PLADIS: Pushing the Limits of Attention in Diffusion Models at Inference Time by Leveraging Sparsity

Supplementary Material

A. Supplementary Section

In this supplementary document, we present the following:

- Theoretical background on Hopfield energy networks and sparse Hopfield energy networks, the proof of the noise robustness in the intermediate cases, and the error bound of PLADIS in Section B.
- Detailed description of the evaluation metrics and implementation in Section C.
- Further detail and results of the user preference study in Section D.
- Results for other backbone models including Stable Diffusion 1.5 and SANA, and combination with FreeU in Section E.
- Results from one-step sampling with a guidance-distilled model in Section F.
- Additional ablation studies, including attention temperature, cross-attention maps, the effect of layer selection, extrapolation strategy, and only sparse attention in Section G.
- Additional qualitative results, including interactions with existing guidance sampling approaches, the guidance-distilled model, and further ablation studies in Section H.

B. Theoretical Background

Notations. For $a \in \mathbb{R}$, $a_+ := \max\{0, a\}$. For $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$, $\langle \mathbf{z}, \mathbf{z}' \rangle = \mathbf{z}^\intercal \mathbf{z}'$ is the inner product of two vectors. For $\mathbf{z} = (z_1, \dots, z_d) \in \mathbb{R}^d$, we denote the sorted coordinates of $\mathbf{z}$ as $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(d)}$, that is, $z_{(\nu)}$ is the ν'th largest element among z_i's. $\Delta^M := \{\mathbf{p} \in \mathbb{R}^M | p_i \geq 0, \sum p_i = 1\}$, $(M-1)$-dimensional simplex.

In this section, we provide the concept of modern Hopfield network and its sparse extension in simple form, to make readers fully understand the motivation and intuition of our method and encourage further research upon our works.

Initially, a Hopfiled model was introduced as an associative memory that can store binary patterns[23]. The model is optimized to store patterns in the local minima of associated energy function. Then, given query input, the closest local minimum point of the energy function is retrieved. There were many extensions of the classic model to improve stability and capacity of the model, such as exponential energy functions or continuous state models[2, 9, 29].

Ramsauer et al. proposed modern Hopfield network that can be integrated into deep learning layers [43]. The network is equipped with a new energy function E and retrieval dynamics $\mathcal{T}$ that are differentiable and retrieve patterns after one update:

$$E_{\texttt{Dense}} : \mathbb{R}^d \to \mathbb{R}, \mathbf{x} \mapsto -\texttt{lse}(\beta, \mathbf{\Xi}^\top \mathbf{x}) + \frac{1}{2}\langle \mathbf{x}, \mathbf{x} \rangle, \tag{16}$$

$$\mathcal{T}_{\texttt{Dense}} : \mathbb{R}^d \to \mathbb{R}^d, \mathbf{x} \mapsto \mathbf{\Xi}\texttt{Softmax}(\beta \mathbf{\Xi}^\top \mathbf{x}) \tag{17}$$

where $\mathbf{x} \in \mathbb{R}^d$ represents a query input, $\mathbf{\Xi} = [\boldsymbol{\xi}_1 \dots \boldsymbol{\xi}_M] \in \mathbb{R}^{d \times M}$, $\boldsymbol{\xi}_i \in \mathbb{R}^d$ denotes a pattern stored, $\texttt{lse}(\beta, \mathbf{z}) := \log\left(\sum_{i=1}^M \exp(\beta z_i)\right)/\beta$ is log-sum-exponential function for $\beta > 0$ and $\texttt{Softmax}(\mathbf{z}) := \frac{1}{\sum_{i=1}^d \exp(z_i)}(\exp(z_1), \dots, \exp(z_d))$, for $\mathbf{z} \in \mathbb{R}^M$. Theoretical results about the energy function and the retrieval dynamics including convergence, properties of states were proposed [43].

Connection with attention of the Transformer Interesting connection between the update rule and self-attention mechanism used in transformer and BERT models was also proposed [43]. Specifically, we provide the detail derivation of this connection by following [43]. Firstly, we extend $\mathcal{T}_{\texttt{Dense}}$ in Eq. 17 to multiple queries $\mathbf{X} := \{\mathbf{x}_i\}_{i \in [N]}$. Given any raw query $\mathbf{R}$ and memory matrix $\mathbf{Y}$ that are input into Hopfield model, we calculate $\mathbf{X}$ and $\mathbf{\Xi}$ as $\mathbf{X}^\top = \mathbf{R}\mathbf{W}_Q := \mathbf{Q}, \mathbf{\Xi}^\top = \mathbf{Y}\mathbf{W}_K := \mathbf{K}$, using weight matrices, $\mathbf{W}_Q, \mathbf{W}_K$. Therefore, we rewrite $\mathcal{T}_{\texttt{Dense}}$ as $\mathbf{K}^\top \texttt{Softmax}(\beta \mathbf{K}\mathbf{Q}^\top)$.

Then, by taking transpose and projecting $\mathbf{K}$ to $\mathbf{V}$ with $\mathbf{W}_V$, we have

$$\mathcal{T}_{\texttt{Dense}} : \mathbf{X} \mapsto \texttt{Softmax}(\beta \mathbf{Q}\mathbf{K}^\top)\mathbf{K}\mathbf{W}_V = \texttt{Softmax}(\beta \mathbf{Q}\mathbf{K}^\top)\mathbf{V}, \tag{18}$$

which is exactly transformer self-attention with $\beta = 1/\sqrt{d}$. In other words, we obtain by employing the notations in the Eq. (12),

$$\mathcal{T}_{\texttt{Dense}} : \mathbf{X} \mapsto \texttt{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V} := \texttt{At}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{At}(\mathbf{W}_Q\mathbf{X}, \mathbf{W}_K\mathbf{X}, \mathbf{W}_V\mathbf{X}) \tag{19}$$

However, we can extend the interpretation to a cross-attention mechanism:

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

We find similarity in the above cross-attention formula with inputs $\mathbf{X}, \mathbf{Y}$ and weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$. As discussed in lines of this paper, we focus on this extension into the cross-attention mechanism.

In terms of modern Hopefield network, the input query is processed with additional transformation $\mathbf{W}_Q$ to increase complexity of network and inner product are computed with stored (learned) $\mathbf{W}_K\mathbf{Y}$ patterns (keys). Then, the retrieved patterns (values) for next layers are computed. Different layers can have different patterns, so hierarchical patterns are stored and retrieved in deep layers. Note that while Hopfield network outputs one pattern, the attention yields multiple patterns, so attention corresponds to stack of outputs of Hopfield network. Hence, the attention is multi-level and multi-valued Hopfield network.

Sparse Hopfield Network Later, sparse extensions of the modern Hopfield network are proposed [24, 60]. The energy function was modified to make sparse the computation of retrieval dynamics:

$$E_\alpha : \mathbb{R}^d \to \mathbb{R}, \mathbf{x} \mapsto -\mathbf{\Psi}_\alpha^\star(\beta, \mathbf{\Xi}^\top\mathbf{x}) + \frac{1}{2}\langle\mathbf{x}, \mathbf{x}\rangle, \tag{20}$$

$$\mathcal{T}_\alpha : \mathbb{R}^d \to \mathbb{R}^d, \mathbf{x} \mapsto \mathbf{\Xi}\alpha\texttt{-Entmax}(\beta\mathbf{\Xi}^\top\mathbf{x}), \tag{21}$$

and $\mathbf{\Psi}_\alpha^\star$ is the convex conjugate of Tsallis entropy [56], $\mathbf{\Psi}_\alpha, \alpha\texttt{-Entmax}(\mathbf{z})$, represents the probability mapping:

$$\Psi_\alpha(\mathbf{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)}\sum_{i=1}^M (p_i - p_i^\alpha), & \alpha \neq 1, \\ -\sum_{i=1}^M (p_i - \log p_i), & \alpha = 1, \end{cases} \tag{22}$$

$$\alpha\texttt{-Entmax}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^M}{\arg\max}[\langle\mathbf{p}, \mathbf{z}\rangle - \Psi_\alpha(\mathbf{p})], \tag{23}$$

where $\mathbf{p} \in \mathbb{R}^M$. Here, α controls the sparsity. When $\alpha = 1$, it is equivalent to a dense probability mapping, $1\texttt{-Entmax} = \texttt{Softmax}$, and as α increases towards 2, the outputs of $\alpha\texttt{-Entmax}$ become increasingly sparse, ultimately converging to $2\texttt{-Entmax} \equiv \texttt{Sparsemax}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^M}{\arg\min} \|\mathbf{p} - \mathbf{z}\|$ [38]. Notably, when $\alpha = 1$, $\mathcal{T}_\alpha$ becomes equivalent to $\mathcal{T}_{\texttt{Dense}} \equiv \mathcal{T}_1$ [58].
We have simple formula for $\alpha\texttt{-Entmax}$[38]. There is a unique threshold function $\tau : \mathbb{R}^M \to \mathbb{R}$ that satisfies

$$\alpha\texttt{-Entmax}(\mathbf{z}) = [(\alpha - 1)\mathbf{z} - \tau(\mathbf{z})\mathbf{1}]_+^{1/(\alpha-1)}. \tag{24}$$

From this formula, we know that the entries less than $\tau/(\alpha - 1)$ map to zero, so sparsity is achieved. We will denote the number of nonzero entries in $\alpha\texttt{-Entmax}$ as $\kappa(\mathbf{z})$ for later use to derive theoretical results. For $\alpha = 2$, the exact solution can be efficiently computed using a sorting algorithm [15, 39]. For $1 < \alpha < 2$, inaccurate and slow iterative algorithm was used for computing $\alpha\texttt{-Entmax}$ [36]. Interestingly, for $1.5\texttt{-Entmax}$, an accurate and exact solution are derived in a simple form [41].

