DIRECTIONAL ATTENTION GUIDANCE FOR DIFFUSION MODELS VIA SPARSE HOPFIELD RETRIEVAL

Anonymous authors
Paper under double-blind review

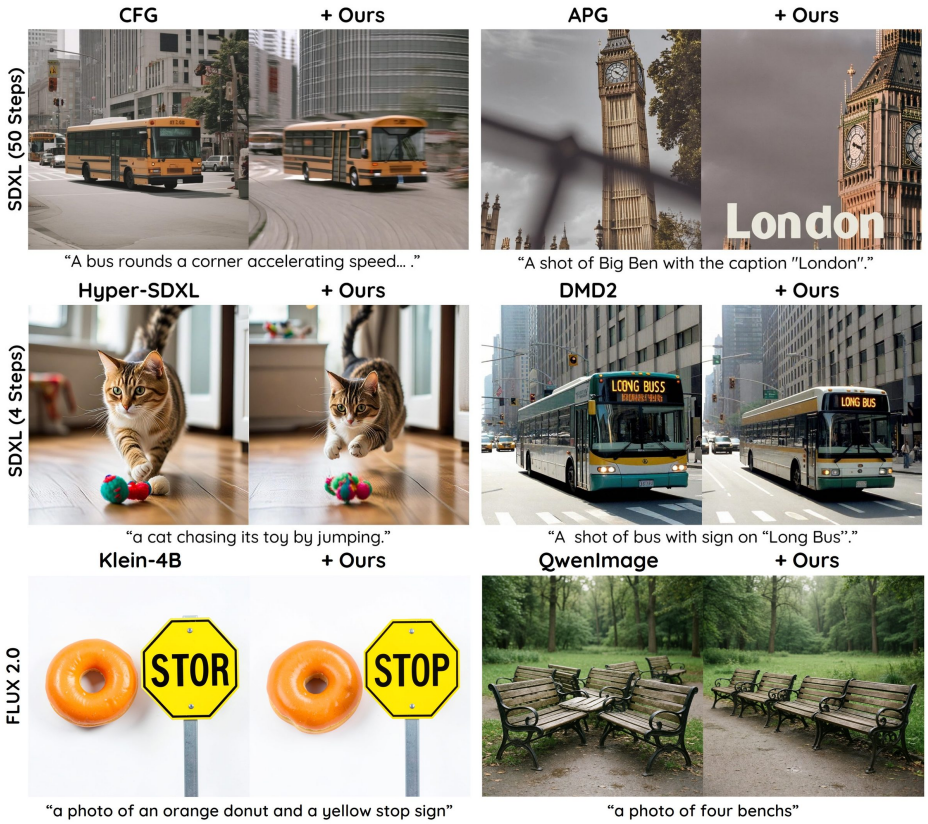


Figure 1: DAG on guidance methods (CFG, APG; top), step-distilled models (Hyper-SDXL, DMD2; middle), and MMDiT backbones (FLUX.2, Qwen-Image; bottom): a training-free, single-pass rule that improves text–image alignment and detail at a small added cost per image.

ABSTRACT

Classifier-Free Guidance (CFG) improves sample quality in diffusion models, but its dual-pass inference and null-condition training limit its use in few-step regimes. Attention-space guidance is a single-pass alternative, yet why sparse-versus-dense attention guidance works remains elusive. We analyze attention extrapolation through Modern Hopfield dynamics and establish two directional properties of the sparse–dense discrepancy under shared conditioning that identify it as an acceleration direction toward the dominant stored pattern. Building on this, we propose **Directional Attention Guidance (DAG)**, a training-free, single-pass rule that amplifies only the component of the discrepancy aligned with the retrieval direction; stability follows from a weak contraction property, and an analysis inside real cross-attention layers shows that this alignment is observed throughout sampling. DAG generalizes across UNet and MMDiT backbones and multi-step and few-step regimes, improving text alignment and human-preference scores on every backbone tested, including FLUX.1, FLUX.2, and Qwen-Image, at a small added cost per image over the base sampler.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

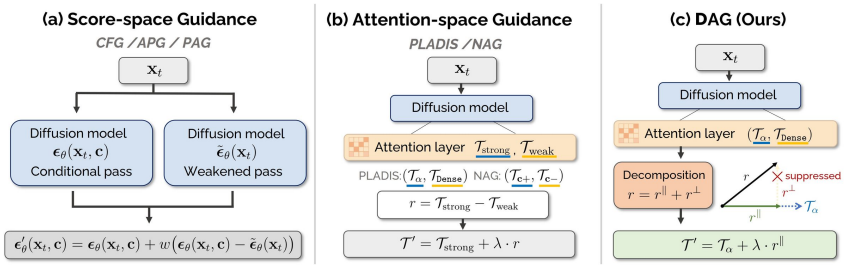


Figure 2: **Conceptual comparison.** (a) Score-space guidance needs two forward passes per step; (b) attention-space guidance is single-pass but amplifies the full residual; (c) DAG amplifies only the convergence-aligned $r^{\parallel}$.

1 Introduction

Text-to-image (T2I) diffusion models (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024) now drive image (Xie et al., 2025; Black Forest Labs, 2024) and video (Blattmann et al., 2023; Chen et al., 2024; Yang et al., 2025; Wan Team, 2025) generation, and because vanilla sampling in diffusion (Ho et al., 2020) and flow-matching (Lipman et al., 2023) models is often suboptimal, guidance is standard practice. Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) extrapolates between conditional and unconditional score estimates at the cost of dual-pass inference and null-condition training, which hinders its use in few-step distilled or single-step models (Lin et al., 2024; Sauer et al., 2024; Wang et al., 2024); training-free alternatives (Ahn et al., 2024; Hong, 2024; Rajabi et al., 2025; Kim, 2026) remove the null-condition training but still need two passes.

To address this cost, attention-space extrapolation has emerged as an alternative operating inside the attention layer rather than across forward passes: PLADIS (Kim & Sim, 2025) contrasts *sparse and dense* attention from the same prompt, and Normalized Attention Guidance (NAG) (Chen et al., 2025) contrasts attention from *positive and negative* prompts. Both extrapolate from the discrepancy between two attention operators, yet why this discrepancy is a valid update direction remains unexplained.

Score-space guidance, in contrast, has been refined through *directional* analysis, with Adaptive Projected Guidance (APG) (Sadat et al., 2025) projecting the conditional–unconditional contrast onto parallel and orthogonal components and Angle Domain Guidance (ADG) (Jin et al., 2025) aligning its angle. We therefore give a directional analysis of attention-space extrapolation: when a discrepancy *accelerates* convergence to the semantic target, when it stays *aligned* with it, and when the update it prescribes is stable.

We address this by modeling cross-attention through Modern Hopfield dynamics (Ramsauer et al., 2021; Hu et al., 2023; Wu et al., 2024), in the shared-conditioning sparse-vs-dense case (PLADIS-style) where the two operators differ only in the probability mapping (α−`Entmax` vs. `Softmax`). For their discrepancy $r(\mathbf{x})$ we establish two *directional properties*: sparse retrieval reaches the dominant stored pattern at least as fast as dense retrieval, and $r(\mathbf{x})$ is non-negatively aligned with the residual to that pattern, so $r(\mathbf{x})$ is an acceleration direction toward it; the NAG-style case with different conditionings lies outside this regime, a failure mode we predict and verify.

Building on these properties, we propose **Directional Attention Guidance (DAG)**, which decomposes $r(\mathbf{x})$ along the sparse retrieval direction and amplifies only the convergence-aligned component (Figure 2); as Figure 1 shows, the effect carries from classical guidance and step-distilled models to the latest MMDiT/DiT backbones. Stability follows from a weak contraction property (Alber & Guerre-Delabriere, 1997; Rhoades, 2001), and the rule admits a reading as first-order Anderson Acceleration (AA) (Walker & Ni, 2011) in attention space. Our contributions are as follows:

- We propose Directional Attention Guidance (DAG), a training-free, single-pass rule that amplifies only the component of the sparse–dense attention residual aligned with sparse retrieval. Across UNet and MMDiT backbones and multi-step and few-step regimes, DAG improves text alignment and human-preference scores on every backbone tested.
- To explain why an attention residual is a valid update direction, we give a directional analysis of attention-space extrapolation under shared conditioning, showing when the

residual accelerates and aligns with retrieval and reading the stability of the resulting update through a weak contraction property.

- To connect the analysis to practice, we conduct an in-depth empirical analysis inside real cross-attention layers, locating them relative to the regime the analysis assumes and showing that each ingredient of the rule contributes.

2 PRELIMINARY

2.1 DIFFUSION MODELS AND GUIDANCE SAMPLING

Diffusion models (Ho et al., 2020; Song et al., 2021b) learn to generate data by reversing a progressive Gaussian noise process. The forward process perturbs $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ as $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$, with closed-form marginal $q(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$ where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. The reverse process p_θ is trained via denoising score matching (Vincent, 2011); for conditional generation (Rombach et al., 2022) with context $\mathbf{c}$:

$$\min_\theta \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},\mathbf{c}}\left[\|\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, t, \mathbf{c}) - \boldsymbol{\epsilon}\|_2^2\right]. \tag{1}$$

Sampling follows the DDIM (Song et al., 2021a) update with the denoised estimate $\hat{\mathbf{x}}_0(t) := (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}))/\sqrt{\bar{\alpha}_t}$ from Tweedie's formula (Efron, 2011; Kim & Ye, 2021).

Score-space guidance. Guidance biases the score toward the conditional distribution by extrapolating the conditional estimate $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) := \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ away from a weakened counterpart $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t)$:

$$\boldsymbol{\epsilon}'_\theta(\mathbf{x}_t, \mathbf{c}) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) + w\big(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - \tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t)\big), \tag{2}$$

with guidance scale w; anchoring at the conditional estimate matches (3) and equals the CFG convention with $w = w_{\text{CFG}} - 1$. CFG (Ho & Salimans, 2022) uses the unconditional estimate $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \varnothing)$, weak-model guidance perturbs the architecture or self-attention (Ahn et al., 2024; Hong, 2024; Kim, 2026), and APG (Sadat et al., 2025) and ADG (Jin et al., 2025) refine the direction of the contrast. All require dual forward passes, which limits their use in step-distilled frameworks.

Attention-space guidance. To remove the dual-pass overhead while retaining the extrapolation principle, recent studies (Kim & Sim, 2025; Chen et al., 2025) apply the same contrast inside the attention layer rather than across forward passes, contrasting a strong attention $\texttt{At}$ with a weakened counterpart $\tilde{\texttt{At}}$:

$$\texttt{At}'(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \texttt{At}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) + \lambda\big(\texttt{At}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) - \tilde{\texttt{At}}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t)\big), \tag{3}$$

with extrapolation scale λ. PLADIS (Kim & Sim, 2025) pairs sparse and dense attention; NAG (Chen et al., 2025) contrasts attention from positive and negative prompts. This work provides a directional analysis of such contrasts.

2.2 ENERGY-BASED INTERPRETATIONS OF ATTENTION

Attention has been linked to Energy-Based Models through Hopfield energy functions: a query $\mathbf{x} \in \mathbb{R}^d$ is associated with a pattern matrix $\boldsymbol{\Xi} = [\xi_1, \ldots, \xi_M] \in \mathbb{R}^{d\times M}$ by minimizing an energy $E(\mathbf{x})$ through retrieval dynamics $\mathcal{T}$, and Modern Hopfield Networks (MHN) (Ramsauer et al., 2021) and Sparse Hopfield Networks (SHN) (Hu et al., 2023; Wu et al., 2024) define

$$E(\mathbf{x})_{\texttt{Dense}} := -\texttt{lse}(\beta, \boldsymbol{\Xi}^\top\mathbf{x}) + \tfrac{1}{2}\langle\mathbf{x},\mathbf{x}\rangle, \qquad \mathcal{T}_{\texttt{Dense}}(\mathbf{x}) := \boldsymbol{\Xi}\,\texttt{Softmax}(\beta\boldsymbol{\Xi}^\top\mathbf{x}), \tag{4}$$

$$E_\alpha(\mathbf{x}) := -\boldsymbol{\Psi}^\star_\alpha(\beta, \boldsymbol{\Xi}^\top\mathbf{x}) + \tfrac{1}{2}\langle\mathbf{x},\mathbf{x}\rangle, \qquad \mathcal{T}_\alpha(\mathbf{x}) := \boldsymbol{\Xi}\,\alpha\text{-}\texttt{Entmax}(\beta\boldsymbol{\Xi}^\top\mathbf{x}), \tag{5}$$

where $\texttt{lse}(\beta, \mathbf{z}) := \frac{1}{\beta}\log\sum_{i=1}^{M}\exp(\beta z_i)$ and $\boldsymbol{\Psi}^\star_\alpha$ is the convex conjugate of the Tsallis entropy (Tsallis, 1988) Ψ_α. The $\alpha\text{-}\texttt{Entmax}$ operator $\arg\max_{\mathbf{p}\in\Delta^M}[\langle\mathbf{p},\mathbf{z}\rangle - \Psi_\alpha(\mathbf{p})]$ controls sparsity, recovering softmax at $\alpha = 1$ and growing sparser toward $\alpha = 2$, with exact, efficient solutions at $\alpha \in \{1.5, 2\}$ (Held et al., 1974; Michelot, 1986; Peters et al., 2019); SHN reduces retrieval errors and is more robust to noise than dense dynamics. With $\beta = 1/\sqrt{d}$, the dynamics map directly to attention: $\texttt{At}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$ and $\texttt{At}_\alpha(\mathbf{Q},\mathbf{K},\mathbf{V}) = \alpha\text{-}\texttt{Entmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$.

We focus on **cross-attention**, where text tokens are the stored patterns and sparsification helps, whereas sparsifying self-attention degrades generation (Kim & Sim, 2025); the faster, noise-robust retrieval of $\texttt{At}_\alpha$ is what PLADIS (Kim & Sim, 2025) exploits as the strong component in (3).

3 Main Contribution: DAG

We revisit MHN retrieval as a fixed-point iteration (Sec. 3.1), establish two directional properties of the discrepancy (Sec. 3.2), derive the update rule (Sec. 3.3) and its stability (Sec. 3.4), and connect it to Anderson Acceleration (Sec. 3.5); proofs are in the appendix.

3.1 Retrieval Dynamics as Fixed-Point Iteration

We revisit the retrieval dynamics $\mathcal{T}$ of MHNs (Ramsauer et al., 2021; Hu et al., 2023), which are Picard iterations $\mathbf{x}_{k+1} = \mathcal{T}(\mathbf{x}_k)$ toward a stationary point $\mathbf{x}_\star$ of the energy, that is, a stored memory; for a contraction, the iteration converges to its unique fixed point (Banach, 1922).

Proposition 1 (Generalized Sparse Hopfield Retrieval Dynamics (Ramsauer et al., 2021; Hu et al., 2023; Wu et al., 2024))**.** *The retrieval dynamics of the generalized sparse Hopfield model is a monotonic one-step update:*

$$\mathbf{x}_{k+1} = \mathcal{T}(\mathbf{x}_k) := \boldsymbol{\Xi}\, \alpha\text{-}\mathit{Entmax}(\beta \boldsymbol{\Xi}^\top \mathbf{x}_k). \tag{6}$$

This implies that attention dynamics are the mechanism for finding fixed points $\mathbf{x}_\star$ in MHNs, where the standard attention layer performs exactly one iteration of the update rule: with $\beta = 1/\sqrt{d}$, $\mathcal{T}_{(\cdot)}(\mathbf{x}) \to \mathtt{At}_{(\cdot)}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, where $\mathbf{x}$ is the query and $(\mathbf{K}, \mathbf{V})$ are the stored memories. This fixed-point view lets us compare two retrieval operators that share the same memories but differ in the probability mapping.

3.2 Directional Properties of the Sparse–Dense Discrepancy

Under this fixed-point view, we ask when the discrepancy between the two operators serves as an acceleration direction.

