BEYOND POSITIONAL ADAPTATION: CALIBRATING ATTENTION FOR IMAGE RESOLUTION EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion transformers have rapidly advanced image generation, yet training them at higher resolutions remains expensive because dense attention scales quadratically with the number of image tokens. A prominent approach to training-free resolution extrapolation adapts rotary positional embeddings to spatial coordinates beyond the training range. However, higher resolution also expands the pool of tokens competing within each softmax, which can dilute attention to informative tokens even after positional adaptation. We introduce Cardinality-Calibrated Attention (CCA), a plug-and-play correction that compensates for this change in token count. CCA rescales attention logits by $g = \log K / \log K_0$, where K and K_0 are the current and native-reference key counts of each attention pool. The rule retains the chosen positional mapping and requires neither retraining nor fitted calibration parameters. It exactly recovers the native attention operator at the reference key count. Experiments on FLUX and Qwen-Image across multiple benchmarks demonstrate that CCA improves image quality in training-free resolution extrapolation.

1 INTRODUCTION

Modern diffusion transformers generate images with rich visual detail and complex compositions (Chen et al., 2024; Labs, 2024; Wu et al., 2025). Extending these capabilities to higher resolutions through additional training, however, remains expensive. Under a fixed tokenization scheme, image-token count grows with image area, and dense attention has quadratic computational cost in that count. Training-free resolution extrapolation offers a practical way to reuse pretrained generators at larger image sizes without further high-resolution training (Jin et al., 2023; Du et al., 2024a; Issachar et al., 2026).

A prominent approach to this task adapts the model's positional representation to an enlarged image grid (Lu et al., 2024; Zhuo et al., 2024; Issachar et al., 2026). Larger grids can introduce spatial coordinates and relative offsets outside the training range, disrupting the positional relationships on which generation relies. Rotary positional embeddings (RoPE) encode relative offsets through rotations of queries and keys. Position interpolation, NTK-aware scaling, and YaRN adapt their coordinate or frequency mappings, while DyPE further varies the adaptation over denoising (Issachar et al., 2026). These methods address the spatial relationships that the model must represent at larger resolutions.

However, a larger image grid also expands the attention pools containing image keys, increasing the number of tokens competing within each softmax. Even after positional adaptation, this increase in competition can alter the attention assigned to informative tokens. When a query needs to attend selectively to a limited set, additional competitors can dilute the mass assigned to that set unless score separation grows accordingly. We refer to this effect as *attention dilution*. Its effect on feature aggregation motivates treating *positional geometry and attention-pool size as complementary factors in resolution extrapolation*.

YaRN (Peng et al., 2024) studies attention scaling for language-model context extension, while entropy-based scaling has also been explored in variable-sized diffusion (Jin et al., 2023). Temperature settings developed for language models need not transfer optimally to DiT resolution extrapolation. Increasing spatial resolution expands image-only and joint image–text key pools, while fixed-prompt text self-attention and cross-attention with text keys retain the same key counts. Adapting attention

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.

scaling to a frozen DiT therefore requires specifying which pools change, how each pool is anchored to the model's native reference, and how calibration composes with the chosen positional mapping.

We introduce Cardinality-Calibrated Attention (CCA), a calibration rule tailored to these differences in DiT attention pools. We base each pool's correction on its actual visible key count K_q and its reference count $K_{0,q}$ at a declared native image grid, under the same prompt and mask convention. We choose a logarithmic dependence on key count and require exact recovery of the unit-scale attention operator at that reference. Within this family, the constraint determines $g_q = \log K_q / \log K_{0,q}$, adapting the correction to each pool; pools with unchanged key counts retain unit scale. CCA retains the chosen positional mapping and replaces any existing attention-temperature correction, allowing spatial adaptation and attention calibration to be applied together without retraining or fitted calibration parameters.

Our analysis links this design to count competition and characterizes its limits. For a fixed selection budget and bounded score spread, we show that maintaining attention mass requires a logit scale that grows at least logarithmically with pool size. A two-population model with fixed support and score levels then gives the exact multiplier that preserves native attention mass and an explicit finite-count residual for CCA. This residual identifies when CCA under- or overcompensates within the model, distinguishing exact native operator recovery from attention-mass preservation in expanded pools.

Experiments on FLUX.1-Krea-dev and Qwen-Image across DrawBench and Aesthetic-4K evaluate this calibration in image generation. We hold positional mappings fixed to isolate its effect and compare CCA with complete positional baselines, a fitted constant temperature, and a fitted count-dependent strength. Evaluations across resolutions and backbones test transfer of the same unfitted rule using each backbone's own native reference. The results show that CCA improves on DY-YaRN's original temperature correction and remains competitive with fitted alternatives, with gains in image quality and local detail.

Our contributions are:

- We develop CCA for resolution extrapolation in pretrained DiTs, calibrating their distinct attention pools through native-reference key counts and integrating with positional adaptation without retraining or fitted parameters.
- We determine the calibration coefficient within the logarithmic scaling family by exact native operator recovery, and characterize its finite-count mass discrepancy and failure boundaries in a two-population model.
- We evaluate CCA through controlled positional and temperature comparisons and transfer across resolutions and diffusion backbones, demonstrating improvements in image quality without refitting.

2 Related Work

Diffusion transformers. DiT establishes transformers over latent image patches as scalable diffusion backbones (Peebles & Xie, 2023). PixArt-α extends this approach to efficient text-to-image generation (Chen et al., 2024), while Stable Diffusion 3, FLUX, and Qwen-Image develop multimodal backbones for image–text interaction (Esser et al., 2024; Labs, 2024; Wu et al., 2025). Across these models, image self-attention, text cross-attention, and joint image–text attention provide different conditioning layouts. Their growing capability motivates extending pretrained models to higher resolutions, where dense attention becomes increasingly expensive.

High-resolution image generation. Existing methods pursue higher resolutions through progressive sampling, scale fusion, model adaptation, and super-resolution. DemoFusion combines progressive upscaling with skip residuals and dilated sampling (Du et al., 2024a), while FreCaS uses frequency-aware cascaded sampling (Zhang et al., 2024). DiffuseHigh uses low-resolution images to guide high-resolution structure (Kim et al., 2025), and FreeScale fuses frequency components from different receptive scales (Qiu et al., 2025). For transformer-based generators, I-Max introduces projected flow to stabilize resolution extrapolation (Du et al., 2024b). UltraPixel uses cascaded diffusion with implicit upsampling (Ren et al., 2024), while Diffusion-4K employs wavelet-based fine-tuning on 4K images (Zhang et al., 2025). A complementary route pairs image generation with

a separate super-resolution stage, as in the FLUX + BSRGAN pipeline evaluated by DyPE (Zhang et al., 2021; Issachar et al., 2026).

Position extrapolation. Static schemes use a fixed positional mapping for a given target resolution. Position interpolation (PI) compresses position indices into the pretrained range (Chen et al., 2023). NTK-aware scaling adjusts the RoPE base to preserve high-frequency information, while YaRN blends interpolated and unchanged rotary frequencies across spectral bands (Peng et al., 2024). FiT and FiTv2 adapt these schemes to image grids (Lu et al., 2024; Wang et al., 2024). Dynamic schemes further vary positional adaptation during denoising. Lumina-Next introduces time-aware scaling between PI and NTK-aware mappings (Zhuo et al., 2024), and DyPE modulates extrapolation parameters to follow the coarse-to-fine progression of image formation (Issachar et al., 2026). SigMa further explores sigmoid modulation for position extrapolation in ultra-high-resolution diffusion (Zhao et al., 2026). CCA complements these spatial adaptations by retaining the chosen positional mapping and calibrating attention logits according to actual attention-pool growth.

