BETTER POSITIONS OR DIFFERENT VALUES? RETHINKING CONFIDENCE RANKING IN GUIDED MASKED IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Confidence ranking in masked image generation commits the positions whose candidate tokens are most probable, and its advantage over random positions depends on the sampling temperature. Hayakawa et al. (2026) explained it as implicit cooling on an unconditional image model. With assumptions about complete generation, we turn this cooling interpretation into a temperature rule for text-to-image models with classifier-free guidance. We develop the rule on one model and test it on three more. As the rule predicts, confidence ranking helps at temperatures above the best tested temperature for random positions but brings no significant improvement at that temperature. On MMaDA-8B, confidence ranking improves Fréchet Inception Distance (FID) by up to 51.8 at a high temperature, whereas the released confidence ranking worsens FID by 14.1 at the best tested temperature for random positions. After each sampler receives its own best tested temperature, confidence ranking has no significant FID advantage in three model families at released step counts. We extend the single-step analysis to many simultaneous commits and to position scores computed from probabilities at temperature one. This improves predictions of committed values and provides a starting point for tuning ranking strength. A controlled change to position scores shows that selection can weaken the guidance reflected in committed values even when candidate-value sampling stays fixed. These findings call for temperature-calibrated baselines and guidance controls when the scoring rule changes effective guidance.

1 INTRODUCTION

Masked image generators draw a candidate token at each masked position from the model's predicted distribution, or *proposal*, and commit a subset of positions at each step (Chang et al., 2022; 2023). Most released samplers prioritize positions whose sampled tokens have the highest model probabilities. This confidence ranking comes from MaskGIT and is used by aMUSEd, Meissonic and LaViDa-O (Patil et al., 2024; Bai et al., 2025; Li et al., 2026). Because these scores depend on sampled candidates, selecting positions also changes the distribution of committed values. However, the advantage of confidence ranking over selecting positions at random depends on sampling temperature: on MMaDA-8B, ranking improves FID by up to 51.8 at a high temperature, whereas the released ranking worsens FID by 14.1 at the best tested temperature for random positions. LaViDa-O shows a similar pattern in mean CLIP-H gains for image–text alignment (Figure 1). These observations raise three questions: How does temperature govern the gains from confidence ranking, and do they persist after each sampler chooses its own temperature? (Q1) How can we quantify the effect on committed token values under practical sampling conditions and use it to choose ranking strength? (Q2) Can position scoring alter guidance toward the prompt even when the guidance for sampling values is fixed? (Q3)

Hayakawa et al. (2026) show that confidence ranking favors high-probability values and thus acts as implicit cooling. On a single unconditional MAGE model (Li et al., 2023), lowering the temperature for random positions recovers most of its benefit. Their image experiments do not cover classifier-free guidance (CFG). Moreover, their quantitative description assumes sparse commits and identical probabilities for value sampling and position ranking. Released samplers can violate both assumptions, changing the concentration of committed values and potentially their guidance toward the prompt.

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

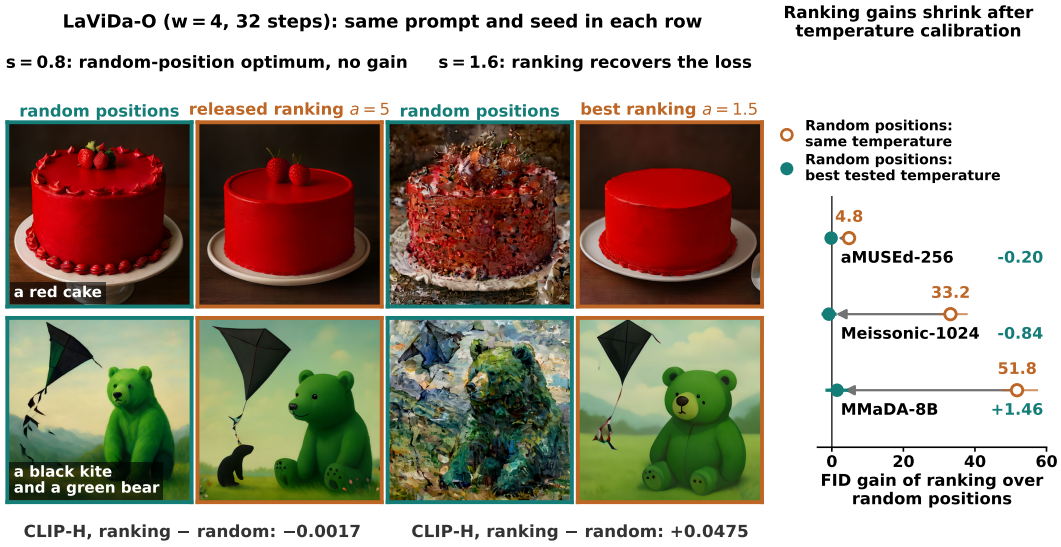


Figure 1: What confidence ranking gains depends on the temperature. *Left:* Results of LaViDa-O, with each row sharing the same prompt and seed. A smaller noise multiplier a gives stronger confidence ranking. At temperature s=0.8 (the best tested temperature for random positions), the released confidence ranking (a=5) gives no significant CLIP-H improvement. At s=1.6, stronger confidence ranking (a=1.5) recovers what random positions lose. The values below the images are mean CLIP-H gains (ranking minus random positions) across 241 prompts. *Right:* the FID gain of confidence ranking over random positions on three other models. Each row fixes the ranking setting (Table 13 in Appendix E). Open circles compare it with random positions at the same temperature, and filled dots with random positions at their own best tested temperature. The filled-dot gains are -0.20, -0.84 and $+1.46$, and none is significant.

To test the cooling explanation in complete generation, we formulate a *temperature rule* under assumptions motivated by our aMUSEd observations. Ranking can help only above the random-position optimum, with weaker ranking preferred as temperature approaches that optimum. Tests on three further CFG models include a prospective MMaDA-8B study (Yang et al., 2025). After each sampler receives its own best tested temperature, evaluations on prompts excluded from tuning find no significant FID advantage for confidence ranking in three model families at released step counts (Table 2).

Choosing ranking strength requires quantifying this cooling. As a step commits a larger fraction of positions, high-confidence candidates become more likely to be selected, and their selection probabilities approach one. The sparse *power form* of Hayakawa et al. (2026) misses this saturation and can overstate the concentration of committed values. Our *finite-gate law* accounts for saturation and distinct value-sampling and position-scoring probabilities. On aMUSEd states committing half the remaining positions, its mean total-variation error is about 0.006, versus up to 0.22 for the power form (Section 3). Its MMaDA strength recommendation improves on the power form's, although neither selects the best tested setting (Section 4.5).

To test whether these value changes explain the reversal, we keep positions random and draw values from the finite-gate law, with commit counts fixed. This reproduces MMaDA's temperature-dependent reversal without ranking positions (Section 4.4).

Beyond cooling, position scoring can also change the prompt's influence on committed values, or their *effective guidance*. Our COND probe leaves candidate sampling unchanged but scores positions with conditional probabilities without CFG amplification. Under the power form, it has the same effective temperature as confidence ranking but lower guidance. Sparse-selection tests with double-precision CFG support this relation (Section 5.1). In two aMUSEd samplers with random positions and identical temperatures, applying the predicted guidance schedule improves FID by 7.0 over keeping guidance fixed, showing an effect beyond cooling (Section 5.2).

By addressing these three questions, we make the following contributions:

- We test the temperature rule across four guided image models and reassess ranking gains after separate temperature calibration.
- We derive a finite-gate approximation that improves committed-value predictions and ranking-strength recommendations beyond the sparse power form.
- We derive the guidance change induced by conditional scoring and demonstrate its image-quality effect at matched temperatures.