Setup. Let $\mathcal{T}_\alpha$ and $\mathcal{T}_{\mathtt{Dense}}$ be instantiated from the same model with identical $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ and conditioning $\mathbf{c}$, differing only in the probability mapping (α-$\mathtt{Entmax}$ vs. $\mathtt{Softmax}$). Both outputs lie in $\mathrm{span}(\mathbf{V})$, and we define the discrepancy

$$r(\mathbf{x}) \;:=\; \mathcal{T}_\alpha(\mathbf{x}) - \mathcal{T}_{\mathtt{Dense}}(\mathbf{x}) \;=\; \mathbf{V}\big(\mathbf{p}_\alpha(\mathbf{x}) - \mathbf{p}_{\mathtt{dense}}(\mathbf{x})\big), \tag{7}$$

where $\mathbf{p}_\alpha, \mathbf{p}_{\mathtt{dense}} \in \Delta^M$. We write $r(\mathbf{x})$ for this discrepancy throughout.

Theorem 1 (Directional Surrogate Property)**.** *Under the setup above, let $\boldsymbol{\xi}_\mu$ be the dominant stored pattern, let $\Delta(\mathbf{x}) := \langle \boldsymbol{\xi}_\mu, \mathbf{x}\rangle - \max_{j\neq\mu}\langle \boldsymbol{\xi}_j, \mathbf{x}\rangle > 0$ be its logit margin at $\mathbf{x}$, and let $\alpha \in (1, 2]$. Then $r(\mathbf{x})$ satisfies:*

(i) ***Convergence rate separation.*** *The retrieval-error bound of sparse retrieval is no larger than that of dense retrieval (Hu et al., 2023; Wu et al., 2024), and for $\beta \geq \beta_0(\mathbf{x}) := 1/((\alpha-1)\Delta(\mathbf{x}))$ sparse retrieval is exact, $\mathcal{T}_\alpha(\mathbf{x}) = \boldsymbol{\xi}_\mu$, so that*

$$\big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \;=\; 0 \;\leq\; \big\|\mathcal{T}_{\mathtt{Dense}}(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|. \tag{8}$$

(ii) ***Directional alignment.*** *At every $\mathbf{x}$,*

$$\big\langle r(\mathbf{x}),\, \boldsymbol{\xi}_\mu - \mathcal{T}_{\mathtt{Dense}}(\mathbf{x})\big\rangle \;\geq\; \tfrac{1}{2}\Big(\|r(\mathbf{x})\|^2 - \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|^2\Big). \tag{9}$$

Hence $r(\mathbf{x})$ is non-negatively aligned with the residual $\boldsymbol{\xi}_\mu - \mathcal{T}_{\mathtt{Dense}}(\mathbf{x})$ whenever the discrepancy is at least as large as the sparse retrieval error, $\|r(\mathbf{x})\| \geq \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|$; for $\beta \geq \beta_0(\mathbf{x})$ the sparse error vanishes and $r(\mathbf{x}) = \boldsymbol{\xi}_\mu - \mathcal{T}_{\mathtt{Dense}}(\mathbf{x})$.

Well-separation thus enters through the logit margin alone: above the threshold $(\alpha-1)\beta\Delta(\mathbf{x}) \geq 1$ the discrepancy is exactly the residual to the dominant pattern (Appendix C); below it, (ii) is conditional on the discrepancy exceeding the sparse retrieval error, and Sec. 5.2 measures where real layers stand.

Theorem 1 yields a structural corollary on the geometric soundness of extrapolating along $r(\mathbf{x})$.

Corollary 1 (Directional Consistency)**.** *Let $\mathcal{T}'(\mathbf{x}) := \mathcal{T}_\alpha(\mathbf{x}) + \lambda\, r(\mathbf{x})$ with $\lambda \geq 0$. Under the setup of Theorem 1:*

(a) ***Direction preservation.*** *$r(\mathbf{x}) \in \mathrm{span}(\mathbf{V})$ and, under the condition of Theorem 1(ii), is non-negatively aligned with $\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$, so $\mathcal{T}'(\mathbf{x}) \in \mathrm{span}(\mathbf{V})$ with no off-subspace drift.*

(b) ***Common attractor at $\boldsymbol{\xi}_\mu$.*** *Above the collapse threshold of Theorem 1(i), $r(\mathbf{x}) \to \mathbf{0}$ as $\mathbf{x} \to \boldsymbol{\xi}_\mu$, so $\mathcal{T}'(\mathbf{x}) \to \boldsymbol{\xi}_\mu$.*

Property (a) is immediate from (7) and Theorem 1(ii); (b) follows from the retrieval-error bounds in Proposition 3 (full proof in Appendix D). Along real sampling trajectories, the orthogonal component stays the smaller one (Sec. 5.2; Figure 8c).

Remark 1 (Failure mode: non-shared conditioning)**.** Theorem 1 requires $\mathcal{T}_\alpha$ and $\mathcal{T}_{\texttt{Dense}}$ to share $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ and $\mathbf{c}$. NAG (Chen et al., 2025) contrasts attention under different conditioning ($\mathbf{c}_{\text{pos}}$ vs. $\mathbf{c}_{\text{neg}}$), so the two operators receive different $\mathbf{V}$, share no span, and have distinct attractors; the residual is then not guaranteed to align with any single $\boldsymbol{\xi}_\mu$, and composing DAG with NAG degrades accordingly (Figure 4b).

3.3 Directional Attention Guidance

Building on the directional properties of Sec. 3.2, we reformulate attention extrapolation as

$$\mathcal{T}'(\mathbf{x}) = \mathcal{T}_\alpha(\mathbf{x}) + \lambda\, r(\mathbf{x}), \tag{10}$$

which recovers PLADIS (Kim & Sim, 2025) for a suitable λ. To analyze how $r(\mathbf{x})$ acts on the sampling trajectory, we decompose it relative to the sparse retrieval direction $\mathcal{T}_\alpha(\mathbf{x})$, the direction that Theorem 1(i) certifies as the closer one to the target.

Definition 1 (Geometric Decomposition of the Discrepancy)**.** Let $\mathbf{P}_\alpha, \mathbf{P}_{\texttt{dense}}$ be the attention tensors of a layer (all heads and queries) and $\boldsymbol{\Delta} := \mathbf{P}_\alpha - \mathbf{P}_{\texttt{dense}}$. The parallel and orthogonal components of the discrepancy with respect to the sparse retrieval are the projection of $\boldsymbol{\Delta}$ onto $\mathbf{P}_\alpha$ and its remainder,

$$\boldsymbol{\Delta}^{\parallel} := \frac{\langle \boldsymbol{\Delta}, \mathbf{P}_\alpha \rangle_F}{\|\mathbf{P}_\alpha\|_F^2}\, \mathbf{P}_\alpha, \qquad \boldsymbol{\Delta}^{\perp} := \boldsymbol{\Delta} - \boldsymbol{\Delta}^{\parallel}, \tag{11}$$

and $r^{\parallel}(\mathbf{x})$, $r^{\perp}(\mathbf{x})$ are the rows of $\mathbf{V}\boldsymbol{\Delta}^{\parallel}$, $\mathbf{V}\boldsymbol{\Delta}^{\perp}$ at query $\mathbf{x}$, so that $r^{\parallel}(\mathbf{x}) \in \mathrm{span}\{\mathcal{T}_\alpha(\mathbf{x})\}$; the projection is taken over the layer's attention tensor, as APG (Sadat et al., 2025) does in output space.

In the regime of Theorem 1, $r(\mathbf{x})$ is non-negatively aligned with the residual $\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$ along $\mathrm{span}(\mathbf{V})$. The parallel component $r^{\parallel}(\mathbf{x})$ therefore captures the convergence-accelerating part of $r(\mathbf{x})$, while the orthogonal component $r^{\perp}(\mathbf{x})$ is not aligned with the retrieval direction. APG (Sadat et al., 2025) contrasts operators with distinct attractors, whereas in our shared-attractor setting (Corollary 1(b)) off-direction components carry less guidance. Since the stored pattern $\boldsymbol{\xi}_\mu$ is not identified at inference, we anchor on $\mathcal{T}_\alpha(\mathbf{x})$, which Theorem 1(i) places closer to $\boldsymbol{\xi}_\mu$ than $\mathcal{T}_{\texttt{Dense}}(\mathbf{x})$; among alternative anchors, it also performs best (Sec. 5.1, Table 16 in Appendix L).



Figure 3: **Components of r.** Orthogonal-only fails, the full residual interferes, parallel-only is best.

Effect of each component. Figure 3 isolates each component during PLADIS-based sampling (prompts: a red dog; four dogs; a "Hello World" storefront): orthogonal-only extrapolation fails to recover semantic structure, the full residual is degraded by $r^{\perp}$ relative to parallel-only, and parallel-only yields the most faithful generations (finer ζ-sweep in Figure 4c).

Motivated by the directional analysis above, we propose **Directional Attention Guidance (DAG)**:

$$\mathcal{T}_\alpha^{\lambda}(\mathbf{x}) = \mathcal{T}_\alpha(\mathbf{x}) + \underbrace{\lambda \min\left(1, \frac{\eta}{\|\boldsymbol{\Delta}\|_F}\right) \left[r^{\parallel}(\mathbf{x}) + \zeta\, r^{\perp}(\mathbf{x})\right]}_{\text{rescaled geometry-aware term}}, \tag{12}$$

where $\zeta \in [0, 1]$ controls orthogonal suppression and η clamps the norm of the layer's discrepancy before the projection, following the norm rescaling of APG (Sadat et al., 2025). Setting $\zeta = 0$ uses only the convergence-aligned component, justified by Theorem 1; the clamp bounds the magnitude of the update in the high-λ regime.

3.4 Stability via Orthogonal Weak Contraction

Let $\mathbf{x}^*$ be an attractor of $\mathcal{T}_\alpha$ and $\Pi_\perp$ the orthogonal projector onto $(\mathrm{span}\{\mathcal{T}_\alpha(\mathbf{x}^*)\})^\perp$, fixed along the iteration.

Assumption 1 (Orthogonal Weak Contraction (Rhoades, 2001; Alber & Guerre-Delabriere, 1997))**.** There exists a strictly increasing forcing function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\phi(0) = 0$ such that, for $\mathcal{T}_\alpha^\lambda$ with $\zeta = 0$,

$$\left\|\Pi_\perp(\mathcal{T}_\alpha^\lambda(\mathbf{x}) - \mathbf{x}^*)\right\| \;\leq\; \|\Pi_\perp(\mathbf{x} - \mathbf{x}^*)\| - \phi\big(\|\Pi_\perp(\mathbf{x} - \mathbf{x}^*)\|\big). \tag{13}$$

Proposition 2 (Asymptotic Convergence of the Orthogonal Error)**.** *Suppose Assumption 1 holds, and let $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ be generated by the iterated dynamics $\mathbf{x}_{t+1} = \mathcal{T}_\alpha^\lambda(\mathbf{x}_t)$ with $\zeta = 0$. Then the orthogonal error magnitude $u_t := \|\Pi_\perp(\mathbf{x}_t - \mathbf{x}^*)\|$ satisfies $\lim_{t\to\infty} u_t = 0$.*

Remark 2 (Stability and robustness)**.** Proposition 2 concerns the iterated Hopfield dynamics, whereas an attention layer performs a single step (Proposition 1), so we read it as a property of the update direction: with $\zeta = 0$ the update adds no orthogonal component and Assumption 1 can drive the orthogonal error down, whereas $\zeta = 1$ re-injects it at every step. Sec. 5.2 tests the observable consequence, that $r^\perp$ is harmful through its direction rather than its scale (Figure 7).

3.5 Connection to Anderson Acceleration

Beyond its geometric motivation, our framework provides a basis for one interpretation of attention-space extrapolation in classical fixed-point acceleration. For an operator $\mathcal{F}$ on a real Hilbert space, Anderson Acceleration (AA) (Walker & Ni, 2011) extrapolates from the last $m + 1$ iterates, $\mathbf{x}_{k+1} = \sum_{i=0}^{m} \omega_i^{(k)} \mathcal{F}(\mathbf{x}_{k-m+i})$ with $\sum_i \omega_i^{(k)} = 1$ chosen to minimize the residual norm; AA is a standard acceleration scheme (Walker & Ni, 2011), and the m=1 case retains local convergence for contractive $\mathcal{F}$ (Toth & Kelley, 2015). With $m = 1$, the update reduces to

$$\mathbf{x}_{k+1} = \mathcal{F}(\mathbf{x}_k) + \omega\big(\mathcal{F}(\mathbf{x}_k) - \mathcal{F}(\mathbf{x}_{k-1})\big). \tag{14}$$

Attention layers realize m=1 through a proxy. A single attention layer is one Picard step of MHN dynamics (Proposition 1) without history of previous iterates, so $m\geq 2$ AA is unrealizable and the missing $\mathcal{F}(\mathbf{x}_{k-1})$ must be approximated by a *proxy*. For PLADIS (10), the discrepancy $r(\mathbf{x})$ is this proxy by the faster convergence of sparse retrieval (Theorem 1), and NAG (Chen et al., 2025) admits a parallel reading with the positive/negative-prompt contrast; each method thus admits a reading as an *implicit* m=1 AA step in attention space, and an explicit m=2 surrogate brings no gain (Appendix F).

4 Experiments

Setup and metrics. To demonstrate the broad applicability of our method, we employ SDXL (Podell et al., 2024), FLUX.1 (Black Forest Labs, 2024), FLUX.2 (Klein-4B/9B) (Black Forest Labs, 2026), and Qwen-Image (2512) (Qwen Team, 2025) under CFG, CFG+PAG, and APG (Sadat et al., 2025), and the step-distilled SDXL-DMD2 (Yin et al., 2024), Hyper-SDXL (Ren et al., 2024), and SDXL-Lightning (Lin et al., 2024), with default schedulers and step counts. We report GenEval (GE) (Ghosh et al., 2023), FID on MS-COCO (Lin et al., 2014), CLIPScore (CS) (Hessel et al., 2021), and the human-preference proxies ImageReward (IR) (Xu et al., 2023), PickScore (PS) (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023) on 10K MS-COCO captions (details in Appendix G).

Guidance sampling. To evaluate DAG under standard multi-step guidance, we compare against CFG, APG, and CFG+PAG at 50 steps (Table 1, left). DAG improves GenEval and the human-preference metrics over both the base guidance and PLADIS in every configuration, with PickScore under APG the one exception, and every configuration surpasses a GenEval of 0.6, a level that SDXL does not reach with guidance alone (CFG: 0.545). Whereas PAG typically enhances aesthetic quality at the expense of text alignment, and APG on its own improves alignment only modestly (GE 0.557), DAG

Table 1: Quantitative comparison on GenEval and MS-COCO with SDXL. Left: high-NFE regimes; right: low-NFE step-distilled models.

Method	Steps	GE ↑	FID ↓	CS ↑	IR ↑	PS ↑	HPSv2 ↑
CFG	50	0.545	20.22	26.45	0.604	22.31	0.266
+ PLADIS	50	0.596	15.81	26.97	0.799	22.32	0.268
+ DAG (Ours)	50	**0.607**	**14.63**	**27.26**	**0.812**	**22.40**	**0.271**
APG	50	0.557	18.53	26.78	0.702	**22.36**	0.269
+ PLADIS	50	0.608	14.51	27.36	0.808	21.92	0.261
+ DAG (Ours)	50	**0.613**	**13.22**	**27.55**	**0.851**	22.24	**0.274**
CFG+PAG	100	0.528	19.59	25.98	0.640	22.30	0.277
+ PLADIS	100	0.598	15.43	26.05	0.751	22.33	0.279
+ DAG (Ours)	100	**0.619**	**14.32**	**26.11**	**0.822**	**22.36**	**0.281**

Method	Steps	GE ↑	FID ↓	CS ↑	IR ↑	PS ↑	HPSv2 ↑
SDXL-DMD2	4	0.581	19.83	26.69	0.888	22.64	0.297
+ PLADIS	4	0.585	18.78	27.06	**0.965**	22.65	0.297
+ DAG (Ours)	4	**0.598**	**17.49**	**27.42**	0.952	**22.66**	**0.298**
Hyper-SDXL	4	0.556	21.23	26.09	0.959	22.87	0.313
+ PLADIS	4	0.589	19.83	27.11	1.060	22.91	0.315
+ DAG (Ours)	4	**0.594**	**19.12**	**27.55**	**1.071**	**22.93**	**0.317**
SDXL-Light	4	0.535	24.42	26.05	0.725	22.62	0.293
+ PLADIS	4	0.570	23.23	26.70	0.835	22.62	0.290
+ DAG (Ours)	4	**0.579**	**22.85**	**26.79**	**0.843**	**22.66**	**0.294**

Table 2: (Left) FLUX.1 (Schnell and Dev). (Middle) FLUX.2 and Qwen-Image on GenEval; all metrics in Table 11 in Appendix J. (Right) Per-image wall-clock and peak memory on one H100 (1024^2, batch 1); all three methods use our implementation and identical execution; details in Appendix I.