Attention scaling. Jin et al. (2023) derive an asymptotic entropy relation for Gaussian scores and use a square-root logarithmic count ratio in diffusion attention. YaRN fits attention temperature for LLM context extension (Peng et al., 2024). Scalable-Softmax (Nakanishi, 2025) uses $\mathrm{softmax}((c \log K)\mathbf{s})$ and tests replacement after pretraining, initializing c from the average log training context size. CCA fixes c by exact native recovery for each DiT attention pool.

3 Method

We analyze count competition and exact native-mass compensation, then derive CCA by imposing native recovery within a logarithmic scaling family. An exact residual relates the two calibration criteria at finite pool sizes.

3.1 Cardinality Competition

For query q, let $\mathcal{K}_q$ be the keys admitted by its mask and $K_q = |\mathcal{K}_q|$. After QK normalization and RoPE, write $s_{qj} = \langle \mathbf{q}^{\mathrm{rot}}, \mathbf{k}_j^{\mathrm{rot}} \rangle / \sqrt{d_h}$. A positive effective logit multiplier g defines

$$a_{qj}(g) = \frac{e^{g s_{qj}}}{\sum_{k \in \mathcal{K}_q} e^{g s_{qk}}}, \qquad \mathbf{o}_q(g) = \sum_{j \in \mathcal{K}_q} a_{qj}(g) \mathbf{v}_j. \tag{1}$$

$K_{0,q} > 1$ is the key count at the native reference grid under the same prompt, module, and mask convention. We suppress q below and study $K \geq K_0$. For a nonempty selected set $\mathcal{A}$ of $b < K$ keys, let $M_{K,\mathcal{A}}(g)$ be its mass and $\Delta_{\mathcal{A}}$ the difference between the selected and background log-mean-exp scores. The softmax denominator gives

$$\log \frac{M_{K,\mathcal{A}}(g)}{1 - M_{K,\mathcal{A}}(g)} = \underbrace{\Delta_{\mathcal{A}}(g; K)}_{\text{score contrast}} - \underbrace{\log \frac{K-b}{b}}_{\text{count competition}}. \tag{2}$$

For fixed b, expansion increases the count penalty. If the full score spread is uniformly bounded and $b(K) = \Theta(K^{\beta})$ with $0 \leq \beta < 1$, preserving a fixed positive top-b mass requires $g = \Omega(\log K)$ (Appendix A.1). This necessary order leaves the coefficient undetermined.

3.2 Exact Compensation of Native Attention

To isolate count growth, consider a fixed set of $1 \leq b < K_0$ selected keys with score $\delta > 0$ and background keys with score zero after a common shift. Holding these score levels fixed as K grows gives

$$m_K(g) = \mathrm{sigmoid}\left(g\delta - \log \frac{K-b}{b} \right), \qquad \rho = m_{K_0}(1). \tag{3}$$

Our calibration target is to preserve this native mass. Solving $m_K(g_*) = \rho$ gives the unique exact multiplier

$$g_*(K) = 1 + \frac{1}{\delta} \log \frac{K-b}{K_0 - b}. \tag{4}$$

Exact compensation depends on the support b and score gap δ; key counts alone cannot determine it for every attention distribution.

3.3 Cardinality-Calibrated Attention

Exact native anchoring. The logarithmic order in Section 3.1 motivates choosing the logarithmic scaling family (Nakanishi, 2025),

$$\mathbf{P}_K^{(c)}(\mathbf{s}) = \mathrm{softmax}((c \log K)\mathbf{s}), \qquad c > 0, \tag{5}$$

where c is independent of the current count K for a given native reference. We require the attention operator at K_0 to agree with the pretrained unit-scale operator for every score vector:

$$\mathbf{P}_{K_0}^{(c)}(\mathbf{s}) = \mathrm{softmax}(\mathbf{s}) \quad \text{for all } \mathbf{s} \in \mathbb{R}^{K_0}. \tag{6}$$

Softmax pairwise odds satisfy $\log(P_i/P_j) = c \log K_0 (s_i - s_j)$. For any nonzero score difference, Equation (6) therefore holds if and only if $c \log K_0 = 1$. Substitution gives the exact rule

$$\boxed{\lambda(K) = \frac{\log K}{\log K_0}, \qquad a_{K,j}^{CCA} = \frac{e^{\lambda(K) s_j}}{\sum_k e^{\lambda(K) s_k}}.} \tag{7}$$

CCA recovers unit scale at $K = K_0$, preserves score rankings, and requires neither estimates of b, δ nor a fitted calibration strength. The coefficient is unique within Equation (5); native recovery alone does not select a scaling family. Appendix A.2 also gives an exact entropy-regularized characterization of the resulting operator.

Relation to exact mass compensation. In the two-population model, the native mass gives the exact identity

$$\delta = \log K_0 + \xi + \log(1 - b/K_0), \qquad \xi = \mathrm{logit}\, \rho - \log b. \tag{8}$$

Writing $r_K = \log((1 - b/K)/(1 - b/K_0))$, we can decompose Equation (4) without dropping either correction:

$$g_*(K) = 1 + \frac{\log K_0}{\delta}\big(\lambda(K) - 1\big) + \frac{r_K}{\delta}. \tag{9}$$

Thus the difference between CCA and exact mass preservation is fully specified:

$$\mathrm{logit}\, m_K(\lambda) - \mathrm{logit}\, \rho = (\lambda - 1)(\delta - \log K_0) - r_K. \tag{10}$$

CCA preserves native mass exactly if and only if this residual vanishes. The logarithmic compensation slope in Equation (9) is $\log K_0/\delta$, with the separate finite-count correction r_K/δ. To examine sensitivity to the slope, we introduce the strength family

$$g_\alpha(K) = 1 + \alpha\big(\lambda(K) - 1\big), \qquad \alpha \geq 0. \tag{11}$$

Here α scales the added compensation $\lambda(K) - 1$, so $g_\alpha(K_0) = 1$ for every α. Unit scale and CCA correspond to $\alpha = 0$ and $\alpha = 1$, respectively; larger α strengthens compensation when $K > K_0$. In generation experiments, a shared α serves as a strength control, without estimating individual score gaps. Section 4.2 uses this family to assess sensitivity and the benefit of fitting; CCA itself fixes $\alpha = 1$ without fitting.

The native anchoring derivation is exact for arbitrary finite scores. Its mass-preservation error is model-dependent; the native-pool limit varies K_0 and does not assert unlimited extrapolation of a fixed backbone. Finite-count error bounds and failure boundaries appear in Appendices A.3 and A.4.

3.4 Applying CCA to Diffusion Transformers

Restore query indices and set $\lambda_q = \log K_q / \log K_{0,q}$. With full image visibility, packed image-token counts N, N_0, and L_q visible text keys,

$$(K_q, K_{0,q}) = \begin{cases} (N, N_0), & \text{image-only attention,} \\ (N + L_q, N_0 + L_q), & \text{joint image–text attention,} \\ (L_q, L_q), & \text{fixed-prompt text-key attention.} \end{cases} \tag{12}$$

The last case covers text self-attention and cross-attention with text keys. General masks require row-wise visible counts, including unmasked padding; counts at most one retain the backbone convention. The native grid is a declared inference reference.

Scaling post-normalization, post-RoPE queries by λ_q implements CCA through the existing attention kernel when the additive bias only masks keys. CCA replaces any existing temperature correction while retaining the positional mapping and values. Appendix B.1 and Algorithm 1 specify the effective-scale accounting.

4 Experiments

We first evaluate CCA across positional mappings and backbones, then ablate the calibration rule and strength to assess transfer and the benefit of fitting. Visual comparisons and human preference examine the resulting image quality. The native anchor fixes the coefficient within the chosen logarithmic family, and the two-population model characterizes its mass discrepancy. These experiments test utility in generation and transfer, without assuming that every attention head follows that model.

