\mathrm{s}}_{ij}$ is Hebbian and the skew-symmetric part $\mathbf{J}^{\mathrm{as}}_{ij}$ introduces asymmetry. Singh et al. (1995) analytically counted attractors in this setting and reported that adding an asymmetric component causes an exponential decrease in the total number of attractors, suggesting a mechanism for suppressing metastable states while preserving retrieval when the asymmetry is modest.

Attention mechanisms model interactions among a sequence of feature representations and have become a central building block of modern neural architectures (Vaswani et al., 2017). In language models, sequence positions typically correspond to text tokens (Brown et al., 2020; Touvron et al., 2023), whereas in vision-generative backbones they often correspond to image patches, or flattened latent positions. Attention explicitly parameterizes interactions among these positions, making it a natural target for controlling generation behavior through architectural design or inference-time modulation (Chen et al., 2024; Hong, 2024; Kim & Sim, 2025). These studies suggest that attention can serve as a handle for modulating generation dynamics.

Viewing attention as associative retrieval bridges memory-based dynamics and transformer attention (Vaswani et al., 2017). Ramsauer et al. (2021) formalize self-attention as a retrieval step in a continuous-state modern Hopfield network, where softmax implements an exponential Gibbs weighting over stored patterns. From a dynamical perspective, D'Amico & Negri (2024) reinterpret self-attention through an energy-based lens, emphasizing attractor-like behavior induced by attention updates. Complementing these activation-centric views, Bietti et al. (2023) offers a parameter-centric perspective, interpreting transformer *weight matrices* as associative memories that store embedding pairs as weighted outer products. However, these connections are typically framed either as token-level retrieval dynamics (Ramsauer et al., 2021; D'Amico & Negri, 2024) or as static memories residing in the parameters (Bietti et al., 2023). Consequently, the role of the underlying *feature interactions* instantiated in the $\mathbf{QK}^\top$ remains underexplored.

3. Hopfield Interpretation of Attention Matrix

To analyze the internal structure of attention (Vaswani et al., 2017), we view the input feature map $\mathbf{X} \in \mathbb{R}^{L\times d_{\mathrm{in}}}$ as a collection of d_{in} *real-valued* features, denoted by

$$\boldsymbol{x}^{(i)} \triangleq [\mathbf{X}]_{:,i} \in \mathbb{R}^L, \qquad i = 1, \ldots, d_{\mathrm{in}}. \tag{2}$$

Let the query and key projections be

$$\mathbf{Q} \triangleq \mathbf{X}\mathbf{W}_Q, \qquad \mathbf{K} \triangleq \mathbf{X}\mathbf{W}_K, \tag{3}$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_{\mathrm{in}}\times d_k}$. The pre-softmax attention matrix $\mathbf{QK}^\top$ is then

$$\mathbf{QK}^\top = \mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top. \tag{4}$$

For notational convenience, define the interaction weight matrix

$$\mathbf{W} \triangleq \mathbf{W}_Q\mathbf{W}_K^\top \in \mathbb{R}^{d_{\mathrm{in}}\times d_{\mathrm{in}}}, \tag{5}$$

so that the attention matrix admits the compact factorization

$$\mathbf{QK}^\top = \mathbf{X}\mathbf{W}\mathbf{X}^\top. \tag{6}$$

This expansion shows that $\mathbf{QK}^\top$ is a weighted superposition of rank-one outer products, analogous in form to classical Hopfield-style constructions (Personnaz et al., 1986):

$$\mathbf{QK}^\top = \sum_{i}^{d_{\mathrm{in}}} \underbrace{W_{ii}\,\boldsymbol{x}^{(i)}\big(\boldsymbol{x}^{(i)}\big)^\top}_{\text{self association}} + \sum_{i\neq j}^{d_{\mathrm{in}}} \underbrace{W_{ij}\,\boldsymbol{x}^{(i)}\big(\boldsymbol{x}^{(j)}\big)^\top}_{\text{hetero association}}. \tag{7}$$

This formulation establishes $\mathbf{QK}^\top$ as an associative memory encoding pairwise feature interactions, dynamically constructed from $\mathbf{X}$ as a weighted superposition of *self-association* and *hetero-association* terms (Figure 2c), with interaction strengths governed by the coefficient W_{ij}.

Hopfield retrieval dynamics. Given the attention matrix defined in Equation (8) as

$$\mathbf{M}(\mathbf{X}) \triangleq \mathbf{QK}^\top = \mathbf{X}\mathbf{W}\mathbf{X}^\top \in \mathbb{R}^{L\times L}, \tag{8}$$

and for each index $a \in \{1, \ldots, L\}$, the *local field* corresponds to the a-th row slice of $\mathbf{M}(\mathbf{X})$, viewed as a column vector:

$$\boldsymbol{m}_a(\mathbf{X}) \triangleq [\mathbf{M}(\mathbf{X})]^\top_{a,:} \in \mathbb{R}^L. \tag{9}$$

Since the local field vectors $\boldsymbol{m}_a(\mathbf{X})$ are real-valued and generally unbounded, we apply a normalization that (i) produces nonnegative, unit-sum mixing weights for *retrieval* and (ii) preserves the ranking induced by the local field. Accordingly, we map each local field vector $\boldsymbol{m}_a(\mathbf{X})$ to simplex-valued coefficients via $\phi : \mathbb{R}^L \to \boldsymbol{\Delta}^{L-1}$, where $\mathbf{1} \in \mathbb{R}^L$ denotes the all-ones vector and

$$\boldsymbol{\Delta}^{L-1} \triangleq \Big\{\boldsymbol{\kappa} \in \mathbb{R}^L : \boldsymbol{\kappa} \geq 0,\ \mathbf{1}^\top\boldsymbol{\kappa} = 1\Big\}, \tag{10}$$

yielding a normalized weighting over the L spatial positions. In the spirit of classical Hopfield retrieval (Hopfield, 1982), we further require ϕ to be monotone with respect to the local field: for any $\boldsymbol{m} \in \mathbb{R}^L$ and any j, k,

$$[\boldsymbol{m}]_j \geq [\boldsymbol{m}]_k \implies \big[\phi(\boldsymbol{m})\big]_j \geq \big[\phi(\boldsymbol{m})\big]_k. \tag{11}$$

which ensures that such normalization does not alter the preference ordering established by the energy landscape.

We extend ϕ row-wise to the matrix operator $\Phi : \mathbb{R}^{L\times L} \to \mathbb{R}^{L\times L}$ for any reference matrix $\mathbf{A} \in \mathbb{R}^{L\times L}$ via

$$\big[\Phi(\mathbf{A})\big]_{a,:} \triangleq \phi\big(\mathbf{A}^\top_{a,:}\big)^\top, \quad \text{for all } a \in \{1, \ldots, L\}, \tag{12}$$

and define the *Hopfield retrieval operator* (Ramsauer et al., 2021)

$$\mathbf{H}_{\mathbf{X}} \triangleq \Phi\big(\mathbf{M}(\mathbf{X})\big) = \Phi\big(\mathbf{X}\mathbf{W}\mathbf{X}^\top\big). \tag{13}$$

The retrieved features are then obtained by mixing input features according to $\mathbf{H}_{\mathbf{X}}$:

$$\Xi \triangleq \mathbf{H}_{\mathbf{X}}\,\mathbf{X} \in \mathbb{R}^{L\times d_{\text{in}}}, \quad \xi^{(i)} \triangleq [\Xi]_{:,i}. \tag{14}$$

Interpreting self-attention as Hopfield retrieval. A particular choice of Φ recovers the standard self-attention retrieval. In particular, with row-wise softmax,

$$\mathbf{H}_{\mathbf{X}} \triangleq \operatorname{softmax}\big(\mathbf{M}(\mathbf{X})\big), \tag{15}$$

the retrieved features Ξ become

$$\Xi \triangleq \mathbf{H}_{\mathbf{X}}\mathbf{X} = \operatorname{softmax}\big(\mathbf{X}\mathbf{W}\mathbf{X}^\top\big)\,\mathbf{X}. \tag{16}$$

Applying a value projection $\mathbf{W}_V \in \mathbb{R}^{d_{\text{in}}\times d_k}$ to the retrieved feature Ξ transforms the mixture into the output representation, yielding the standard update:

$$\operatorname{Attn}(\mathbf{X}) = \Xi\,\mathbf{W}_V. \tag{17}$$

4. Energy-based Stability Measures

Under a Hopfield-style lens, the attention mechanism can exhibit *metastable states* that are not captured by analyses that treat the attention matrix as symmetric, since $\mathbf{Q}\mathbf{K}^\top$ is generally asymmetric. To disentangle these effects, we decompose $\mathbf{Q}\mathbf{K}^\top$ into symmetric and skew components.