Similar to $\mathcal{T}_{\texttt{Dense}}$, $\mathcal{T}_\alpha$ can be extended to attention mechanisms, establishing a strong connection with sparse attention. In other words, by following the derivation as provided in Eq. (18), and Eq. (19), we can obtain

$$\mathcal{T}_\alpha : \mathbf{X} \mapsto \alpha\texttt{-Entmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V} := \texttt{At}_\alpha(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \tag{25}$$

Furthermore, similar to the dense attention mechanism, we can also extend into a cross-attention mechanism with inputs $\mathbf{X}$ and $\mathbf{Y}$:

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

Noise robustness of sparse Hopfield network In SHN, sparsity reduces retrieval errors and provide faster convergeness compared to dense retrieval dynamics [24, 60]. While the sparse extension is an efficient counterpart of dense Hopfield network, it has been discovered that there is more advantages to use sparse one besides efficiency [24, 60].

Definition 1 (Pattern Stored and Retrieved)**.** *Suppose every pattern $\boldsymbol{\xi}_\mu$ is contained in a ball B_μ. We say that $\boldsymbol{\xi}_\mu$ is stored if there is a single fixed point $\mathbf{x}_i^* \in B_\mu$, to which all point $\mathbf{x} \in B_\mu$ converge, and B_μ's are disjoint. We say that $\boldsymbol{\xi}_\mu$ is retrieved for an error ϵ if $||\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu|| \leq \epsilon$ for all $\mathbf{x} \in B_\mu$*

For following theorems, $m := \max_\nu ||\boldsymbol{\xi}_\nu||$.

Theorem 3 (Retrieval Error)**.** *[24, 43, 60] Let $\mathcal{T}_\alpha$ be the retrieval dynamics of Hopfield model with α-Entmax.*

$$\textit{For } \alpha = 1, \qquad ||\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu|| \leq 2m(M-1)\exp\left\{-\beta\left(\langle \boldsymbol{\xi}_\mu, \mathbf{x}\rangle - \max_\nu \langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu\rangle\right)\right\}. \tag{26}$$

$$\textit{For } \alpha = 2, \qquad ||\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu|| \leq m + m\beta\left[\kappa\left(\max_\nu \langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa)}\right) + \frac{1}{\beta}\right]. \tag{27}$$

$$\textit{For } \alpha > \alpha', \qquad ||\mathcal{T}_\alpha(x) - \boldsymbol{\xi}_\mu|| \leq ||\mathcal{T}_{\alpha'} - \xi||. \tag{28}$$

You can find the result Eq. (26) in [43], Eq. (27) in [24], and Eq. (28) in [24, 60].

Corollary 3.1. *(Noise-Robustness) [24, 60]. In case of noisy patterns with noise $\boldsymbol{\eta}$, i.e. $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta}$ (noise in query) or $\tilde{\xi}_\mu = \xi_\mu + \boldsymbol{\eta}$ (noise in memory), the impact of noise $\boldsymbol{\eta}$ on the sparse retrieval error $||\mathcal{T}_2(\mathbf{x}) - \boldsymbol{\xi}_\mu|$ is linear, while its effect on the dense retrieval error $||\mathcal{T}_1(\mathbf{x}) - \boldsymbol{\xi}_\mu||$ is exponential.*

where ξ_μ is memory pattern and to be considered stored at a fixed point of $\mathcal{T}$. This theorem suggests that under noisy conditions, sparse attention mechanisms governed by $\mathcal{T}_\alpha$ with $\alpha > 1$ exhibit superior noise robustness compared to standard dense attention. Critically, increasing sparsity (via higher α) further diminishes retrieval errors.

We propose a new theoretical result that completes above theorem by providing error estimation for all intermediate cases that was not given.

Theorem 4 (Retrieval Error 2)**.** *Let $\mathcal{T}_\alpha$ be the retrieval dynamics of Hopfield model with α-Entmax.*

$$\textit{For } 1 < \alpha \leq 2, \qquad ||\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu|| \leq m + m\kappa\left[(\alpha-1)\beta\left(\max_\nu \langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa+1)}\right)\right]^{\frac{1}{\alpha-1}}, \tag{29}$$

Here, we abuse the notation $[\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(M+1)} := [\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(M)} - M^{1-\alpha}/(\alpha-1)$.

Thanks to this new theorem, we can estimate the impact of noise on the sparse retrieval error for all $1 < \alpha < 2$.

Corollary 4.1. *(Noise-Robustness) In case of noisy patterns with noise $\boldsymbol{\eta}$, the impact of noise $\boldsymbol{\eta}$ on the retrieval error $||\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu||$ is polynomial of order $\frac{1}{\alpha-1}$ for $1 < \alpha \leq 2$.*

Remark The proposed theorem includes the case $\alpha = 2$. In that case, the right hand side becomes

$$m\beta\left[\kappa\left(\max_\nu \langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa+1)}\right)\right].$$

Therefore, by combining with previous result, we obtain tighter bound:

$$||\mathcal{T}_2(\mathbf{x}) - \boldsymbol{\xi}_\nu|| \leq m\beta\left[\kappa \max_\nu \langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle + \min\left\{-\kappa[\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa+1)}, -\kappa[\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa)} + \frac{1}{\beta}\right\}\right]$$

proof of Thm. 4.

$$||\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\xi}_\mu|| = \left\|\boldsymbol{\Xi}\alpha\text{-Entmax}\left(\beta\boldsymbol{\Xi}^\intercal \mathbf{x}\right) - \boldsymbol{\xi}_\mu\right\| = \left\|\sum_{\nu=1}^{\kappa} \boldsymbol{\xi}_{(\nu)}\left[\alpha\text{-Entmax}\left(\beta\boldsymbol{\Xi}^\intercal \mathbf{x}\right)\right]_{(\nu)} - \boldsymbol{\xi}_\mu\right\| \tag{30}$$

$$\leq ||\boldsymbol{\xi}_\mu|| + \sum_{\nu=1}^{\kappa}\left\|\boldsymbol{\xi}_{(\nu)}\right\|\left[\alpha\text{-Entmax}\left(\beta\boldsymbol{\Xi}^\intercal \mathbf{x}\right)\right]_{(\nu)} \tag{31}$$

$$\leq m + m\sum_{\nu=1}^{\kappa}\left[(\alpha-1)\left([\beta\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\nu)} - [\beta\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa+1)}\right)\right]^{\frac{1}{\alpha-1}} \tag{32}$$

$$\leq m + m\kappa\max_\nu\left[(\alpha-1)\beta\left(\langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa+1)}\right)\right]^{\frac{1}{\alpha-1}}. \tag{33}$$

For Eq. (32), we use the following lemma. □

Lemma 1. *For* $\mathbf{z} \in \mathbb{R}^M$ *and* $\nu \le \kappa(\mathbf{z})$, $[\alpha\text{-}\mathtt{Entmax}(\mathbf{z})]_{(\nu)} \le [(\alpha-1)(z_{(\nu)} - z_{(\kappa+1)})]^{1/(\alpha-1)}$.

Proof.