Method	Steps	GE ↑	CS ↑	IR ↑	PS ↑	HPSv2 ↑
FLUX.1-Schnell	4	0.669	26.28	1.059	22.62	0.295
+ PLADIS	4	0.712	26.45	1.066	22.31	0.295
+ DAG (Ours)	4	**0.738**	**26.56**	**1.114**	**22.80**	**0.301**
FLUX.1-Dev	28	0.666	25.71	1.093	23.00	0.304
+ PLADIS	28	0.691	**25.80**	1.115	23.03	0.305
+ DAG (Ours)	28	**0.737**	25.76	**1.127**	**23.14**	**0.309**

Method	GE ↑
FLUX.2 (4B)	0.824
+ PLADIS	0.826
+ DAG	**0.843**
FLUX.2 (9B)	0.849
+ PLADIS	0.865
+ DAG	**0.878**
Qwen-Image	0.824
+ PLADIS	0.828
+ DAG	**0.842**

Method	Steps	Time (s) ↓	Mem. (GiB) ↓
SDXL+CFG	50	2.36	8.98
+ PLADIS	50	2.93	8.98
+ DAG	50	3.15	8.98
FLUX.1-Schnell	4	0.88	33.83
+ PLADIS	4	0.99	33.83
+ DAG	4	1.00	33.83
FLUX.2-4B	4	0.51	17.35
+ PLADIS	4	0.54	17.35
+ DAG	4	0.55	17.35

raises both dimensions when combined with either, so the directional update adds alignment without the usual cost in visual quality (qualitative comparisons in Appendix S).

Step-distilled models. To test the regime where dual-pass guidance is unavailable, we apply DAG to DMD2, Hyper-SDXL, and SDXL-Lightning at 4 steps (Table 1, right). DAG outperforms both the baseline and PLADIS on every metric but one, improving both text alignment and visual quality; the single-pass update is therefore robust to the step budget, transferring from 50-step samplers to 4-step distilled models without retuning.

Generalizability across backbones. To establish that DAG is not specific to UNets, we evaluate the MMDiT backbones FLUX.1-Schnell and FLUX.1-Dev, both guidance-distilled (Table 2, left). Despite the difficulty of guiding pre-distilled models, DAG reaches GenEval scores of 0.738 and 0.737 with consistent gains in the human-preference metrics, and the same trend appears on FLUX.2 (4B/9B) and Qwen-Image (Table 2, middle). DAG therefore generalizes across guidance mechanisms, distillation regimes, and backbone families; as the analysis predicts, the gains follow the cross-attention structure it addresses rather than any one architecture.

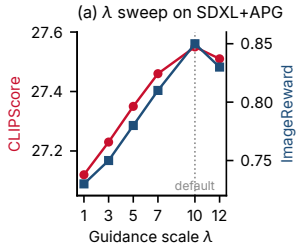
Beyond text-to-image. On DDIM inversion-reconstruction (Song et al., 2021a), where the attractor is the input image itself, sparse retrieval reconstructs more faithfully than dense and DAG refines it further (SSIM $0.746 \to 0.755 \to 0.757$; Table 22 in Appendix Q), a direct test of Theorem 1(i).

5 Analysis

5.1 Ablation Study and Component Analysis

Scale λ and NAG comparison. Sweeping $\lambda \in [1, 12]$ on SDXL+APG (5K samples; Figure 4a), GenEval, CLIPScore, and ImageReward improve up to the default λ=10 and remain stable beyond it; for MMDiT backbones, the sweep on FLUX.2-4B selects λ=5 (Table 6 in Appendix H). NAG contrasts positive and negative prompts and needs dual-pass inference per step; even at that budget it remains slightly below PLADIS, whereas 4-step DAG is the best of the three and improves further at 8 steps (Figure 4b). Composing DAG with NAG degrades performance, validating Remark 1.

Decomposition analysis (ζ ablation). We sweep $\zeta \in \{0, 0.25, 0.5, 0.75, 1\}$ on the same 10K MS-COCO prompts with the clamp on throughout (Figure 4c). Every metric improves monotonically as $\zeta \to 0$, with parallel-only (ζ=0) the best and orthogonal-only the weakest. This matches Theorem 1:



(b) NAG comparison on Hyper-SDXL

Method	NFE	CS ↑	IR ↑	PS ↑
Hyper-SDXL	4	26.09	0.959	22.87
+PLADIS	4	27.11	1.060	22.91
+DAG	4	**27.55**	**1.071**	**22.93**
+NAG	>4	27.10	1.046	22.86
+DAG (inc.)	>4	26.64	0.971	22.88
+PLADIS	8	27.21	1.181	23.10
+DAG	8	**27.78**	**1.213**	**23.13**

(c) ζ ablation on SDXL

ζ	CS ↑	IR ↑	PS ↑
0.00 (Par.)	**27.25**	**0.811**	**22.39**
0.25	27.13	0.789	22.31
0.50	27.01	0.763	22.15
0.75	26.82	0.755	22.11
1.00 (Full)	26.63	0.735	22.10
Ortho.	25.25	0.415	21.79

Figure 4: (a) λ sweep on SDXL+APG. (b) DAG with NAG (inc.: DAG on top of NAG at both default scales) degrades, validating Remark 1. (c) ζ sweep: parallel-only is best. Full metrics in Appendix J.

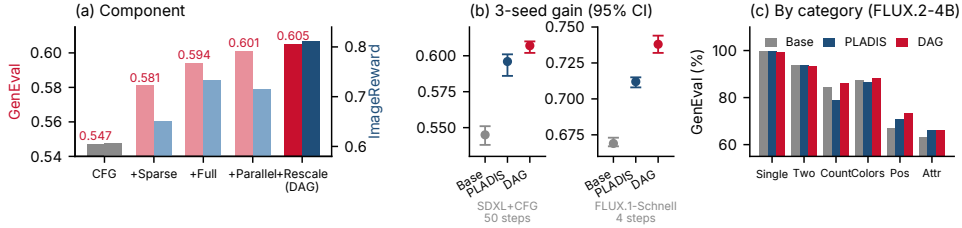


Figure 5: (a) Component isolation on SDXL+CFG. (b) Three-seed means with 95% CIs. (c) GenEval by category on FLUX.2-4B: gains concentrate on compositional categories.

the parallel component along $\mathcal{T}_\alpha(\mathbf{x})$ carries the alignment between $r(\mathbf{x})$ and $\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$, while the orthogonal component acts as off-manifold noise, motivating $\zeta = 0$ as our default (Proposition 2); the rescaling in (12) then bounds the magnitude of that direction at high λ (Appendix M).

Component isolation. To separate what is new relative to PLADIS from what is adapted, Figure 5(a) isolates the ingredients under one setup (SDXL+CFG, λ=10). GenEval rises with each added component. The projection raises GenEval, CLIPScore, and PickScore over the full residual while ImageReward dips slightly, and the norm clamp adapted from APG recovers ImageReward and adds the largest CLIPScore gain; the two are complementary (Table 13 in Appendix K).

Robustness across seeds. To confirm that the margins over PLADIS are not seed effects, we regenerate SDXL+CFG and FLUX.1-Schnell with three seeds (Figure 5b); the CFG and Schnell rows of Tables 1 and 2 report these means. DAG's bootstrap 95% intervals separate from the baseline's in every metric and from PLADIS's in all but one cell, so the improvements are statistically significant (full intervals in Table 14 in Appendix K).

Table 3: Three-way user study (95% CIs).

Method	Text Align.	Visual Quality
Baseline	10.7% [7.9−13.8]	16.7% [13.1−20.5]
PLADIS	12.9% [9.8−16.2]	31.4% [27.1−36.0]
DAG	76.4% [72.1−80.5]	51.9% [47.1−56.7]

Computational overhead. Measured end-to-end on one H100 with the same implementation and execution for all methods, DAG adds 0.79 s per image to the base sampler on SDXL (50 steps) and 0.12 s and 0.04 s on FLUX.1-Schnell and FLUX.2-4B (4 steps), with peak memory unchanged on FLUX (Table 2, right). Our implementation computes α−Entmax with a Triton kernel, realizes the ζ=0 projection as a scalar rescaling, and never materializes the joint attention on FLUX, which makes it 1.4–1.8× faster than the public PLADIS implementation (Appendix I).

Gains by prompt category. By GenEval category on FLUX.2 (4B/9B), the saturated single-object, two-object, and color categories barely move while counting, position, and attribute binding improve (Figure 5c; Table 15 in Appendix K), so the gains concentrate where several objects and attributes must be resolved.

User study. In a three-way blind preference study (31 participants, 50 random prompts; Table 3), DAG's preference rates for text alignment and visual quality have 95% bootstrap intervals that do not overlap either baseline's (protocol in Appendix N).

Projection-target ablation. Theorem 1 certifies the residual, while the projection onto $\mathcal{T}_\alpha$ in (11) is a design choice. To validate it, we compare against value-centered and dominant-token alternatives that change only the projection target; $\mathcal{T}_\alpha$ is best on every metric (Table 16 in Appendix L).

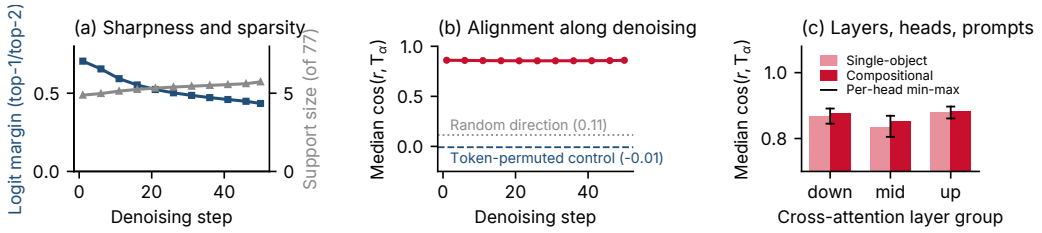


Figure 6: SDXL cross-attention relative to Theorem 1: (a) sharp but not collapsed retrieval, (b) aligned residual versus controls, (c) alignment across layers, heads, and prompts (full: Figure 8).

5.2 Empirical Analysis of Real Cross-Attention Layers

In this section, we analyze where real cross-attention layers stand relative to the regime of Theorem 1 and whether the mechanism behind Proposition 2 is observed. We measured SDXL over 250 stratified GenEval prompts, all 70 cross-attention layers and heads, and all 50 denoising steps (protocol and further panels in Appendix O).

Retrieval is sharp but below the collapse threshold (Theorem 1(i)). Since $\beta = 1/\sqrt{d}$ is fixed, sharpness is governed by the logit margin and the α-Entmax support. The top-1 pattern leads by $1.8\times$ early and $1.5\times$ late while the support stays at 4–7 of 77 tokens (Figure 6a): a sharpened regime below the single-token collapse of Theorem 1, in which retrieval weakens along denoising (margin -38%, support $+17\%$) but remains sparse and top-1 dominant, so alignment is a matter of measurement.

The residual points along retrieval (Theorem 1(ii)). Since the stored pattern $\boldsymbol{\xi}_\mu$ is not identified inside a trained layer, we measure the alignment with the sparse retrieval direction $\mathcal{T}_\alpha$, which (i) places closer to $\boldsymbol{\xi}_\mu$. The median $\cos(r, \mathcal{T}_\alpha)$ is 0.86 in probability space at every step with no non-positive measurement, and 0.75–0.81 in retrieval space with 0.08% non-positive (Figure 6b; Appendix O), against 0.11 for a random direction and -0.01 for a token-permuted control that keeps the sparsity pattern but scrambles where retrieval points. The alignment thus reflects where retrieval points rather than sparsity, and it stays flat ($0.861 \to 0.860$) while sharpness decays. This is a consistency check on (ii) rather than a test of its condition, which involves $\boldsymbol{\xi}_\mu$.

Alignment is uniform across layers, heads, and prompt types (Theorem 1(ii)). The alignment in (ii) is positive in every layer group and every head (Figure 6c), compositional prompts align slightly better than single-object prompts (Mann–Whitney $p < 10^{-4}$), and $r^{\parallel}$ dominates $r^{\perp}$ throughout, by a factor of 1.50–1.88 (Figure 8c).

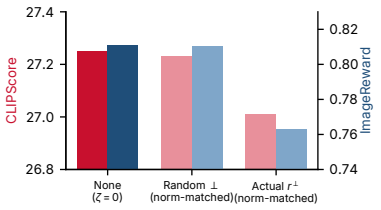


Figure 7: Norm-matched injection of $r^{\perp}$ on SDXL.

Direction versus scale of $r^{\perp}$ (Proposition 2). Proposition 2 attributes the stability of $\zeta{=}0$ to Assumption 1, whose attractor is unobservable; the observable test is whether the harm of $r^{\perp}$ lies in its direction or its scale. A random vector orthogonal to $\mathbf{P}_\alpha$ at the norm of $r^{\perp}$ leaves every metric unchanged, while the actual $r^{\perp}$ at the same norm degrades all of them (Figure 7; Table 21 in Appendix P), so $\zeta{=}0$ removes a harmful direction rather than excess magnitude (Remark 2). Taken together, real cross-attention operates below the collapse threshold of Theorem 1(i), the alignment of Theorem 1(ii) is observed throughout sampling, and $r^{\perp}$ is harmful through its direction, as Proposition 2 describes.

6 Conclusion

We proposed DAG, a single-pass attention guidance rule derived from a directional analysis of the sparse–dense attention discrepancy under Modern Hopfield dynamics, and analyzed the retrieval regime in which real cross-attention layers operate. Across UNet and MMDiT backbones and multi-step and few-step regimes, DAG improves text alignment and human-preference scores on every backbone tested at a small added cost per image, and the directional view extends to any setting where two retrieval operators share their memories, including inversion and editing.

AI USE STATEMENT

We used a large language model for copy-editing of the manuscript, for formatting BibTeX entries, and for drafting the reference-verification scripts. All theoretical results, experiments, code, and reported numbers were produced and verified by the authors, and we take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Hyperparameters and per-backbone settings are listed in Appendix H, the measurement protocol of the diagnostics in Appendix O, and the implementation details and exact versions in Appendix I. Complete proofs are given in Appendices C–E. Code will be released upon publication.

ETHICS STATEMENT

DAG is an inference-time method on publicly released text-to-image models and inherits their known risks of synthetic-media misuse, copyright, and biased outputs; mitigation rests with the model providers' safety practices. The user study used unpaid volunteers recruited through an anonymous link, with no personal data collected.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance. In *European Conference on Computer Vision (ECCV)*, pp. 1–17. Springer, 2024. doi: 10.1007/978-3-031-73464-9_1.

Ya I Alber and Sylvie Guerre-Delabriere. Principle of weakly contractive maps in Hilbert spaces. In *New results in operator theory and its applications: The Israel M. Glazman memorial volume*, pp. 7–22. Springer, 1997.

Luca Ambrogioni. In Search of Dispersed Memories: Generative Diffusion Models are Associative Memory Networks. *arXiv preprint arXiv:2309.17290*, 2023.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Adriano Barra, Matteo Beccaria, and Alberto Fachechi. A new mechanical approach to handle generalized Hopfield neural networks. *Neural Networks*, 106:205–222, 2018.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2. https://github.com/black-forest-labs/flux2, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized Attention Guidance: Universal Negative Guidance for Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. arXiv:2505.21179.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.14135.

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

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.03206.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Nuno Gonçalves, Marcos Treviso, and André F. T. Martins. AdaSplash: Adaptive Sparse Flash Attention. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.12082.