Decomposition of attention matrix. We begin by decomposing the attention matrix into symmetric and skew-symmetric components:

$$\mathbf{Q}\mathbf{K}^\top = \mathbf{M}_{\text{sym}}(\mathbf{X}) + \mathbf{M}_{\text{skew}}(\mathbf{X}), \text{ where}$$
$$\mathbf{M}_{\text{sym}}(\mathbf{X}) \triangleq \tfrac{\mathbf{Q}\mathbf{K}^\top + (\mathbf{Q}\mathbf{K}^\top)^\top}{2}, \quad \mathbf{M}_{\text{skew}}(\mathbf{X}) \triangleq \tfrac{\mathbf{Q}\mathbf{K}^\top - (\mathbf{Q}\mathbf{K}^\top)^\top}{2}. \tag{18}$$

Equivalently, it suffices to decompose the learned interaction weight matrix $\mathbf{W}$ as:

$$\mathbf{S} \triangleq \frac{\mathbf{W}+\mathbf{W}^\top}{2}, \quad \mathbf{N} \triangleq \frac{\mathbf{W}-\mathbf{W}^\top}{2}. \tag{19}$$

Substituting Equation (19) into Equation (6) yields the induced decomposition of the associative memory structure:

$$\mathbf{Q}\mathbf{K}^\top = \mathbf{X}\mathbf{W}\mathbf{X}^\top = \overbrace{\mathbf{X}\mathbf{S}\mathbf{X}^\top}^{\triangleq \mathbf{M}_{\text{sym}}(\mathbf{X})} + \overbrace{\mathbf{X}\mathbf{N}\mathbf{X}^\top}^{\triangleq \mathbf{M}_{\text{skew}}(\mathbf{X})}. \tag{20}$$

This decomposition allows us to separately analyze how the symmetric and skew components of the attention matrix contribute to the denoising process. Figure 3 qualitatively illustrates this separation: the symmetric component preserves global object-level structure, while the skew component captures fine-grained irregular details.

Energy of attention matrix. Since $\mathbf{M}_{\text{sym}}(\mathbf{X})$ is symmetric, it defines a valid Hopfield-style energy of features. For a *real-valued* feature $\xi \in \mathbb{R}^L$, we define the quadratic energy (Hopfield, 1982; Amit et al., 1985) induced by the symmetric component as:

$$E_{\mathbf{X}}(\xi) \triangleq -\frac{1}{2}\,\xi^\top \mathbf{M}_{\text{sym}}(\mathbf{X})\,\xi. \tag{21}$$

Lower energy (i.e., more negative $E_{\mathbf{X}}$) corresponds to a feature ξ that is more *strongly supported* by the associative memory constructed from $\mathbf{X}$ and the learned symmetric interaction rule $\mathbf{S}$.[1]

Figure 3. **Visualization of samples generated via decomposed components.** Samples generated through the sym. component encapsulate the underlying global structure, whereas those generated via the Skew component manifest fine-grained, irregular details.

In contrast, $\mathbf{M}_{\text{skew}}(\mathbf{X})$ is skew-symmetric and therefore contributes *no quadratic energy* for real-valued states:

$$\xi^\top \mathbf{M}_{\text{skew}}(\mathbf{X})\,\xi = (\mathbf{X}^\top\xi)^\top \mathbf{N}(\mathbf{X}^\top\xi) = 0 \tag{22}$$

since $\mathbf{N} = -\mathbf{N}^\top$ implies

$$\mathbf{u}^\top\mathbf{N}\mathbf{u} = 0, \quad \forall \mathbf{u} \in \mathbb{R}^{d_{\text{in}}}. \tag{23}$$

Hence, the skew-symmetric component serves to drive the circulation dynamics.

4.1. From Global Energy to Local Stability

Equation (21) provides a global measure quantifying how strongly a state ξ is supported by the symmetric interaction component $\mathbf{M}_{\text{sym}}(\mathbf{X})$. However, identifying metastable mixtures requires pinpointing *where* structural incoherence manifests across the L spatial positions; a single scalar energy is insufficient for this purpose.

We therefore complement the global energy with *local stability measures*. These metrics analyze the alignment between the state ξ and its driving local field, thereby exposing the localized conflicts that underlie metastability.

Local field and local stability. Importantly, the symmetric component $\mathbf{M}_{\text{sym}}(\mathbf{X})$ is itself a *weighted superposition* of rank-one feature associations,

$$\mathbf{M}_{\text{sym}}(\mathbf{X}) = \sum_{i=1}^{d_{\text{in}}}\sum_{j=1}^{d_{\text{in}}} S_{ij}\,\boldsymbol{x}^{(i)}\big(\boldsymbol{x}^{(j)}\big)^\top, \tag{24}$$

so its effect on a state ξ is mediated by the induced symmetric local field

$$\begin{aligned}\boldsymbol{h}_{\mathbf{X}}(\xi) &\triangleq \mathbf{M}_{\text{sym}}(\mathbf{X})\,\xi \in \mathbb{R}^L,\\ &= \sum_{i=1}^{d_{\text{in}}}\sum_{j=1}^{d_{\text{in}}} S_{ij}\,\boldsymbol{x}^{(i)}\,\big\langle \boldsymbol{x}^{(j)},\,\xi\big\rangle.\end{aligned} \tag{25}$$

[1] For notational clarity, we omit the $\sqrt{d_k}$ scaling in $\mathbf{Q}\mathbf{K}^\top$, which can be absorbed into $\mathbf{W}$ as an overall multiplicative factor.

Table 1. **Correlation between evaluation metrics and stability measures.** We report Spearman Rank correlation ρ between sample evaluation metrics (set: **A**) and three Hopfield-style stability measures computed from attention retrieval at each stage (set: **B**). Specifically, for each generated sample, we correlate the final external metric score against the internal stability values averaged over the retrieved features $\boldsymbol{\xi}$ within the specified block range. indicates that higher metric values co-occur with higher stability, while indicates an association with increased conflict or misalignment. SDXL UNet$_{[s-e]}$ denotes the layer range (s: start, e: end).