(i) $\kappa < M$
From the definition of κ, we have following properties.

$$\alpha\text{-}\mathtt{Entmax}(\mathbf{z})_{(\kappa+1)} = 0.$$

$$z_{(\kappa+1)} \le \tau(\mathbf{z})/(\alpha-1).$$

Keep the last inequality, and now consider the ν'th largest coordinate of Eq. (24), but we can omit $+$ since it is strictly positive.

$$\begin{aligned}
\alpha\text{-}\mathtt{Entmax}(\mathbf{z})_{(\nu)} &= [(\alpha-1)z_{(\nu)} - \tau(\mathbf{z})]_+^{1/(\alpha-1)} \\
&= [(\alpha-1)z_{(\nu)} - \tau(\mathbf{z})]^{1/(\alpha-1)} \\
&\le [(\alpha-1)z_{(\nu)} - (\alpha-1)z_{(\kappa+1)}]^{1/(\alpha-1)}
\end{aligned}$$

(ii) $\kappa = M$
We use Hölder inequality

$$\left(\sum |a_i|^p\right)^{1/p} \left(\sum |b_i|^q\right)^{1/q} \ge \sum |a_i b_i| \quad \text{for } p, q \in (1,\infty), 1/p + 1/q = 1$$

to estimate a lower bound of τ for $\alpha \neq 2$. By substituting $a_i = (\alpha-1)z_i - \tau, b_i = 1, p = 1/(\alpha-1), q = 1/(2-\alpha)$,

$$\left(\sum |(\alpha-1)z_i - \tau|^{1/(\alpha-1)}\right)^{\alpha-1} \left(\sum 1\right)^{2-\alpha} \ge \sum |(\alpha-1)z_i - \tau|.$$

We know that all entries are positive $(\alpha-1)z_i - \tau > 0$ since $\kappa = M$. Moreover,

$$\sum [(\alpha-1)z_i - \tau]^{1/(\alpha-1)} = 1$$

since the left hand side is the sum of the coordinates of $\alpha\text{-}\mathtt{Entmax}$ output. Therefore,

$$\begin{aligned}
M^{2-\alpha} &\ge (\alpha-1)\sum z_i - M\tau \\
\frac{\tau}{\alpha-1} &\ge \frac{1}{M}\sum z_i - \frac{M^{1-\alpha}}{\alpha-1} \\
&\ge \min z_i - \frac{M^{1-\alpha}}{\alpha-1} = z_{(M)} - \frac{M^{1-\alpha}}{\alpha-1}
\end{aligned}$$

We remain the case $\alpha = 2$. We directly sum up the entries of $2\text{-}\mathtt{Entmax}$:

$$\begin{aligned}
1 = \sum |z_i - \tau| &= \sum z_i - M\tau \\
&\ge M \min z_i - M\tau \\
\therefore \tau &\ge z_{(M)} - \frac{1}{M} = z_{(M)} - \frac{M^{1-\alpha}}{\alpha-1}
\end{aligned}$$

□

We further estimate the retrieval error of retrieval dynamics defined in PLADIS. We use the notation:

$$\mathcal{T}_\alpha^\lambda(\mathbf{x}) := \lambda\mathcal{T}_\alpha(\mathbf{x}) + (1-\lambda)\mathcal{T}_1(\mathbf{x}).$$

Then, we have following result for the retrieval error of $\mathcal{T}_\alpha^\lambda$.

Theorem 5 (Retrieval Error 3)**.** *Consider the retrieval dynamics* $\mathcal{T}_\alpha^\lambda$

$$||\mathcal{T}_\alpha^\lambda(\mathbf{x}) - \boldsymbol{\xi}_\mu|| \leq |\lambda|m + |\lambda|m\kappa \left[(\alpha-1)\beta\left(\max_\nu \langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle - [\boldsymbol{\Xi}^\intercal \mathbf{x}]_{(\kappa+1)}\right)\right]^{\frac{1}{\alpha-1}} \tag{34}$$

$$+ |1-\lambda|2m(M-1)\exp\left\{-\beta\left(\langle \boldsymbol{\xi}_\mu, \mathbf{x} - \max_\nu \langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu\rangle\right)\right\}. \tag{35}$$

Proof.

$$\begin{aligned}||\mathcal{T}_\alpha^\lambda(\mathbf{x}) - \boldsymbol{\xi}_\nu|| &= ||\lambda\mathcal{T}_\alpha(\mathbf{x}) + (1-\lambda)\mathcal{T}_1(\mathbf{x}) - \boldsymbol{\xi}_\nu|| \\ &\leq |\lambda|||\mathcal{T}_\alpha(\mathbf{x}) + \boldsymbol{\xi}_\nu|| + |1-\lambda|||\mathcal{T}_1(\mathbf{x}) - \boldsymbol{\xi}_\nu||\end{aligned}$$

and apply Eq. (26) and Eq. (29). □

This theorem suggests that the retrieval dynamics given in PLADIS have the error bound of mixture of polynomial and exponential terms.

C. Metrics and Implementation Detail

For image sampling in Table 2, sampling without CFG guidance is conducted using 30,000 randomly selected text prompts from the MSCOCO validation dataset. Conversely, sampling with CFG is performed with uniformly selected values of w in the range (3,5). In both cases, the PAG and SEG scales are fixed at 3.0, following the recommended settings from the corresponding paper.

For Tables 3 and 4, we use 200 prompts from Drawbench [49], 400 prompts from HPD [61], and 500 prompts from the test set of Pick-a-pic [28], generating 5 images per prompt. Additionally, for the ablation study in Table 5, we generate 5,000 images from the MSCOCO validation set with CFG and PAG guidance. As with Table 2, the CFG scale is uniformly selected within the range of (3,5), while the PAG scale remains set at 3.0.

D. User Preference Study

As presented in Fig. 7, we employ human evaluation and do not rely solely on automated evaluation metrics such as FID, CLIPScore, ImageReward, etc. Our aim is to assess whether PLADIS truly improves image quality and prompt coherence. To rigorously evaluate these aspects, we categorized caess into two groups: interaction with guidance sampling including CFG [19], PAG [1], SEG [21], and interaction with guidance-distilled models such as SDXL-Turbo [50], SDXL-Lightening [33], DMD2 [64], and Hyper-SDXL [44]. We evaluate all models based on 20 selected prompts from the randomly selected Drawbench [49], HPD [61], and Pick-a-pic [28]. For the guidance-distilled model, we select half from one-step sampling results and the other half from four-step sampling results. Human evaluators, who are definitely blind and anonymous, are restricted to participating only once. Evaluators are shown two images from model outputs with and without PLADIS based on the same text prompt and measure images with two questions: for image quality, "Which image is of higher quality and visually more pleasing?" and for prompt alignment, "Which image looks more representative of the given prompt." The order of prompts and the order between models are truly randomized. In Fig. 7, we averaged all of the results related to the guidance-distilled model due to limited space. Further presenting in detail, we present a user preference study for each guidance-distilled model as shown in Fig. 9. As similar to guidance sampling, guidance-distilled models with PLADIS outperform both image quality and prompt alignment, validating the practical effectiveness of PLADIS.

E. Application on Other Backbone

To demonstrate the robustness of our proposed method, we perform experiments using additional backbones, including Stable Diffusion v1.5 (SD1.5) and SANA [62]. SANA is a recently introduced text-to-image diffusion model that uses linear attention, enabling faster image generation. It is based on the Diffusion Transformer (DiT) architecture. We generate 30K samples from randomly selected MS COCO validation set images and evaluate them using FID, CLIPScore, and ImageReward, as shown in Table 8. For SD1.5, we use CFG, while SANA is tested with its default configuration without modifications.

Interestingly, we observe that both SD1.5 and SANA, when integrated with our PLADIS method, consistently improve performance across all metrics. A visual comparison is provided in Fig. 13 and Fig. 14. As shown in the figures, the generation with our PLADIS provides more natural and pleasing images and precise matching between images and text prompts on both backbones. As seen in other experiments, our PLADIS enhances both generation quality and text alignment with the given prompts. By confirming these improvements with SD1.5 and SANA, we demonstrate that PLADIS is robust across different backbones, particularly transformer-based architectures.