Section 3 addresses Q2 through committed-value analysis, motivating the tests of Q1 in Section 4. Further tests ask whether changing values alone reproduces the reversal and whether the analysis guides ranking strength. Section 5 addresses Q3 by analyzing guidance changes.

2 SETUP

We compare confidence ranking with random position selection, both at a shared temperature and after each sampler chooses its own temperature.

Sampler. A run has T steps. At a given step, m positions remain masked. The current token sequence, including which positions remain masked, is the *sampling state*. At position i, the model predicts logits $z_{c,i}$ with the prompt and $z_{u,i}$ without it, both conditioned on this state. At guidance w and temperature s, classifier-free guidance (Ho & Salimans, 2022) gives the candidate proposal

$$q_{w,s,i} = \mathrm{softmax}\big([(1+w)z_{c,i} - w\, z_{u,i}]/s\big). \tag{1}$$

Setting $w = 0$ removes CFG amplification and gives the conditional distribution $q_{0,s,i}$, which still uses the prompt. aMUSEd and Meissonic use $g = 1 + w$ as their guidance parameter. Standard confidence ranking, CONF, commits k positions in three moves.

(S1) Draw a candidate value $X_i \sim q_{w,s,i}$ at every masked position.
(S2) Score its position using the probability of that candidate value:

$$S_i = \log q_{w,s,i}(X_i) + \tau G_i, \qquad \tau = r\alpha_t, \tag{2}$$

where G_i is standard Gumbel noise drawn independently at each position (Kool et al., 2019), and α_t is a factor that the release's schedule sets for step t.
(S3) Commit the k positions with the largest scores, retaining their candidate values.

Temperature and ranking strength. Temperature s controls the spread of the candidate proposal, while ranking noise τ perturbs position scores. In (S1)–(S3), $\tau \to 0$ gives greedy confidence ranking, while $\tau \to \infty$ gives *random positions*: uniformly selected positions retaining the candidates from (S1). The *ranking strength* is $1/\tau$, so larger noise means weaker ranking. The noise multiplier is r for aMUSEd, Meissonic and MMaDA, and a for LaViDa-O. Releases differ in α_t and their scoring probabilities: LaViDa-O samples values from $q_{w,s}$ but ranks positions with temperature-one probabilities (Table 5 in Appendix A).

Two comparisons. Confidence ranking uses candidate values to select positions and thus changes the committed-value distribution. Both comparisons keep guidance and step count fixed. A *same-temperature comparison* uses one temperature for both samplers. A *calibrated comparison* uses each sampler's best tested temperature. Scores that alter effective guidance additionally require a guidance control (Section 5).

Experimental settings. We evaluate aMUSEd, Meissonic and MMaDA on COCO captions (Lin et al., 2014) using FID and CLIP. LaViDa-O's alignment tests use GenEval prompts, as do the compositional-accuracy evaluations on the other models. Runs follow the released re-masking policy unless stated otherwise. After developing the rule on aMUSEd-256, we fixed settings, metrics, interval methods and decision rules before subsequent tests. MMaDA predictions preceded any ranking or temperature results.

Evaluation. We report FID_n from n generated images (lower is better), and CLIP and GenEval (higher is better). GenEval checks prompted objects, counts, colors and spatial relations. Positive gains favor the compared sampler. Unless stated otherwise, intervals are 95% paired-prompt bootstraps, and a difference is significant when its interval excludes zero.

3 How confidence ranking changes committed values

At a fixed sampling state, let $\widetilde{q}_i$ denote the distribution of X_i conditional on position i being committed. Selection based on the sampled value changes this distribution. For standard confidence scores, the likelihood tilt of Hayakawa et al. (2026) gives

$$\widetilde{q}_i(x) = \frac{q_{w,s,i}(x)\,\pi_i\big(q_{w,s,i}(x)\big)}{Z_i}, \tag{3}$$

Here $\pi_i(\rho)$ is the probability of committing position i given a candidate of proposal probability ρ, and Z_i is the probability of committing i. For sparse commits and $\tau > 0$, the *power form* of Hayakawa et al. (2026) approximates $\pi_i(\rho) \propto \rho^{1/\tau}$, giving

$$\widetilde{q}_i(x) \propto q_{w,s,i}(x)^{1+1/\tau}, \qquad s' = \frac{s\tau}{1+\tau}. \tag{4}$$

Thus $\widetilde{q}_i$ is the proposal $q_{w,s',i}$ at effective temperature $s' < s$ and unchanged guidance w (Appendix B).

A saturating approximation for denser commits. The sparse-commit assumption becomes less appropriate as generation progresses. For example, aMUSEd-256 commits 3 of 256 masked positions at its first step, but 33 of the remaining 66 at step 10 of 12, indexed from zero. As the committed fraction grows, high-probability candidates are increasingly likely to be retained. Their commit probabilities approach one, whereas the power form has no saturation.

To model saturation, write q_i for the value proposal and h_i for the probabilities used to score positions. For $1 \le k < m$, a candidate is retained if its score exceeds the kth largest score among the other $m-1$ positions. For $\tau > 0$, we approximate this random competing threshold by one fixed value $\log\theta$ shared across the state. This is most plausible when competing thresholds concentrate near a common value. For scoring probability $\rho = h_i(x)$ and standard Gumbel noise G, the resulting commit probability is

$$f(\rho) = \Pr(\log\rho + \tau G > \log\theta) = 1 - \exp\left[-\left(\frac{\rho}{\theta}\right)^{1/\tau}\right].$$

This fixed-threshold approximation is motivated by classical sampling without replacement (Rosén, 1972). We combine this gate with the likelihood tilt to obtain the *finite-gate law*, with θ set to give k commits in expectation:

$$\widetilde{q}_i(x) = \frac{q_i(x) f(h_i(x))}{Z_i}, \qquad Z_i = \sum_x q_i(x) f(h_i(x)), \qquad \sum_{i=1}^{m} Z_i = k. \tag{5}$$

Here Z_i approximates the probability of committing position i. When $(\rho/\theta)^{1/\tau}$ is small, $f(\rho) \approx (\rho/\theta)^{1/\tau}$ and the gate recovers the power form. As this quantity grows, $f(\rho)$ approaches one. If every position is committed, the value distribution is simply q_i.

Unlike the power form with standard confidence scores, the finite-gate law need not belong to the temperature–guidance family. Its effective temperature and guidance at each state and position are the parameters of the closest proposal in KL divergence, within the domain used in Appendix B.2:

$$(s_i', w_i') \in \operatorname*{arg\,min}_{s',\,w'} \mathrm{KL}(\widetilde{q}_i \,\|\, q_{w',s',i})\,.$$

We compare both approximations with Monte Carlo estimates of the committed-value distribution on 20 aMUSEd-256 states per step and guidance, with $s = 1$, $r = 6$ and $T = 12$ (Appendix B).

Averaged over states and positions, both approximations have total-variation error below 0.006 at steps 0 and 5. At step 10, the power-form error is 0.22 at $g = 4$ and 0.17 at $g = 10$, versus about 0.006 for the finite-gate law (Appendix B).

The power form's tendency to overstate concentration also follows from an entropy bound. For standard confidence scores, $h_i = q_i$, the entropy of a committed value lies between the power-form and proposal entropies when candidates are drawn independently (Proposition 4 in Appendix B).

Different probabilities for values and positions. The distinction between q_i and h_i also matters for sparse commits. LaViDa-O draws values from $q_{w,s}$ but scores positions with $q_{w,1}$, temperature-one probabilities (Table 5 in Appendix A). The power form gives

$$\widetilde{q}_i(x) \propto q_{w,s,i}(x)\, q_{w,1,i}(x)^{1/\tau}, \qquad s' = \frac{s\tau}{s+\tau}.$$

Guidance stays w, but the temperature relation differs from Eq. (4) unless $s = 1$ (Appendix B). Thus, implicit cooling depends on both the scoring probabilities and ranking noise.