	Correlation ρ btw A and B (B:)	Down (SDXL UNet$_{[0-47]}$)			Mid (SDXL UNet$_{[48-67]}$)			Up (SDXL UNet$_{[68-139]}$)			All		
		$-E_{\mathbf{X}}$	$r_{\mathbf{X}}$	$\mathbf{Align}_{\mathbf{X}}$	$-E_{\mathbf{X}}$	$r_{\mathbf{X}}$	$\mathbf{Align}_{\mathbf{X}}$	$-E_{\mathbf{X}}$	$r_{\mathbf{X}}$	$\mathbf{Align}_{\mathbf{X}}$	$-E_{\mathbf{X}}$	$r_{\mathbf{X}}$	$\mathbf{Align}_{\mathbf{X}}$
A:	Aesthetic Score	**+0.181**	**-0.162**	**+0.151**	**+0.207**	**-0.229**	**+0.204**	**+0.255**	**-0.255**	**+0.280**	**+0.265**	**-0.273**	**+0.296**
	LPIPS Diversity	-0.074	**+0.192**	**-0.194**	**-0.336**	**+0.283**	**-0.250**	**-0.270**	**+0.237**	**-0.238**	**-0.279**	**+0.283**	**-0.297**
	CLIPScore	+0.040	**+0.155**	**-0.202**	**-0.158**	+0.088	-0.042	-0.006	-0.073	**+0.142**	-0.010	+0.030	-0.014
	ImageReward	**+0.129**	**-0.161**	**+0.146**	**-0.168**	**+0.102**	-0.090	**-0.122**	**+0.114**	**-0.192**	-0.074	+0.046	-0.074

Figure 4. **Qualitative comparison from samples sorted by Alignment Score.** For three prompts, we group baseline generations into *Stable* (top row) and *Unstable* (bottom row) subsets according to $\mathbf{Align}_{\mathbf{X}}$. Stable samples show coherent, object-centric structures, whereas unstable samples exhibit diverse but less coherent mixtures. White labels indicate the corresponding $\mathbf{Align}_{\mathbf{X}}$ values.

This expansion explicitly characterizes the mixing mechanism: the field $\boldsymbol{h}_{\mathbf{X}}(\xi)$ is generally a mixture of input feature $\{\boldsymbol{x}^{(i)}\}_{i=1}^{d_{\text{in}}}$, with weights determined by both the symmetric interaction coefficients S_{ij} and the alignment $\langle \boldsymbol{x}^{(j)}, \xi \rangle$. In other words, the response of a retrieved feature is determined by a *superposition of feature* associations supported by the symmetric component. A consistent superposition reinforces the current state across spatial locations, whereas incompatible associations produce coordinate-wise conflicts.

To pinpoint *where* this mixing manifests across the L spatial positions, we measure the *coordinate-wise agreement* between the current state ξ and its driving field $\boldsymbol{h}_{\mathbf{X}}(\xi)$:

$$\boldsymbol{\lambda}_{\mathbf{X}}(\xi) \triangleq \xi \odot \boldsymbol{h}_{\mathbf{X}}(\xi) \in \mathbb{R}^{L}, \tag{26}$$

under which the symmetric energy decomposes exactly as

$$E_{\mathbf{X}}(\xi) = -\frac{1}{2}\mathbf{1}^{\top}\boldsymbol{\lambda}_{\mathbf{X}}(\xi) = -\frac{1}{2}\sum_{a=1}^{L}[\boldsymbol{\lambda}_{\mathbf{X}}(\xi)]_a. \tag{27}$$

Thus, the scalar value $[\boldsymbol{\lambda}_{\mathbf{X}}(\xi)]_a$ indicates where the local field reinforces the current state ($[\boldsymbol{\lambda}_{\mathbf{X}}(\xi)]_a > 0$) versus where it conflicts with it ($[\boldsymbol{\lambda}_{\mathbf{X}}(\xi)]_a < 0$), providing a direct, spatially resolved view of retrieval stability. We summarize this conflict as the *instability fraction*

$$r_{\mathbf{X}}(\xi) \triangleq \frac{1}{L}\sum_{a=1}^{L}\mathbb{I}([\boldsymbol{\lambda}_{\mathbf{X}}(\xi)]_a < 0)\,. \tag{28}$$

Finally, to quantify the *global* directional agreement between ξ and its induced field, we define the alignment score via cosine similarity, which provides a scale-insensitive summary of whether the retrieved state and its induced field point in a consistent direction:

$$\mathbf{Align}_{\mathbf{X}}(\xi) \triangleq \cos(\xi, \boldsymbol{h}_{\mathbf{X}}(\xi)). \tag{29}$$

Figure 5 provides a schematic summary of these three stability measures.

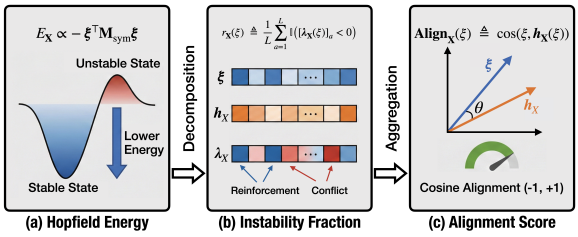

Figure 5. **Hopfield-style stability measures.** We characterize the stability of the retrieved state through three complementary lenses: (a) **Hopfield Energy** $E_{\mathbf{X}}$ measuring overall self-consistency; (b) **Instability Fraction** $r_{\mathbf{X}}$ identifying local reinforcement or conflict and (c) **Alignment Score** $\mathbf{Align}_{\mathbf{X}}$ measuring the global directional agreement between the ξ and its induced field.

4.2. Retrieval Stability and Perceptual Correlations

Having defined the Hopfield-style stability measures (Equations (21), (28) and (29)), we now examine how these measures relate to externally perceived sample quality and diversity across the generation process.

Evaluation metrics and protocol. We compare our Hopfield-style stability measures to three widely used, human-trained metrics that assess distinct dimensions of generation quality: the Aesthetic Score Predictor (Schuhmann et al., 2022) (visual preference), CLIPScore (Hessel et al., 2021) (text–image alignment), and ImageReward (Xu et al., 2023) (preference signals aggregated from curated human feedback). We also report LPIPS (Zhang et al., 2018) diversity as a reference-free proxy for perceptual variation across seeds (Lee et al., 2018). All results use SDXL (Podell et al., 2024) with classifier-free guidance (Ho & Salimans, 2021) $\omega = 5.0$ and 30 sampling steps, generating 1K random-seed samples for each of 10 COCO2014 (Lin et al., 2014) captions (10K total samples).

Table 2. **Perceptual quality stratification by Alignment Score.** For each baseline sample, we compute the Alignment Score $\mathbf{Align}_{\mathbf{X}}(\xi)$. We define **Stable/Unstable** regimes as the top/bottom 20% quantiles of $\mathrm{Align}_{\mathbf{X}}(\xi)$ over the full prompt set. For each external metric (ImageReward, AES, CLIPScore), we report the subset mean. The stable subset consistently achieves higher quality scores, whereas the unstable subset shows substantial degradation, suggesting that low stability indicates structural incoherence.

1K diverse prompts	$\mathbf{Align}_{\mathbf{X}}$	ImageReward	Aesthetic	CLIP
1K samples	0.669	0.546	5.643	0.263
▶ (High-Alignment)	0.690	0.692	5.906	0.270
Top-20% quantile	(+0.021)	(+0.146)	(+0.263)	(+0.004)
▶ (Low-Alignment)	0.650	-0.045	5.472	0.244
Bottom-20% quantile	(-0.019)	(-0.591)	(-0.171)	(-0.019)

Figure 6. **Qualitative visualization of the stability spectrum.** Baseline samples are sorted by their Alignment Score $\mathbf{Align}_{\mathbf{X}}(\xi)$. High-Alignment samples (Stable) exhibit *structural coherence* and consistent object-centric compositions. In contrast, Low-Alignment samples (Unstable) display *fragmented structures* and incompatible texture mixtures, indicating *metastable entrapment*.

Fidelity–Diversity trade-off via stability measures. Table 1 shows a consistent association between the proposed stability measures and external evaluation metrics. Stability indicators correlate positively with Aesthetic Score, while showing negative correlations with LPIPS diversity. This suggests that highly stable retrieval is associated with visually coherent generations, whereas lower stability is associated with greater perceptual variation across samples.

The qualitative results in Figure 4 provide a complementary view of this trend. Samples with high $\mathbf{Align}_{\mathbf{X}}(\xi)$ exhibit cleaner structure and fewer hallucinations, but often converge to similar viewpoints or repeated salient features. Conversely, samples with low $\mathbf{Align}_{\mathbf{X}}(\xi)$ show more diverse compositions, while also exhibiting increased structural inconsistencies and artifacts.