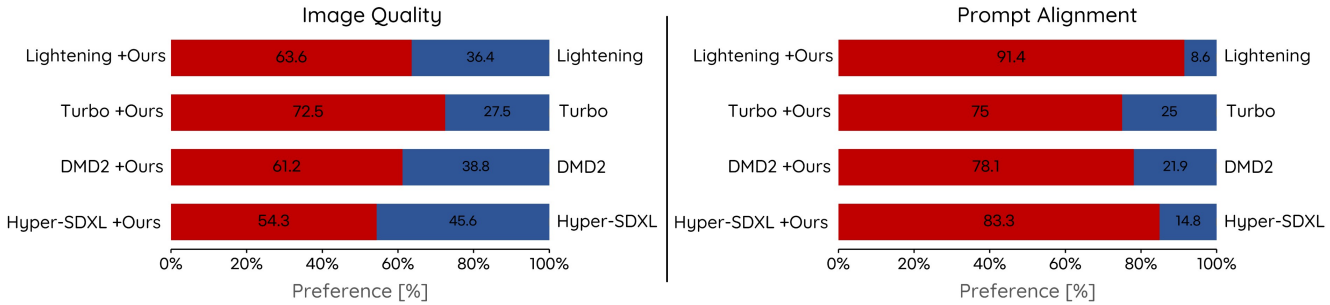

Figure 9. User preference study for PLADIS in the context of guidance-distilled models. We evaluate the two aspects of model output with and without PLADIS such as image quality and prompt alignment.

Table 7. Quantitative comparison across various datasets using 1-steps sampling with the guidance-distilled model.

	Drawbench [49]			HPD [61]			Pick-a-pic [28]		
Method	CLIPScore ↑	PickScore ↑	ImageReward ↑	CLIPScore ↑	PickScore ↑	ImageReward ↑	CLIPScore ↑	PickScore ↑	ImageReward ↑
Turbo [50]	27.19	21.67	0.305	28.45	21.85	0.479	26.89	21.16	0.346
+ Ours	**27.56** (+0.37)	**21.68** (+0.01)	**0.390** (+0.08)	**28.78** (+0.33)	**21.86** (+0.01)	**0.517** (+0.04)	**27.10** (+0.21)	**21.17** (+0.01)	**0.378** (+0.04)
Light [33]	26.08	21.86	0.428	27.37	22.05	0.730	25.73	21.34	0.585
+ Ours	**26.66** (+0.58)	**21.94** (+0.08)	**0.558** (+0.13)	**28.42** (+1.05)	**22.24** (+0.19)	**0.830** (+0.10)	**26.63** (+0.90)	**21.46** (+0.12)	**0.680** (+0.10)
DMD2 [64]	27.91	22.04	0.651	29.95	22.18	0.888	28.14	21.57	0.770
+ Ours	**28.09** (+0.19)	**22.05** (+0.01)	**0.662** (+0.01)	**30.21** (+0.26)	**22.20** (+0.02)	**0.902** (+0.01)	**28.38** (+0.43)	**21.58** (+0.01)	**0.794** (+0.02)
Hyper [44]	27.41	22.27	0.662	29.09	22.61	0.912	27.29	21.91	0.812
+ Ours	**27.80** (+0.39)	**22.30** (+0.03)	**0.674** (+0.01)	**29.42** (+0.33)	**22.65** (+0.04)	**0.932** (+0.02)	**27.85** (+0.56)	**21.92** (+0.01)	**0.832** (+0.02)

Table 8. Application on other BackBone Model on MS COCO validation set and Comparison results for another extrapolation strategy and combination with FreeU [51]. SD1.5 and SANA indicate that Stable Diffusion version 1.5 and SANA 1.6 B model, respectively.

Resolution	BackBone	FID ↓	CLIPScore ↑	ImageReward ↑
512 × 512	SD1.5	23.88	24.11	-0.368
	+ PLADIS (Ours)	22.41(-1.48)	25.09 (+0.98)	-0.08 (+0.360)
1024 × 1024	SANA [62]	28.01	26.61	0.867
	+ PLADIS (Ours)	27.53(-0.48)	26.83 (+0.21)	0.883(+0.016)
Resolution	Method	FID ↓	CLIPScore ↑	ImageReward ↑
1024 × 1024	SDXL (CFG)	32.68	25.90	0.425
	+ Ours (Prediction)	29.48	26.60	0.619
	+ Ours (In-model)	**28.50**	**26.61**	**0.626**
1024 × 1024	SDXL + FreeU	35.66	25.96	0.425
	+ PLADIS (Ours)	**28.79**	**26.93**	**0.626**

Table 9. Ablation study on layer group which is replaced with PLADIS on MS COCO validation dataset.

Layer	FID ↓	CLIPScore ↑	ImageReward ↑
Baseline	33.76	25.41	0.478
Up	29.78(-3.98)	25.78 (+0.37)	0.624(+0.15)
Mid	31.76(-2.00)	25.46 (+0.05)	0.496(+0.02)
Down	31.46(-2.30)	25.43 (+0.02)	0.501(+0.02)
Up, Mid	30.76(-3.00)	25.46 (+0.05)	0.548(+0.07)
Up, Down	28.46(-5.30)	26.12 (+0.71)	0.658(+0.18)
Mid, Down	31.36(-2.40)	25.52 (+0.11)	0.498(+0.02)
All (Ours)	27.87(-5.89)	26.41 (+1.00)	0.726(+0.25)

F. Comparison Results on One-Step Sampling

As discussed in Section 5, we found that our proposed method, PLADIS, is also effective for one-step sampling with a guidance-distilled model. Following the experimental settings in Table 4, we generate images from text prompts in human preference datasets such as Drawbench [49], HPD [61], and Pick-a-pick [28]. The generated images are evaluated using CLIPScore, ImageReward, and PickScore, as presented in Table 7. Our method consistently yields performance improvements, particularly in text alignment and human preference, across all baselines. This demonstrates the robustness of our approach for denoising steps and highlights its potential as a generalizable boosting solution.

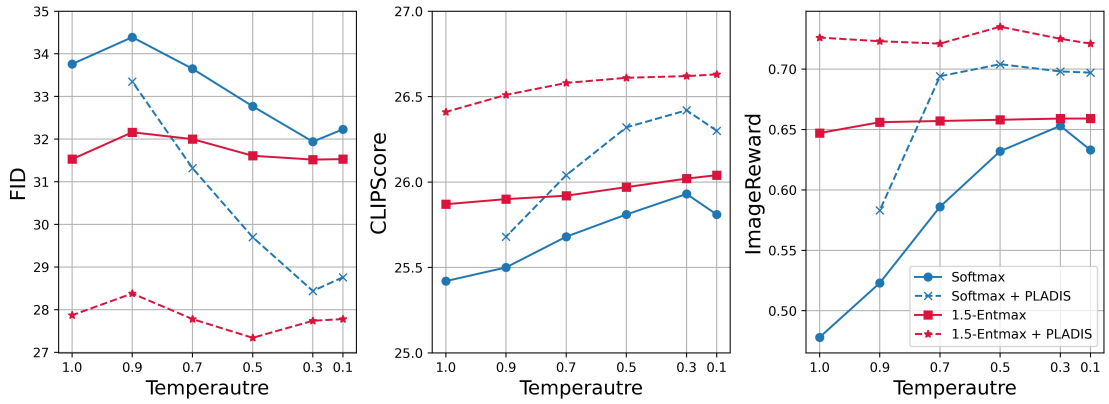

Figure 10. Comparison results for various temperatures, with and without PLADIS, are presented, including the baseline (Softmax) and 1.5−Entmax. While lower temperatures with the baseline offer benefits in both cases, our proposed method (α = 1.5), with and without PLADIS, outperforms across all temperature settings.

G. Additional Ablation Study

G.1. Comparison with Attention Temperature

In the field of NLP, to improve existing attention mechanisms, temperature scaling [32], also known as inverse temperature, has been extensively studied to adjust the sharpness of attention. It is defined as follows:

$$\texttt{At}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}*\tau}) \tag{36}$$

where τ denotes the temperature, which controls the softness of the attention. A lower temperature results in sharper activations, creating a more distinct separation between values. Importantly, it is closely related to the β in α−Entmax. In common attention mechanisms, β is typically set to the square root of the dimension, $\sqrt{d}$, which corresponds to $\tau = 1.0$. In modern sparse Hopfield energy functions, β serves as a scaling factor for the energy function, influencing the sharpness of the energy landscape and thereby controlling the dynamics [24]. Hu et al. argue that high β values, corresponding to low temperatures ($\tau < 1$), help maintain distinct basins of attraction for individual memory patterns, facilitating easier retrieval.

As discussed in the main paper, we provide an ablation study on the hyperparameter τ (which is equivalent to β) by varying τ from 0.9 to 0.1 for Softmax, alongside our default configuration (1.5−Entmax). Similar to the previous ablation study, we generate 5K images from randomly selected samples in the MS-COCO validation set under CFG and PAG guidance with our PLADIS, as shown in Fig. 10.