4 Temperature predictions and full-generation tests

The entropy bound supports a cooling interpretation even when the power form overstates its magnitude. We test how this local cooling affects image quality over a complete run as the sampling temperature changes.

4.1 Temperature reverses the effect of confidence ranking

On aMUSEd at guidance $g = 4$ and 12 steps, random positions work best at the tested temperature $s = 0.8$. Confidence ranking helps at higher temperatures, but the preferred strength changes: as s falls from 1.25 to 1.0, the best tested noise multiplier rises from $r = 2$ to $r = 6$, meaning weaker ranking (Appendix C.1).

Assumptions behind the temperature rule. Weaker ranking near the optimum is consistent with less cooling being needed. To test the full-generation trends predicted by the sparse map in Eq. (4), we make two assumptions at fixed guidance:

1. Performance under the chosen metric improves as temperature approaches a single optimum and deteriorates beyond it.
2. The sparse cooling map approximately governs full-run performance, as if ranking were replaced by random positions at its predicted temperature.

Qualitative predictions of implicit cooling. Under these assumptions, ranking helps only when its cooling moves the sampler toward the optimum. Writing $s^\star$ for that temperature gives three predictions (Appendix B):

- **L1:** Above $s^\star$, a suitable ranking strength improves performance.
- **L2:** At or below $s^\star$, no ranking strength improves performance.
- **L3:** The optimal strength decreases as temperature approaches $s^\star$ from above.

We approximate $s^\star$ by the best tested random-position temperature.

We tested these predictions on Meissonic-1024, LaViDa-O and MMaDA-8B, comparing ranking with random positions at their best tested temperature and at a hotter one (Table 1). Guidance and step count stay fixed. For MMaDA, we specified the predictions before any temperature or ranking results were available. The same ranking noise, $r = 16$, changes from 6.1 FID worse than random positions at $s = 1.0$ to 16.4 better at $s = 1.4$. Meissonic shows the same reversal, while LaViDa's alignment results support gains at the hotter temperature. The aMUSEd sweep tests the strength trend L3, while the other models test the temperature endpoints (Appendix C).

4.2 No significant FID gain remains after calibration

The reversals above compare ranking with random positions at the same temperature. To assess gains after both samplers can adjust temperature, we compare them at their own best tested temperatures and tune ranking noise while fixing guidance, step count and metric (Table 2).

The calibrated comparisons answer Q1's second part: no significant FID gain remains for ranking at released step counts. This conclusion holds at both tested Meissonic guidance levels and after expanding the MMaDA search over temperature and ranking noise. On Meissonic at $g = 6$, random positions are significantly better by 0.83 FID. Retuning narrows the aMUSEd and Meissonic deficits (Appendix E.2), but calibration renders their large same-temperature gains nonsignificant.

Table 1: Temperature-dependent ranking gains across four models. Guidance and steps are fixed per model, and $s^\star$ is the best tested random-position temperature. The aMUSEd sweep tests the strength trend. The others compare ranking at $s^\star$ and a hotter temperature. See Appendix C for sweeps and criteria.

Model (guidance, steps)	Metric / $s^\star$	Temperatures	Result
aMUSEd-256 ($g = 4, 12$)	FID / 0.8	$1.25 \to 1.0$	Best r: $2 \to 6$, giving weaker ranking nearer $s^\star$
Meissonic-1024 ($g = 3, 64$)	FID / 1.0	$1.0,\ 1.4$	All three strengths hurt at 1.0 but help at 1.4
LaViDa-O ($w = 4, 32$)	CLIP-H / 0.8	$0.8,\ 1.6$	No significant released-ranking gain at 0.8. All five strengths help at 1.6
MMaDA-8B ($w = 3.5, 15$)	FID / 1.0	$1.0,\ 1.4$	No significant gain at 1.0. Three of four strengths help at 1.4

Table 2: Confidence ranking after separate temperature calibration. Guidance and step count are fixed within each comparison. Temperatures and ranking noise are selected on prompts excluded from evaluation. Positive FID gains favor confidence ranking. Intervals are 95%, except †97.5%. For *, nine ranking settings are searched on 200 of the preceding 1,400 prompts and evaluated on the other 1,200, so the two MMaDA evaluations overlap. Full comparisons, tuning records, and interval protocols are in Appendix E. aMUSEd uses eight independent sampling blocks on the same 2,000 captions, with two images per caption and FID_{4000} per block. Its gain is the block mean, with a Student-t interval.

Model (guidance, steps)	Prompts	Conf. (s, r)	Rand. s	FID gain [interval]
aMUSEd ($g = 4, 12$)	$8 \times 2{,}000$	$(1, 6)$	0.8	$-0.20\ [-0.53, 0.13]$
Meissonic ($g = 3, 64$)	700	$(1.4, 3)$	1	$-0.84\ [-2.43, 0.89]$
Meissonic ($g = 6, 64$)	5,000	$(1.2, 8)$	1	$-0.83\ [-1.22, -0.46]$
MMaDA† ($w = 3.5, 15$)	1,400	$(1.4, 16)$	1	$+0.13\ [-1.75, 1.38]$
MMaDA* ($w = 3.5, 15$)	1,200	$(1.2, 16)$	0.9	$-1.49\ [-2.85, 0.15]$

4.3 Random positions improve on the released samplers

The temperature rule also explains why the released ranking can hurt at its default settings. At fixed temperature, replacing confidence ranking with random positions improves FID by 14.1 on MMaDA and 13.9 on Meissonic (Appendices C.3 and E.3).

Further lowering guidance improves Meissonic's FID but reduces GenEval, whereas switching to random positions alone gives no significant GenEval change (Table 16 in Appendix E).

4.4 Changing the committed values alone reproduces the reversal

The temperature reversal in Section 4.1 could arise from changes in committed values, preferred positions or even commit counts. In the released MMaDA sampler, ties between bfloat16 (BF16) scores can change how many tokens a step commits. We repeated the comparison on 700 further prompts with exactly matched counts at every step. Confidence ranking remains harmful at $s = 1.0$ and helpful at $s = 1.4$.

To test whether this reversal requires ranking the positions, we keep the same counts but choose positions uniformly at random and draw each committed value from Eq. (5) instead of the proposal. This value-only sampler reproduces both directions. Its FID gains over confidence ranking are $+0.60\ [-1.62, 2.66]$ at $s = 1.0$ and $-1.61\ [-3.71, 0.67]$ at $s = 1.4$ (Appendix C.3). Changing the committed values is sufficient to reproduce the reversal in this experiment, without a detected FID difference from ranking.

4.5 The finite-gate law gives a starting point for ranking strength

If the value changes behind this reversal act mainly through cooling, ranking strength should bring the effective temperature close to the random-position optimum. We compare recommendations from the power form and finite-gate law.

For each noise multiplier, we average power-form temperatures over steps using commit counts. For the finite-gate law, we average committed-value entropy over positions using commit probabilities, then over steps using commit counts. Matching this entropy to the random-position entropy curve gives a temperature for the run. Each method recommends the multiplier whose temperature is closest to the random-position optimum. Entropy matching summarizes a whole run, whereas the KL projection in Section 3 fits each state and position.

On MMaDA-8B at $s = 1.7$, we compute both recommendations from random-position trajectories on 64 earlier prompts, using matched prompts and seeds across temperatures. We fix the recommendations before evaluation on 700 new prompts.

Table 3: Ranking-strength recommendations on MMaDA-8B. At $s = 1.7$, $w = 3.5$ and 15 steps, positive gains favor the listed sampler over its reference on 700 evaluation prompts. Intervals are 97.5%.