Generalization across diverse prompts. To validate this relationship under a broader distribution, we extend the analysis to 1,000 COCO2014 captions. Table 2 and Figure 6 confirm that stratifying baseline samples by the Alignment Score $\mathbf{Align}_{\mathbf{X}}(\xi)$ induces consistent shifts in external quality metrics. Specifically, *Stable* (high-alignment) samples exhibit strong structural coherence and object-centricity (often at the expense of diversity), yielding higher perceptual ratings. Conversely, *Unstable* (low-alignment) samples display broader visual variation but suffer from fragmented structures and incoherent feature mixtures, leading to significant quality degradation.

5. Methods

Building on the correlation established in Section 4, we propose a training-free mechanism that modulates the attention matrix $\mathbf{QK}^\top$. Our goal is to provide a tunable control over the Hopfield retrieval dynamics, exposing a controllable trade-off between *stability* and *circulation*.

Modulating circulation via the skew component. Inspired by classical observations on asymmetric Hopfield networks (Singh et al., 1995; Chengxiang et al., 2000), we utilize the skew-symmetric component of $\mathbf{QK}^\top$ as a lever to control the *circulation dynamics*. This approach is intrinsic to self-attention, since the retrieval operator is constructed from the full attention matrix, which inherently comprises a symmetric part and a skew part:

$$\mathbf{H}_{\mathbf{X}} \;=\; \Phi\big(\mathbf{X}\mathbf{S}\mathbf{X}^\top + \mathbf{X}\mathbf{N}\mathbf{X}^\top\big). \tag{30}$$

Since the retrieved features are obtained by applying the retrieval operator to this full matrix and mixing the input features (as in Equation (14)), controlling the skew component provides a direct handle to modulate $\mathbf{H}_{\mathbf{X}}$, thereby influencing the trajectory of the retrieved features $\{\xi^{(i)}\}_{i=1}^{d_{\text{in}}}$ without altering the underlying energy landscape.

5.1. Skew Scaling of the Attention Matrix

Recall the classical observation that increasing the asymmetric component leads to an exponential decrease in the total number of stable attractors (Singh et al., 1995). We leverage this property by *scaling* the skew interaction component within the $\mathbf{QK}^\top$. Specifically, we modulate the skew-induced term via a scalar control parameter α:

$$\mathbf{X}\mathbf{S}\mathbf{X}^\top + \mathbf{X}\mathbf{N}\mathbf{X}^\top \;\longrightarrow\; \mathbf{X}\mathbf{S}\mathbf{X}^\top + \alpha\,\mathbf{X}\mathbf{N}\mathbf{X}^\top, \tag{31}$$

which yields the perturbed retrieval operator

$$\begin{aligned} &\mathbf{H}_{\mathbf{X}}^{(\alpha)} \;\triangleq\; \Phi\big(\mathbf{X}\mathbf{S}\mathbf{X}^\top + \alpha\cdot\mathbf{X}\mathbf{N}\mathbf{X}^\top\big), \\ &\text{yielding} \quad \Xi_\alpha \;\triangleq\; \mathbf{H}_{\mathbf{X}}^{(\alpha)}\mathbf{X} \in \mathbb{R}^{L\times d_{\text{in}}}, \end{aligned} \tag{32}$$

where $\mathbf{H}_{\mathbf{X}}^{(\alpha)}$ denotes the retrieval operator in which the circulation is scaled by α.

5.2. Blending of Retrieved Features

The circulation-scaled operator $\mathbf{H}_{\mathbf{X}}^{(\alpha)}$ induces an alternative retrieval state Ξ_α that facilitates perturbation of *metastable mixtures* (Singh et al., 1995; Chengxiang et al., 2000), yet may introduce excessive *state wandering* if the circulation is too strong. To balance these dynamics, we compute the difference vector induced by the perturbation:

$$\Delta \;\triangleq\; \Xi_\alpha - \Xi, \tag{33}$$

and leverage it to form the blended retrieval:

$$\Xi_{\text{blended}} \;\triangleq\; \Xi + \beta\,\Delta, \tag{34}$$

followed by a *normalization step* that matches the baseline feature scale, ensuring that improvements reflect the blending dynamics rather than changes in feature magnitude.

Table 3. **Skew-Symmetric Attention Perturbation exhibits an operating-curve on average while selectively repairing failure subsets.** **(a)** MSCOCO–1K reports absolute mean scores over the full prompt set (1K), together with internal Hopfield-style stability measures ($-E_{\mathbf{X}}$, $r_{\mathbf{X}}$, $\mathbf{Align}_{\mathbf{X}}$), as we sweep control strengths. **(b)** Low-quality subset blocks report *paired* mean changes Δ relative to the baseline for the worst 20% baseline samples by each target verifier; gray entries denote side effects on non-target verifiers.

(a) All prompts (avg.).

Metrics	Baseline	Proposed Methods on SDXL 1K samples							
		$\alpha = 1.05$			$\alpha = 1.10$			$\alpha = 1.15$	
		$\beta = 5$	$\beta = 6$	$\beta = 7.5$	$\beta = 4$	$\beta = 5$	$\beta = 6$	$\beta = 3$	$\beta = 4$
MSCOCO–1K									
Aesthetic Score (↑)	*5.6436*	5.6657	5.6750	5.6834	5.6971	5.7172	**5.7345**	5.7042	5.7335
ImageReward (↑)	*0.5460*	**0.5756**	0.5573	0.5172	0.4992	0.4417	0.3533	0.4449	0.3383
CLIPScore (↑)	***0.2638***	0.2632	0.2626	0.2612	0.2605	0.2593	0.2573	0.2597	0.2576
$-E_{\mathbf{X}}$	3248.17	3174.35	3153.39	3118.93	3105.30	3060.02	2991.55	3086.97	2989.29
$r_{\mathbf{X}}$	*0.2314*	0.2398	0.2416	0.2443	0.2527	0.2416	0.2453	0.2489	0.2526
$\mathbf{Align}_{\mathbf{X}}$	*0.6693*	0.6540	0.6506	0.6456	0.6297	0.6504	0.6435	0.6366	0.6300

(b) Low-quality subset (Δ values against baseline).

Metrics	Baseline	Proposed Methods on Low-quality SDXL subset							
		$\alpha = 1.05$			$\alpha = 1.10$			$\alpha = 1.15$	
		$\beta = 5$	$\beta = 6$	$\beta = 7.5$	$\beta = 4$	$\beta = 5$	$\beta = 6$	$\beta = 3$	$\beta = 4$
▶ **MSCOCO–Low-quality subset**: worst 20% sorted by Aesthetic									
Δ Aesthetic	–	**+0.166**	**+0.183**	**+0.243**	**+0.273**	**+0.327**	**+0.331**	**+0.294**	**+0.353**
Δ ImageReward	–	+0.043	+0.030	-0.038	-0.015	-0.111	-0.112	-0.075	-0.123
Δ CLIPScore	–	+0.004	+0.001	-0.001	-0.002	-0.005	-0.006	-0.002	-0.009
▶ **MSCOCO–Low-quality subset**: worst 20% sorted by ImageReward									
Δ Aesthetic	–	+0.022	+0.004	+0.017	+0.012	+0.040	+0.006	+0.018	+0.010
Δ ImageReward	–	**+0.453**	**+0.526**	**+0.483**	**+0.518**	**+0.430**	**+0.454**	**+0.450**	**+0.382**
Δ CLIPScore	–	+0.004	+0.004	+0.002	+0.001	+0.001	+0.000	+0.002	+0.000
▶ **MSCOCO–Low-quality subset**: worst 20% sorted by CLIPScore									
Δ Aesthetic	–	+0.019	+0.004	-0.013	-0.015	+0.039	+0.035	+0.001	+0.044
Δ ImageReward	–	+0.116	+0.099	-0.005	+0.026	-0.043	-0.073	-0.051	-0.079
Δ CLIPScore	–	**+0.0065**	**+0.0070**	**+0.0075**	**+0.0070**	**+0.0067**	**+0.0072**	**+0.0078**	**+0.0075**

Figure 7. **Qualitative results of feature blending.** *Perturbation on unstable sample* (left): perturbation breaks spurious mixture configurations and yields a cleaner, *object-centric* reconstruction. *Perturbation on stable sample* (right): perturbation injects variation (texture/background/composition) and may introduce drift, illustrating the operating-point trade-off.