Michael Held, Philip Wolfe, and Harlan P Crowder. Validation of subgradient optimization. *Mathematical programming*, 6:62–88, 1974.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. arXiv:2104.08718.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Susung Hong. Smoothed Energy Guidance: Guiding Diffusion Models with Reduced Energy Curvature of Attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2408.00760.

Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in Plain Sight: Surveying the Uncanny Resemblances of Associative Memories and Diffusion Models. *arXiv preprint arXiv:2309.16750*, 2023.

John J. Hopfield. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On Sparse Modern Hopfield Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2309.12673.

Cheng Jin, Zhenyu Xiao, Chutao Liu, and Yuantao Gu. Angle Domain Guidance: Latent Diffusion Requires Rotation Rather Than Extrapolation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025. arXiv:2506.11039.

Kwanyoung Kim. Toward the Frontiers of Reliable Diffusion Sampling via Adversarial Sinkhorn Attention Guidance. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)*, 2026. doi: 10.1609/aaai.v40i7.37488.

Kwanyoung Kim and Byeongsu Sim. 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 (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.01507.

Kwanyoung Kim and Jong Chul Ye. Noise2Score: Tweedie's Approach to Self-supervised Image Denoising Without Clean Images. *Advances in Neural Information Processing Systems*, 34: 864–874, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.02747.

Jun Liu and Jieping Ye. Efficient euclidean projections in linear time. In *Proceedings of the 26th annual international conference on machine learning*, pp. 657–664, 2009.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Christian Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50:195–200, 1986.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse Sequence-to-Sequence Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1504–1519, 2019. doi: 10.18653/v1/P19-1146.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J. Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to Generalization: Emergence of Diffusion Models from Associative Memory. *arXiv preprint arXiv:2505.21777*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.01952.

Qwen Team. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token Perturbation Guidance for Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. arXiv:2506.10036.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2008.02217.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-SD: Trajectory Segmented Consistency Model for Efficient Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2404.13686.

B. E. Rhoades. Some theorems on weakly contractive maps. *Nonlinear Analysis: Theory, Methods & Applications*, 47(4):2683–2693, 2001. doi: 10.1016/S0362-546X(01)00388-1.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02416.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation. In *European Conference on Computer Vision (ECCV)*, pp. 87–103. Springer, 2024. doi: 10.1007/978-3-031-73016-0_6.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Alex Toth and C. T. Kelley. Convergence analysis for Anderson acceleration. *SIAM Journal on Numerical Analysis*, 53(2):805–819, 2015. doi: 10.1137/130919398.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Homer F Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

Wan Team. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased Consistency Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2405.18407.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Dennis Wu, Jerry Yao-Chieh Hu, Weijian Li, Bo-Yu Chen, and Han Liu. STanHop: Sparse Tandem Hopfield Model for Memory-Enhanced Time Series Prediction. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2312.17346.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10629.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.06072.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2405.14867.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Directional Attention Guidance for Diffusion Models via Sparse Hopfield Retrieval

Supplementary Material

A Overview

This supplementary document collects theoretical details, implementation specifics, and extended experimental results, ordered to follow the main text:

- **Theoretical Foundations** (Appendix B): background on (Sparse) Modern Hopfield Networks, their connection to (cross-)attention, the retrieval-error bounds used by the directional analysis, and the noise robustness of α-Entmax.
- **Proofs** (Appendices C, D, E): Theorem 1, Corollary 1, and Proposition 2.
- **Anderson Acceleration m=2 Surrogate** (Appendix F): an explicit m=2 surrogate and its comparison with the implicit m=1 form of Sec. 3.5.
- **Implementation Details and Settings** (Appendices G, H): sampling protocol, metrics, baselines, per-backbone hyperparameters, and the λ and η sweeps.
- **Computational Overhead** (Appendix I): per-component wall-clock and the optimized implementation.
- **Complete Metrics** (Appendix J): all metrics behind Figure 4 and the FLUX.2/Qwen-Image block of Table 2.
- **Robustness and Ablations** (Appendices K, L, M): component isolation, three-seed confidence intervals, GenEval by category, the projection-target ablation, and the rescaling ablation.
- **User Study Protocol** (Appendix N).
- **Alignment Diagnostics** (Appendices O, P): additional diagnostic panels behind Sec. 5.2, per-layer and per-head tables, and the norm-matched injection test.
- **DDIM Inversion-Reconstruction** (Appendix Q): setup and qualitative reconstructions for the test summarized in Sec. 4.
- **Extended Discussion** (Appendix R): limitations, Hopfield–diffusion connections, and the choice of cross-attention.
- **Additional Qualitative Results** (Appendix S): visual comparisons across guidance methods, step-distilled models, and MMDiT/DiT backbones.

B Theoretical Background

This section provides a self-contained background on (Sparse) Modern Hopfield Networks, the connection to (cross-)attention, and the retrieval-error bounds used in the proofs of Theorem 1, Corollary 1, and Proposition 2.

Notations. For $a \in \mathbb{R}$, $a_+ := \max\{0, a\}$. For $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$, $\langle \mathbf{z}, \mathbf{z}' \rangle = \mathbf{z}^\top \mathbf{z}'$ denotes the inner product. For $\mathbf{z} = (z_1, \dots, z_d) \in \mathbb{R}^d$, the sorted coordinates are written $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(d)}$, where $z_{(\nu)}$ is the ν-th largest entry. We write $\Delta^{M-1} := \{\mathbf{p} \in \mathbb{R}^M : p_i \geq 0, \sum_i p_i = 1\}$ for the $(M-1)$-dimensional simplex.

Modern Hopfield Networks. The classical Hopfield model (Hopfield, 1982) stores binary patterns as local minima of an associated energy function and retrieves the closest minimum given a query input; numerous extensions have been proposed to improve stability and capacity (Demircigil et al., 2017; Krotov & Hopfield, 2016; Barra et al., 2018). Ramsauer et al. (2021) introduced the *Modern Hopfield Network* (MHN), a continuous-state, differentiable variant equipped with an energy function

E and retrieval dynamics $\mathcal{T}$ that achieve pattern retrieval in a single update:

$$E_{\texttt{Dense}}(\mathbf{x}) := -\texttt{lse}(\beta, \mathbf{\Xi}^\top \mathbf{x}) + \tfrac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle, \tag{15}$$

$$\mathcal{T}_{\texttt{Dense}}(\mathbf{x}) := \mathbf{\Xi}\,\texttt{Softmax}(\beta\,\mathbf{\Xi}^\top \mathbf{x}), \tag{16}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the query, $\mathbf{\Xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M] \in \mathbb{R}^{d\times M}$ collects the stored patterns $\boldsymbol{\xi}_i \in \mathbb{R}^d$, and $\texttt{lse}(\beta, \mathbf{z}) := \frac{1}{\beta}\log\sum_i \exp(\beta z_i)$ is the log-sum-exponential function. Convergence and state properties of the energy and retrieval dynamics are established in (Ramsauer et al., 2021).

Connection with Transformer attention. A key observation of Ramsauer et al. (2021) is that the MHN update rule coincides with one Picard step of self-attention. Following (Ramsauer et al., 2021), $\mathcal{T}_{\texttt{Dense}}$ is first extended to multiple queries $\mathbf{X} := \{\mathbf{x}_i\}_{i\in[N]}$. With raw query $\mathbf{R}$ and memory matrix $\mathbf{Y}$, define the query/key matrices via $\mathbf{X}^\top = \mathbf{R}\mathbf{W}_Q := \mathbf{Q}$ and $\mathbf{\Xi}^\top = \mathbf{Y}\mathbf{W}_K := \mathbf{K}$. Taking the transpose and projecting $\mathbf{K}$ to $\mathbf{V}$ via $\mathbf{W}_V$ in (16), we obtain

$$\mathcal{T}_{\texttt{Dense}}(\mathbf{X}) = \texttt{Softmax}(\beta\,\mathbf{Q}\mathbf{K}^\top)\,\mathbf{K}\mathbf{W}_V = \texttt{Softmax}(\beta\,\mathbf{Q}\mathbf{K}^\top)\,\mathbf{V}, \tag{17}$$

which corresponds to standard scaled dot-product attention with $\beta = 1/\sqrt{d}$. Adopting the notation of (Ramsauer et al., 2021),

$$\mathcal{T}_{\texttt{Dense}}(\mathbf{X}) = \texttt{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\,\mathbf{V} := \texttt{At}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \tag{18}$$

This interpretation extends naturally to cross-attention with inputs $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathcal{T}_{\texttt{Dense}}(\mathbf{X}, \mathbf{Y}) = \texttt{Softmax}\big(\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{Y}^\top/\sqrt{d}\big)\mathbf{Y}\mathbf{W}_V = \texttt{At}(\mathbf{W}_Q\mathbf{X},\, \mathbf{W}_K\mathbf{Y},\, \mathbf{W}_V\mathbf{Y}).$$

Our analysis focuses on this cross-attention setting, in which text tokens act as stored memory patterns and image tokens as queries. Since different layers may store distinct patterns, MHNs naturally facilitate hierarchical storage and retrieval; whereas a basic Hopfield network outputs a single pattern, the attention mechanism yields multiple patterns simultaneously, so attention can be viewed as a stack of multi-valued Hopfield retrievals.

Sparse Hopfield Network and α-Entmax. A sparse extension of MHN is introduced in (Hu et al., 2023; Wu et al., 2024), replacing the softmax mapping with $\alpha\texttt{-Entmax}$:

$$E_\alpha(\mathbf{x}) := -\mathbf{\Psi}^\star_\alpha(\beta, \mathbf{\Xi}^\top \mathbf{x}) + \tfrac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle, \tag{19}$$

$$\mathcal{T}_\alpha(\mathbf{x}) := \mathbf{\Xi}\,\alpha\texttt{-Entmax}(\beta\,\mathbf{\Xi}^\top \mathbf{x}), \tag{20}$$

where $\mathbf{\Psi}^\star_\alpha$ is the convex conjugate of the negative Tsallis entropy (Tsallis, 1988) Ψ_α, defined by

$$\Psi_\alpha(\mathbf{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)}\sum_{i=1}^M (p_i^\alpha - p_i), & \alpha \neq 1, \\ \sum_{i=1}^M p_i \log p_i, & \alpha = 1, \end{cases} \tag{21}$$

and the probability mapping $\alpha\texttt{-Entmax}(\mathbf{z}) := \arg\max_{\mathbf{p}\in\Delta^{M-1}} \big[\langle \mathbf{p}, \mathbf{z}\rangle - \Psi_\alpha(\mathbf{p})\big]$. The mapping admits the closed form (Peters et al., 2019)

$$\alpha\texttt{-Entmax}(\mathbf{z}) = \big[(\alpha-1)\mathbf{z} - \tau(\mathbf{z})\,\mathbf{1}\big]_+^{1/(\alpha-1)}, \tag{22}$$

for a unique threshold function $\tau : \mathbb{R}^M \to \mathbb{R}$, so entries below $\tau(\mathbf{z})/(\alpha-1)$ map to zero, yielding sparsity. Throughout, $\kappa(\mathbf{z})$ denotes the number of nonzero entries. The parameter α controls sparsity: $\alpha = 1$ recovers softmax (and hence $\mathcal{T}_1 \equiv \mathcal{T}_{\texttt{Dense}}$ (Peters et al., 2019)); $\alpha = 2$ corresponds to $\texttt{sparsemax}$ (Martins & Astudillo, 2016), computable via sorting (Held et al., 1974; Michelot, 1986) or bisection (Liu & Ye, 2009); and $\alpha = 1.5$ admits an exact sorting-based algorithm (Peters et al., 2019). For general $1 < \alpha < 2$, a bisection scheme applies (Peters et al., 2019).

Extension to sparse attention. Analogously to (17)–(18), the sparse retrieval dynamics extend directly to attention. For multiple queries,

$$\mathcal{T}_\alpha(\mathbf{X}) = \alpha\texttt{-Entmax}\big(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}\big)\,\mathbf{V} := \texttt{At}_\alpha(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \tag{23}$$

and for cross-attention,

$$\mathcal{T}_\alpha(\mathbf{X}, \mathbf{Y}) = \alpha\texttt{-Entmax}\big(\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{Y}^\top/\sqrt{d}\big)\,\mathbf{Y}\mathbf{W}_V = \texttt{At}_\alpha(\mathbf{W}_Q\mathbf{X},\, \mathbf{W}_K\mathbf{Y},\, \mathbf{W}_V\mathbf{Y}).$$

This extension lets the network selectively retrieve relevant patterns by zeroing out unrelated keys, while inheriting the theoretical properties of sparse associative memories.

Pattern stored / retrieved. For the retrieval-error statements that follow, we recall the standard pattern-stored definition (Hu et al., 2023; Wu et al., 2024).

Definition 2 (Pattern stored and retrieved (Ramsauer et al., 2021; Hu et al., 2023; Wu et al., 2024))**.** Suppose every pattern $\boldsymbol{\xi}_\mu$ lies in a ball B_μ. We say $\boldsymbol{\xi}_\mu$ is **stored** if there exists a unique fixed point $\mathbf{x}^\star_\mu \in B_\mu$ to which all $\mathbf{x} \in B_\mu$ converge, with the $\{B_\mu\}$ disjoint. We say $\boldsymbol{\xi}_\mu$ is **retrieved** with error ϵ if $\big\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \le \epsilon$ for all $\mathbf{x} \in B_\mu$.

Setting $m := \max_\nu \|\boldsymbol{\xi}_\nu\|$, the next proposition collects the retrieval-error bounds from the Hopfield literature that we use throughout the proofs.

Proposition 3 (Retrieval Error (Ramsauer et al., 2021; Hu et al., 2023; Wu et al., 2024; Kim & Sim, 2025))**.** *Let $\mathcal{T}_\alpha$ be the Hopfield retrieval with α-*`Entmax`*. Then:*

$$\alpha = 1: \qquad \big\|\mathcal{T}_1(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \le 2m(M-1)\exp\Big\{-\beta\big(\langle\boldsymbol{\xi}_\mu, \mathbf{x}\rangle - \max_{\nu\neq\mu}\langle\boldsymbol{\xi}_\nu, \mathbf{x}\rangle\big)\Big\}, \tag{24}$$

$$1 < \alpha \le 2: \qquad \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \le m + m\,\kappa\Big[(\alpha-1)\beta\big(\max_\nu\langle\boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\top\mathbf{x}]_{(\kappa+1)}\big)\Big]^{\frac{1}{\alpha-1}}, \tag{25}$$

$$\alpha = 2: \qquad \big\|\mathcal{T}_2(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \le m + d^{1/2}\,m\beta\Big[\kappa\big(\max_\nu\langle\boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\top\mathbf{x}]_{(\kappa)}\big) + 1/\beta\Big], \tag{26}$$

$$\alpha > \alpha': \qquad \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \le \big\|\mathcal{T}_{\alpha'}(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| \quad \textit{for } \mathbf{x} \in B_\mu. \tag{27}$$

The bounds in (24), (25), (26), (27) are established, respectively, in (Ramsauer et al., 2021) (in the form of Hu et al. (2023)), (Kim & Sim, 2025), (Hu et al., 2023), and (Hu et al., 2023; Wu et al., 2024) (for $\alpha = 2$ and for α-`Entmax`*, respectively). We adopt the convention $[\boldsymbol{\Xi}^\top\mathbf{x}]_{(M+1)} := [\boldsymbol{\Xi}^\top\mathbf{x}]_{(M)} - M^{1-\alpha}/(\alpha-1)$.*

Theorem 1(i) uses the ordering of the bounds (24)–(26), which Hu et al. (2023) and Wu et al. (2024) show to be tighter for sparse retrieval, together with the exact retrieval above the collapse threshold; the pointwise ordering (27) is stated in those works on the storage ball and is not used in our proofs.