Sampler	Noise r	FID gain	Reference
Finite-gate recommendation	4	-1.22 $[-4.40, 1.21]$	Best tested ($r = 16$)
Power-form recommendation	256	-17.67 $[-20.97, -13.81]$	Best tested ($r = 16$)
Best tested	16	0	—
Finite-gate recommendation	4	$+16.44$ $[12.41, 19.50]$	Power form ($r = 256$)

The finite-gate law recommends $r = 4$ and the power form $r = 256$, while the best tested multiplier is $r = 16$ (Table 3). Neither selects the best multiplier. The finite-gate choice is 1.2 FID worse, with an interval that includes zero, whereas the power-form choice is 17.7 worse. A broader prediction made before evaluation correctly placed the best multiplier in $\{4, 16\}$. The finite-gate recommendation provides a starting point for the strength search posed in Q2, with random positions retained in the sweep.

5 How the scoring probability changes effective guidance

Beyond concentration, position scoring can also change the prompt's influence on committed values. For standard confidence scores, the power form scales the conditional and unconditional logits equally, preserving guidance. Scoring with conditional probabilities alone changes that balance by adding weight only to the conditional logits.

We study this effect with the controlled probe Cond. It draws candidate values from the same guided proposal $q_{w,s}$ as Conf, but changes only the log-probability term in (S2) to $\log q_{0,s,i}(X_i)$. Here $q_{0,s} = \mathrm{softmax}(z_c/s)$ uses the prompt without CFG amplification. Under the power form, the value retained at a fixed position has distribution

$$\begin{aligned}\widetilde{q}_{\mathrm{Cond}}(x) &\propto q_{w,s}(x)\, q_{0,s}(x)^{1/\tau} \\ &\propto \exp\left(\frac{(1 + w + 1/\tau)z_c(x) - w z_u(x)}{s}\right).\end{aligned}$$

Matching these logit coefficients with the proposal in Eq. (1) gives the effective parameters.

Proposition 1 (Effective guidance of the conditional score)**.** *Fix a sampling state, a position and $\tau > 0$. If the commit probability of* Cond *is proportional to $q_{0,s}(x)^{1/\tau}$, its committed value has distribution $q_{w',s'}$ with*

$$s' = \frac{s\tau}{1 + \tau}, \qquad w' = \frac{w\tau}{1 + \tau}. \tag{6}$$

Under the power form, Cond matches Conf's implicit cooling while lowering effective guidance (Appendix B).

5.1 Controlled checks of effective guidance

At fixed states, we fit one temperature and guidance shared across positions to tokens retained by Gumbel top-1 selection. Increasing the number of competing positions while keeping one winner

makes selection sparser. On saved MMaDA-8B logits at $w = 3.5$, $s = 1$ and $\tau = 0.3$, double-precision (FP64) CFG gives fitted guidance 0.8104, close to the prediction 0.8077, when each position has 64 independently sampled copies, all competing in one pool for a single winner. The ideal prediction remains unconfirmed under the release's BF16 arithmetic, where both rounding and estimation error matter (Table 7 in Appendix B.3).

Separate fits at each position test whether this coefficient mismatch affects prediction. We compare fits by held-out log likelihood: the average log probability assigned to committed tokens from independent draws not used for fitting. At 1% selection in BF16, fixing guidance at 0.81 and refitting temperature costs only 0.012 nats per token relative to fitting both parameters at each position (Appendix B.2).

For denser commits, projecting the finite-gate law onto the temperature–guidance family gives held-out likelihoods as high as or higher than fitting both parameters from repeated selections on MMaDA and aMUSEd states (Table 6 in Appendix B.2).

5.2 The predicted guidance change improves FID at matched temperatures

These fixed-state guidance changes may also affect complete generation. To separate guidance from temperature, we compare two random-position samplers in a 12-step aMUSEd-256 loop without re-masking. At each non-final step, RAND+T uses the temperature predicted for CONF and keeps the original guidance. RAND+T+G uses the same temperature but the guidance predicted for COND in Eq. (6). Both draw the remaining values from the original proposal at the final step, so only guidance differs between them. The source settings are $s = 1$ and $r = 6$, and each of 2,000 prompts has two draws. Here the release uses $g = 1 + w$ (Appendix D).

At original guidance $g = 10$, RAND+T+G improves FID by 7.0 over RAND+T (Table 4b). Since their temperatures agree at every step, this gain identifies a guidance effect beyond implicit cooling.

RAND+T+G and COND also outperform the best tested random-position temperature at this guidance. Ordinary guidance tuning remains a strong alternative: constant $g = 4$ beats the predicted schedule on the matched development comparison (Table 11 in Appendix D).

Table 4: Separating temperature and guidance effects on aMUSEd-256. Positive gains favor the first sampler: FID_{4000} of the second minus the first, or CLIP-B of the first minus the second. FID intervals are 97.5%, except 95% at $g = 4$. In (b), the temperature baseline is the best random-position sampler among five constant temperatures and RAND+T, at original guidance $g = 10$. Section 5.2 gives the shared generation settings.

(a) FID gains across guidance g

g	Conf over Rand+T	Cond over Conf
1.25	-1.74 $[-3.36, 0.25]$	not run
2	-0.93 $[-2.08, 0.19]$	-1.62 $[-2.61, -0.64]$
3	-0.35 $[-1.16, 0.56]$	-0.80 $[-1.57, 0.02]$
4	not run	$+0.04$ $[-0.58, 0.74]$
10	-1.48 $[-2.60, -0.42]$	$+5.03$ $[4.07, 6.23]$

(b) Guidance intervention at g=10

Comparison	FID gain	CLIP-B gain
Rand+T+G over Rand+T (same temperature)	$+7.02$ $[6.12, 8.22]$	$+0.0034$ $[0.0026, 0.0042]$
Rand+T+G over temperature baseline	$+6.30$ $[5.38, 7.39]$	$+0.0032$ $[0.0025, 0.0040]$
Cond over temperature baseline	$+2.83$ $[1.82, 4.06]$	$+0.0004$ $[-0.0004, 0.0013]$

5.3 Conditional scoring across guidance levels

The gain from lowering effective guidance should reverse when the original guidance falls below the optimum. On aMUSEd, COND follows this pattern: it is significantly worse than CONF at $g = 2$ and better at $g = 10$ (Table 4a). In the same experiments, RAND+T is never significantly worse than CONF at the tested guidance levels.

Conditional scoring also improves FID at high guidance on Meissonic. On MMaDA, the change between low and high guidance has the predicted direction but is not significant, and no tested

guidance lies above its FID optimum (Table 12 in Appendix D). These results reinforce the need to control guidance when the position score changes it.

6 Related work

Implicit temperature and softened ranking. Hayakawa et al. (2026) explain confidence ranking as implicit cooling and recover most of its MAGE benefit by adjusting the random-position temperature. Olausson et al. (2026) introduce separate token and position temperatures to improve diffusion-language-model diversity. Their tempered low-confidence rule uses the same weighted-without-replacement selection family as Gumbel ranking. The sparse cooling relation does not describe saturation when many positions are committed at once. Our finite-gate analysis accounts for this saturation and for distinct probabilities in value sampling and position scoring.

Confidence-based decoding in masked models. LLaDA retains high-confidence candidates (Nie et al., 2025), while Fast-dLLM uses confidence thresholds for adaptive parallel decoding (Wu et al., 2025). EB-Sampler constrains the entropy of parallel commits (Ben-Hamu et al., 2025), KLASS uses KL-based stability estimates (Kim et al., 2025), and ReMDM allows decoded tokens to be re-masked (Wang et al., 2025). ReMDM also tunes softmax temperatures separately in image experiments. On images, Halton (Besnier et al., 2025), LPD (Zhang et al., 2026) and attention-based UNCAGE (Kang et al., 2025) alter position selection. For candidate-based scores, changing the selection rule can alter both selected positions and committed values. Our matched-count and value-only controls examine this distinction in guided image generation.