Here, α governs the intensity of the circulation perturbation, while β regulates the injection of these dynamics into the baseline retrieval. Together, α and β provide a controllable trade-off between *stability* and *diversity*.

Algorithm 1 Skew-symmetric perturbation blending

Require: Input $\mathbf{X}$, association components $\mathbf{M}_{\text{sym/skew}}(\mathbf{X})$
Require: Circulation scale α, injection scale β

Input: Initial query/state $\mathbf{X}$
#1. Standard Retrieval
$\mathbf{A} \leftarrow \mathbf{M}_{\text{sym}}(\mathbf{X}) + \mathbf{M}_{\text{skew}}(\mathbf{X})$
$\mathbf{\Xi} \leftarrow \Phi(\mathbf{A})\,\mathbf{X}$
#2. Circulation-Scaled Retrieval
$\mathbf{A}_\alpha \leftarrow \mathbf{M}_{\text{sym}}(\mathbf{X}) + \alpha \cdot \mathbf{M}_{\text{skew}}(\mathbf{X})$
$\mathbf{\Xi}_\alpha \leftarrow \Phi(\mathbf{A}_\alpha)\,\mathbf{X}$
#3. Perturbation via Blending
$\mathbf{\Delta} \leftarrow \mathbf{\Xi}_\alpha - \mathbf{\Xi}$
$\mathbf{\Xi}_{\text{blended}} \leftarrow \mathbf{\Xi} + \beta \cdot \mathbf{\Delta}$
Return: $\mathbf{\Xi}_{\text{blended}}$

6. Results & Discussion

For Table 3 and Figure 7, we apply the proposed method to self-attention retrieval within the UNet by replacing the baseline retrieval Ξ with the blended feature Ξ_{blended}. We follow the experimental protocol in Section 4.2: SDXL with ω=5.0 and 30 steps, using the same evaluation metrics. We evaluate on 1,000 COCO2014 prompts (Lin et al., 2014).

Regime-dependent impact of circulation injection. Recall that Table 2 and Figure 6 stratified baseline generations by the Alignment Score $\mathbf{Align}_{\mathbf{X}}(\xi)$ into a *Stable* regime and an *Unstable* regime. This separation implies that the utility of circulation injection depends on baseline stability. Therefore, we evaluate whether our method yields the *state-dependent correction* effect suggested by Figure 1: controlled circulation should resolve metastable states while potentially disrupting coherent configurations if excessive.

Table 3b supports this hypothesis through paired, case-conditional evaluation. On the lowest-performing 20% of

Table 4. **Stability disruption cost on High-performance baselines.** To complement the rectification results, we evaluate the impact of *circulation injection* on *already high-performing baseline samples*. For each metric, we define a *High-performance subset* by selecting the *top-20% quantile* of baseline samples and report the *paired* mean change on the same prompts.

Proposed Methods on High-quality SDXL subset	Baseline	$\alpha = 1.05$		
		$\beta = 5$	$\beta = 6$	$\beta = 7.5$
MSCOCO–high-performance subset: top-20% quantile of Aesthetic				
Δ Aesthetic	–	-0.104	-0.092	-0.094
MSCOCO–high-performance subset: top-20% quantile of ImageReward				
Δ ImageReward	–	-0.096	-0.131	-0.190
MSCOCO–high-performance subset: top-20% quantile of CLIPScore				
Δ CLIPScore	–	-0.0073	-0.0095	-0.0115

baseline samples under each metric, the proposed perturbation yields consistent improvements. Qualitatively, Figure 7 illustrates the same regime dependence: on an *Unstable* baseline, circulation injection suppresses incoherent mixture artifacts and produces a cleaner, object-centric reconstruction. In contrast, on a *Stable* baseline, it tends to inject local variation (texture, background, composition) which may lead to unintended deviation.

Performance trade-offs and cost on high-quality samples. Aggregate behavior (Table 3a) reflects a trade-off where increasing the circulation parameters (α, β) raises Aesthetic Score but can reduce ImageReward and CLIPScore. This state dependence becomes explicit on *High-Performance* baselines. Table 4 reports paired changes on the *top-20% quantile*, revealing that for samples already scoring highly under each external metric, circulation injection produces *degradation* of the corresponding metric. Combined with the substantial gains on the complementary *bottom-20% quantile* (Table 3b), this result suggests that circulation injection perturbs metastable states when baseline retrieval is trapped in poor configurations, yet can disrupt coherent, high-quality configurations when applied excessively.

6.1. Operating Regime of Asymmetric Retrieval Dynamics and Adaptive Control

The subset analyses in Tables 3 and 4 suggest that the same circulation perturbation can have different effects depending on the retrieval state. We further interpret this state dependence through the attractor-regime perspective of asymmetric neural networks. As a representative example, Hwang et al. (2019) study deterministic recurrent neural networks and show that the *degree of symmetry* in the connectivity controls the structure of attractors, including fixed points and limit cycles. They quantify this degree of symmetry as

$$\eta_{\mathrm{H}} = \langle J_{ij} J_{ji} \rangle / \langle J_{ij}^2 \rangle, \tag{35}$$

where $\eta_{\mathrm{H}} = 1$ corresponds to symmetric connectivity, while smaller values indicate increasing asymmetry. In symmetric or near-symmetric regimes, the dynamics are more closely tied to fixed-point-like retrieval, whereas increasing asymmetry can induce cyclic attractors with longer periods. This perspective motivates treating the sym–skew balance not merely as a static property of $\mathbf{QK}^\top$, but as an operating parameter that can shift the retrieval dynamics between stable convergence and circulation-driven exploration.

Table 5. **Functional symmetry regimes on SDXL.** We stratify samples by ImageReward and report the realized symmetry index η_M. Low-performance samples occupy a slightly lower-symmetry regime than average and high-performance samples. Circulation control moves low-performance samples toward this band, but further perturbing already high-performance samples pushes them away from their favorable operating point and reduces quality.

Skew perturbation	**ImageReward** ↑	$\boldsymbol{\eta_M}$
MSCOCO–low-performance subset: low-20% quantile of ImageReward:		
×	-1.289	0.655
✓	$-0.819_{\Delta:\,+0.470}$	0.659
MSCOCO–average-performance subset: avg-20% quantile of ImageReward:		
×	0.512	0.663
MSCOCO–high-performance subset: top-20% quantile of ImageReward:		
×	1.814	0.666
✓	$1.716_{\Delta:\,-0.098}$	0.669

Functional symmetry of realized attention. To quantify this operating regime at the sample level, we measure the symmetry–circulation balance of the realized attention interaction. Given the decomposed attention matrix (Equation (18)), we define the functional symmetry index

$$\eta_{\mathbf{M}}(\mathbf{X}) = \frac{\|\mathbf{M}_{\mathrm{sym}}(\mathbf{X})\|_F^2 - \|\mathbf{M}_{\mathrm{skew}}(\mathbf{X})\|_F^2}{\|\mathbf{M}_{\mathrm{sym}}(\mathbf{X})\|_F^2 + \|\mathbf{M}_{\mathrm{skew}}(\mathbf{X})\|_F^2}. \tag{36}$$

The index is close to 1 when the realized attention interaction is dominated by the symmetric component and decreases (e.g., $\eta_{\mathbf{M}} \to -1$) as the skew-symmetric component becomes stronger. Thus, $\eta_{\mathbf{M}}(\mathbf{X})$ summarizes the relative dominance of energy-supported retrieval and circulation-driven dynamics for the current retrieval state.