Corollary 2 (Noise Robustness (Hu et al., 2023; Wu et al., 2024; Kim & Sim, 2025))**.** *For noisy queries $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta}$ or noisy memories $\tilde{\boldsymbol{\xi}}_\mu = \boldsymbol{\xi}_\mu + \boldsymbol{\eta}$, the noise impact on $\big\|\mathcal{T}_2(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|$ is **linear**; for $1 < \alpha \le 2$ it scales **polynomially** with order $\frac{1}{\alpha-1}$; whereas for $\alpha = 1$ it grows **exponentially** in the noise.*

These bounds quantify why sparse retrieval is more noise-robust than dense retrieval, a property that matches the diffusion setting, in which queries carry residual Gaussian noise across denoising steps.

C Proof of Theorem 1 (Directional Surrogate Property)

Setup. As in the main text, $\mathcal{T}_\alpha(\mathbf{x}) = \mathbf{V}\,\mathbf{p}_\alpha(\mathbf{x})$ and $\mathcal{T}_{\texttt{Dense}}(\mathbf{x}) = \mathbf{V}\,\mathbf{p}_{\texttt{dense}}(\mathbf{x})$ share $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ and conditioning $\mathbf{c}$, so the discrepancy $r(\mathbf{x}) := \mathcal{T}_\alpha(\mathbf{x}) - \mathcal{T}_{\texttt{Dense}}(\mathbf{x}) = \mathbf{V}\,\boldsymbol{\delta}(\mathbf{x})$ lies in $\mathrm{span}(\mathbf{V})$, with $\boldsymbol{\delta}(\mathbf{x}) := \mathbf{p}_\alpha(\mathbf{x}) - \mathbf{p}_{\texttt{dense}}(\mathbf{x}) \in \mathbb{R}^M$ a difference of two simplex points. We write $z_j := \langle\boldsymbol{\xi}_j, \mathbf{x}\rangle$ for the logits, so that $\Delta(\mathbf{x}) = z_\mu - \max_{j\neq\mu} z_j$.

(i) Convergence rate separation. *Step 1 (bounds).* The retrieval-error bounds of Proposition 3 are tighter for sparse retrieval than for dense retrieval (Hu et al., 2023; Wu et al., 2024); with $\alpha' = 1$ (the dense case, since 1-`Entmax` $\equiv$ `Softmax` (Peters et al., 2019)) and $\alpha \in (1, 2]$ (the sparse case used in PLADIS-style guidance), this gives the ordering of the bounds in (i).

Step 2 (exact retrieval above the collapse threshold). α-`Entmax` has the closed form (Peters et al., 2019)

$$(\mathbf{p}_\alpha)_i \;=\; \big[(\alpha-1)\beta z_i - \tau\big]_+^{1/(\alpha-1)},$$

with τ chosen so that the entries sum to one. The support is $\{\mu\}$ if and only if $(\alpha-1)\beta z_\mu - \tau = 1$ and $(\alpha-1)\beta z_j - \tau \le 0$ for all $j \neq \mu$, i.e., $(\alpha-1)\beta(z_\mu - z_j) \ge 1$ for all $j \neq \mu$, which is $\beta \ge \beta_0(\mathbf{x}) = 1/((\alpha-1)\Delta(\mathbf{x}))$. In that case $\mathbf{p}_\alpha = \mathbf{e}_\mu$ and $\mathcal{T}_\alpha(\mathbf{x}) = \boldsymbol{\xi}_\mu$.

Therefore, for $\beta \ge \beta_0(\mathbf{x})$, $\big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\| = 0 \le \big\|\mathcal{T}_{\texttt{Dense}}(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|$, which is (8).

(ii) Directional alignment. *Step 1 (identity).* For any a, b, c in a Hilbert space,

$$\langle a - b,\, c - b\rangle \;=\; \tfrac{1}{2}\big(\|a-b\|^2 + \|c-b\|^2 - \|a-c\|^2\big).$$

With $a = \mathcal{T}_\alpha(\mathbf{x})$, $b = \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$, and $c = \boldsymbol{\xi}_\mu$, we have $a - b = r(\mathbf{x})$ and $c - b = \boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$, so

$$\big\langle r(\mathbf{x}),\, \boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x}) \big\rangle \;=\; \tfrac{1}{2}\Big(\|r(\mathbf{x})\|^2 + \big\|\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})\big\|^2 - \big\|\boldsymbol{\xi}_\mu - \mathcal{T}_\alpha(\mathbf{x})\big\|^2 \Big).$$

Step 2 (inequality). Dropping the non-negative term $\big\|\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})\big\|^2$ yields

$$\big\langle r(\mathbf{x}),\, \boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x}) \big\rangle \;\geq\; \tfrac{1}{2}\Big(\|r(\mathbf{x})\|^2 - \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|^2 \Big),$$

whose right-hand side is non-negative whenever $\|r(\mathbf{x})\| \geq \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|$.

Step 3 (collapse regime). For $\beta \geq \beta_0(\mathbf{x})$, $\mathcal{T}_\alpha(\mathbf{x}) = \boldsymbol{\xi}_\mu$ by (i), so $r(\mathbf{x}) = \boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$ and the inner product equals $\big\|\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})\big\|^2$.

Therefore, $r(\mathbf{x})$ is non-negatively aligned with $\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$ whenever $\|r(\mathbf{x})\| \geq \big\|\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu\big\|$, and exactly equals that residual above the collapse threshold, which is (9). □

Remark (coordinate-wise sharpening). Coordinate-wise sharpening of $\mathbf{p}_\alpha$ relative to $\mathbf{p}_{\texttt{dense}}$ ($\delta_\mu \geq 0$ and $\delta_j \leq 0$ for all $j \neq \mu$) does not hold in general: for logits $\beta\mathbf{z} = (3,\, 2.7,\, 1.8)$, $1.5\text{-}\texttt{Entmax}$ moves mass from the third token to both of the others, giving $\boldsymbol{\delta} = (+0.100,\, +0.019,\, -0.119)$. This is why (ii) is stated through retrieval errors rather than through the entries of $\boldsymbol{\delta}$.

Quantitative form of the collapse threshold. For $\alpha = 1.5$ the condition $\beta \geq \beta_0(\mathbf{x})$ reads $\beta\Delta(\mathbf{x}) \geq 2$, a scaled top-1/top-2 logit margin of at least 2. Sec. 5.2 reports the measured margins and support sizes, which lie below this threshold; in that regime the alignment in (ii) is a measured rather than a certified property.

D Proof of Corollary 1 (Directional Consistency)

(a) Direction preservation. Both $\mathcal{T}_\alpha(\mathbf{x})$ and $\mathcal{T}_{\texttt{Dense}}(\mathbf{x})$ live in $\mathrm{span}(\mathbf{V})$, so $r(\mathbf{x}) \in \mathrm{span}(\mathbf{V})$. Since $\mathcal{T}'(\mathbf{x}) = \mathcal{T}_\alpha(\mathbf{x}) + \lambda r(\mathbf{x})$ with $\mathcal{T}_\alpha(\mathbf{x}) \in \mathrm{span}(\mathbf{V})$, we have $\mathcal{T}'(\mathbf{x}) \in \mathrm{span}(\mathbf{V})$. Under the condition of Theorem 1(ii), $r(\mathbf{x})$ is non-negatively aligned with $\boldsymbol{\xi}_\mu - \mathcal{T}_{\texttt{Dense}}(\mathbf{x})$, so the extrapolation does not introduce off-direction drift away from $\boldsymbol{\xi}_\mu$.

(b) Common attractor at $\boldsymbol{\xi}_\mu$. As $\mathbf{x} \to \boldsymbol{\xi}_\mu$ with β sufficiently large and well-separated patterns, the dense retrieval-error bound in Proposition 3 goes to zero and sparse retrieval is exact above the collapse threshold of Theorem 1(i): $\mathcal{T}_\alpha(\mathbf{x}) \to \boldsymbol{\xi}_\mu$ and $\mathcal{T}_{\texttt{Dense}}(\mathbf{x}) \to \boldsymbol{\xi}_\mu$, hence $r(\mathbf{x}) \to \mathbf{0}$. Substituting in $\mathcal{T}'$:

$$\mathcal{T}'(\mathbf{x}) = \mathcal{T}_\alpha(\mathbf{x}) + \lambda r(\mathbf{x}) \;\longrightarrow\; \boldsymbol{\xi}_\mu + \mathbf{0} = \boldsymbol{\xi}_\mu.$$

Thus $\boldsymbol{\xi}_\mu$ is asymptotically a fixed point of $\mathcal{T}'$. □

E Proof of Proposition 2 (Asymptotic Convergence)

Proposition 2 (Asymptotic Convergence of the Orthogonal Error)**.** *Suppose Assumption 1 holds, and let $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ be generated by the iterated dynamics $\mathbf{x}_{t+1} = \mathcal{T}_\alpha^\lambda(\mathbf{x}_t)$ with $\zeta = 0$. Then the orthogonal error magnitude $u_t := \|\Pi_\perp(\mathbf{x}_t - \mathbf{x}^*)\|$ satisfies $\lim_{t\to\infty} u_t = 0$.*

Proof. Since $\Pi_\perp$ is fixed, applying Assumption 1 at $\mathbf{x} = \mathbf{x}_t$ gives

$$u_{t+1} \;\leq\; u_t - \phi(u_t), \qquad \forall t \geq 0, \tag{28}$$

with ϕ strictly increasing and $\phi(0) = 0$. Hence $u_{t+1} < u_t$ whenever $u_t > 0$, so $\{u_t\}$ is monotone non-increasing and bounded below by 0. By the Monotone Convergence Theorem, $u_\infty := \lim_t u_t$ exists. Taking limits in (28) gives $\phi(u_\infty) \leq 0$, which forces $u_\infty = 0$. □

F Anderson Acceleration: Explicit $m = 2$ Surrogate

The implicit $m = 1$ form realized by the sparse–dense contrast can be extended to an explicit $m = 2$ surrogate by drawing two distinct sparsity levels $\alpha_1 < \alpha_2$ and forming

$$\mathbf{x}_{k+1} \;=\; \omega_0\, \mathcal{T}_{\alpha_2}(\mathbf{x}_k) + \omega_1\, \mathcal{T}_{\alpha_1}(\mathbf{x}_k) + \omega_2\, \mathcal{T}_{\texttt{Dense}}(\mathbf{x}_k), \qquad \sum_i \omega_i = 1,$$

which approximates a 2-step AA history along the sparsity axis. Coefficients are obtained by minimizing the residual norm over Δ^2. Table 4 reports results with $\alpha_2 = 1.5$, $\alpha_1 = 1.25$ and least-squares-determined ω. The $m = 2$ surrogate does not exceed the $m = 1$ form, in line with the one-step retrieval picture.

Table 4: m=1 **vs.** m=2 **AA in attention space (SDXL).** An explicit $m = 2$ surrogate, constructed from a multi-α history, does not improve over the implicit $m = 1$ form, consistent with the one-step MHN retrieval picture.

Method	GE ↑	CS ↑	IR ↑
CFG	0.547	26.46	0.606
Attn-AA (m=1)	**0.597**	**26.99**	**0.801**
Attn-AA (m=2, approx.)	0.556	26.83	0.753

G Implementation Details

Default schedulers and step counts are used throughout (50/28/4 steps for SDXL/FLUX.1-Dev/FLUX.1-Schnell; 4 steps for the step-distilled models), on a single NVIDIA H100 with batch size 1. For SDXL evaluations in Table 1 (left), we sample 10,000 prompts from the validation set of MS-COCO (three seeds for the CFG and DAG rows). GenEval results use the official benchmark and report the mean overall score. For CFG-based comparisons we use scale $w = 5.0$; for PAG and APG we use the recommended scales 3.0 and 15.0, respectively. For step-distilled models (Hyper-SDXL, DMD2) and FLUX.1 variants we follow the official baseline configurations. For DAG, we set the rescaling threshold $\eta = 15$ and use $\zeta = 0$ throughout. Per-backbone settings, the selection rule, and sensitivity sweeps are in Appendix H.

MMDiT backbones. FLUX and Qwen-Image use joint attention over the concatenated text and image tokens. DAG is applied to the text-to-image block of the joint attention map only (image queries attending to text keys), where the shared-memory setting of Theorem 1 holds; the block is then rescaled so that each query keeps the mass that the dense softmax assigns to the text tokens, following the official PLADIS implementation (Kim & Sim, 2025). The remaining blocks are left as they are.

H Hyperparameter Settings and Sensitivity

Table 5 lists the settings used for every backbone. Two sweeps were run in total, one per architecture family and none per backbone: λ=10 for the SDXL family comes from the sweep in Figure 4(a), and λ=5 for every MMDiT backbone (FLUX.1, FLUX.2, Qwen-Image) comes from the sweep on FLUX.2-4B in Table 6. α=1.5 follows PLADIS, ζ=0 follows Theorem 1, and η=15 is the value used throughout: results are insensitive to η over $[1, 15]$ (Table 7) and degrade without the clamp (Table 17). PLADIS uses its recommended scale (λ=2.0) with α=1.5 and no clamp. Unlike the heuristic layer selection of PAG and SEG, DAG is applied to all cross-attention layers, following PLADIS.

Table 5: Per-backbone settings.

Backbone	λ	η	α	ζ	Layers
SDXL (and distilled variants)	10	15	1.5	0	all cross-attention
FLUX.1 (Dev, Schnell)	5	15	1.5	0	all cross-attention
FLUX.2 (4B, 9B)	5	15	1.5	0	all cross-attention
Qwen-Image	5	15	1.5	0	all cross-attention

Table 6: λ sweep (FLUX.2-4B, GenEval).

Configuration	GE ↑
Baseline (softmax)	0.824
DAG, λ=3	0.832
DAG, λ=5 (selected)	**0.843**
DAG, λ=7	0.834
DAG, λ=10	0.834

Table 7: η sweep (SDXL+CFG, λ=10).

η	GE ↑	CS ↑	IR ↑
1	0.598	27.11	0.805
5	0.602	27.22	0.808
10	0.605	27.25	0.811
15 (default)	0.605	27.25	0.811

I Computational Overhead and Practical Implementation

Table 2 (right) reports wall-clock and peak memory at identical execution for all methods. Table 8 compares the public reference implementations of PLADIS and DAG with our accelerated implementation. All numbers are end-to-end per image (text encoding, denoising, VAE decode) on one NVIDIA H100 80GB, 1024×1024, batch 1, with two warm-up and five timed runs (seven for SDXL); SDXL runs with CFG (scale 5.0, batch of 2), and the FLUX models are guidance-distilled and run without CFG. Peak memory is the maximum allocated over the run, weights included. The released implementation is the accelerated one.

What the accelerated implementation changes. (i) α−`Entmax` is computed with the root-finding Triton kernel of AdaSplash (Gonçalves et al., 2025) instead of the sort-based reference (two kernels instead of 28). (ii) The ζ=0 projection is rank one, so $r^{\|}$ is computed as a per-sample scalar rescaling of $\mathcal{T}_\alpha(\mathbf{x})$ rather than an explicit projection with fp64 temporaries (agreement with the explicit form 2.2×10^{-16} in fp64). (iii) On the FLUX family the guided weights are linear in the sparse and dense maps, so the image-to-text block is recomposed from flash attention (Dao et al., 2022) and its log-sum-exp without materializing the 4608×4608 joint attention matrix and its fp32 copy (4.5 GiB; agreement 8.8×10^{-16}). (iv) SDXL is launch-bound rather than GPU-bound (48% GPU-busy under standard execution), so the all-zero bias and unit-temperature operations are skipped and the UNet forward is replayed from a CUDA graph; this is applied to the base sampler, PLADIS, and DAG alike, and latents are bit-identical to standard execution. The accelerated path is algebraically identical to the reference except for $\mathbf{QK}^\top$ in fp16 on SDXL and two entmax iterations.

Table 8: Reference versus our implementation, per image on one H100. Base is the softmax sampler.

		Time (s) ↓			Peak memory (GiB) ↓		
Model	Implementation	Base	PLADIS	DAG	Base	PLADIS	DAG
SDXL, 50 steps	Reference	2.36	5.23	6.21	8.98	8.98	8.98
	Ours	2.36	2.93	3.15	8.98	8.98	8.98
FLUX.1-Schnell, 4 steps	Reference	0.88	1.82	1.94	33.83	36.44	36.44
	Ours	0.88	0.99	1.00	33.83	33.83	33.83
FLUX.2-4B, 4 steps	Reference	0.51	0.80	0.87	17.35	19.96	19.96
	Ours	0.51	0.54	0.55	17.35	17.35	17.35

On SDXL the sparse attention accounts for about two thirds of the added cost and the projection and clamp for the rest; on FLUX the projection and clamp add no measurable cost beyond the sparse attention (within ± 0.02 s).