Guidance and evaluation. In text generation, Zheng et al. (2025) trace sampler gains to a precision fault that lowers effective temperature, and Frković et al. (2026) show that re-masking benefits depend on decoding settings. For images, Xie et al. (2026) establish the importance of guidance calibration, while Rojas et al. (2025) study guidance normalization and schedules. These interventions change the candidate proposal, whereas position scoring can change effective guidance with the proposal fixed. We analyze this selection-induced effect and test it at matched temperatures.

7 Discussion and limitations

Together, the answers to Q1–Q3 show why temperature and guidance should be controlled when evaluating ranking. The finite-gate analysis improves committed-value predictions and strength recommendations, while value-only sampling reproduces the MMaDA reversal.

Scope of the conclusions. The fixed-state relations describe single committed values, with an approximate gate for top-k selection. Full-generation predictions require the temperature-rule assumptions. Value-only residuals mix position preference, dependence between commits and approximation error, preventing isolation of position effects (Appendix B.1).

The rule holds guidance fixed. Treating guidance as a proxy for distance to the temperature optimum gave incorrect predictions in all three strength sweeps (Appendix C.2).

Our calibrated comparisons use released step counts and do not establish sampler equivalence. Few-step runs lacked independently calibrated random-position temperatures. The guidance intervention lacks a matched-mean linear-decay control and does not establish superiority over joint temperature and guidance tuning.

These FID findings concern candidate-based ranking. Candidate-independent Halton and LPD have effects that vary with guidance and model (Appendix F). Other metrics need separate calibration (Appendix E). FID retains finite-sample bias (Chong & Forsyth, 2020), LaViDa-O is evaluated only for alignment, and human preference is unmeasured.

Reproducibility statement

We will publicly release the code, experiment configurations and analysis scripts upon acceptance. The appendix provides sampler settings, evaluation protocols and proofs.

AI use statement

Generative AI tools assisted with language editing, searching for relevant literature, debugging and executing analysis code, and checking mathematical arguments. The authors independently verified the cited literature, source code, experimental outputs, numerical results, and mathematical arguments. The authors take full responsibility for the final content of the paper.

References

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.08261.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking, 2025. URL https://arxiv.org/abs/2505.24857v1. Version 1, May 30, 2025.

Victor Besnier, Mickael Chen, David Hurych, Eduardo Valle, and Matthieu Cord. Halton scheduler for masked generative image transformer. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.17076.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022. URL https://arxiv.org/abs/2202.04200.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 4055–4075, 2023. URL https://proceedings.mlr.press/v202/chang23b.html.

Min Jin Chong and David Forsyth. Effectively unbiased FID and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1911.07023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. URL https://image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Stipe Frković, Metod Jazbec, Dan Zhang, Christian A. Naesseth, Ilija Bogunovic, and Eric Nalisnick. Re-evaluating confidence remasking in masked diffusion language models. *arXiv preprint arXiv:2606.12232*, 2026. URL https://arxiv.org/abs/2606.12232v1. Version 1, June 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Demystifying MaskGIT sampler and beyond: Adaptive order selection in masked diffusion. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2510.04525v2. Analyzed version: arXiv:2510.04525v2, October 2025.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. URL https://github.com/mlfoundations/open_clip.

Wonjun Kang, Byeongkeun Ahn, Minjae Lee, Kevin Galim, Seunghyuk Oh, Hyung Il Koo, and Nam Ik Cho. UNCAGE: Contrastive attention guidance for masked generative transformers in text-to-image generation. *arXiv preprint arXiv:2508.05399*, 2025. URL https://arxiv.org/abs/2508.05399v1. Version 1, August 2025.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. KLASS: KL-guided fast inference in masked diffusion models, 2025. URL https://arxiv.org/abs/2511.05664v2. Version 2, revised March 5, 2026.

Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508, 2019. URL https://proceedings.mlr.press/v97/kool19a.html.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. LaViDa-O: Elastic large masked diffusion models for unified multimodal understanding and generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/52d7e1f59be8a140bd5c3b0b5b8bdee2-Abstract-Conference.html.

Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. MAGE: MAsked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023. URL https://arxiv.org/abs/2211.09117.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. URL https://arxiv.org/abs/2502.09992v3.

Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in PyTorch, 2020. URL https://github.com/toshas/torch-fidelity.

Theo X. Olausson, Metod Jazbec, Xi Wang, Armando Solar-Lezama, Christian A. Naesseth, Stephan Mandt, and Eric Nalisnick. A tale of two temperatures: Simple, efficient, and diverse sampling from diffusion language models. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2604.09921v2.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://github.com/GaParmar/clean-fid.

Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. aMUSEd: An open MUSE reproduction. *arXiv preprint arXiv:2401.01808*, 2024. URL https://arxiv.org/abs/2401.01808.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Kevin Rojas, Ye He, Chieh-Hsin Lai, Yuhta Takida, Yuki Mitsufuji, and Molei Tao. Improving classifier-free guidance in masked diffusion: Low-dim theoretical insights with high-dim impact. *arXiv preprint arXiv:2507.08965*, 2025. URL https://arxiv.org/abs/2507.08965v2. Version 2, March 2026.

Bengt Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement, I and II. *The Annals of Mathematical Statistics*, 43(2–3):373–397, 748–776, 1972.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling, 2025. URL https://arxiv.org/abs/2503.00307v4. Version 4, revised February 7, 2026.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. MaskBit: Embedding-free image generation via bit tokens. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=NYe2JuN3v3.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding, 2025. URL https://arxiv.org/abs/2505.22618v3. Version 3, revised July 3, 2025.

Dian Xie, Shitong Shao, Lichen Bai, Zikai Zhou, Bojun Cheng, Shuo Yang, Jun Wu, and Zeke Xie. Guidance matters: Rethinking the evaluation pitfall for text-to-image generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/559a0998fab1d19b80e7e43a5852401c-Abstract-Conference.html.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/caa934a507a952698d54efb24845fc4b-Abstract-Conference.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1801.03924.

Zhuoyang Zhang, Luke J. Huang, Chengyue Wu, Shang Yang, Kelly Peng, Yao Lu, and Song Han. Locality-aware parallel decoding for efficient autoregressive image generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/02687e7b22abc64e651be8da74ec610e-Abstract-Conference.html.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.02908.

A SETUP

Models and samplers. Our samplers follow the release-specific noise and commit schedules in Table 5, using the listed step counts, guidance, temperature and ranking noise as defaults. Every model combines its logits as in Eq. (1), with the guidance parameter of the release.

In the native aMUSEd, Meissonic and MMaDA loops, committed tokens keep a large finite score, so strong ranking noise can re-mask them. Some studies instead disable re-masking to keep committed tokens fixed throughout generation.

In the released LaViDa-O code, the branch for zero ranking noise multiplies zero by $-\infty$ in bfloat16, which turned 48 ranking scores into NaN across eight steps of a replayed example. Our LaViDa-O sampler uses the log probability directly at zero noise and is otherwise identical to the release, and a first LaViDa-O study that ran the released branch is excluded from the evidence.

Table 5: Sampler settings of each release. N is the number of image tokens, T the number of steps, $t = 0, \dots, T-1$ the step, and n_t the number of positions still masked before step t. Released values are those of the public code. A dash marks a setting varied and specified in our studies. Position scores use either the sampling probabilities $q_{w,s}$ (Eqs. (1) and (2)) or probabilities at temperature 1, $q_{w,1}$. The latter gives $s' = s\tau/(s+\tau)$ under the power-form approximation (Section 3). Every release commits all remaining positions in the last step.