We observed that lowering the temperature (increasing β) consistently improved generation performance in both transformations, such as Softmax and 1.5−Entmax. In the case without PLADIS, Softmax with a lower temperature improved all metrics, but its performance still remained inferior to sparse attention (α = 1.5). When using PLADIS, the trend was similar: Softmax with a lower temperature benefited from PLADIS, but it still did not outperform the 1.5−Entmax configuration with PLADIS.

Furthermore, 1.5−Entmax with a lowered temperature consistently improves generation quality in terms of visual quality and text alignment, ultimately converging to similar performance. Notably, very low temperatures with Softmax result in nearly identical sparse transformations, but with larger-than-zero intensities. This suggests that lowering the temperature benefits all transformations in α−Entmax for $1 \leq \alpha \leq 2$. However, dense alignment with a lowered temperature is insufficient, and sparse attention remains necessary in both cases, with and without PLADIS. Additionally, adjusting other hyperparameters is time-consuming, but our PLADIS with 1.5−Entmax does not require finding the optimal hyperparameter τ, thanks to the convergence of performance across various τ values. Therefore, these results demonstrate that the noise robustness of sparse cross-attention in diffusion models (DMs) is crucial for generation performance.

G.2. Analysis on Cross-Attention Map

To analyze the effect of our proposed method in the cross-attention module, we directly visualize the cross-attention maps, as shown in Fig. 11. Each word in the prompt corresponds to an attention map linked to the image, showing that the information

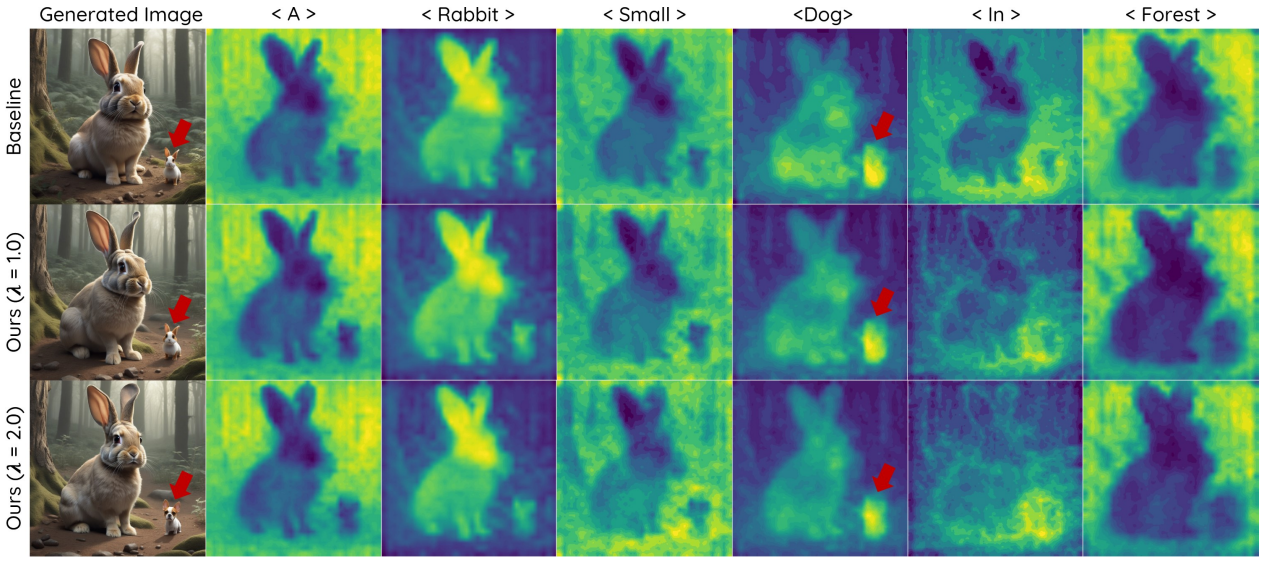

Figure 11. Qualitative comparison of cross-attention average maps across all time steps. Top: Baseline. Middle: PLADIS (with $\lambda = 1$) represent only use $\alpha-\texttt{Entmax}$ transformation. Bottom: PLADIS (with $\lambda = 2.0$). Our PLADIS with $\lambda = 2.0$ provides a more sparse and sharp correlation with each text prompt, especially "rabbit" and "dog." Furthermore, other approaches yield incorrect attention maps that highlight the space between the dog prompt and rabbit space. However, our method provides an exact attention map.

related to the word appears in specific areas of the image. We observe that the baseline (dense alignment with softmax) produces blurrier attention maps for the related words. Moreover, the generated image does not accurately reflect the text prompt of a "small dog," instead generating a "small rabbit." The cross-attention map highlights the small rabbit and a large rabbit nearby, associated with the dog prompt, resulting in poor text alignment.

When replacing the cross-attention with a sparse version, the maps become more sparse but still generate a "small rabbit" and incorrect attention maps. In contrast, our PLADIS produces both sparse and sharp attention maps compared to the baseline, and correctly aligns the attention maps with the given text prompts. As a result, PLADIS consistently improves text alignment and enhances the quality of generated samples across various interaction guidance sampling techniques and other distilled models.

G.3. The Effect of Layer Group Selection

To apply PLADIS in the cross-attention module, we incorporate it into all layers, including the down, mid, and up groups in the UNet. In SDXL, each group contains multiple layers; for example, the mid group has 24 layers, while the up group has 36 layers. To examine the effect of layer group selection, we focus on groups like the mid and up, instead of studying each layer *ex.* the first layer in the up group. We conduct experiments by varying the groups for the application of PLADIS in the cross-attention module, as shown in Tab 9.

Similar to previous ablation studies, we generate 5K samples from randomly selected data in the MS COCO validation set under CFG and PAG guidance. We observe that when applied to a single group, the up group has the most significant impact compared to others. However, in all cases, the use of PLADIS improves both generation quality and text alignment, as measured by FID and CLIPScore. Finally, combining all groups yields the best performance, confirming that no heuristic search for the target layer is necessary and validating our default configuration choice.

G.4. Two Extrapolation Strategies

To validate our design choice, we investigate two types of extrapolation strategies using different attention mechanisms: in-model extrapolation and output-based extrapolation. For in-model extrapolation, we test perturbations using sparse attention, the identity matrix (PAG), and blurred attention maps (SEG). We observe that only sparse attention consistently improves performance under extrapolation, while other variants yield semantically meaningless outputs even under minor extrapolation (Fig. 12). This suggests that sparse attention operates as a valid energy landscape under Modern Hopfield dynamics, whereas identity or blurred attention matrices may affect diffusion outputs but fail to define coherent attention dynamics, ultimately leading to degraded generation quality.

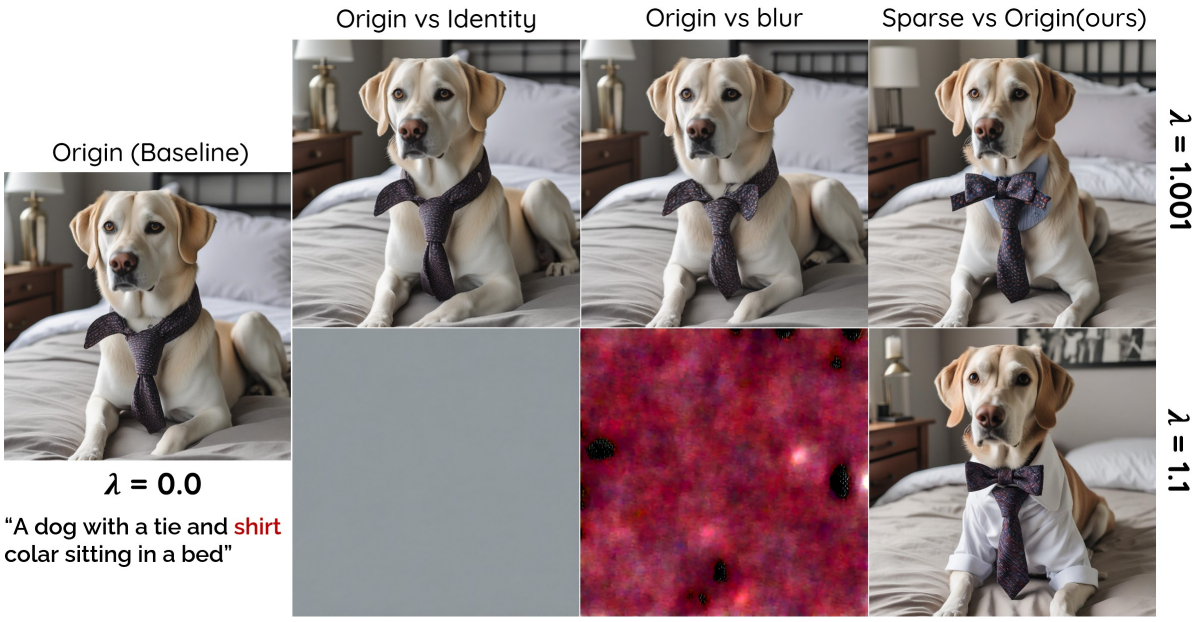

Figure 12. In-model extrapolation results. Other perturbation approaches result in semantically degraded outputs even under minor extrapolation, whereas our method consistently improves generation quality.