Evaluation protocol. The evaluation uses FLUX.1-Krea-dev, a 1024^2 inference reference, the 28-step DyPE configuration, 200 DrawBench prompts and 195 Aesthetic-4K image–prompt pairs (Issachar et al., 2026; Saharia et al., 2022; Zhang et al., 2025). Within each controlled positional or temperature comparison, checkpoint, prompt, initial latent, guidance, precision, and the complete sigma schedule are identical.

ImageReward is the primary endpoint; CLIPScore and aesthetic score are complementary (Xu et al., 2023; Hessel et al., 2021; Schuhmann, 2022). The small-reference FID results remain supplementary (Heusel et al., 2017; Chong & Forsyth, 2020). Per-image scores are averaged over the prompt sets specified for each experiment. Method comparisons use matched prompts within each evaluation cohort.

Appendix B specifies the comparison design, sample accounting, metric preprocessing, and score aggregation.

4.1 Effectiveness and Backbone Transfer

Complete positional baselines. Table 1 groups results by resolution, placing the complete positional baselines above their CCA replacements. Benchmark subcolumns expose alignment, aesthetics, and distributional quality. Table 4 supplements these square-resolution results with rectangular outputs, including unit-temperature positional components. The results show that gains depend on the benchmark: replacing full DY-YaRN's temperature with CCA improves ImageReward by $+0.16/+0.19$ at 3K and $+0.06/+0.17$ at 4K on DrawBench/Aesthetic-4K. Point estimates alone do not establish significance. There are also regressions: at 2048×3072, DY-NTK's Aesthetic-4K ImageReward changes from 1.00 to 0.98. Beyond ImageReward, the 4K DY-YaRN replacement improves CLIPScore by $1.15/1.22$ and aesthetic score by $0.30/0.41$ on DrawBench/Aesthetic-4K; Aesthetic-4K FID decreases from 188.91 to 176.02. These are point estimates. At 3K, DY-NTK + CCA has a lower FID than DY-YaRN-pos + CCA (171.22 versus 174.15), so no single configuration wins every metric.

Transfer without refitting. Qwen-Image is evaluated on the complete DrawBench and Aesthetic-4K datasets (200 and 195 prompts, respectively), using its own native grid and packed counts. Under each positional mapping, the comparison retains the same inference settings and uses CCA with $\alpha = 1$, without refitting or importing FLUX's count anchor. The resolution sweep compares NTK, NTK + CCA, YaRN-pos, and YaRN-pos + CCA from 2K to 5K in 0.5K increments.

Figure 1 shows how ImageReward changes with resolution. The corresponding numerical results are reported in Appendix C.2, Table 7. CCA improves YaRN-pos by $+0.31/+0.09$ at 4K and $+0.48/+0.31$ at 5K on DrawBench/Aesthetic-4K. The corresponding 2K changes are $-0.01/+0.12$. The ranking also depends on the benchmark and resolution: at 4K on Aesthetic-4K, NTK + CCA scores 0.41, compared with 0.38 for YaRN-pos + CCA. These results support transfer without refitting, while showing that CCA does not improve every configuration at every resolution.

Table 1: **Resolution extrapolation with complete positional baselines.** Two resolution groups and a separated CCA block follow the presentation of Issachar et al. (2026), Table 3. YaRN baselines retain their original temperature; CCA replaces it. Bold/underlined values are the best/second distinct point values within each column, without implying significance. Results at rectangular resolutions, including unit-temperature positional components, appear in Table 4.

Method	3072 × 3072							4096 × 4096						
	DrawBench			Aesthetic-4K				DrawBench			Aesthetic-4K			
	CLIP↑	IR↑	Aesth↑	FID↓	CLIP↑	IR↑	Aesth↑	CLIP↑	IR↑	Aesth↑	FID↓	CLIP↑	IR↑	Aesth↑
Vanilla	25.43	-0.42	5.12	185.12	28.91	0.56	6.27	16.14	-2.10	3.16	198.85	25.18	-0.84	5.30
NTK	26.39	-0.03	5.16	179.37	28.97	1.07	6.57	17.20	-2.01	3.44	207.24	24.55	-0.65	5.38
DY-NTK	27.35	0.41	5.47	176.89	29.12	1.21	6.69	21.21	-1.34	4.13	186.56	27.75	0.70	6.31
YaRN	28.26	0.52	5.48	181.20	29.56	1.25	6.79	25.43	-0.45	4.74	195.27	28.27	0.77	6.36
DY-YaRN	28.46	0.78	5.67	176.73	30.06	1.35	6.82	26.67	0.05	5.06	188.91	28.99	1.01	6.57
Vanilla + CCA	26.51	-0.38	5.41	181.83	30.22	0.70	6.59	16.98	-1.93	3.34	187.12	26.37	-0.77	5.61
NTK + CCA	27.65	-0.02	5.48	172.91	30.28	1.15	6.98	18.25	-1.82	3.68	198.46	25.86	-0.58	5.74
DY-NTK + CCA	28.52	0.45	5.76	**171.22**	30.34	1.28	7.08	22.38	-1.23	4.38	182.35	28.98	0.75	6.69
YaRN-pos + CCA	29.38	0.57	5.82	178.75	30.74	1.34	7.21	26.55	-0.41	5.03	183.62	29.48	0.83	6.77
DY-YaRN-pos + CCA	**29.68**	**0.94**	**6.01**	174.15	**31.28**	**1.54**	**7.24**	**27.82**	**0.11**	**5.36**	**176.02**	**30.21**	**1.18**	**6.98**

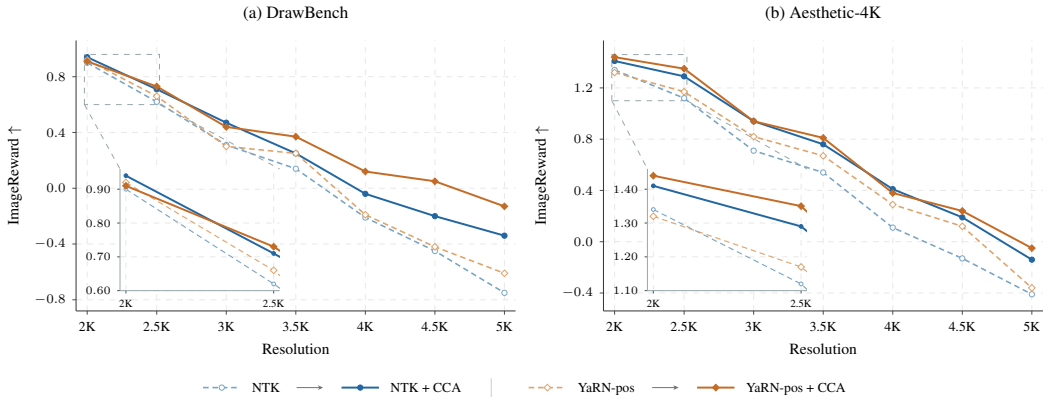

Figure 1: **Resolution scaling on Qwen-Image without refitting.** ImageReward on the complete DrawBench (200 prompts) and Aesthetic-4K (195 prompts) datasets from 2K to 5K, in 0.5K increments. In the shared legend, arrows connect each baseline to its CCA counterpart. Dashed lines with open markers show the baselines; solid lines with filled markers show CCA. Markers indicate measured scores, with straight lines connecting adjacent measurements. Insets magnify the 2K and 2.5K points. Panels use separate y-axis ranges. CCA uses Qwen-Image's own native key counts with $\alpha = 1$.

The next experiment isolates count dependence from a fixed temperature choice.

4.2 Ablations of the Calibration Rule and Strength

Controls and selection. With DY-YaRN-pos fixed, we compare unit scale, original `mscale`, the unfitted square-root-log rule of Jin et al. (2023), CCA, and two controls chosen only on 20 disjoint development prompts at 2K and 3K:

$$g_c(K) = \begin{cases} 1, & K = K_0, \\ c^\star, & K > K_0, \end{cases} \qquad g_{\alpha^\star}(K) = 1 + \alpha^\star(\lambda(K) - 1). \tag{13}$$

Both controls share native identity and leave fixed text cross-attention unchanged. The fitted-slope control selects a member of Equation (11), retaining CCA's count dependence while fitting its strength. The complete DY-YaRN and CCA entries reuse the corresponding outputs from Table 1; the additional controls distinguish the benefit of the count rule from that of changing temperature alone.