(a) Released parameters and position scores

Model	N	T	Guidance, temperature, noise	Score probabilities
aMUSEd-256 (Patil et al., 2024)	256	12	g=10, s=1, r=1	$q_{w,s}$
Meissonic-1024 (Bai et al., 2025)	4,096	64	g=9, s=1, r=1	$q_{w,s}$
MMaDA-8B (Yang et al., 2025)	1,024	15	w=3.5, s=1, r=1	$q_{w,s}$
LaViDa-O (Li et al., 2026)	4,096	32	w: —, s=0.8, a=5	$q_{w,1}$
H-MaskGIT-B, -S (Besnier et al., 2025)	256	—	w: —, s=1, τ_0=4.5	$q_{w,s}$
MaskBit (Weber et al., 2024)	512	—	w: —, s=1, r=8.2	$q_{w,1}$ (s=1 only)

(b) Step schedules

Model	Ranking noise τ_t	Positions kept masked after step t
aMUSEd, Meissonic	$2r\,(1 - t/(T-1))$	$\lfloor N \cos(\pi(t+1)/2T) \rfloor$, clipped to $[1, n_t - 1]$
MMaDA-8B	$r \prod_{j=1}^{t+1}(1 - j/T)$	Same cosine schedule and clipping as aMUSEd
LaViDa-O	$a\,\alpha_t$, where $\alpha_t = 0.5 + 0.5\cos(\pi t/2(T-1))$	Released shift schedule with shift 5: few tokens committed early and most late
H-MaskGIT-B, -S	$\tau_0\,(1 - t/(T-1))$	N minus $\max\{\lfloor N(1 - \frac{2}{\pi}\arccos\frac{t}{T-1})\rfloor, t+1\}$
MaskBit	$r\,(1 - (t+1)/T)$	$\lfloor N\frac{2}{\pi}\arccos\frac{t+1}{T}\rfloor$, clipped to $[1, n_t - 1]$

The aMUSEd-512 studies use the aMUSEd sampler with N=1,024. Our LaViDa-O runs use guidance 4 unless stated, and the factor α_t is read from the recorded runs. For H-MaskGIT, ranking noise denotes the first-step value τ_0. MaskBit has 256 positions with two codebook splits each.

Metrics. FID uses torch-fidelity's `inception-v3-compat` features and clean-fid's PIL bicubic resizing to 299^2 (Heusel et al., 2017; Obukhov et al., 2020; Parmar et al., 2022), with the pinned `weights-inception-2015-12-05-6726825d.pth` checkpoint. Each experiment keeps one fixed reference, which for text-to-image models is 30,000 COCO images (Lin et al., 2014), and no reference is rebuilt between the settings of a comparison. FID from a finite number of images keeps a sampling bias (Chong & Forsyth, 2020). CLIP-B and CLIP-L are cosine image–text alignments with ViT-B/32 and ViT-L/14 (Radford et al., 2021). CLIP-H uses OpenCLIP 3.3.0 ViT-H/14 (Ilharco et al., 2021) with the `laion2b_s32b_b79k` checkpoint and its validation transform.[1] Diversity is the cosine distance between the CLIP embeddings of repeated images of one prompt. Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) is measured between such repeats. GenEval uses its object detector on CPU with pinned thresholds (Ghosh et al., 2023), and its overall score weights six task means equally. All metrics read decoded, clipped and quantized PNG files. Generation ran in bf16 on RTX PRO 6000 Blackwell Max-Q and RTX 6000 Ada cards with torch 2.7.0 and CUDA 12.8, and scoring ran on the same cards with TF32 matrix multiplication disabled.

Intervals. Text-to-image studies use paired-prompt bootstraps that keep all images of a prompt together, with basic intervals for FID and percentile intervals for CLIP, except as specified here. The

[1] Frozen checkpoint: revision `1c2b8495b281`.

MMaDA value-only and expanded-tuning comparisons use percentile FID intervals (Appendix C.3), as do predicted contrasts and residuals in the new-temperature test (Table 10). H-MaskGIT studies use paired seeds with order-statistic intervals for the median, for example $[d_{(4)}, d_{(13)}]$ over 16 seeds at 97.87%. The eight-block aMUSEd evaluation uses Student-t_7 intervals over its blocks. Multiplicity adjustments are specified with each comparison. The exploratory guidance control uses unadjusted intervals (Table 11). No correction is made across studies, and inference conditions on the fixed prompts, reference and evaluator.

How the studies were run. The temperature rule was developed on aMUSEd-256. For every later study, the settings, the metric, the interval method and the decision rule were fixed before its images were generated, and the program that applies the rule was fixed before scoring. All changes to study plans preceded inspection of that study's results, except one centering of logits in a development readout of Appendix B.

Some studies reuse the prompts or images of others, and some choose and evaluate their settings on the same prompts (Tables 14 and 9).

Several earlier predictions outside the temperature rule also failed and are not used here. A sampler that draws values from the finite-gate law at random positions met three of its four criteria, and its extension to low ranking noise failed all three of its sets of predictions. A prediction that ordering information explains order effects was not supported, and the predicted responses to the step count failed (Appendix C.2).

B THEORY: STATEMENTS AND PROOFS

Setting. Fix one sampling state with m masked positions, whose earlier commits stay fixed. The proposals q_i have finite support, and the candidates $X_i \sim q_i$ and the standard Gumbel variables G_i are independent across positions. Scores are $S_i = \log q_i(X_i) + \tau G_i$ with a common $\tau > 0$, and the sampler commits the k highest scores. Continuous scores give zero probability to ties, and every statement concerns the support of each proposal.

Proposition 2 (Likelihood tilt, Hayakawa et al., 2026)**.** *Let $\pi_i(\rho)$ be the probability of committing position i when its candidate has proposal probability ρ. Conditional on committing i, its token value has distribution*

$$\widetilde{q}_i(x) = \frac{q_{w,s,i}(x)\,\pi_i\big(q_{w,s,i}(x)\big)}{Z_i},$$

where π_i is nondecreasing and $Z_i = \sum_x q_{w,s,i}(x)\,\pi_i(q_{w,s,i}(x))$ is the probability of committing i.

Proof of Proposition 2. Raising $q_i(X_i)$ raises S_i, so the probability $\pi_i(\rho)$ that position i is committed, given a candidate of probability ρ, is nondecreasing. Bayes' rule gives

$$\Pr(X_i = x \mid i \text{ committed}) = \frac{q_i(x)\Pr(i \text{ committed} \mid X_i = x)}{\Pr(i \text{ committed})} = \frac{q_i(x)\,\pi_i(q_i(x))}{Z_i},$$

so Z_i is the probability that position i is committed, and $\sum_i Z_i = k$. This tilt is the starting point of the implicit-temperature analysis of Hayakawa et al. (2026). Positions still differ through Z_i, and the committed tokens share the dependence that a subset of fixed size induces. □

The exact gate and the finite-gate law. For $1 \le k < m$, let $W_{i,k}$ be the kth largest score among the other $m-1$ positions. It is independent of X_i and G_i, and the standard Gumbel distribution function $F_G(z) = \exp(-e^{-z})$ gives

$$\pi_i(\rho; k) = \Pr(\log\rho + \tau G_i > W_{i,k}) = \mathbb{E}\Big[1 - \exp\Big\{-\exp\Big(\frac{\log\rho - W_{i,k}}{\tau}\Big)\Big\}\Big]. \tag{7}$$

For $k = m$ every candidate is committed. Replacing $W_{i,k}$ by a constant t gives the fixed-threshold gate

$$f_t(\rho) = 1 - e^{-(\rho/\theta)^{1/\tau}}, \qquad \theta = e^t, \tag{8}$$

and one threshold per state matches the expected number of commits,

$$\sum_i \sum_x q_i(x) f_t(q_i(x)) = k. \tag{9}$$

The left side falls continuously from m to 0 as t grows, so the threshold is unique for $0 < k < m$. Applying the likelihood tilt to this fixed-threshold gate gives the *finite-gate law*, $\widetilde{q}_i(x) \propto q_i(x) f_t(q_i(x))$. Conditional on the candidates, write $\rho_i = q_i(X_i)$. Committing the top k of $\log \rho_i + \tau G_i = \tau(\log \rho_i^{1/\tau} + G_i)$ is successive sampling without replacement with weights $\rho_i^{1/\tau}$ (Kool et al., 2019). Classical asymptotic analyses of successive sampling (Rosén, 1972) motivate this fixed-threshold approximation. Here we use the gate derived in Eq. (8) and assess its error on actual sampling states.