J Complete Metrics for Figure 4 and Table 2

This section lists every metric behind Figure 4 and the FLUX.2/Qwen-Image block of Table 2: Table 9 (NAG comparison), Table 10 (ζ ablation), Table 11 (FLUX.2 and Qwen-Image), and Table 12 (λ sweep on SDXL).

Table 9: NAG comparison on Hyper-SDXL, all metrics.

Method	NFE	GE ↑	CS ↑	IR ↑	PS ↑	HPSv2 ↑	FID ↓
Hyper-SDXL	4	0.556	26.09	0.959	22.87	0.313	21.23
+ PLADIS	4	0.589	27.11	1.060	22.91	0.315	19.83
+ DAG	4	**0.594**	**27.55**	**1.071**	**22.93**	**0.317**	**19.12**
+ NAG	>4	0.588	27.10	1.046	22.86	0.315	19.92
+ DAG (inc., on top of NAG)	>4	0.571	26.64	0.971	22.88	0.312	20.83
+ PLADIS	8	0.592	27.21	1.181	23.10	0.317	19.35
+ DAG	8	**0.602**	**27.78**	**1.213**	**23.13**	**0.319**	**18.81**

Table 10: ζ ablation (SDXL+CFG, λ=10, clamp on for all rows), all metrics; FID and HPSv2 for ζ=0 are the Table 1 values.

ζ	FID ↓	CS ↑	IR ↑	PS ↑	HPSv2 ↑
0.00 (parallel)	**14.63**	**27.25**	**0.811**	**22.39**	**0.271**
0.25	15.35	27.13	0.789	22.31	0.268
0.50	15.56	27.01	0.763	22.15	0.270
0.75	15.71	26.82	0.755	22.11	0.265
1.00 (full)	15.92	26.63	0.735	22.10	0.262
Orthogonal only	19.23	25.25	0.415	21.79	0.260

Table 11: FLUX.2 and Qwen-Image, all metrics.

Method	GE ↑	CS ↑	IR ↑	PS ↑	HPSv2 ↑
FLUX.2 (4B)	0.824	27.53	1.156	23.01	0.311
+ PLADIS	0.826	27.52	1.168	23.03	0.311
+ DAG	**0.843**	**27.73**	**1.173**	**23.10**	**0.313**
FLUX.2 (9B)	0.849	27.68	1.213	23.15	0.315
+ PLADIS	0.865	27.81	1.223	23.14	0.313
+ DAG	**0.878**	**27.84**	**1.235**	**23.21**	**0.316**
Qwen-Image	0.824	27.63	1.123	23.13	0.309
+ PLADIS	0.828	27.63	1.156	23.15	0.310
+ DAG	**0.842**	**27.79**	**1.171**	**23.19**	**0.312**

Table 12: λ sweep on SDXL+APG (5K samples; Figure 4a).

λ	1	3	5	7	10	12
CS ↑	27.12	27.23	27.35	27.46	**27.55**	27.51
IR ↑	0.73	0.75	0.78	0.81	**0.85**	0.83
PS ↑	22.37	22.40	22.46	22.44	22.68	**22.71**

K Robustness Across Seeds and Prompts

Table 13: Component isolation under one setup (SDXL+CFG, λ=10, single seed; the CFG and DAG rows of Table 1 are three-seed means).

Configuration	GE ↑	CS ↑	IR ↑	PS ↑
CFG	0.547	26.46	0.606	22.31
+ sparse attention	0.581	26.49	0.651	22.29
+ full residual	0.594	26.54	0.732	22.11
+ parallel only	0.601	26.68	0.715	22.34
+ rescaling (= DAG)	**0.605**	**27.25**	**0.811**	**22.39**

Table 14: Means over three seeds; 95% intervals from a bootstrap over images pooled across seeds.

Setting	Method	GE ↑	CS ↑	IR ↑	PS ↑
SDXL+CFG, 50	Base	0.545 (0.538–0.551)	26.45 (26.41–26.53)	0.604 (0.601–0.610)	22.31 (22.21–22.33)
	+ PLADIS	0.596 (0.586–0.601)	26.97 (26.81–27.03)	0.799 (0.797–0.802)	22.32 (22.23–22.34)
	+ DAG	**0.607 (0.602–0.610)**	**27.26 (27.11–27.33)**	**0.812 (0.806–0.815)**	**22.40 (22.38–22.42)**
FLUX.1-Schnell, 4	Base	0.669 (0.667–0.673)	26.28 (26.21–26.31)	1.059 (1.051–1.063)	22.62 (22.55–22.64)
	+ PLADIS	0.712 (0.708–0.715)	26.45 (26.38–26.52)	1.066 (1.062–1.074)	22.31 (22.23–22.38)
	+ DAG	**0.738 (0.732–0.744)**	**26.56 (26.51–26.63)**	**1.114 (1.110–1.121)**	**22.80 (22.77–22.84)**

Table 15: GenEval category breakdown (553 prompts $\times$ 4 images, 4 steps).

Backbone	Method	Single	Two	Count	Colors	Pos	Attr	Overall
FLUX.2-4B	Base	**99.69**	**93.69**	84.38	87.23	66.75	63.00	82.46
	+ PLADIS	**99.69**	**93.69**	79.00	86.48	70.75	66.00	82.60
	+ DAG	99.38	93.18	**85.94**	**88.03**	**73.25**	**66.25**	**84.34**
FLUX.2-9B	Base	**99.69**	95.96	79.06	90.96	71.50	72.75	84.99
	+ PLADIS	99.38	**96.72**	83.44	91.22	75.75	73.00	86.58
	+ DAG	**99.69**	96.21	**87.19**	**91.76**	**76.75**	**75.20**	**87.80**

Failure cases. DAG modifies only the retrieval on the same trajectory, so its characteristic failure is the absence of improvement rather than a large degradation: when the baseline already retrieves the prompt correctly, the residual is small and the output is unchanged, and the rare degradations we observed are of the same kind, a slight over-sharpening of an already correct retrieval at high λ, which the clamp in (12) bounds.

L Projection-Direction Ablation

Theorem 1 certifies the full residual $r(\mathbf{x})$; the projection onto $\mathcal{T}_\alpha(\mathbf{x})$ is a choice, and we compare it with alternatives that change only the projection target (SDXL, GenEval 553 prompts, identical seeds, λ=3 for this controlled ablation): value-centered subtracts the mean value vector, $\mathcal{T}_\alpha - \bar{\mathbf{V}}$; dominant token uses the value vector of the most-attended token, either globally or per query. Each alternative replaces $\mathbf{P}_\alpha$ in the coefficient of Definition 1 by the probability-space preimage of the stated direction, so all four rows differ only in the projection target. Our choice is best on all four metrics (Table 16).

Table 16: Projection-direction ablation (SDXL, GenEval 553-prompt subset, λ=3; not comparable with Table 1).

Projection direction	GE ↑	CS ↑	IR ↑	PS ↑
$\mathcal{T}_\alpha$ (ours)	**0.615**	**28.99**	**0.833**	**22.85**
value-centered ($\mathcal{T}_\alpha - \bar{\mathbf{V}}$)	0.612	28.91	0.821	22.82
dominant token (global)	0.603	28.92	0.815	22.84
dominant token (per query)	0.596	28.83	0.786	22.56

M Rescaling Ablation

At high guidance scale ($\lambda = 10$), the parallel component $r_\parallel(\mathbf{x})$ tends to grow in magnitude, so the extrapolated update can drift outside the typical norm range of the underlying attention output. Following the norm rescaling of APG (Sadat et al., 2025), the discrepancy of a layer is clamped to Frobenius norm η before the projection, $\mathbf{\Delta} \leftarrow \min(1, \eta/\|\mathbf{\Delta}\|_F)\,\mathbf{\Delta}$ in (12); since the projection is linear, this scales $r^\parallel$ and $r^\perp$ by the same factor and preserves their directions and the decomposition of Definition 1. The clamp acts on the norm of the whole attention tensor of a layer, not on a per-query quantity, so η is not comparable with the per-query magnitudes of Appendix O; within $[1, 15]$ the results are insensitive to its value (Table 7), whereas removing it lets the update grow without bound at high λ and costs most of the ImageReward gain (Table 17).

Table 17 reports the effect on SDXL at $\lambda = 10$ across CLIP score (CS), ImageReward (IR), and PickScore (PS). Without rescaling, DAG already improves all three metrics over CFG; adding rescaling yields a further gain across the board, with the largest improvement on IR ($0.715 \rightarrow 0.811$) and a sizable gain on CS ($26.68 \rightarrow 27.25$). PS shifts only marginally, indicating that rescaling primarily curbs occasional over-amplification at high λ rather than altering the operating point. We use $\eta = 15$ throughout the main experiments.

Table 17: Effect of the rescaling factor at high guidance scale ($\lambda = 10$).

Method	CS ↑	IR ↑	PS ↑
CFG	26.46	0.606	22.31
DAG (w/o rescaling)	26.68	0.715	22.34
DAG (w/ rescaling)	**27.25**	**0.811**	**22.39**

N User Study Protocol

The study was a blinded three-way comparison with 31 participants on 50 random prompts (SDXL and FLUX). Each trial showed the three images (baseline, PLADIS, DAG) simultaneously, labeled A/B/C with randomized assignment, and participants picked the best image per criterion (text alignment, visual quality). Participants were unpaid volunteers recruited through an anonymous institutional link, blind to the paper and the methods, and were not screened. Preference rates are reported with 95% bootstrap confidence intervals over trials. The study is small (31 unpaid, unscreened participants; 50 prompts) and is reported as supporting evidence alongside the automatic metrics rather than as a primary result.

O Alignment Diagnostics

Protocol. SDXL-base-1.0, 250 GenEval prompts stratified over the six categories (125 single-object, 125 compositional), seed 42, 50 steps, CFG scale 5.0, all 70 cross-attention layers and all heads. Measurements are taken in observe mode: the sampled trajectory is the baseline softmax trajectory and the logged quantities do not intervene (pixel-identical outputs verified). Per query we record $\cos(r, \mathcal{T}_\alpha)$ with $r = \mathcal{T}_\alpha - \mathcal{T}_{\texttt{Dense}}$, the top-1 versus top-2 logit margin, the α-$\texttt{Entmax}$ support size (number of tokens with nonzero attention), and $\|r^\parallel\| / \|r^\perp\|$; tables report medians over queries and then over prompts, layers, heads, and steps within each cell. In total 16.25M (prompt×layer×head×step) measurements. The same analysis in retrieval space ($\mathcal{T} = \mathbf{pV}$) gives medians of 0.75–0.81 with 0.08% non-positive values; we report the probability-space quantity

because it is the one DAG operates on. Tables 18–20 report the alignment by denoising band and layer group, the per-head medians, and the logit margins.

Assumptions in multi-head cross-attention. The analysis of Sec. 3.2 is stated for one retrieval operator; in a diffusion backbone every head of every cross-attention layer is one such operator, with the text-token values as stored patterns and each spatial query as the input state, and $\beta = 1/\sqrt{d}$ fixed by the architecture. The projection of Definition 1 is taken over the attention tensor of a layer, whereas the diagnostics are resolved per head, so they test whether each head is in a sharp, well-separated regime, which is what the logit-margin and support-size measurements report.

Table 18: Median $\cos(r, \mathcal{T}_\alpha)$ by denoising-step band and layer group (single-object / compositional prompts).

Denoising steps	down	mid	up
early (1–17)	0.869 / 0.879	0.836 / 0.856	0.879 / 0.883
middle (18–34)	0.865 / 0.874	0.834 / 0.853	0.876 / 0.879
late (35–50)	0.867 / 0.875	0.832 / 0.851	0.882 / 0.881

Table 19: Per-head median $\cos(r, \mathcal{T}_\alpha)$ by layer group. No head has a non-positive median; within any single cell of Table 18 the most- and least-aligned heads differ by at most 0.021.

Layer group	median across heads	per-head min–max
down	0.875–0.877	0.845–0.891
mid	0.845–0.848	0.805–0.869
up	0.877–0.882	0.861–0.897

Table 20: Median top-1 versus top-2 logit margin by denoising-step band and layer group (single-object / compositional prompts). Its exponential is the top-1/top-2 probability ratio.

Denoising steps	down	mid	up
early (1–17)	0.578 / 0.520	0.409 / 0.352	0.367 / 0.338
middle (18–34)	0.439 / 0.396	0.369 / 0.308	0.322 / 0.295
late (35–50)	0.405 / 0.359	0.346 / 0.288	0.297 / 0.274

Summary statistics. Of the 16.25M measurements, 0.000% have $\cos \leq 0$ and 2.5% have $\cos < 0.7$. Per-layer medians span 0.767 to 0.924; the lower values concentrate in the high-resolution blocks and the higher values in the low-resolution blocks, a resolution-dependent pattern rather than noise. The prompt-level median is 0.874 for compositional and 0.868 for single-object prompts (Mann–Whitney $p = 2 \times 10^{-5}$); the strongest categories are attribute binding (0.884) and two-object (0.877). Along the trajectory the pooled margin falls by 38% (top-1/top-2 ratio $1.78 \to 1.50$) and the support grows by 17% ($4.9 \to 5.7$ of 77 tokens) while the median alignment moves from 0.861 to 0.860. The magnitude ratio $\|r^{\parallel}\| / \|r^{\perp}\|$ stays within 1.50–1.88. A token-permuted control, which keeps the sparsity pattern but permutes which tokens it points to, gives a median alignment of -0.008; a random direction in $\mathbb{R}^{77}$ has expected $|\cos| \approx 1/\sqrt{77} \approx 0.11$.

On Assumption 1. The exact orthogonal error u_t requires the attractor $\mathbf{x}^*$, which is not observable during sampling; the observable proxy $\|r^{\perp}\|$ shows no trend along the trajectory ($0.1345 \to 0.1334$). We therefore do not claim direct verification of the weak contraction hypothesis; the injection test of Appendix P supports reading it as a constraint on the direction of the orthogonal error rather than on its scale.

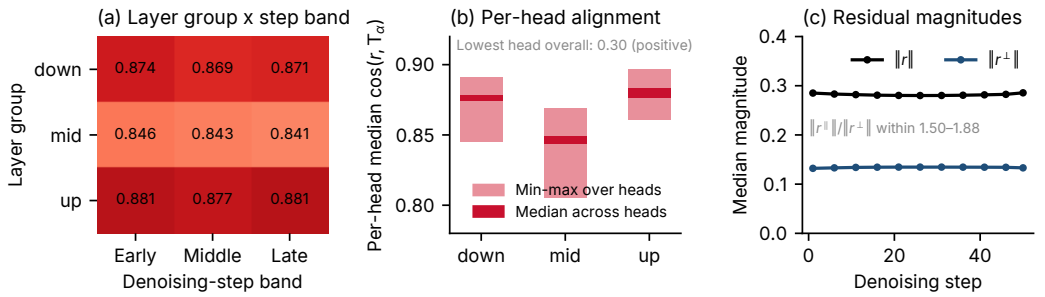


Figure 8: Additional diagnostics relative to Theorem 1: (a) alignment by layer group and denoising band, (b) per-head range of the median alignment, (c) residual magnitudes $\|r\|$ and $\left\|r^{\perp}\right\|$ along denoising.

P Orthogonal Component: Direction versus Scale

To separate the direction of $r^{\perp}$ from its scale, we run a norm-matched injection at $\zeta{=}0.5$ on SDXL, replacing the actual $r^{\perp}$ by a random vector orthogonal to $\mathbf{P}_{\alpha}$ in the space of Definition 1, at the norm of $r^{\perp}$ (Table 21). The random injection is nearly a no-operation, while the actual $r^{\perp}$ at identical norm degrades all three metrics: the harm comes from the direction of the orthogonal component, which is what the $\zeta{=}0$ projection removes.