Table 10. Quantitative comparison on Geneval. Rows denote different methods, and columns denote guidance/backbone combinations.

Method	SDXL (CFG)	SDXL (CFG + PAG)	SDXL (CFG + SEG)	FLUX (schnell)	FLUX (dev)
Baseline	0.547	0.553	0.551	0.671	0.666
Only Sparse	0.581	0.571	0.582	0.694	0.676
Ours (Extrapolation)	0.594	**0.598**	**0.601**	**0.713**	**0.691**

We also explore output-based extrapolation using both sparse and dense attention variants. Although the output-based version of our method yields better performance than the baseline, it still underperforms compared to our in-model extrapolation while incurring higher inference costs (Tab. 8). These findings further support the efficiency and efficacy of our in-model extrapolation approach.

Our design is grounded in a principled integration of Hopfield retrieval dynamics and diffusion guidance. Specifically, we reinterpret extrapolation as a guidance process between a strong and a weak attention module inside the model. While diffusion-level extrapolation using blurred or identity attention may produce plausible outputs, such attention forms cannot act as valid components of attention dynamics. In contrast, sparse attention preserves the energy-based retrieval structure required for stable and interpretable in-model extrapolation.

G.5. Comparison with Sparse Attention Only

To isolate the benefit of our extrapolation design beyond simply applying sparse attention, we conduct experiments on the Geneval benchmark, a reliable dataset for evaluating both text-image coherence and visual quality. As shown in Table 10, across various guidance settings and even with a more challenging backbone (MMDiT), our method—extrapolation between sparse and dense attention—consistently outperforms both the baseline and the version using only sparse attention. These results further validate the effectiveness of our design choice.

H. Additional Qualitative Results

In this section, we present additional qualitative results to highlight the effectiveness and versatility of our proposed method, PLADIS, across various generation tasks and in combination with other approaches.

Comparison of Guidance Sampling with Our Method Fig. 15, 16, and 17 provide qualitative results demonstrating interactions with existing guidance methods such as CFG, PAG, and SEG, respectively. By combining PLADIS with these guidance approaches, we observe a significant enhancement in image plausibility, particularly in text alignment and coherence with the given prompts, including improvements in visual effects and object counting. Through various examples of this joint usage, we demonstrate that PLADIS improves generation quality without requiring additional inference steps.

Comparison of Guidance-Distilled Models with Ours Fig. 18 and 19 present qualitative results from applying our method, PLADIS, to guidance-distilled models such as SDXL-Turbo [50], SDXL-Lightening [33], DMD2 [64], and Hyper-SDXL [44], for both 1-step and 4-step cases. Notably, PLADIS significantly enhances generation quality, removes unnatural artifacts, and improves coherence with the given text prompts, all while being nearly cost-free in terms of additional computational overhead.

Ablation Study on Scale λ Fig. 20 shows a visual example of conditional generation with controlled scale λ. We generate samples using a combination of CFG and PAG, or CFG and SEG. For the ablation study, all other guidance scales are fixed, and only our scale λ is adjusted. Consistent with the results shown in Sec 6, a scale λ of 2.0 produces the best results in terms of visual quality and text alignment, which leads to our default configuration.

Ablation Study on α in α-`Entmax` As discussed in Sec. 6, PLADIS offers two options for choosing α: 1.5 or 2. Fig. 21 provides a qualitative comparison between the baseline, $\alpha = 1.5$, and $\alpha = 2$. Empirically, we adopt $\alpha = 1.5$ as our default configuration. While PLADIS with $\alpha = 2$ improves generation quality and text alignment compared to the baseline (dense cross-attention), PLADIS with $\alpha = 1.5$ offers a more stable and natural enhancement in sample quality.

Figure 13. Qualitative evaluation of Stable Diffusion 1.5 using our PLADIS method: PLADIS significantly boosts generation quality, strengthens alignment with the given text prompt, and generates visually compelling images.

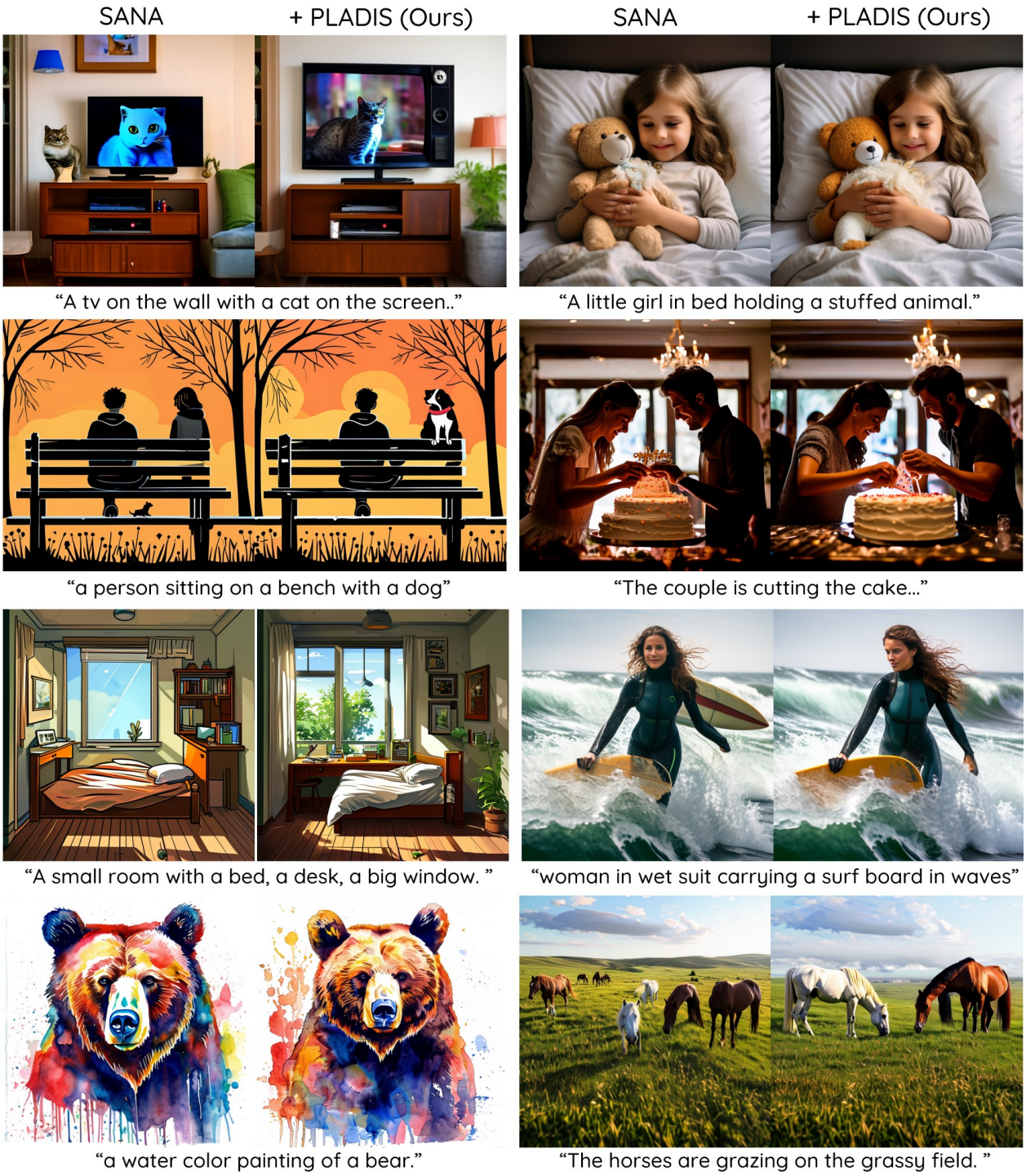

Figure 14. Qualitative assessment of SANA [62] with and without our PLADIS method: PLADIS notably improves generation quality, strengthens alignment with the provided text prompt, and produces visually striking images.

Figure 15. Qualitative evaluation of the joint usage CFG [19] with our method: CFG with PLADIS generates more plausible images with significantly improved text alignment based on the text prompt, without requiring additional inference.