When $(\rho/\theta)^{1/\tau}$ is small, $f_t(\rho) \approx (\rho/\theta)^{1/\tau}$, recovering the power form $q^{1+1/\tau}$ of Proposition 3. As $\tau \to 0$ at fixed k, the gate approaches a threshold rule, with fractional inclusion at the limiting threshold when needed to satisfy Eq. (9). As $\tau \to \infty$ at a fixed committed fraction, the gate becomes that fraction, consistent with uniform position selection. We give no error bound for the fixed-threshold approximation at finite m.

We check the approximation on aMUSEd-256 confidence trajectories with $T = 12$, $s = 1$, $r = 6$ and $g \in \{4, 10\}$. For each guidance, we capture 20 states at steps 0, 5 and 10, indexed from zero. The committed fractions are $3/256$, $22/203$ and $33/66$, with $\tau = 12$, 6.545 and 1.091. At each state, we evaluate eight fixed, evenly spaced indices among the masked positions. We draw candidates and Gumbel noise 2,048 times to estimate each position's competing-score threshold distribution, then integrate over that position's full vocabulary and its own Gumbel noise. The result estimates its token-value distribution conditional on being committed. We compute total variation from each approximation and average over the eight positions and 20 states within each guidance and step.

At steps 0 and 5, both approximations have mean error below 0.006. At step 10, the power-form errors are 0.2176 and 0.1739 at $g = 4$ and 10, and the finite-gate errors are 0.00572 and 0.00557. The corresponding split-half Monte Carlo discrepancies are 0.00099 and 0.00092. These compare estimates from the two halves of the 2,048 draws and are diagnostics, not error bounds.

Relation to an existing bound. Theorem 7 of Hayakawa et al. (2026) bounds the total variation between the joint distributions of ordered committed positions and their values under confidence ranking and the moment sampler. The bound is $5\sqrt{x}\,(1 + \sqrt{\log^+(1/x)})$, where $x = k^2|S|^{1/\tau}/N$, $|S|$ is the vocabulary size and N the number of masked positions. At the first step of aMUSEd-256, $k = 3$, $N = 256$, $|S| = 8192$ and $\tau = 12$ give $x = 0.0745$ and a bound of 3.56, exceeding the maximum TV distance of one. The bound is uninformative at this setting. The empirical TV errors above concern each position's token distribution conditional on that position being committed, so they are not directly comparable to this joint-distribution bound.

Proposition 3 (Implicit temperature, Hayakawa et al., 2026)**.** *At a fixed sampling state and position with $\tau > 0$, a power-form commit probability gives the committed-value distribution $\widetilde{q} = q_{w,s'}$ with*

$$s' = \frac{s\tau}{1+\tau}.$$

The guidance w is unchanged.

Proof of Propositions 3 and 1. Fix the state and position, and suppress the position index. Write $q_{0,s} \propto e^{z_c/s}$ and $q_{w,s} \propto e^{[(1+w)z_c - wz_u]/s}$. Under the stated gate assumption, Bayes' rule gives $\widetilde{q}(x) \propto q_{w,s}(x) e^{a(x)/\tau}$. For CONF, $a = \log q_{w,s}$, so

$$\widetilde{q}_{\mathrm{Conf}}(x) \propto q_{w,s}(x)^{1+1/\tau} \propto \exp\!\left(\frac{(1+w)z_c(x) - wz_u(x)}{s'}\right), \qquad s' = \frac{s\tau}{1+\tau}.$$

Thus guidance remains w. For COND, $a = \log q_{0,s}$, so

$$\widetilde{q}_{\mathrm{Cond}}(x) \propto q_{w,s}(x) q_{0,s}(x)^{1/\tau} \propto \exp\!\left(\frac{(1+w+1/\tau)z_c(x) - wz_u(x)}{s}\right).$$

Matching the coefficients with $[(1+w')z_c - w'z_u]/s'$ gives $s' = s\tau/(1+\tau)$ and $w' = w\tau/(1+\tau)$. The normalizing constants depend on the position but not on the token value. When scoring uses probabilities at temperature 1, as in LaViDa-O, the same argument gives $s' = s\tau/(s+\tau)$ for both rules. Guidance remains w for CONF and becomes $w\tau/(s+\tau)$ for COND. These formulas agree with Eqs. (4) and (6) at $s = 1$. □

The normalizing constants are where the choice of positions enters. For either score a_i, the unnormalized mass is $M_i = \sum_x q_{w,s,i}(x) e^{a_i(x)/\tau}$. Two rules with identical normalized value distributions can still have different masses across positions, hence different position preferences. Matching per-position values therefore does not match the distribution of positions.

Proposition 4 (Entropy bounds for committed values)**.** *Let* $q_i^{\mathrm{pow}}(x) \propto q_i(x)^{1+1/\tau}$. *Both the exact gate in Eq.* (7) *and its fixed-threshold approximation satisfy*

$$H(q_i^{\mathrm{pow}}) \le H(\widetilde{q}_i) \le H(q_i). \tag{10}$$

Proof of Proposition 4. Let $b = 1/\tau$, write $q = (q_1, \dots, q_d)$ in decreasing order, and set $f(\rho) = \mathbb{E}[1 - e^{-A\rho^b}]$ for a positive random variable A independent of this position's candidate. Define $p_j \propto q_j f(q_j)$ and $v_j \propto q_j^{1+b}$. First, f is increasing, so p_j/q_j decreases in j. Two normalized distributions whose probability ratio decreases cross once, and the one with the larger leading ratio puts more mass on every leading prefix. Thus $p \succeq q$, where $\succeq$ denotes majorization. Second,

$$\frac{f(\rho)}{\rho^b} = \mathbb{E}\left[A \frac{1 - e^{-A\rho^b}}{A\rho^b}\right]$$

decreases in ρ, because $(1 - e^{-z})/z$ decreases for $z > 0$. Consequently $v_j/p_j \propto q_j^b / f(q_j)$ decreases in j, which gives $v \succeq p$. Shannon entropy is Schur-concave, so $H(v) \le H(p) \le H(q)$. The proof covers a fixed A and the exact leave-one-out threshold. It also covers $k = m$ by taking $p = q$. □

The bound concerns the entropy of one committed token. Image quality and diversity are properties of whole runs, which the experiments measure separately.