Table 21: Norm-matched orthogonal injection (SDXL+CFG).

Injection	CS ↑	IR ↑	PS ↑
none ($\zeta{=}0$, reference)	**27.25**	**0.811**	**22.39**
random orthogonal, norm-matched	27.23	0.810	22.39
actual $r^{\perp}$	27.01	0.763	22.15

Q DDIM Inversion-Reconstruction

This section gives the setup and results of the DDIM inversion-reconstruction test summarized in Sec. 4 (Table 22).

Experimental setup. We sample 200 prompts at random from the MS-COCO validation set and use the corresponding real images as ground-truth attractors $\mathbf{x}^*$. All experiments are conducted on SDXL (Podell et al., 2024). The pipeline has two stages:

- **Inversion ($\mathbf{x}^* \to \mathbf{z}$).** For each real image $\mathbf{x}^*$, we obtain the inverted noise latent $\mathbf{z}$ via 50-step DDIM inversion (Song et al., 2021a) *without* Classifier-Free Guidance (CFG); this isolates the deterministic ODE trajectory and ensures that the reconstruction error reflects only the retrieval operator under test.
- **Reconstruction ($\mathbf{z} \to \hat{\mathbf{x}}$).** Starting from the same inverted latent $\mathbf{z}$, we run 50-step DDIM denoising with each retrieval operator: Dense ($\alpha{=}1.0$), Sparse ($\alpha{=}1.5$), and DAG (geometric decomposition with $\zeta{=}0$, $\eta{=}15$). All other hyperparameters and schedules are held fixed across the three settings.

We score each reconstruction $\hat{\mathbf{x}}$ against the original $\mathbf{x}^*$ using SSIM (Wang et al., 2004) (higher is better) and LPIPS (Zhang et al., 2018) (lower is better), averaged over the 200 prompts. Because $\mathbf{x}^*$ is known by construction, this directly tests the convergence-rate ordering predicted by Theorem 1(i): $\|\mathcal{T}_{\alpha}(\mathbf{x}) - \mathbf{x}^*\| \le \|\mathcal{T}_{\texttt{Dense}}(\mathbf{x}) - \mathbf{x}^*\|$.

Results. Quantitatively (Table 22), Sparse improves SSIM/LPIPS over Dense ($0.7461/0.2091 \to 0.7545/0.1900$), and DAG further refines the reconstruction ($0.7565/0.1838$), consistent with the theoretical ordering and indicating that the parallel/orthogonal decomposition recovers the convergence-aligned component of $r(\mathbf{x})$. Qualitatively, Figure 9 shows representative reconstructions; DAG

preserves fine structure (e.g., text, sharp edges, small objects) more faithfully than Dense or Sparse retrieval alone.

Method	SSIM ↑	LPIPS ↓
Dense (α=1.0)	0.7461	0.2091
Sparse (α=1.5)	0.7545	0.1900
DAG	**0.7565**	**0.1838**

Table 22: DDIM inversion-reconstruction (SDXL).

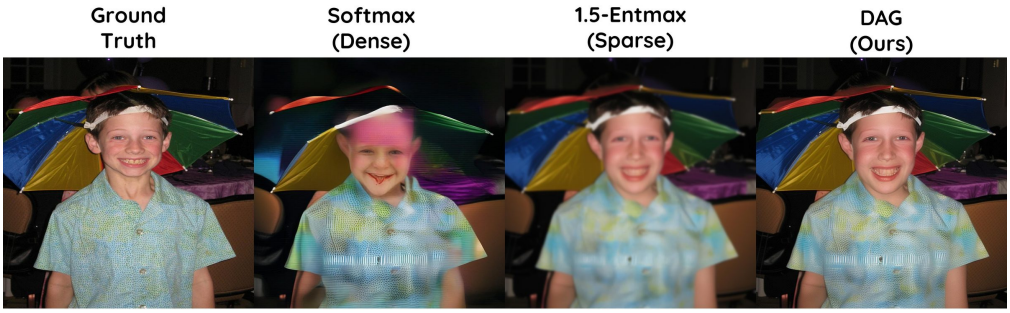


Figure 9: DDIM inversion-reconstruction on SDXL: ground-truth (GT) vs. Dense (α=1.0) vs. Sparse (α=1.5) vs. DAG. All reconstructions start from the same 50-step DDIM-inverted latent (no CFG during inversion); each operator runs 50-step DDIM denoising from that latent.

R Extended Discussion

Limitations. DAG applies to cross-attention under shared conditioning, and its cost over the base sampler (0.79 s per image at 50 steps on SDXL, 0.04–0.12 s at 4 steps on FLUX; Table 2) may matter in latency-sensitive deployment. α=1.5 and the all-layer placement are inherited from PLADIS and were not re-tuned. The diagnostics of Sec. 5.2 cover SDXL; MMDiT backbones are left for future measurement.

Model-level vs. layer-level Hopfield analyses. A growing line of work connects diffusion models with associative memory at the *model level*, treating the entire denoising network or score function as an associative-memory system. Ambrogioni (2023) interprets diffusion sampling as energy descent in a Hopfield network whose patterns are dispersed across the data manifold. Hoover et al. (2023) survey the structural resemblances between dense associative memories and diffusion models, reading the denoising network as an energy-based memory whose stored patterns are the training data. Pham et al. (2025) show that diffusion models trained on small datasets behave as dense associative memories that store training samples as fixed points, and describe the emergence of generalization as the dataset grows. In contrast, Ramsauer et al. (2021) operates at the *layer level*: a single Transformer attention layer is shown to coincide with one Picard step of Modern Hopfield Network (MHN) retrieval, with keys/values playing the role of stored patterns and queries acting as input states. Our analysis builds on this layer-level reading: DAG is not a model-level Hopfield interpretation of diffusion, but a directional analysis of one specific retrieval pair ($\mathcal{T}_\alpha$ vs. $\mathcal{T}_{\texttt{Dense}}$) inside cross-attention. The directional surrogate property (Theorem 1) is therefore complementary to model-level work and does not depend on whether the full diffusion network is itself viewable as a Hopfield system.

Why we focus on cross- and not self-attention. For text-to-image generation, cross-attention layers are the locus of text–image alignment and are well captured by the MHN reading: each text token becomes a stored pattern, and the query image token retrieves the most relevant pattern via α-Entmax. Self-attention, in contrast, is image-to-image and encodes spatial structure; aggressive sparsification there has been reported to degrade generation quality (Kim & Sim, 2025), since sparse self-attention can cut off long-range spatial dependencies that are necessary for coherent layout. Our directional analysis (shared $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, shared conditioning) cleanly applies to PLADIS-style

cross-attention pairs but is not designed for self-attention; extending it to that setting is left as future work, as noted in the Limitations.

S Additional Qualitative Results

This section provides extended qualitative evaluations to illustrate the behavior of DAG across diverse generation tasks and in combination with existing guidance frameworks. All comparisons within a row use identical prompts and seeds, so visual differences reflect the guidance scheme rather than sampling stochasticity.

Synergy with existing guidance. Figures 10, 11, and 12 compare CFG (Ho & Salimans, 2022), APG (Sadat et al., 2025), and PAG (Ahn et al., 2024) with their PLADIS- and DAG-augmented counterparts. Relative to PLADIS, DAG yields visibly better text–image alignment and detail preservation on complex prompts (object counting, visual-effect rendering), consistent with the directional analysis in Sec. 3.2 and adding no inference steps.

Step-distilled models. Figures 13 and 14 report results on DMD2 (Yin et al., 2024) and Hyper-SDXL (Ren et al., 2024) in the 4-step regime. DAG reduces visible artifacts and improves prompt–image coherence over PLADIS at nearly the same cost, fitting step-distilled pipelines where dual-pass guidance is unavailable.

Backbones. Figures 15 and 16 extend the analysis to FLUX.1-Schnell (4 steps) and FLUX.1-Dev (28 steps), showing that DAG transfers from UNet-based (SDXL) to MMDiT-based (FLUX.1) backbones across both few-step and multi-step regimes.

Comparison with NAG. Figure 17 compares NAG (Chen et al., 2025), PLADIS, and DAG on Hyper-SDXL (4 steps). NAG improves over the baseline but at additional inference cost; DAG attains comparable text alignment without extra forward passes. This matches the failure-mode analysis in Sec. 3.2: NAG uses $\mathbf{c}_{\text{pos}} \neq \mathbf{c}_{\text{neg}}$ (distinct $\mathbf{V}$), violating the shared-conditioning regime under which the directional guarantees hold.

Latest MMDiT/DiT backbones. Figures 18, 19, and 20 show qualitative comparisons (baseline vs. DAG) on FLUX.2 Klein-4B, FLUX.2 Klein-9B, and Qwen-Image, the most recent generation of MMDiT/DiT backbones. Across all three, DAG yields visibly improved text–image alignment and detail without modifying weights or adding forward passes, complementing the GenEval gains reported in Table 2 (middle).

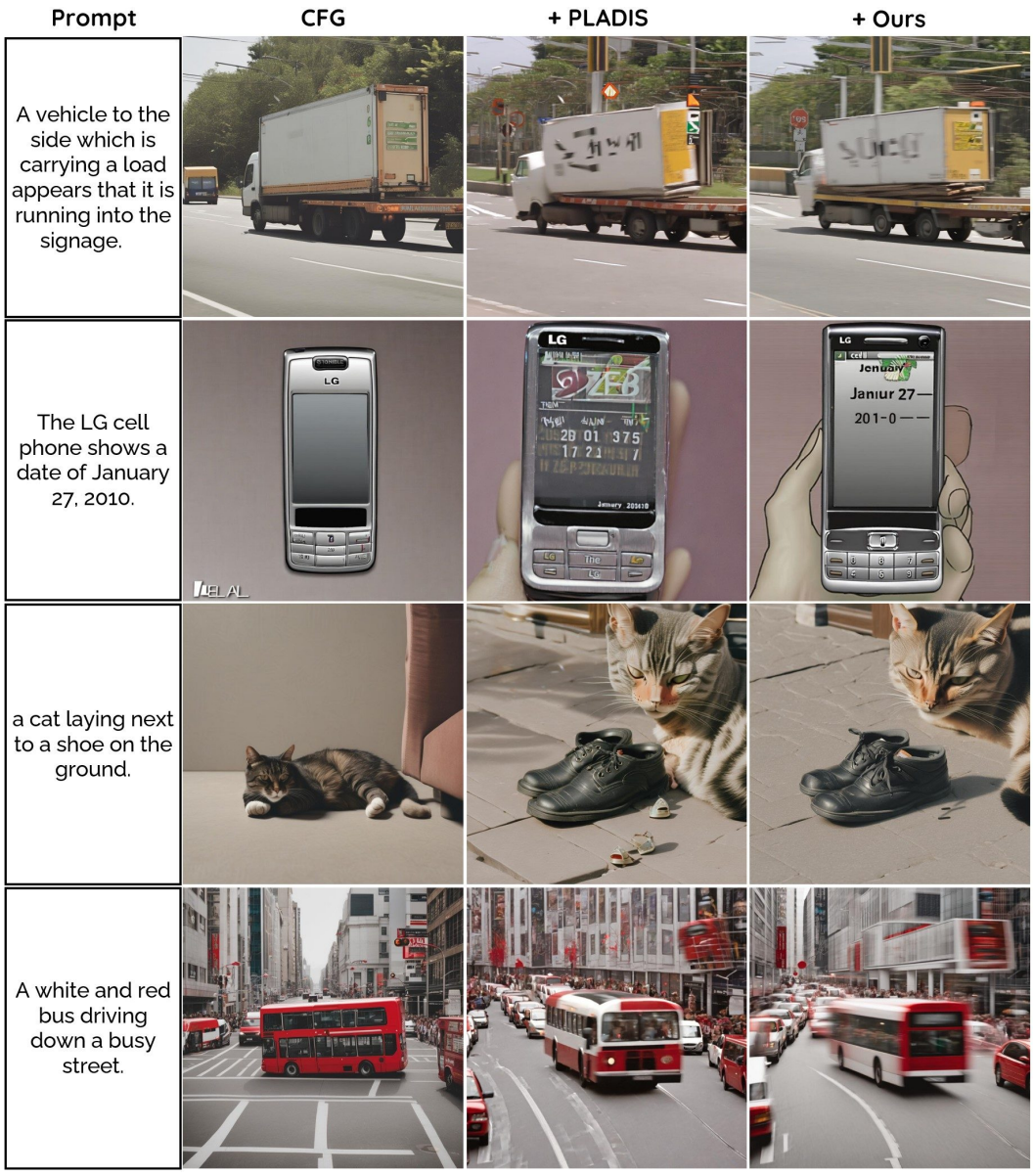


Figure 10: Qualitative comparison of CFG (Ho & Salimans, 2022) alone, CFG with PLADIS, and CFG with DAG.

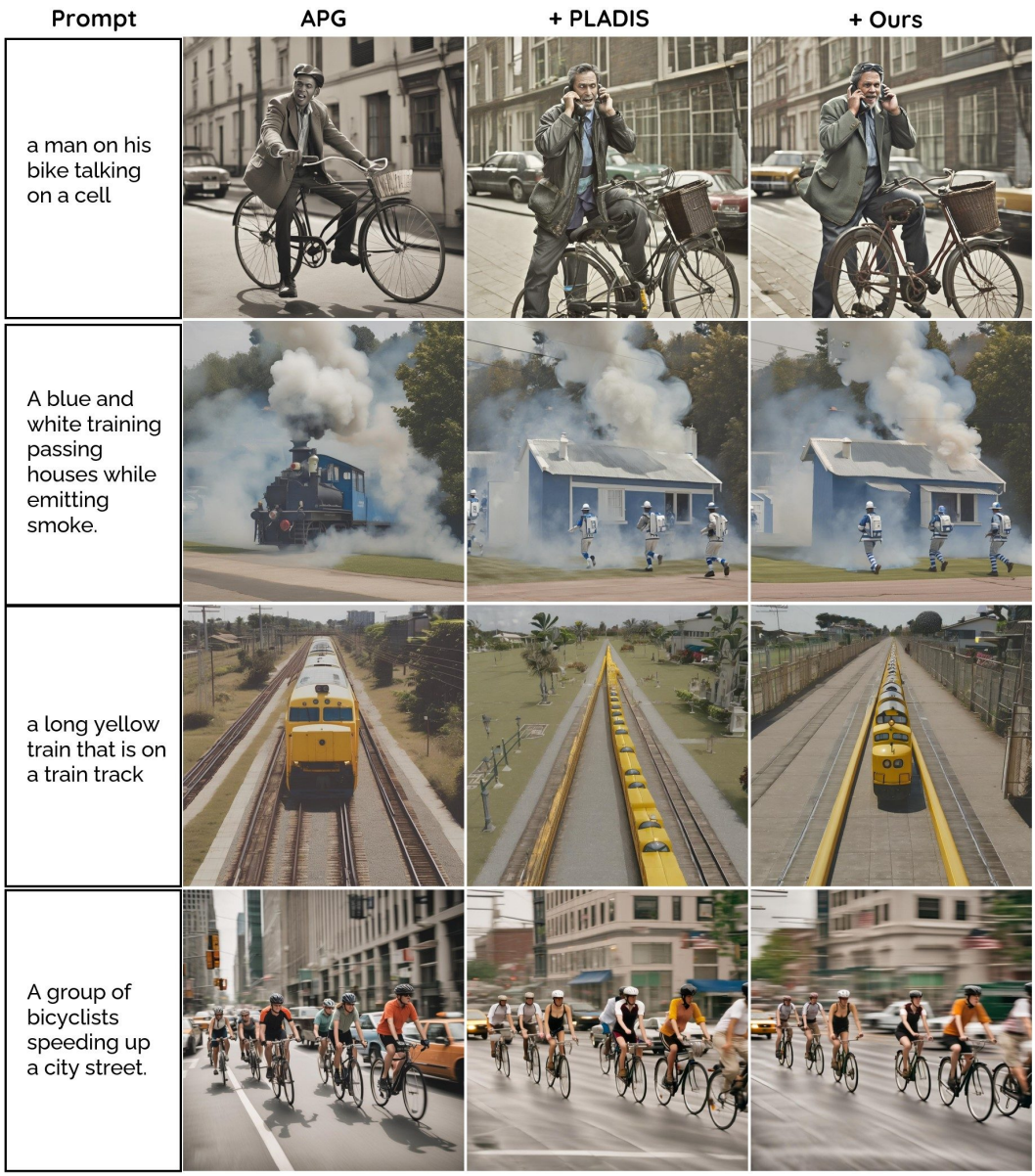


Figure 11: Qualitative comparison of APG (Sadat et al., 2025) alone, APG with PLADIS, and APG with DAG.