Functional symmetry band. We next examine whether $\eta_{\mathbf{M}}(\mathbf{X})$ reflects the state-dependent behavior observed in Tables 3 and 4. As shown in Table 5, low-IR samples occupy a slightly lower-symmetry regime than average and high-performing samples. Applying circulation control to the low-performance subset improves IR and moves $\eta_{\mathbf{M}}(\mathbf{X})$ toward the average/high-quality band. However, applying the same perturbation to the high-performance subset increases $\eta_{\mathbf{M}}(\mathbf{X})$ further while decreasing IR.

Thus, the effect of circulation control can be viewed as an under- or over-shift along this operating coordinate: perturbation is beneficial when it moves low-performance retrievals toward the preferred band, but can degrade samples that are already near a favorable regime.

Adaptive circulation control. The operating-band behavior above suggests that a fixed (α, β) is inherently state-dependent. We therefore consider a lightweight adaptive variant that uses $\eta_{\mathbf{M}}(\mathbf{X})$ to modulate the additional circulation injected at test time. In practice, $\eta_{\mathbf{M}}(\mathbf{X})$ is computed per sample and attention head, and we use a single shared scalar for the corresponding attention call:

$$\bar{\eta}_{\mathbf{M}} \leftarrow \mathrm{Agg}_{b,h}[\eta_{\mathbf{M}}(\mathbf{X})_{b,h}], \tag{37}$$

Table 6. **Adaptive circulation control.** We evaluate a lightweight adaptive variant on 350 COCO samples. For the *moderate* setting, adaptive control preserves the gains of static perturbation across preference metrics while maintaining CLIP and Pick. For the *excessive* setting, static perturbation substantially degrades all metrics, whereas adaptive control recovers from this collapse and improves over the baseline on IR, HPS, and AES.

Method	IR ↑	CLIP ↑	HPS ↑	AES ↑	Pick ↑
COCO baseline SDXL subset					
	0.487	0.264	0.2695	5.64	0.224
COCO *moderate* skew-perturbation subset: $(\alpha, \beta) = (1.05, 3)$					
Static	0.546	0.262	0.2730	5.66	0.224
Adaptive	0.522	0.264	0.2723	5.64	0.224
COCO *excessive* skew-perturbation subset: $(\alpha, \beta) = (1.20, 5)$					
Static	-1.486	0.207	0.1570	5.23	0.191
Adaptive	**0.568**	**0.264**	**0.2737**	**5.65**	**0.224**

Figure 8. **Effectiveness of adaptive circulation control.** For the prompt "A *fancy clock* ... with red carpet," moderate circulation improves the baseline structure, whereas excessive static circulation introduces visible distortion. Adaptive control reduces this over-perturbation and preserves a more coherent object structure.

where b and h index the sample and attention head, respectively. We then modulate only the deviation from the baseline circulation scale:

$$\alpha_{\text{eff}} \triangleq (\alpha - 1)\bar{\eta}_{\mathbf{M}}. \tag{38}$$

Equivalently, at the logit level, this corresponds to

$$\mathbf{M}_{\text{adap}}(\mathbf{X}) = \mathbf{M}(\mathbf{X}) + \alpha_{\text{eff}} \cdot \mathbf{M}_{\text{skew}}(\mathbf{X}). \tag{39}$$

The adaptive retrieval state is then

$$\boldsymbol{\Xi}_{\text{adap}} = \Phi(\mathbf{M}_{\text{adap}}(\mathbf{X}))\,\mathbf{X}. \tag{40}$$

The blending coefficient β controls the step size from the baseline retrieval toward the adaptive retrieval. We therefore use a smaller step when the realized attention is more symmetry-dominated, and a larger step when stronger correction is needed:

$$\beta_{\text{eff}} = \beta(1 - \bar{\eta}_{\mathbf{M}}), \tag{41}$$

and form the final blended retrieval by

$$\boldsymbol{\Xi}^{\text{adap}}_{\text{blend}} = \boldsymbol{\Xi} + \beta_{\text{eff}}\,(\boldsymbol{\Xi}_{\text{adap}} - \boldsymbol{\Xi})\,. \tag{42}$$

For stability, the implementation additionally matches the feature norm of the blended retrieval to the reference retrieval with a bounded per-token rescaling.

Table 6 and Figure 8 summarize the effect of adaptive circulation control. Table 6 provides quantitative evidence that adaptive control mitigates the degradation caused by excessive static perturbation. Figure 8 gives a qualitative illustration: a moderate static perturbation improves the baseline structure, whereas excessive static perturbation introduces visible distortion. The adaptive variant preserves the intended circulation correction while reducing the over-perturbation caused by the excessive static setting.

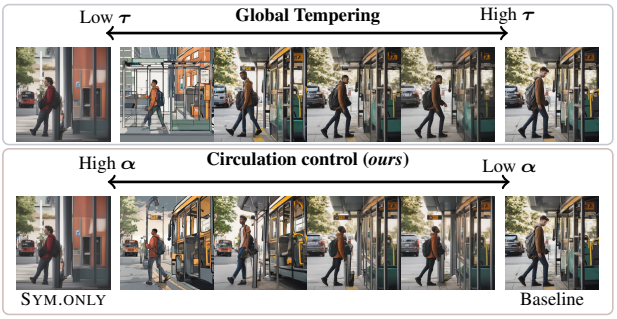

Figure 9. **Ablation against attention temperature τ scaling.** Relative to the SYM.ONLY reference, temperature scaling can introduce unintended structures (e.g., additional leg) due to non-selective strengthening/weakening of interactions across the scene. Instead, our control better preserves strongly supported structure while suppressing weakly supported mixture artifacts.

6.2. Circulation Control and Global Tempering

Having characterized the state-dependent operating regime of circulation control, we next validate whether a simpler global attention manipulation can reproduce the same behavior. To this end, we compare our circulation-based perturbation against an attention temperature baseline that linearly rescales the $\mathbf{QK}^\top$:

$$\mathbf{QK}^\top \;\mapsto\; \mathbf{QK}^\top/\tau, \tag{43}$$

which globally alters the concentration of the attention distribution. As illustrated in Figure 9, this global modification tends to over-sharpen or under-damp interactions that were already well-structured, producing unintended artifacts (e.g., duplicated limbs). In contrast, our approach acts as a *metastable perturbation*: it tends to preserve the dominant structural support governed by the symmetric component $\mathbf{M}_{\text{sym}}$, while leveraging the skew-symmetric component to suppress weakly supported mixture artifacts, producing more coherent refinements than global temperature scaling at comparable intervention strengths.

7. Conclusion, Implications, and Future work

In this work, we propose an associative-memory framework for interpreting self-attention through the structure of $\mathbf{QK}^\top$. By viewing $\mathbf{QK}^\top$ as an association matrix and decomposing it into symmetric and skew components, we derive Hopfield-style stability measures and relate them to the retrieval behavior observed during generation. We further introduce a training-free circulation control mechanism that modulates the skew component using the realized symmetry of attention.

Implications and Future work. Our results suggest a complementary way to analyze attention: not only as a token-mixing operator, but also as an interaction matrix with energy-supported and circulation-driven components. This perspective may provide a useful lens for studying attention dynamics beyond diffusion models, including large language models and other transformer architectures.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00335741), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-25442405, Development of a Self-Learning World Model-Based AGI System for Hyperspectral Imaging), and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism(RS-2024-00345025, International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI).

Impact Statement

This work offers a principled way to diagnose and mitigate spurious feature mixing in attention-based diffusion models, which may improve the reliability and controllability of generative systems. While the method is broadly applicable to image synthesis, it could also increase the fidelity of generated content in ways that may be misused.