Table 2: **Count dependence and resolution transfer.** ImageReward (↑) with DY-YaRN-pos; development-selected $c^\star = 1.2$ and $\alpha^\star = 1.1$. 2K–4K use 200/195 prompts and 6K a locked 50/50-prompt cohort. Subscripts: difference from CCA (↑ higher, ↓ lower).

	DrawBench				Aesthetic-4K			
Rule	2K	3K	4K	6K	2K	3K	4K	6K
Unit	$0.90_{\downarrow 0.05}$	$0.72_{\downarrow 0.22}$	$0.04_{\downarrow 0.07}$	$-0.58_{\downarrow 0.52}$	$1.25_{\downarrow 0.08}$	$1.26_{\downarrow 0.28}$	$0.94_{\downarrow 0.24}$	$0.18_{\downarrow 0.56}$
Original `mscale`	$0.94_{\downarrow 0.01}$	$0.78_{\downarrow 0.16}$	$0.05_{\downarrow 0.06}$	$-0.31_{\downarrow 0.25}$	$1.32_{\downarrow 0.01}$	$1.35_{\downarrow 0.19}$	$1.01_{\downarrow 0.17}$	$0.46_{\downarrow 0.28}$
Square-root log	$0.93_{\downarrow 0.02}$	$0.81_{\downarrow 0.13}$	$0.08_{\downarrow 0.03}$	$-0.20_{\downarrow 0.14}$	$1.31_{\downarrow 0.02}$	$1.42_{\downarrow 0.12}$	$1.11_{\downarrow 0.07}$	$0.57_{\downarrow 0.17}$
Frozen $c^\star$	$0.96_{\uparrow 0.01}$	$0.87_{\downarrow 0.07}$	$0.07_{\downarrow 0.04}$	$-0.24_{\downarrow 0.18}$	$1.34_{\uparrow 0.01}$	$1.46_{\downarrow 0.08}$	$1.10_{\downarrow 0.08}$	$0.53_{\downarrow 0.21}$
Fitted $\alpha^\star$	$0.96_{\uparrow 0.01}$	$0.95_{\uparrow 0.01}$	$0.12_{\uparrow 0.01}$	$-0.08_{\downarrow 0.02}$	$1.34_{\uparrow 0.01}$	$1.56_{\uparrow 0.02}$	$1.19_{\uparrow 0.01}$	$0.73_{\downarrow 0.01}$
CCA (ref.)	**0.95**	**0.94**	**0.11**	**−0.06**	**1.33**	**1.54**	**1.18**	**0.74**

The constant $c^\star$ and slope $\alpha^\star$ are selected independently and frozen for both benchmarks and all test resolutions. No selection uses test prompts.

The primary held-out resolution is 4K on the full benchmark lists. A separate locked 50-prompt cohort per benchmark extends every arm to 6K. Table 2 reports the means, with each comparator's difference from CCA shown beside its score.

Transfer beyond a fixed temperature. The frozen constant is slightly better at 2K. At the held-out resolutions, CCA gains $+0.04/+0.08$ at 4K and $+0.18/+0.21$ at 6K over this control. Thus the strongest evidence for count dependence comes from transfer beyond the development resolutions, rather than the ordering at any one tuning size. All methods use the same prompts within each comparison; the 6K cohort remains separate from the full-benchmark 4K cohort.

CCA also improves over original `mscale` by $+0.06/+0.17$ at 4K and $+0.25/+0.28$ at 6K. Against the unfitted square-root-log rule, the 6K gains are $+0.14/+0.17$; the DrawBench 4K gain is smaller at $+0.03$. These comparisons assess both the existing temperature correction and an alternative count-dependent rule.

The cost of avoiding fitting. CCA differs from the fitted slope by $-0.01/-0.01$ at 4K and $+0.02/+0.01$ at 6K. These small gaps support the practical value of a rule that requires no fitting, without establishing statistical equivalence.

Sensitivity to calibration strength. A 4K sweep of α in Equation (11) on DY-YaRN-pos places the default near the best reported ImageReward means: $\alpha = 1.25$ scores $0.13/1.20$, compared with $0.11/1.18$ for CCA. Increasing strength to $\alpha = 2$ reduces the scores to $0.04/1.10$. On the locked 50-prompt 6K cohorts, CCA trails $\alpha = 1.25$ by 0.01 on both benchmarks and exceeds $\alpha = 2$ by $0.18/0.17$. Thus, stronger compensation does not yield monotonically better results. This test-set sweep describes sensitivity; CCA fixes $\alpha = 1$ without using these results for selection. The full strength results are reported in Appendix C.1, Table 5.

Appendix C.3 examines query placement and reference-count sensitivity. At 4K, actual joint-pool calibration scores $0.11/1.18$ on DrawBench/Aesthetic-4K, exceeding image-only counting by $0.04/0.25$ and image-query-only scaling by $0.03/0.29$ (Table 8). These comparisons support accounting for the joint key pool and calibrating both image and text query rows in this setting, with larger gains on Aesthetic-4K.

Attention concentration and local aggregation. To examine the local response to calibration, we expand key pools on fixed attention states while holding the queries, stored keys and values, masks, and positions unchanged (Appendix C.3). Figure 2(a) shows that CCA remains closer to the reference top-32 mass of 0.42 across the tested expansion ratios. At $K/K_0 = 16$, its mass is 0.32, compared with 0.24 for the frozen constant and 0.16 for Unit. This concentration response is accompanied by lower mean per-head output error: 0.34 for CCA, 0.44 for the frozen constant, and 0.56 for Unit (panel b). The gains vary across reference-entropy groups (panel c). CCA lowers error relative to Unit by 0.26, 0.32, and 0.07 in the lower-, middle-, and upper-entropy groups, respectively; its gains

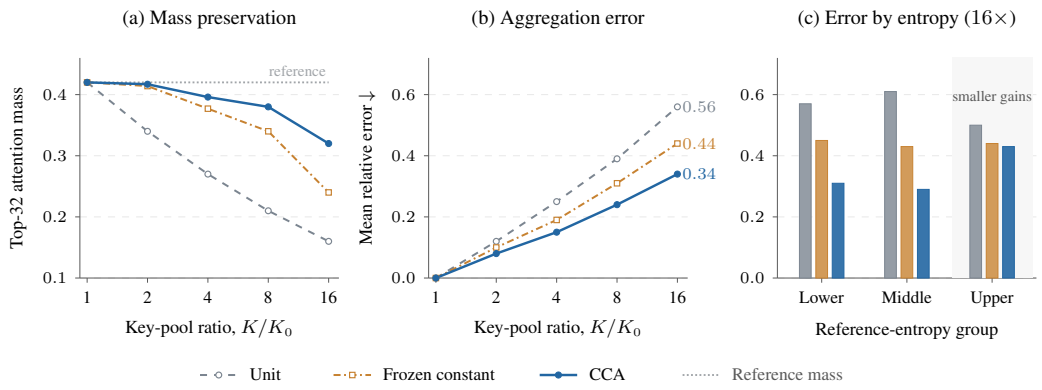

Figure 2: **Count-dependent attention concentration and local aggregation.** Unit, frozen-constant, and CCA scaling on fixed attention states. (a) Top-32 attention mass across key-pool ratios; the dotted line marks the reference-pool mass of 0.42. (b) Mean per-head relative output error against the unit-scale reference-pool output; lower is better. (c) The same error metric by reference-entropy group at $K/K_0 = 16$. CCA has lower mean error in all three groups, with smaller gains in the upper-entropy group. Table 10 lists the group means.