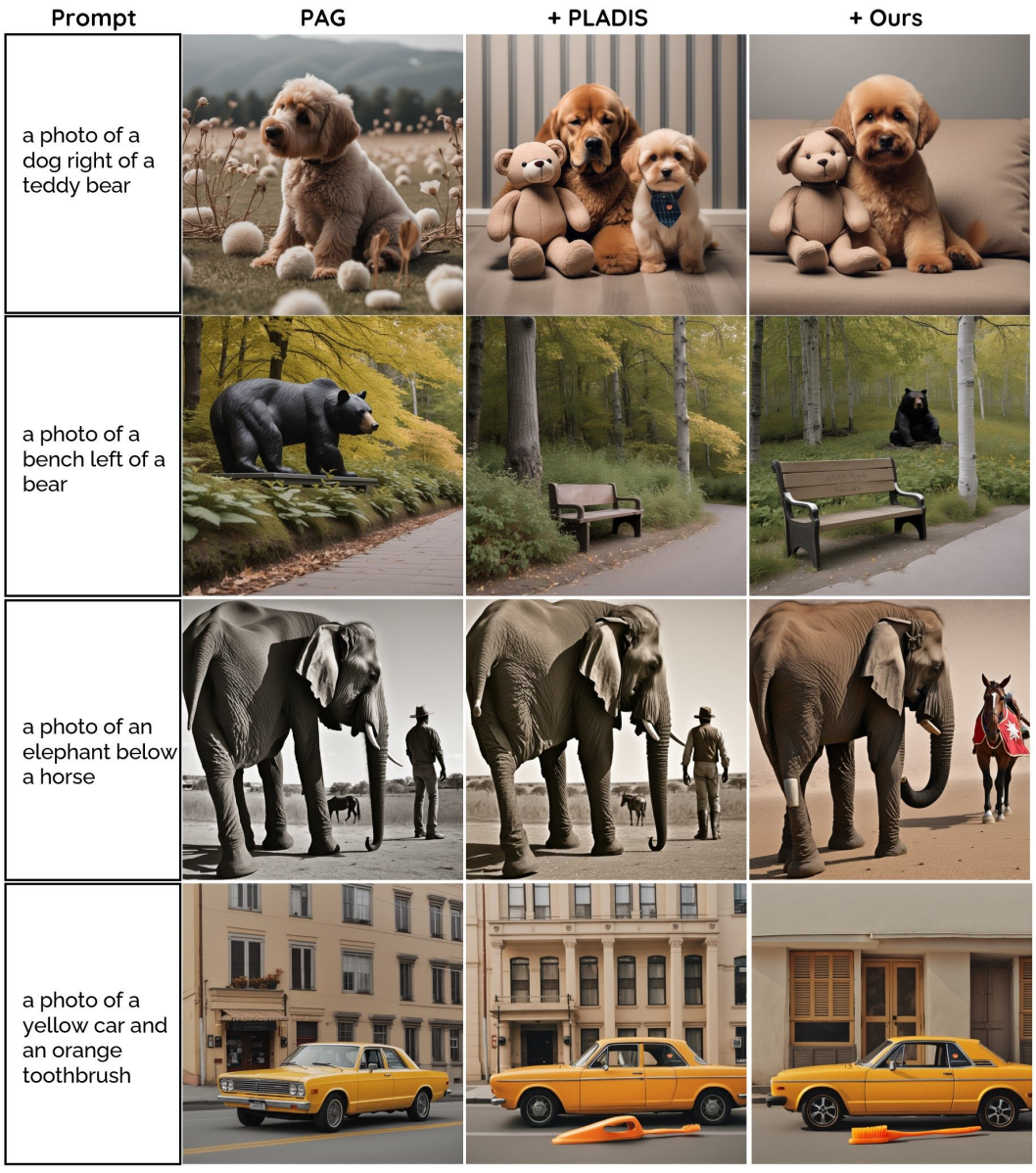


Figure 12: Qualitative comparison combining PAG (Ahn et al., 2024) with PLADIS and with DAG.



Figure 13: Qualitative comparison on Hyper-SDXL (Ren et al., 2024).



Figure 14: Qualitative comparison on DMD2 (Yin et al., 2024).



Figure 15: Qualitative comparison on FLUX.1-Schnell.



Figure 16: Qualitative comparison on FLUX.1-Dev.

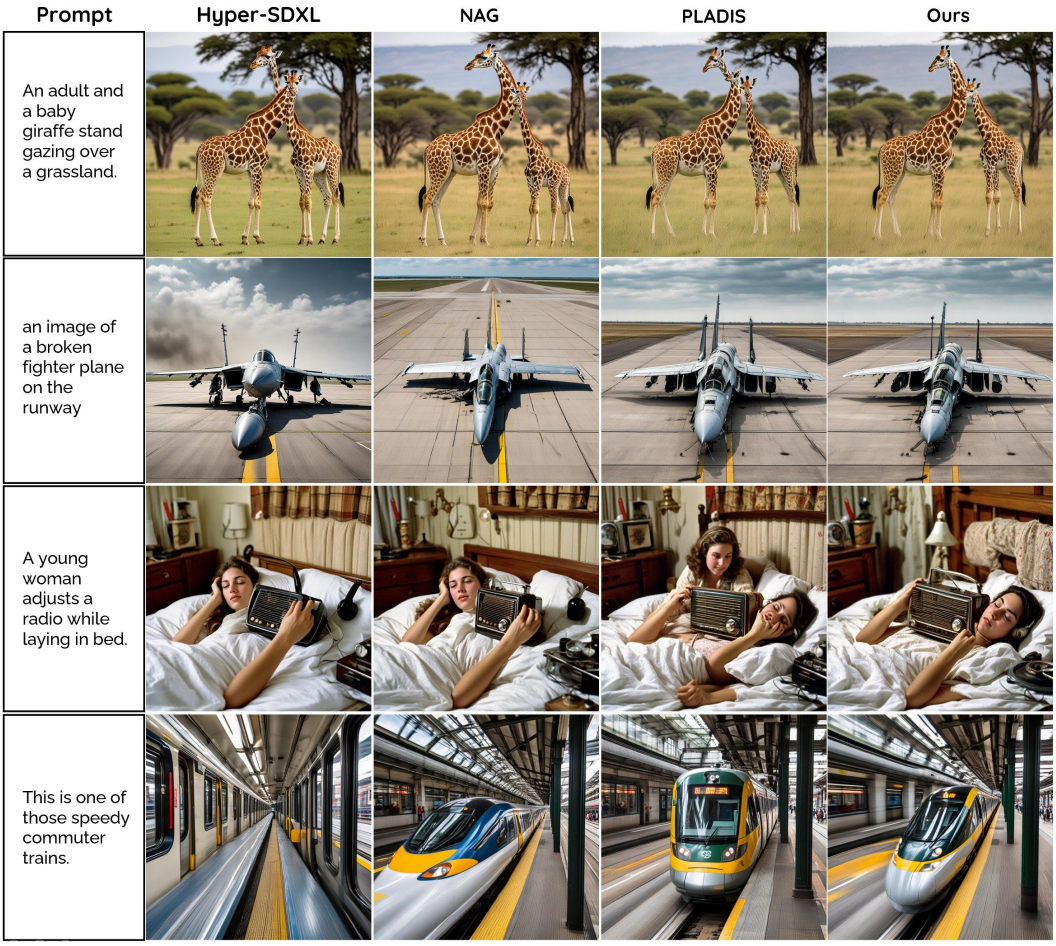


Figure 17: Qualitative comparison of NAG (Chen et al., 2025), PLADIS, and DAG on Hyper-SDXL.



Figure 18: Qualitative comparison on FLUX.2 (Klein-4B) (Black Forest Labs, 2026): baseline vs. DAG.

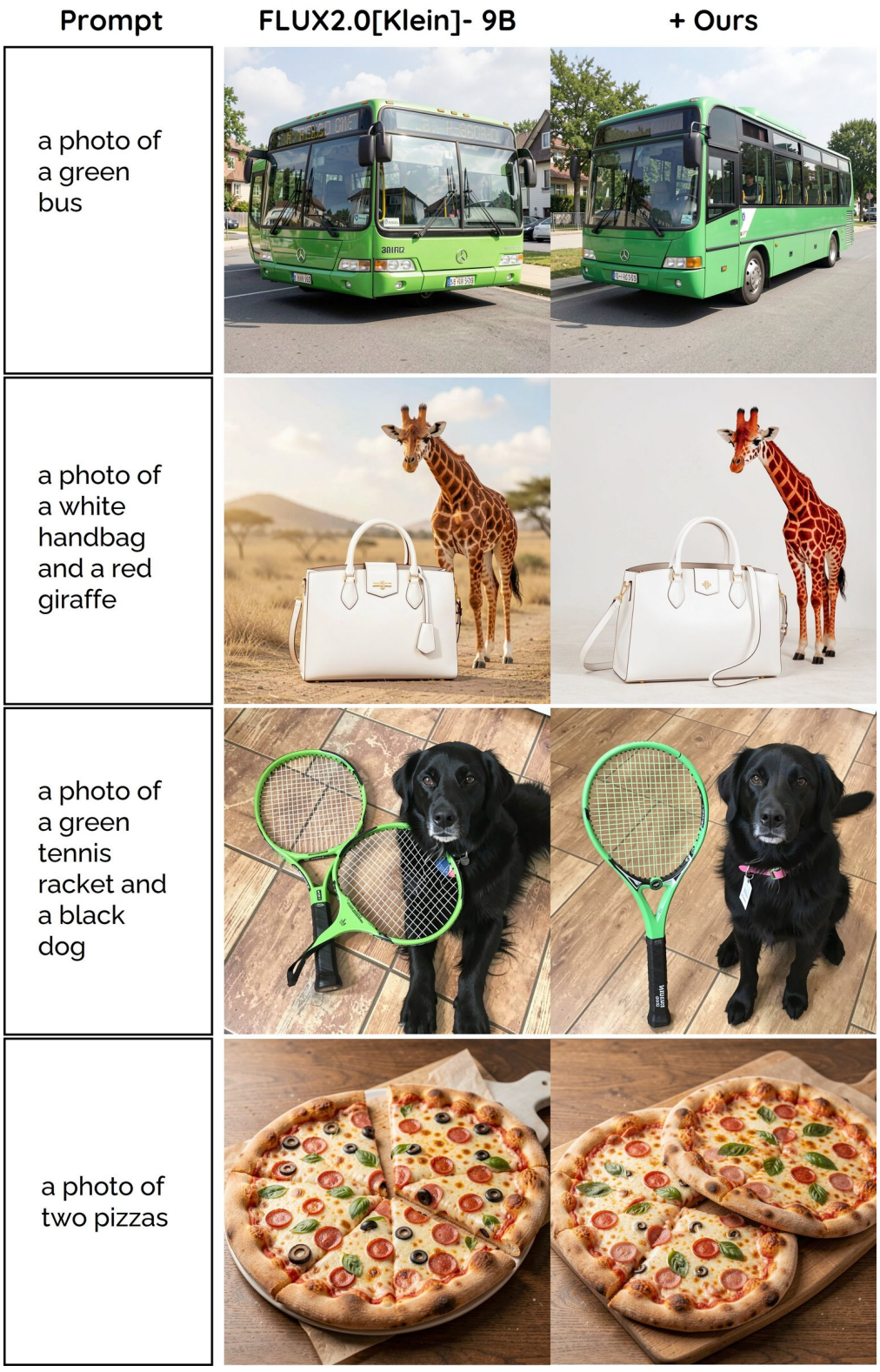


Figure 19: Qualitative comparison on FLUX.2 (Klein-9B) (Black Forest Labs, 2026): baseline vs. DAG.

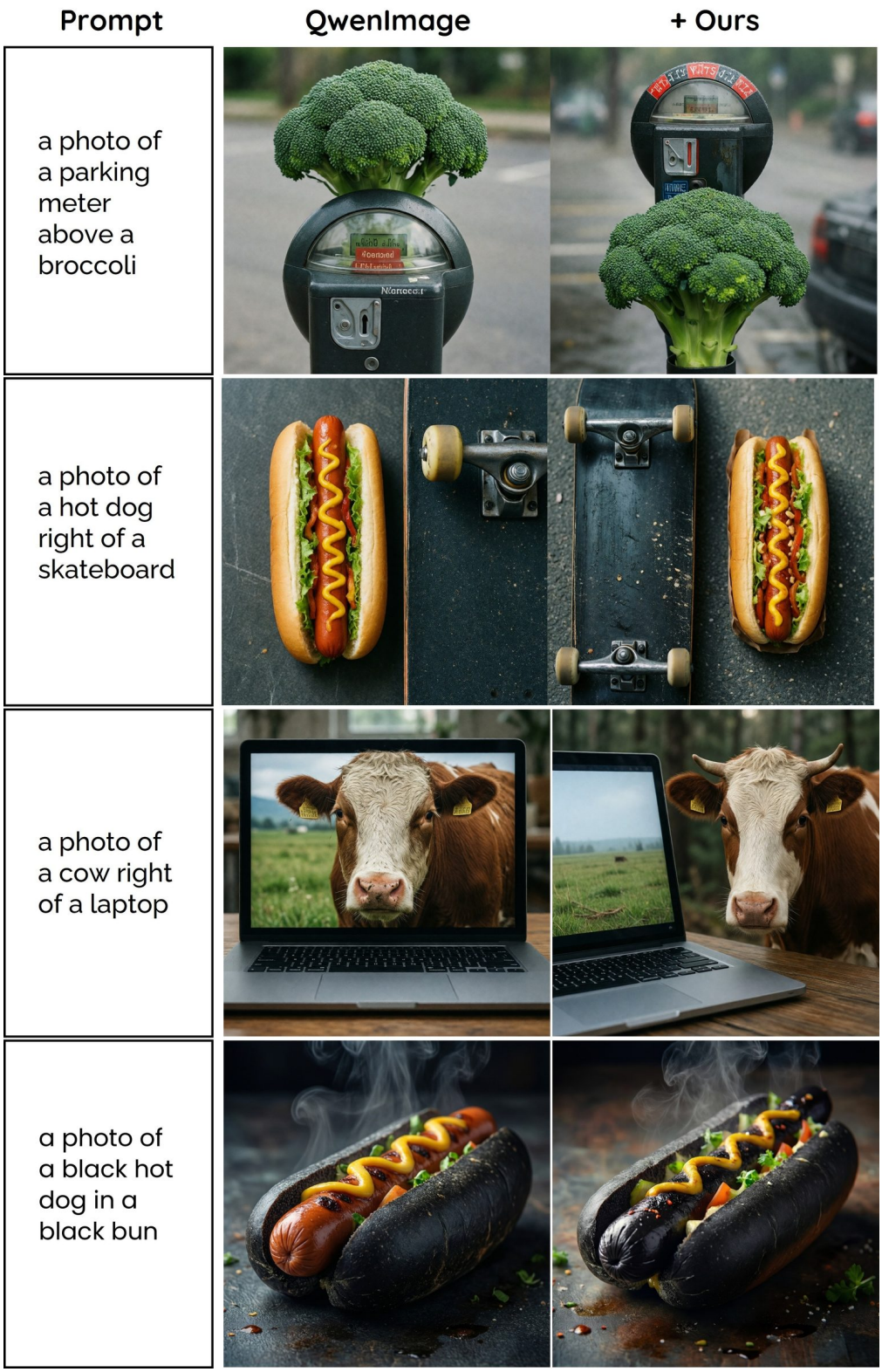


Figure 20: Qualitative comparison on Qwen-Image (Qwen Team, 2025): baseline vs. DAG.