References

Amari, S.-I. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972. doi: 10.1109/T-C.1972.223477.

Ambrogioni, L. In search of dispersed memories: Generative diffusion models are associative memory networks. In *Associative Memory & Hopfield Networks in 2023*, 2023. URL https://openreview.net/forum?id=hkV9CvCOjH.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. Spin-glass models of neural networks. *Phys. Rev. A*, 32: 1007–1018, Aug 1985. doi: 10.1103/PhysRevA.32.1007. URL https://link.aps.org/doi/10.1103/PhysRevA.32.1007.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3X2EbBLNsk.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, X., Liu, N., Zhu, Y., Feng, F., and Tang, J. EDT: An efficient diffusion transformer framework inspired by human-like sketching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MihOCXte41.

Chengxiang, Z., Dasgupta, C., and Singh, M. P. Retrieval properties of a hopfield model with random asymmetric interactions. *Neural Computation*, 12(4):865–880, 2000. doi: 10.1162/089976600300015628.

Derrida, B., Gardner, E., and Zippelius, A. An exactly solvable asymmetric neural network model. *Europhysics Letters*, 4(2):167, jul 1987. doi: 10.1209/0295-5075/4/2/007. URL https://doi.org/10.1209/0295-5075/4/2/007.

D'Amico, F. and Negri, M. Self-attention as an attractor network: transient memories without backpropagation. In *2024 IEEE Workshop on Complexity in Engineering (COMPENG)*, pp. 1–6, 2024. doi: 10.1109/COMPENG60905.2024.10741429.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. CLIPScore: A reference-free evaluation metric for image captioning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach,

H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (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.

Hong, S. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37: 66743–66772, 2024.

Hoover, B., Strobelt, H., Krotov, D., Hoffman, J., Kira, Z., and Chau, D. H. Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories. In *Associative Memory & Hopfield Networks in 2023*, 2023. URL https://openreview.net/forum?id=B1BL9go65H.

Hoover, B., Shi, Z., Balasubramanian, K., Krotov, D., and Ram, P. Dense associative memory with epanechnikov energy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=ZbQ5Zq3zA3.

Hopfield, J. J. 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.

Hwang, S., Folli, V., Lanza, E., Parisi, G., Ruocco, G., and Zamponi, F. On the number of limit cycles in asymmetric neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(5):053402, May 2019. ISSN 1742-5468. doi: 10.1088/1742-5468/ab11e3. URL http://dx.doi.org/10.1088/1742-5468/ab11e3.

Kim, K. and Sim, B. Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16238–16248, 2025.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (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.

Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., and Smith, L. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Little, W. The existence of persistent states in the brain. *Mathematical Biosciences*, 19 (1):101–120, 1974. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(74)90031-5. URL https://www.sciencedirect.com/science/article/pii/0025556474900315.

Liu, X., Gong, C., and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Nakano, K. Associatron-a model of associative memory. *IEEE Trans. Syst. Man Cybern.*, 2:380–388,

1972. URL https://api.semanticscholar.org/CorpusID:38591603.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Oriyad, A. M., Banayeeanzade, M., Abbasi, R., Rohban, M. H., and Baghshah, M. S. Attention overlap is responsible for the entity missing problem in text-to-image diffusion models! *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Xv3ZrFayIO.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Peretto, P. Collective properties of neural networks: A statistical physics approach. *Biological Cybernetics*, 50(1):51–62, February 1984. ISSN 1432-0770. doi: 10.1007/BF00317939. URL https://doi.org/10.1007/BF00317939.

Personnaz, L., Guyon, I., and Dreyfus, G. Collective computational properties of neural networks: New learning mechanisms. *Phys. Rev. A*, 34:4217–4228, Nov 1986. doi: 10.1103/PhysRevA.34.4217. URL https://link.aps.org/doi/10.1103/PhysRevA.34.4217.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory networks. In *New Frontiers in Associative Memories*, 2025. URL https://openreview.net/forum?id=IWZnhP3YgK.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Singh, M. P., Chengxiang, Z., and Dasgupta, C. Fixed points in a hopfield model with random asymmetric interactions. *Phys. Rev. E*, 52:5261–5272, Nov 1995. doi: 10.1103/PhysRevE.52.5261. URL https://link.aps.org/doi/10.1103/PhysRevE.52.5261.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (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.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stein, G., Cresswell, J. C., Hosseinzadeh, R., Sui, Y., Ross, B. L., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=08zf7kTOoh.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JVzeOYEx6d.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7354–7363. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19d.html.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

A. Reproducibility and Implementation Details

Base generator. All experiments in this paper are conducted using *Stable Diffusion XL (SDXL)* (Podell et al., 2024) with classifier-free guidance (Ho & Salimans, 2021) at a guidance weight of $\omega = 5.0$ and 30 sampling steps.

Implementation of Skew-symmetric perturbation blending. During the sampling process, we implement the proposed *circulation-based blending* by intervening on the self-attention retrieval within the UNet layers. Specifically, we replace the baseline retrieval states Ξ with the modulated states Ξ_{blended} as defined in Equation (34). This intervention is applied globally across the UNet architecture to maintain consistency in the resulting feature trajectories.

Compute and infrastructure. Inference is performed on a single NVIDIA GeForce RTX 4090 GPU using 16-bit floating-point (fp16) precision. The experimental framework is implemented using PyTorch (Paszke et al., 2019) and the Hugging Face `diffusers` library (von Platen et al., 2022).

A.1. Code Implementation

Algorithm 2 Code: Skew-symmetric perturbation blending

```
def get_attn_probs(x_q, W, x_k):
   logits = torch.einsum("bti,hij,bsj->bhts", x_q, W, x_k)
   logits = logits * (1.0 / math.sqrt(d))
   attn_probs = torch.softmax(logits, dim=-1)
   return attn_probs

x_q/k/v = hidden_states
Wq/k/v = attn.to_q/k/v.weight

Cq/k = Wq.shape[1]
H = attn.heads
d = Wq.shape[0] // H

Wq/k_h = Wq.view(H, d, Cq/k)

A_h = torch.einsum("hdi,hdj->hij", Wq_h, Wk_h)
S = 0.5 * (A_h + A_h.transpose(-2, -1))
N = 0.5 * (A_h - A_h.transpose(-2, -1))

attn_probs = get_attn_probs(x_q, (S + self.alpha * N), x_k)
attn_probs_org = get_attn_probs(x_q, (S + N), x_k)

Hx    = torch.einsum("bhts,bsc->bhtc", attn_probs, x_v)
Hx_org = torch.einsum("bhts,bsc->bhtc", attn_probs_org, x_v)

beta, eps, r_min, r_max = self.beta, 1e-6, 0.25, 4.0

Hx_new = Hx_org + beta * (Hx - Hx_org)

ref = torch.linalg.vector_norm(Hx, dim=-1, keepdim=True).clamp_min(eps)
cur = torch.linalg.vector_norm(Hx_new, dim=-1, keepdim=True).clamp_min(eps)
ratio = (ref / cur).clamp(r_min, r_max)

Hx = (Hx_new * ratio).to(dtype=Hx.dtype)

Cv = Wv.shape[1]
Wv_h = Wv.view(H, d, Cv)
hidden_states = torch.einsum("bhtc,hdc->bhtd", Hx, Wv_h)
```

B. Broader Experiments

In this section, we provide additional experiments that examine the broader applicability of the proposed symmetric/skew decomposition and circulation-based control. We first evaluate whether the method transfers to a transformer-based diffusion architecture, and then report a larger-scale COCO evaluation with distribution-level metrics.