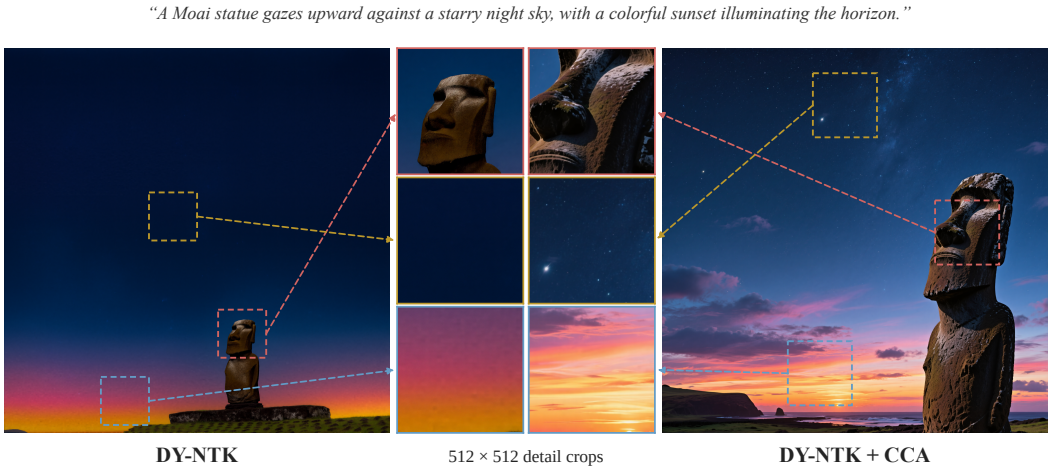

Figure 3: **Qualitative comparison of 4K generations.** Both images were generated at 4096×4096 under the displayed prompt. Left: DY-NTK; right: DY-NTK + CCA. Colored boxes mark the regions enlarged in the center: the statue, night sky, and sunset horizon.

over the frozen constant are 0.14, 0.14, and 0.01. These diagnostics support improved agreement with the reference aggregation over the tested count range, with smaller benefits in the upper-entropy group. The reference is a sampled pool within the same frozen state, so this local comparison does not by itself establish a mechanism for image-quality gains.

4.3 Visual Quality and Human Preference

Figure 3 compares 4K generations from DY-NTK and DY-NTK + CCA under the same prompt. The full images show the resulting scene composition, while enlarged regions highlight differences in statue texture, the star field, and the sunset horizon.

Figure 4 provides a complementary comparison with full DY-YaRN, the baseline family of our primary quantitative evaluation.



Figure 4: **4K qualitative comparison with the full DY-YaRN baseline.** Left: DY-YaRN; right: DY-YaRN-pos + CCA (4096×4096, same prompt). Central 640×640 crops are matched by scene content at equal scale; colored boxes mark source regions.

Table 3: **Blinded preference with original-pixel detail views.** DY-YaRN-pos + CCA is compared with full DY-YaRN on 50 prompts per benchmark, with three raters per pair. Entries report CCA's tie-adjusted preference score (%), followed by win/tie/loss counts in parentheses; 50% is neutral.

Benchmark	Structure	Local detail	Prompt adherence
DrawBench	76.3 (97/35/18)	85.0 (113/29/8)	74.3 (86/51/13)
Aesthetic-4K	75.7 (101/25/24)	82.0 (118/10/22)	78.3 (94/47/9)

Full-image reward models may resize away 4K detail. The visual protocol therefore uses original outputs, equal-source-pixel crops, and a blinded preference study with 50 fixed prompts per benchmark and three raters per pair. The study compares DY-YaRN-pos + CCA with full DY-YaRN and assesses global structure, local detail, and prompt adherence separately. Ties are allowed and presentation sides are randomized.

Table 3 shows observed preference for CCA across all three criteria on both benchmarks. Local detail receives the highest scores ($85.0\%/82.0\%$ on DrawBench/Aesthetic-4K), while prompt adherence scores $74.3\%/78.3\%$. These results support improved perceived quality in the tested comparison.

5 Conclusion and Limitations

CCA calibrates actual DiT attention pools using native-reference key counts while preserving the chosen positional mapping. The rule exactly recovers the native operator at the declared reference; the analysis quantifies its discrepancy from native attention-mass preservation under an explicit model. Controlled FLUX comparisons show setting-dependent quality gains, and held-out-resolution comparisons support count-dependent calibration without fitting its strength. Qwen-Image evaluates backbone transfer, while qualitative comparisons and human preference assess visible changes. Fitted temperatures remain competitive, and the benefits depend on the benchmark and positional baseline.

A single count rule cannot recover changed score populations, preserve every useful support, or prevent text-mass loss. The native anchor is a declared convention, and the pool-count and query-placement ablations cover a single 4K setting. Dense-attention complexity remains unchanged. The analytical example and aggregate ordering alone do not establish statistical or visible benefits; paired quality comparisons and original-pixel evaluation provide the corresponding empirical endpoints.

REPRODUCIBILITY STATEMENT

Appendices A and B provide the assumptions, proofs, deployment algorithm, and attention-probe definitions. Appendix B specifies evaluation settings, temperature selection, metric preprocessing, and score aggregation. The accompanying numerical script checks the native anchor, variational identity, exact mass residual, finite-count error bounds, and replication counterexamples.

ETHICS STATEMENT

The experiments use pretrained generators and benchmark prompts under their applicable terms. Human evaluation requires informed participation and recorded study procedures. The inference modification inherits the underlying models' misuse risks.

AI USE STATEMENT

Yes, to aid or polish writing. Details are described in the paper.

REFERENCES

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078. IEEE, 2020.

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. DemoFusion: Democratising high-resolution image generation with no $$$. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6159–6168, 2024a.

Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. *arXiv preprint arXiv:2410.07536*, 2024b.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Noam Issachar, Guy Yariv, Sagie Benaim, Yossi Adi, Dani Lischinski, and Raanan Fattal. Dype: Dynamic position extrapolation for ultra high resolution diffusion. In *Forty-third International Conference on Machine Learning*, 2026.

Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70847–70860, 2023.

Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4338–4346, 2025.

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

Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*, 2024.

Ken M. Nakanishi. Scalable-Softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025. URL https://arxiv.org/abs/2501.19399.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.00071.

Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16893–16903. IEEE, 2025.

Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. *Advances in Neural Information Processing Systems*, 37:111131–111171, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

ZiDong Wang, Zeyu Lu, Di Huang, Cai Zhou, Wanli Ouyang, et al. Fitv2: Scalable and improved flexible vision transformer for diffusion model. *arXiv preprint arXiv:2410.13925*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23464–23473. IEEE, 2025.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4771–4780. IEEE, 2021.

Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. FreCaS: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024.

Bingxuan Zhao, Qing Zhou, Yu Wang, Chuang Yang, and Qi Wang. σ: Sigmoid modulation for ultra high resolution diffusion. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=47JZSOkw5C.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.

A Supporting Analysis

A.1 Cardinality Competition

For a nonempty proper subset $\mathcal{A}$ of b keys, define

$$\Delta_{\mathcal{A}}(g;K) = \log\left(\frac{1}{b}\sum_{j\in\mathcal{A}} e^{gs_j}\right) - \log\left(\frac{1}{K-b}\sum_{j\notin\mathcal{A}} e^{gs_j}\right). \tag{14}$$

The ratio of the selected and background partition sums is $M_{\mathcal{A}}/(1-M_{\mathcal{A}})$; factoring out their counts proves Equation (2) for arbitrary finite scores.