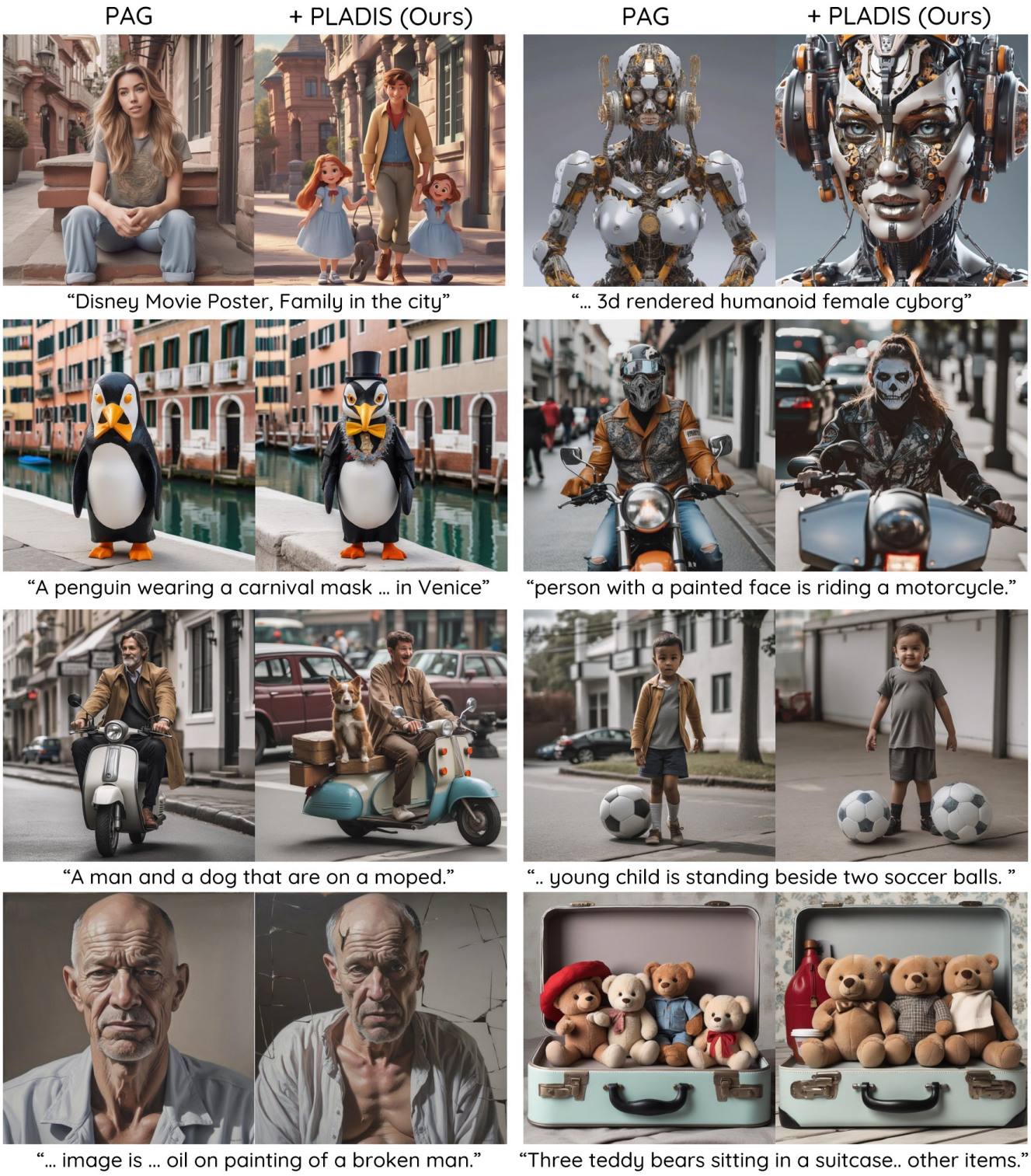

Figure 16. Qualitative evaluation of the joint usage PAG [1] with our method: Integrating PAG with PLADIS produces highly credible images with markedly enhanced correspondence to the text prompt, all achieved without any further inference steps.

Figure 17. Qualitative evaluation of the joint usage SEG [21] with our method: The combination of SEG and PLADIS yields highly convincing image generations with substantially improved alignment to the given text prompt, accomplished without the need for additional inference.

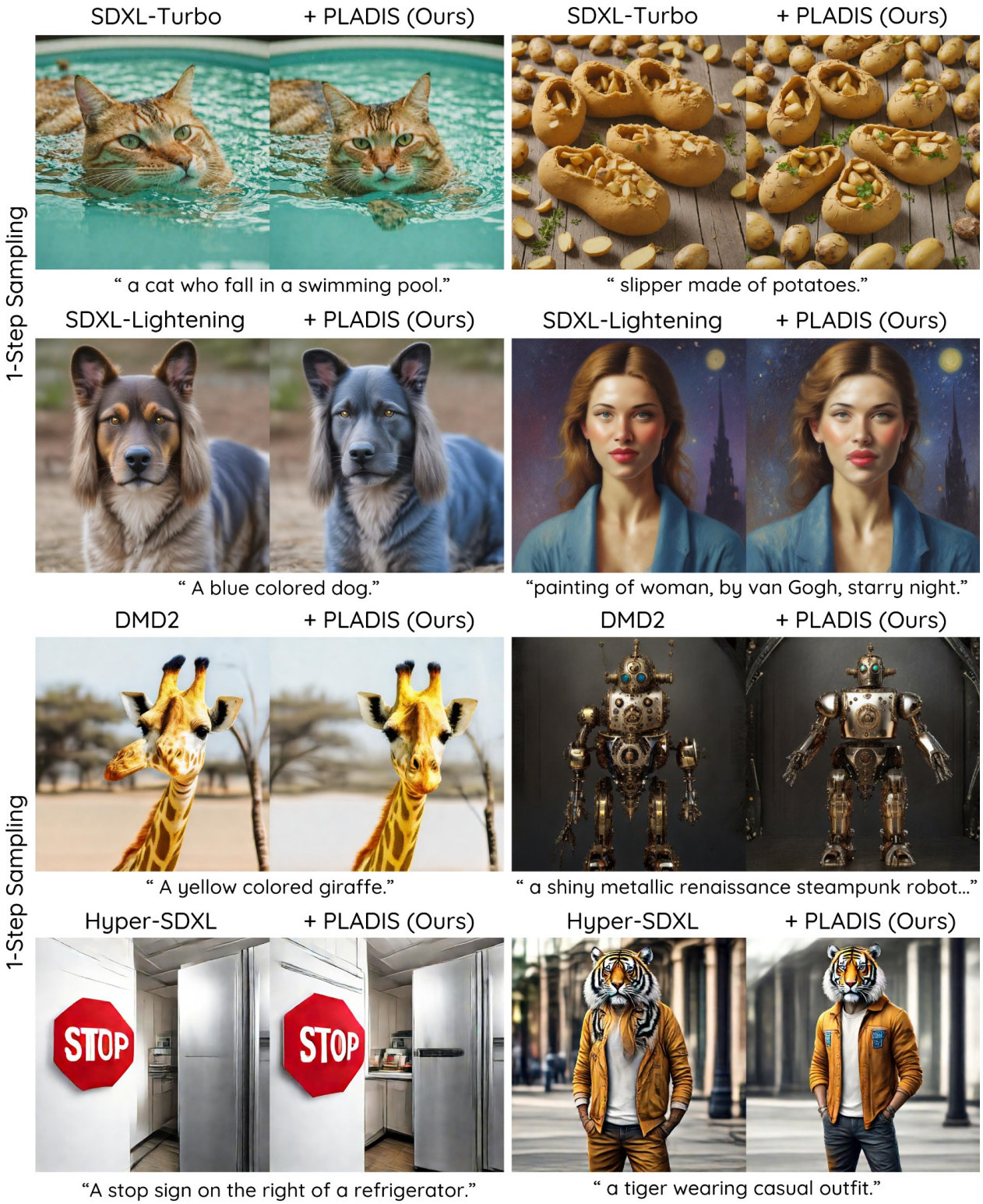

Figure 18. Qualitative comparison of the guidance-distilled model with our PLADIS method for one-step sampling: Even with one-step sampling, our PLADIS enhances generation quality, improves coherence with the given text prompt, and produces visually plausible images.

Figure 19. Qualitative comparison of the guidance-distilled model using our PLADIS method for four-step sampling: In the case of the four-step sampling approach, PLADIS substantially improves generation quality, enhances alignment with the provided text prompt, and produces visually convincing images.