B.1. Generalization to DiT Architectures

We evaluate the proposed circulation control on Stable Diffusion 3, a transformer-based MMDiT architecture (Esser et al., 2024), which follows the broader family of diffusion transformers (Peebles & Xie, 2023). As shown in Table 7, the method transfers beyond the SDXL UNet setting. On the full COCO–1K set, several operating points improve ImageReward (Xu et al., 2023) while keeping Aesthetic Score (Schuhmann et al., 2022) broadly comparable to the baseline. On low-quality subsets, the intervention shows the same regime-dependent pattern observed in the UNet experiments: it improves the target metric, while several settings also yield small or non-negative changes on the other verifiers. These results suggest that the symmetric/skew decomposition and circulation-based control remain meaningful in transformer-based diffusion backbones.

Table 7. **Generalization to SD3 MMDiT.** Full COCO–1K reports absolute scores on Stable Diffusion 3. Low-quality subset blocks report paired changes Δ relative to the baseline.

Metric	Baseline	$(0.95, 3)$	$(0.97, 4)$	$(0.97, 2)$
(a) Full COCO–1K evaluation on SD3 (Esser et al., 2024)				
IR ↑	0.862	0.870	**0.872**	0.861
AES ↑	5.279	5.276	5.277	**5.281**

Metric	$(0.90, 3)$	$(0.95, 3)$	$(0.97, 4)$	$(0.97, 2)$
► **Low-quality subset**: bottom-20% sorted by ImageReward				
ΔIR ↑	**+0.505**	**+0.446**	**+0.439**	**+0.229**
ΔAES ↑	-0.049	+0.018	+0.011	+0.006
ΔCLIP ↑	+0.0043	+0.0025	+0.0015	+0.0025
► **Low-quality subset**: bottom-20% sorted by Aesthetic				
ΔIR ↑	+0.021	+0.056	+0.037	+0.061
ΔAES ↑	**+0.224**	**+0.178**	**+0.149**	**+0.146**
ΔCLIP ↑	-0.0049	+0.0002	+0.0005	-0.0000

B.2. Large-Scale Quantitative Evaluation

We also expand the COCO evaluation from the 1K setting to 10K samples and include distribution-level metrics (Heusel et al., 2017; Stein et al., 2023) in addition to the standard perceptual scores. As shown in Table 8, selected operating points improve ImageReward and Aesthetic Score while keeping CLIP close to the baseline. The distribution-level metrics remain broadly comparable to the baseline, indicating that the intervention changes the retrieval behavior without substantially degrading the overall generated distribution. Together with the SD3 results in Table 7, these experiments support the broader applicability of circulation-based attention control across model scale and architecture.

Table 8. **COCO–10K quantitative evaluation.** COCO–10K reports standard perceptual scores together with distribution-level metrics. Lower values are better for FID, FD-DINOv2, and KD-DINOv2.

Method	IR ↑	AES ↑	CLIP ↑	FID ↓	FD-DINOv2 ↓	KD-DINOv2 ↓
COCO–10K Baseline SDXL (Podell et al., 2024)						
	0.5614	5.6321	**0.2631**	**24.314**	289.352	0.1358
COCO–10K Ours						
$(\alpha, \beta) = (1.03, 3)$	**0.5753**	5.6394	0.2627	24.478	289.623	0.1358
$(\alpha, \beta) = (1.03, 5)$	0.5746	**5.6432**	0.2622	24.607	**289.337**	**0.1354**

B.3. Qualitative Examples of Success and Failure Cases

To complement the quantitative results, we provide paired qualitative examples in Figure 10. All examples use SDXL (Podell et al., 2024) with the proposed circulation control at $(\alpha, \beta) = (1.05, 3)$.

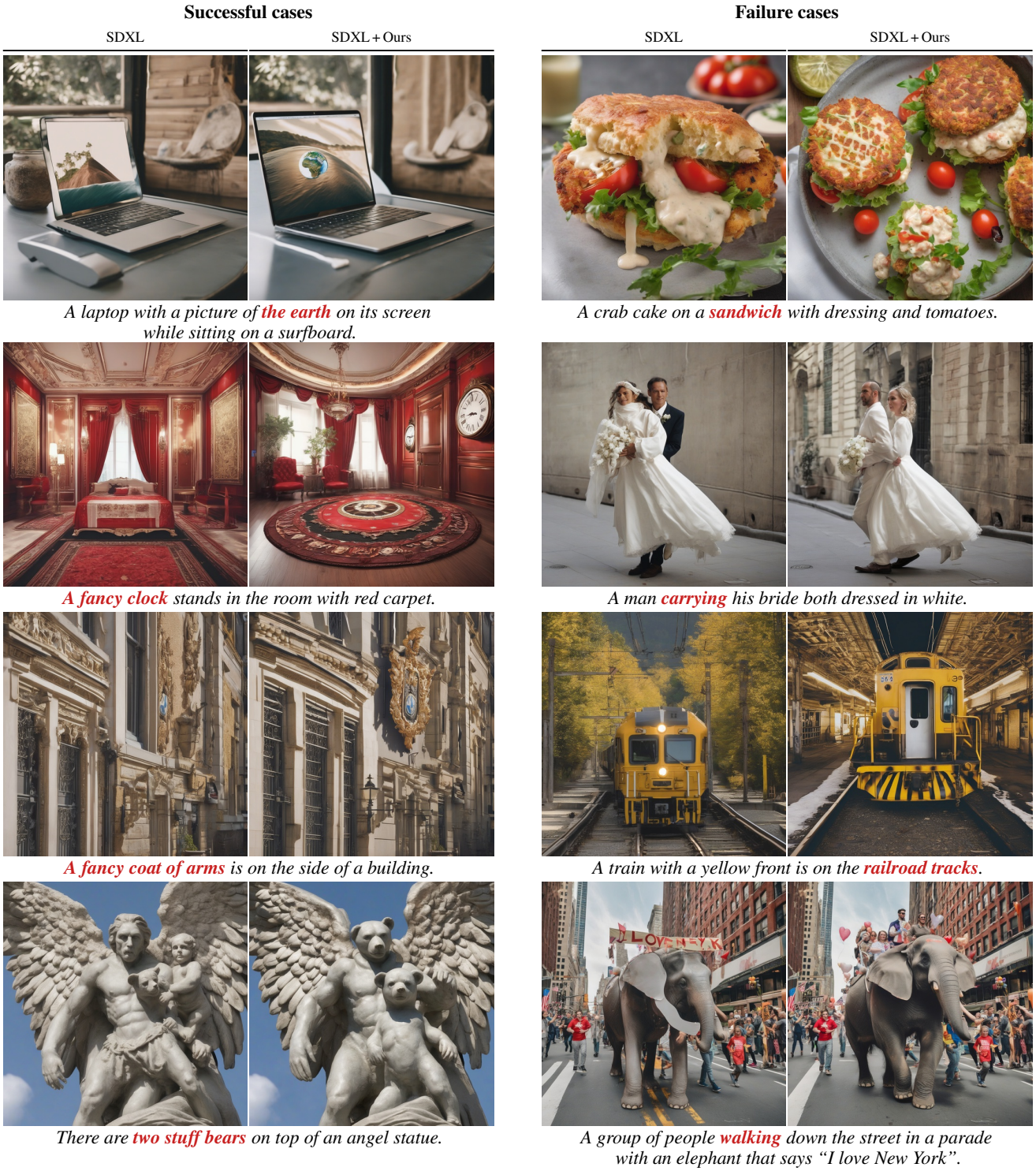

Figure 10. **Qualitative results of SDXL + Ours**. **Left (successful cases):** the **highlighted** concepts are weakly represented or missing in the baseline, and Ours renders them more faithfully (e.g., adding missing objects or correcting on-screen/scene content). **Right (failure cases):** on prompts the baseline already handles well, Ours can *mildly* degrade the **highlighted** aspect (e.g., the sandwich form, the "carrying" pose, staying on the tracks, or walking vs. riding) while overall image quality remains comparable.