Proposition A.1 (Necessary scale for selective attention)**.** *If $s_{\max} - s_{\min} \le \bar{\delta}$ uniformly in K, with $\bar{\delta} > 0$, then retaining top-b mass at least $\rho \in (0,1)$ requires*

$$g\bar{\delta} \ge \operatorname{logit}\rho + \log\frac{K-b}{b}. \tag{15}$$

For $b(K) = \Theta(K^{\beta})$, $0 \le \beta < 1$, this implies $g = \Omega(\log K)$.

Proof. Bounding selected scores above and background scores below gives

$$T_b(g) \le \frac{be^{g\bar{\delta}}}{K-b+be^{g\bar{\delta}}}. \tag{16}$$

Rearrange $T_b(g) \ge \rho$ and use $\log((K-b)/b) = (1-\beta)\log K + O(1)$. □

This requires full score spread, not a percentile spread. Bounded query/key norms suffice: Cauchy–Schwarz gives $\bar{\delta} = 2B_Q B_K/\sqrt{d_h}$, and RoPE preserves norms. If $b/K \to \eta \in (0,1)$, the count penalty instead stays bounded; the result does not identify the semantic support of actual attention.

A.2 Exact Native Anchoring

The declared native reference fixes the coefficient in the logarithmic family of Equation (5). For $K_0 > 1$, native operator recovery for every score vector is equivalent to $c\log K_0 = 1$: necessity follows by comparing the probability ratio of two unequal scores, and sufficiency follows by substitution. Uniform score vectors alone cannot identify c. The uniqueness statement concerns this family with no additive intercept; allowing $g(K) = c\log K + d$ leaves one free parameter after imposing $g(K_0) = 1$.

Exact variational characterization. Let $\mathcal{P}_K = \{\mathbf{p} \in \mathbb{R}^K : p_j \ge 0,\ \sum_j p_j = 1\}$, $H(\mathbf{p}) = -\sum_j p_j \log p_j$ with $0\log 0 = 0$, and $\tau_K = \log K_0/\log K > 0$. CCA is equivalently the unique solution of

$$\mathbf{a}_K^{CCA} = \arg\max_{\mathbf{p}\in\mathcal{P}_K} F_K(\mathbf{p}), \qquad F_K(\mathbf{p}) = \langle \mathbf{p}, \mathbf{s}\rangle + \log K_0 \frac{H(\mathbf{p})}{\log K}. \tag{17}$$

The entropy term uses $H(\mathbf{p})/\log K \in [0,1]$, giving the same maximum regularization contribution $\log K_0$ at every count. At $K = K_0$, the objective reduces to $\langle \mathbf{p}, \mathbf{s}\rangle + H(\mathbf{p})$. To verify the equality, write $\mathbf{a} = \operatorname{softmax}(\mathbf{s}/\tau_K)$ and $Z = \sum_j \exp(s_j/\tau_K)$. For every $\mathbf{p} \in \mathcal{P}_K$,

$$F_K(\mathbf{a}) - F_K(\mathbf{p}) = \tau_K \operatorname{KL}(\mathbf{p}\|\mathbf{a}), \qquad F_K(\mathbf{a}) = \tau_K \log Z. \tag{18}$$

Nonnegativity of KL divergence proves optimality, with equality only at $\mathbf{p} = \mathbf{a}$. This is an exact characterization of the chosen regularizer for any finite score vector. It does not assert invariance of the achieved entropy or selected attention mass. The asymptotic Gaussian-score entropy relation used by Jin et al. (2023) is not needed for this identity.

A.3 Exact Compensation and Finite-Count Error

In the two-population model, the partition sums are $be^{g\delta}$ and $K-b$. Equating the target and native log-odds gives $g_*\delta - \log((K-b)/b) = \delta - \log((K_0-b)/b)$, proving Equation (4). Strict monotonicity in g proves uniqueness. Rearranging the native constraint $\rho = m_{K_0}(1)$ gives Equation (8); a bounded score spread alone does not fix this gap.

Factoring the background-count ratio gives the exact equality

$$\log \frac{K-b}{K_0-b} = \log \frac{K}{K_0} + \log \frac{1-b/K}{1-b/K_0}, \tag{19}$$

which proves Equation (9). Subtracting the native log-odds then proves Equation (10). For $K > K_0$ and $b > 0$, $r_K > 0$: even if $\delta = \log K_0$, the mass residual is $-r_K$, not zero. Exact finite-count mass matching therefore requires the stated residual condition in addition to native anchoring.

Proposition A.2 (Finite-count mass discrepancy)**.** *For $K \geq K_0 > b$ in the stated model, let $\lambda = \log K / \log K_0$ and $\xi = \operatorname{logit}\rho - \log b$. Then*

$$\begin{aligned} D_K &= (\lambda-1)|\xi| - \lambda \log(1-b/K_0), \\ |\lambda - g_*| \leq D_K/\delta, &\qquad |m_K(\lambda) - \rho| \leq D_K/4. \end{aligned} \tag{20}$$

Proof. The log-odds error is

$$e_K = (\lambda - g_*)\delta = (\lambda-1)[\xi + \log(1-b/K_0)] - \log \frac{1-b/K}{1-b/K_0}. \tag{21}$$

Since $0 \leq \log((1-b/K)/(1-b/K_0)) \leq -\log(1-b/K_0)$, the triangle inequality gives $|e_K| \leq D_K$. Divide by δ for the scale bound and use the sigmoid's $1/4$ Lipschitz constant for the mass bound. At $K = K_0$, both exact errors are zero. □

Native-pool limit. Fix b, $\rho \in (0,1)$, and an expansion bound $R > 1$, and set each native gap by Equation (8). Then $\delta = \log K_0 + O(1)$, $\lambda - 1 \leq \log R / \log K_0$, and $-\log(1-b/K_0) = O(b/K_0)$, so

$$\sup_{K_0 \leq K \leq RK_0} |m_K(\lambda) - \rho| = O\big((\log K_0)^{-1} + b/K_0\big) \longrightarrow 0. \tag{22}$$

In contrast, for $K/K_0 \to r > 1$, unit scale gives $m_K(1) \to \rho/[\rho + r(1-\rho)] < \rho$. This limit varies native pools; it does not guarantee unlimited extrapolation of a fixed backbone.

Finite-count accuracy also depends on native selectivity. At $K_0 = 4096$, $K = 65536$, and $b = 8$, CCA uses $\lambda = 1.3333$. For $\rho = 0.5$, exact compensation requires $g_* = 1.4449$ and the mass error is 0.1672; for $\rho = 0.9$, these are 1.3290 and 0.0033. These analytical examples show why small b/K_0 alone is insufficient.

A.4 Scope and Failure Boundaries

Outside the logarithmic family with no additive intercept, native identity admits every $g_\alpha = 1 + \alpha(\lambda - 1)$; it does not establish $\alpha = 1$ as an image-quality optimum. For fixed scores, sharpening increases top-b mass, while text mass can fall. Replicating every key–value pair r times leaves the unit-scale output unchanged, although entropy increases by $\log r$; CCA can change that output. Concentration alone is therefore not evidence of better aggregation. The numerical script checks the analytical identities, bounds, and counterexamples.

B Implementation and Evaluation

B.1 CCA Deployment

Counts follow Equation (12) after packing, using each row's effective mask and including unmasked padding. Joint attention scales both image and text query rows; fixed text-key pools retain unit scale. The native grid is an inference reference. Rows with either count at most one retain the backbone convention.

Separate the positional transform P from its effective temperature g_P: compare the full method (P, g_P), positional-only $(P, 1)$, and replacement (P, λ_q). Remove inherited factors from both Q/K amplitudes and the kernel. Native identity is relative to $(P, 1)$, and also to the full method when $g_P(K_{0,q}) = 1$.