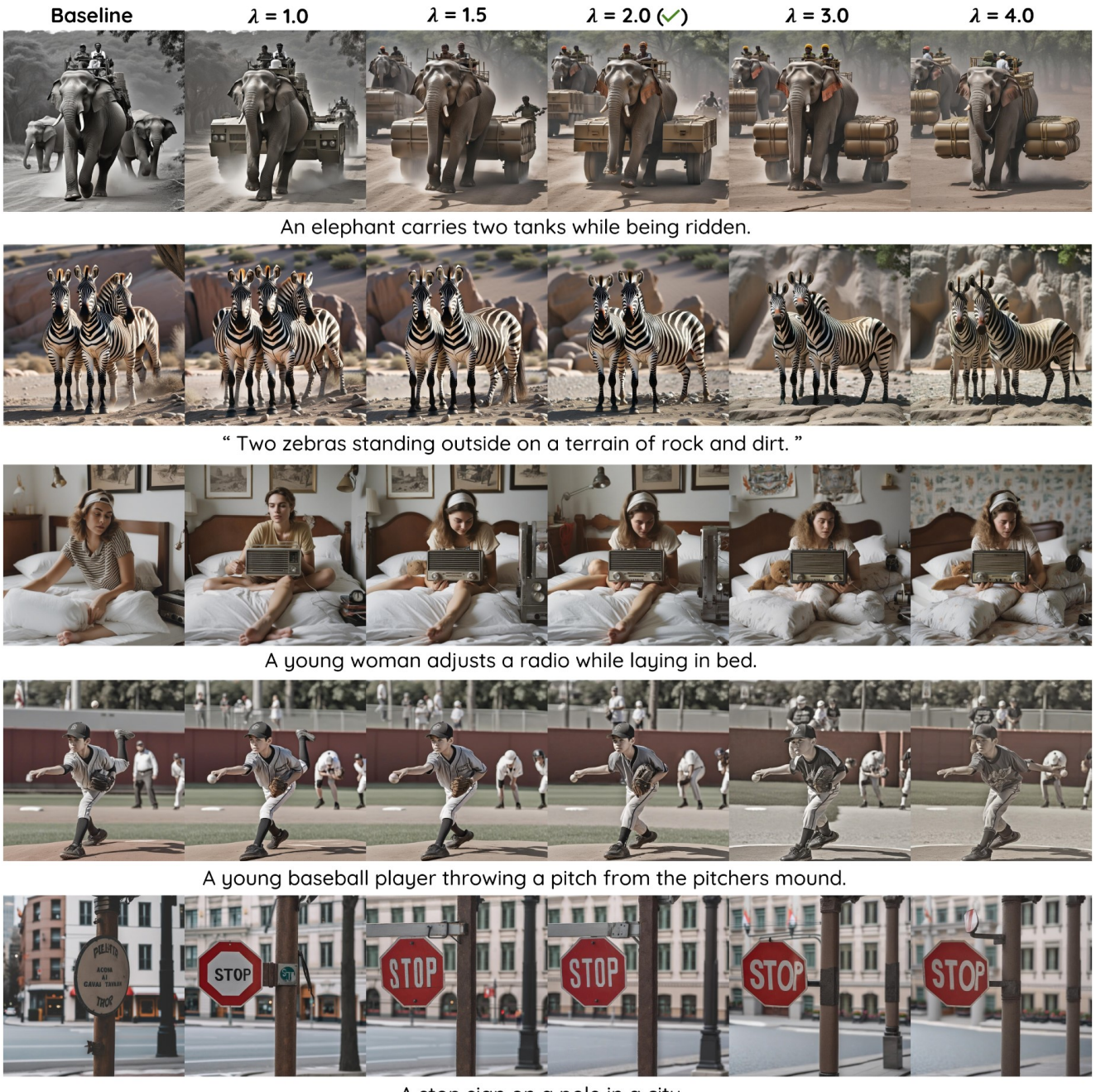

Figure 20. Qualitative comparison by varying the scale λ: As λ increases, the images display greater plausibility and improved text alignment. However, excessively high values lead to smoother textures and potential artifacts, similar to those found in CFG. The first two rows of images are generated using CFG and PAG, while the remaining rows are produced with CFG and SEG. When λ is greater than 1, our PLADIS method is applied. In our configuration, λ is set to 2.0.

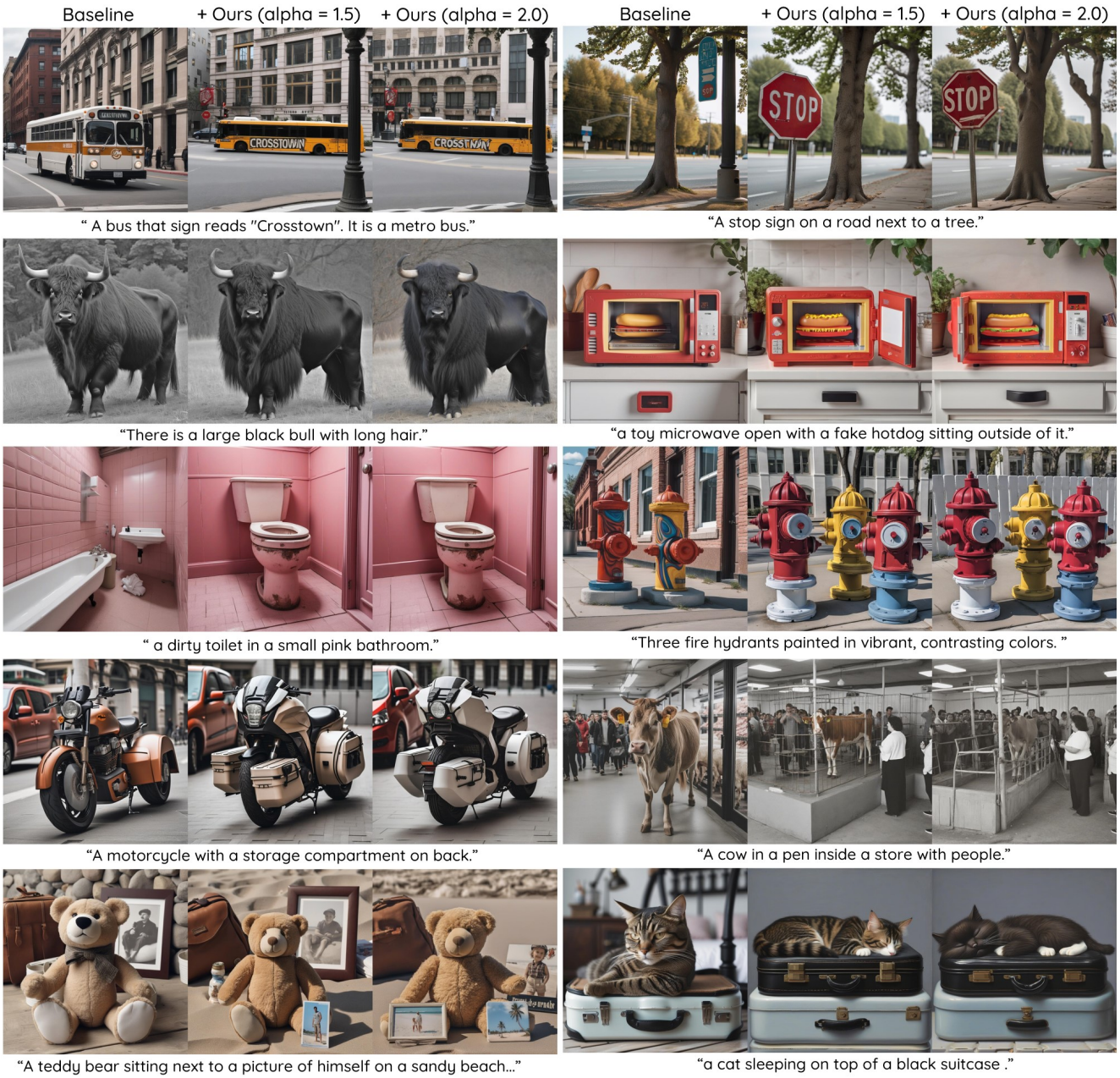

Figure 21. Qualitative comparison by α in PLADIS: Although PLADIS with $\alpha = 2$ also sifgnificantly improves generation quality and text alignment compared to the baseline (dense cross-attention), PLADIS with $\alpha = 1.5$ offers a more robust and coherence given text prompts, leads to our base configuration as $\alpha = 1.5$.