Algorithm 1 CCA attention with the original positional mapping

Require: $\mathbf{Q}, \mathbf{K}$ after normalization and RoPE, with extra temperature factors removed; values $\mathbf{V}$, mask $\mathbf{M}$, visible counts $(K_q, K_{0,q})$.

for all query rows q **do**
  $\lambda_q \leftarrow 1$
  if $K_q > 1$ and $K_{0,q} > 1$ and $K_q \neq K_{0,q}$ **then**
    $\lambda_q \leftarrow \log K_q / \log K_{0,q}$
  end if
  $\widetilde{\mathbf{Q}}_q \leftarrow \lambda_q \mathbf{Q}_q$
end for
return ATTENTIONKERNEL$(\widetilde{\mathbf{Q}}, \mathbf{K}, \mathbf{V}, \mathbf{M}; \text{scale} = 1/\sqrt{d_h})$

The row operations can be vectorized without materializing attention matrices. Query scaling is equivalent to a logit multiplier for masking-only biases; finite score biases need separate treatment. Native identity and unchanged-pool identity follow from $\lambda_q = 1$ under their respective count conditions. These are properties of the defined operator in exact arithmetic. Dense-attention complexity is unchanged; no timing or memory advantage is claimed.

B.2 Evaluation and Score Aggregation

Model settings and matched comparisons follow Section 4. Qwen retains its own native grid, packing, masks, and sampling configuration with $\alpha = 1$. Output reuse is restricted to comparisons with matching checkpoint, prompt, initial latent, resolution, positional mapping, guidance, precision, sigmas, and effective temperature rule. Multiple summaries of those outputs are not independent replications.

ImageReward, CLIPScore, and aesthetic score use their released preprocessing with fixed evaluator versions. These full-image metrics do not measure all native-pixel detail. For per-image metric $y^m_{i,r}$, the paired estimate is

$$\widehat{\Delta}_r = \frac{1}{n} \sum_{i=1}^{n} \left(y^{\text{CCA}}_{i,r} - y^{\text{comparator}}_{i,r} \right). \tag{23}$$

Each comparison averages over its common prompt set. The primary reported count contrasts are against original mscale and the frozen constant at 4K, separately per benchmark; the locked 50-prompt 6K cohort is a transfer endpoint. Table 2 reports the means; its inline annotations use comparator minus CCA, the opposite sign to $\widehat{\Delta}_r$ above.

The 2K–4K results use the full benchmark lists, whereas the 6K results use the locked 50-prompt cohorts. These separate cohorts support within-resolution method comparisons; differences between their mean gains do not estimate a within-prompt change in benefit from 4K to 6K.

For Aesthetic-4K, FID uses 195 generated and 195 reference images with the same square center-crop/resize preprocessing and Inception-v3 evaluator across methods. The small reference set limits these distributional comparisons.

B.3 Temperature Selection

Both selectors use DY-YaRN-pos and 20 disjoint development prompts at 2K/3K. The independent searches cover $c \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$ and $\alpha \in \{0, 0.5, 0.75, 1, 1.1, 1.25, 1.5, 2\}$. Selection maximizes mean ImageReward with equal resolution weights, breaking ties in favor of the smaller value. The selected $c^\star = 1.2$ and $\alpha^\star = 1.1$ are frozen for both benchmarks and all test resolutions. Both controls retain unit scale at native counts and on fixed text-key pools.

The unfitted comparison $g = \sqrt{\log K / \log K_0}$ ports the relative dot-product multiplier of Jin et al. (2023), Equation (7), onto the same pools and positional mapping. Strength sweeps in Tables 5 and 6 describe test sensitivity and do not select CCA's default.

The DY-YaRN-pos 4K sweep uses 200/195 prompts. At 6K, comparisons of the default against $\alpha \in \{1.25, 2\}$ use 50 paired prompts per benchmark and reuse the default outputs. Table 5, panel B, reports $\Delta_\alpha = \overline{y}_{\alpha=1} - \overline{y}_\alpha$ and the analyzed counts. The observed 4K sweep motivated $\alpha = 1.25$, so these 6K comparisons are exploratory.

B.4 Human and Qualitative Evaluation

The preference study compares DY-YaRN-pos + CCA with full DY-YaRN on 50 prompts per benchmark, with three raters per pair. The presentation includes full images at matched display scales and unresized 512-/1024-pixel crops under fixed crop rules. Method identities are hidden, presentation sides are randomized, and ties are allowed. Raters assess the three criteria separately, with local detail as the primary criterion. Win/tie/loss responses from CCA's perspective are encoded as $1, 1/2, 0$. For totals W, T, L, the preference score is $100(W + T/2)/(W + T + L)$, with $W + T + L = 150$ for every criterion and benchmark. With three ratings per prompt, this equals averaging raters within prompts and then prompts within each benchmark, expressed as a percentage. Table 3 reports both the scores and the win/tie/loss totals, with scores rounded to one decimal place.

Figure 4 uses the shared prompt and manually selected 640×640 crops matched by scene content, not pixel alignment. From top to bottom, the crops show moored boats, foreground boat surfaces, and water reflections, with DY-YaRN on the left and CCA on the right. Colored boxes and lines link the equally scaled crops to their source regions. No enhancement is applied. Its script and JSON manifest record source hashes and crop coordinates separately from the preference study.

C Additional Results

C.1 Additional FLUX Results and Strength Sensitivity

Table 4 reports FLUX.1-Krea-dev results at 2048×3072 and 3072×2048, including unit-temperature positional components. The square 3K and 4K baseline comparisons appear in Table 1.

Table 4: **Rectangular resolution extrapolation on FLUX.1-Krea-dev.** We report CLIPScore (CLIP), ImageReward (IR), aesthetic score (Aesth), and, for Aesthetic-4K, FID. Higher is better except for FID. Gray rows apply CCA; within each matched baseline/CCA pair, bold marks the better result. Thin rules separate baseline families. YaRN and DY-YaRN retain their original `mscale` and are additional references.

Method	**2048 × 3072**							**3072 × 2048**						
	DrawBench			Aesthetic-4K				DrawBench			Aesthetic-4K			
	CLIP↑	IR↑	Aesth↑	CLIP↑	IR↑	Aesth↑	FID↓	CLIP↑	IR↑	Aesth↑	CLIP↑	IR↑	Aesth↑	FID↓
Vanilla	26.31	-0.37	5.06	28.32	0.22	6.01	**188.74**	26.85	0.25	5.24	28.45	1.08	**6.27**	144.92
Vanilla + CCA	**26.77**	**-0.21**	**5.11**	**28.42**	**0.34**	**6.12**	188.77	**26.99**	**0.33**	**5.36**	**28.91**	**1.17**	6.23	**144.21**
NTK	27.42	0.12	5.23	**28.84**	0.88	6.40	183.05	27.56	**0.60**	5.30	28.69	1.28	6.35	143.34
NTK + CCA	**27.51**	**0.22**	**5.31**	28.77	**0.94**	**6.51**	**182.91**	**27.66**	0.59	**5.42**	**28.72**	**1.33**	**6.51**	**142.76**
DY-NTK	27.65	0.39	5.45	**28.85**	**1.00**	6.47	178.11	27.72	0.69	5.46	28.83	1.32	6.64	142.99
DY-NTK + CCA	**27.95**	**0.41**	**5.51**	28.77	0.98	**6.86**	**176.68**	**28.40**	**0.77**	**5.58**	**29.37**	**1.41**	**6.75**	**141.87**
YaRN	28.02	0.43	5.54	28.98	0.90	6.51	181.71	28.07	0.71	5.39	28.84	1.35	6.59	143.73
YaRN-pos	27.14	0.39	5.18	27.85	0.84	6.12	189.25	26.93	0.65	5.08	27.61	1.26	6.20	153.89
YaRN-pos + CCA	**29.15**	**0.47**	**5.89**	**30.12**	**0.96**	**6.92**	**174.35**	**29.24**	**0.76**	**5.72**	**30.08**	**1.44**	**7.01**	**137.82**
DY-YaRN	28.18	0.62	5.61	29.15	1.07	6.52	181.68	28.45	0.90	5.59	28.92	1.39	6.61	143.44
DY-YaRN-pos	27.28	0.57	5.25	28.01	0.99	6.13	194.22	27.29	0.83	5.27	27.68	1.30	6.22	159.61
DY-YaRN-pos + CCA	**29.31**	**0.68**	**5.96**	**30.29**	**1.14**	**6.93**	**179.32**	**29.63**	**0.97**	**5.93**	**30.16**	**1.48**	**7.03**	**131.52**

Strength sweep on the primary positional mapping. Table 5, panel A, sweeps α in Equation (11) on DY-YaRN-pos. Here $\alpha = 0$ is unit temperature, $\alpha = 1$ is CCA, and $\alpha = 1.1$ is the slope selected on development prompts. Evaluation uses the full benchmark lists; unit, CCA, and the fitted-slope entries reuse the outputs in Table 2. At 4K, $\alpha = 1.25$ yields the highest reported ImageReward means of $0.13/1.20$, while $\alpha = 2$ reduces them to $0.04/1.10$. This sweep describes test sensitivity and does not select CCA's default.

Table 5: **Default strength and sensitivity.** ImageReward with DY-YaRN-pos. A reports the 4K sweep on 200/195 prompts. B reports the 6K comparisons on locked 50/50-prompt cohorts; Δ is default minus reference-strength score.

A. Reported 4K means

	Strength α							
Benchmark	0	0.5	0.75	1	1.1	1.25	1.5	2
DrawBench	0.04	0.07	0.09	0.11	0.12	0.13	0.10	0.04
Aesthetic-4K	0.94	1.07	1.14	1.18	1.19	1.20	1.19	1.10

B. 6K comparisons

	$\alpha = 1.25$		$\alpha = 2$	
Benchmark	n	Δ	n	Δ
DrawBench	50	-0.010	50	$+0.180$
Aesthetic-4K	50	-0.010	50	$+0.170$

In B, n is the number of paired prompts. Default outputs are reused from Table 2, which also reports comparisons with the development-selected $\alpha = 1.1$.

6K strength comparison. Table 5, panel B, compares the default with $\alpha \in \{1.25, 2\}$ on 50 paired prompts per benchmark. CCA trails $\alpha = 1.25$ by 0.01 on both benchmarks and exceeds $\alpha = 2$ by $0.18/0.17$ on DrawBench/Aesthetic-4K. These results show a small gain from the nearby strength and a decline under stronger compensation. Comparisons with the development-selected $\alpha^\star = 1.1$ remain in Table 2.

Strength sweep on DY-NTK. The DY-NTK sweep in Table 6 has different ImageReward optima across benchmarks ($\alpha = 1.5$ and 2), whereas $\alpha = 1$ has the highest CLIPScore on both.

Table 6: **Calibration-strength ablation at 4096^2.** Fixed DY-NTK positions on FLUX.1-Krea-dev. Unit and CCA (shaded) match Table 1; the remaining rows report the strength sweep. Values are point estimates.

	DrawBench			Aesthetic-4K		
Strength α	CLIP↑	IR↑	Aesth↑	CLIP↑	IR↑	Aesth↑
0 (Unit)	21.21	-1.34	4.13	27.75	0.70	6.31
0.5	21.71	-1.29	4.39	27.99	0.65	6.37
1 (CCA)	22.38	-1.23	4.38	28.98	0.75	6.69
1.5	22.33	-1.21	4.09	28.44	0.73	6.52
2	21.01	-1.48	4.21	28.87	0.77	6.52

C.2 Qwen-Image Resolution Sweep

Table 7 reports the complete ImageReward values for the Qwen-Image resolution sweep in Section 4.1. The two benchmarks share the same seven resolution settings and four methods; CCA uses each pool's Qwen-Image native reference without refitting.

C.3 Count Sensitivity and Entropy-Stratified Diagnostics

Pool definitions and reference anchors. In Table 8, actual joint-pool calibration gives the highest 4K ImageReward: gains over image-only counting and image-query-only scaling are $0.04/0.25$ and $0.03/0.29$, respectively, on DrawBench/Aesthetic-4K. All variants retain native identity and fixed text-key attention; native identity in Table 9 follows each declared count anchor.

Table 7: **Qwen-Image resolution extrapolation without refitting.** ImageReward (↑) on the complete DrawBench and Aesthetic-4K datasets. CCA uses Qwen-Image's own native key counts with $\alpha = 1$. Gray columns apply CCA; bold marks the higher score within each baseline/CCA pair at each resolution, with both scores bold in a tie.

A. DrawBench (200 prompts)

Resolution	NTK	NTK + CCA	YaRN-pos	YaRN-pos + CCA
2K	0.90	**0.94**	**0.92**	0.91
2.5K	0.62	**0.71**	0.66	**0.73**
3K	0.31	**0.47**	0.30	**0.44**
3.5K	0.14	**0.25**	0.25	**0.37**
4K	-0.21	**-0.04**	-0.19	**0.12**
4.5K	-0.45	**-0.20**	-0.42	**0.05**
5K	-0.75	**-0.34**	-0.61	**-0.13**

B. Aesthetic-4K (195 prompts)

Resolution	NTK	NTK + CCA	YaRN-pos	YaRN-pos + CCA
2K	1.34	**1.41**	1.32	**1.44**
2.5K	1.12	**1.29**	1.17	**1.35**
3K	0.71	**0.94**	0.82	**0.94**
3.5K	0.54	**0.76**	0.67	**0.81**
4K	0.11	**0.41**	0.29	**0.38**
4.5K	-0.13	**0.19**	0.12	**0.24**
5K	-0.41	**-0.14**	-0.36	**-0.05**

Table 8: **Pool definitions and query placement.** A 4K comparison on the full benchmark lists. Image-only counts replace $(N + L, N_0 + L)$ by (N, N_0); image-query-only scaling leaves joint text queries at unit scale.

Rule	DrawBench IR	Aesthetic-4K IR
Image count only	0.07	0.93
Actual joint pool	0.11	1.18
Image-query rows only	0.08	0.89

Table 9: **Sensitivity to the reference convention.** DY-NTK results at 4K on the same 50 prompts per benchmark. Only the count anchor changes; the positional mapping stays fixed. These subset means are distinct from the full-benchmark table.

Count anchor	DrawBench IR	Aesthetic-4K IR
768^2	-1.23	0.80
1024^2	-1.19	0.78
1280^2	-1.25	0.74

Table 10: **Relative output error by reference entropy.** Mean per-head error at $K/K_0 = 16$ (Figure 2(c)); lower is better.

Reference-entropy group	Unit	Frozen constant	CCA
Lower	0.57	0.45	0.31
Middle	0.61	0.43	0.29
Upper	0.50	0.44	0.43

Equal-group means round to $0.56/0.44/0.34$ for Unit/frozen constant/CCA. CCA lowers error in every group; its margin over the frozen constant is 0.01 in the upper-entropy group.

This measures agreement with a sampled pool, not a native-resolution or semantic target.

Fixed-state attention diagnostics. Figure 2 compares Unit, a frozen constant, and CCA at $K/K_0 \in \{1, 2, 4, 8, 16\}$ on shared frozen queries, keys, values, masks, and positions. The reference uses unit scale on the sampled K_0 pool. For query q and head h, relative output error is

$$e_{q,h} = \frac{\|\mathbf{o}_{q,h} - \mathbf{o}^{\mathrm{ref}}_{q,h}\|_2}{\|\mathbf{o}^{\mathrm{ref}}_{q,h}\|_2}, \tag{24}$$

where $\mathbf{o}_{q,h}$ is the value aggregation before head concatenation and output projection. Errors are averaged over common reference-norm-valid query/head pairs. Panels (a,b) report absolute top-32 mass and mean relative error; panel (c) and Table 10 report $16\times$ errors for fixed lower-, middle-, and upper-reference-entropy groups.