The conditional prediction for the best ranking noise. Under the assumptions below, the power-form map gives

$$\tau^\star(w, s) = \frac{s^\star(w)}{s - s^\star(w)} \quad \text{if } s > s^\star(w), \qquad \tau^\star = \infty \quad \text{otherwise.} \tag{11}$$

Let $F(w, s)$ be a loss of random positions, such as FID or negative alignment, and assume that at each w it falls strictly up to $s^\star(w)$ and rises strictly afterwards. Assume also that confidence at (w, s, τ) attains $F(w, s\tau/(1+\tau))$, which applies the local identity to the whole run. As τ falls from ∞ to 0, the temperature $s' = s\tau/(1+\tau)$ falls from s to 0. If $s \le s^\star(w)$, every finite τ moves s' further below the optimum, so $\tau^\star = \infty$. If $s > s^\star(w)$, the optimum is reached at $\tau^\star = s^\star(w)/(s - s^\star(w))$. This value rises with $s^\star(w)$ and falls with s. Along guidance at a fixed s, the change from a finite to an infinite $\tau^\star$ lies where $s^\star(w) = s$, and it coincides with the guidance optimum only if $\partial_s F = 0$ there. These assumptions are strong, and Appendix D shows that the whole-run assumption holds only approximately, so the experiments test the qualitative consequences directly.

B.1 Joint commits and the choice of positions

Fix a state h with m unknown positions, and let the sampler commit $1 \le k \le m$ of them. Write $K_h(S, x_S)$ for the joint law of the committed set S and its values, $a_h(S) = \sum_{x_S} K_h(S, x_S)$ for the law of the set, and $b_h(x_S \mid S) = K_h(S, x_S)/a_h(S)$ for the values given the set. Random positions with the same proposals give $K_h^I(S, x_S) = U(S)\, q_S(x_S)$, with $U(S) = \binom{m}{k}^{-1}$ and $q_S = \prod_{i \in S} q_i$. The support of K_h^I contains that of K_h, and the chain rule of the KL divergence gives

$$\mathrm{KL}(K_h \| K_h^I) = \underbrace{\mathrm{KL}(a_h \| U)}_{\text{positions}} + \underbrace{\mathbb{E}_{S \sim a_h} \mathrm{KL}(b_h(\cdot \mid S) \| q_S)}_{\text{kept values}}. \tag{12}$$

For a fixed selected set S with $a_h(S) > 0$, $\mathrm{KL}(b_h(\cdot \mid S) \| q_S)$ splits into dependence among the committed values and the sum of their marginal divergences from q_i. These marginals condition on the entire set S, whereas Proposition 2 conditions only on i being committed and takes a probability-weighted mixture over selected sets containing i. Matching the latter distributions does not generally match the conditional marginals or their dependence.

As the ranking noise grows, the scores divided by τ converge to the Gumbel variables alone. The committed set then becomes uniform and independent of the candidates, and K_h converges to K_h^I in total variation.

Equal values need not mean equal positions. A measurement on real states shows different position inclusion marginals even when the power-form value laws nearly agree. Conditional-score ranking and confidence ranking at the mapped guidance of Appendix D share the same power-form value law. On 120 shared aMUSEd-256 states, with 2,048 competing draws per state, their power-form value distributions agree to between 0.0005 and 0.0018 total variation. Their values under finite top-k selection separate as the committed fraction grows, and they differ by 0.08 at guidance 4 and by 0.17 at guidance 10 when half the remaining tokens are committed. For positions, we compare normalized inclusion probabilities, $p_i = \Pr(i \in S)/k$, using Rao–Blackwell estimates normalized by their total mass. Their TV distance ranges from 0.011 to 0.030 at step 0 and from 0.031 to 0.116 at step 10. These distances concern position-index marginals, not the full selected-set distribution $a_h(S)$ in Eq. (12).

B.2 Committed values on real sampling states

To test these value distributions on real sampling states, we draw many committed tokens from the actual top-k selection, fit candidate laws on one part of the draws, and score them on the other part in nats per committed token. The baseline fits one temperature per position. The power form fixes guidance at the value w' that Proposition 1 predicts and refits only the temperature. The free fit fits temperature and guidance together. The finite-gate projections use no fitted samples. At each position, they minimize $\mathrm{KL}(\widetilde{q}\|p_{a,b})$ over the family $p_{a,b} \propto \exp(az + bd)$, where z is the guided logit and $d = \log q_c - \log q_u$ is the guidance direction. The implied temperature is $1/a$ and the implied guidance is $w + b/a$. At an interior optimum, the first moments match:

$$\mathbb{E}_{p_{a,b}}[z] = \mathbb{E}_{\widetilde{q}}[z], \qquad \mathbb{E}_{p_{a,b}}[d] = \mathbb{E}_{\widetilde{q}}[d]. \tag{13}$$

The free fit maximizes likelihood using the means of the training draws in place of the finite-gate moments. Both optimizations use the same bounded domain, $1/256 \le a \le 256$ and $-3.5a \le b \le 12.5a$. At a boundary optimum, the constrained optimality conditions replace moment matching.

For this comparison, we captured MMaDA-8B states for 32 new prompts at checkpoint 8 of 15, at the ranking noise $\tau = 0.3$ that r=16 uses there, and imposed four committed fractions (Table 6). The criteria for the power form were fixed before these states were captured, and those for the finite-gate projections before any readout. The aMUSEd-256 states come from step 5 of 20 development prompts, and their criteria were fixed before these states were read.

The power form recovers most of the free fit's gain at sparse steps and loses to the free fit at 33%. The finite-gate projections, with nothing fitted, match or beat the free fit at every fraction on both models. Where 12% or more of the positions commit, their guidance agrees with the free fit's to within 0.011. A law with nothing fitted can beat the free fit because the free fit is limited by its training draws. At 1% the finite-gate law itself gains 0.48 nats, against 0.25 for its two implied settings and 0.11 for the free fit.

In this native BF16 experiment, the predicted guidance is not confirmed by the per-position fit. With the standard 2,048 training draws, guidance is weakly identified at 1%, where its median per-position standard error is 1.5. The larger comparison uses 32,768 training draws and 2,048 independent test draws per state for both the free fit and the fit with guidance fixed at 0.81. The free fit settles at 2.04 $[1.72, 2.35]$ instead of the predicted 0.81, and a criterion fixed in advance, that the fit lie closer to the prediction than to the original 3.5, is not met. Fixing guidance at 0.81 and refitting temperature nonetheless costs only 0.012 nats per committed token, because in sparse states the likelihood of the committed tokens is nearly flat along guidance. For confidence, whose predicted guidance is the original 3.5, the free fit gives 3.48.

Earlier readouts on development states of MMaDA-8B led to this design. At the native noise and commit counts, the power form with both settings fixed predicted committed tokens worse than the per-position temperature fit, by 2.21 nats for the conditional score and by 2.32 for confidence, while the finite-gate law predicted them better, by 0.303 $[0.285, 0.320]$ and 0.008 $[0.006, 0.010]$. At $\tau = 0.3$ with the native commit counts, the finite-gate law beat the temperature fit for the conditional score by 0.309, 0.318 and 0.285 nats at checkpoints 3, 8 and 12, and by at most 0.05 for confidence. A first test of the predicted guidance on these states recovered well under half of the finite-gate law's advantage and failed its criteria. Its logits had been centered after a first readout, and every cell was rerun.

Table 6: Predicting committed tokens on real sampling states. Each column is an imposed committed fraction, at ranking noise $\tau = 0.3$. (a) Held-out log-score gains over a per-position temperature fit, in nats per committed token. The power-form guidance is fixed at its prediction and temperature is fitted. The free fit estimates both parameters. The finite-gate projection takes both from the law without fitting the sampled tokens. (b) Guidance parameters from those two-parameter descriptions. (c) The power-form gain divided by the free-fit gain. Ratios can be negative or exceed one. MMaDA uses checkpoint 8 and 32 new prompts, aMUSEd uses step 5 and 20 development prompts. Intervals are 95%, clustered by prompt.

	Committed fraction			
Model and method	1%	4%	12%	33%
(a) Log-score gain (nats/token)				
MMaDA-8B				
Power-form guidance 0.81	$+0.108$ $[0.096, 0.120]$	$+0.109$ $[0.097, 0.121]$	$+0.074$ $[0.062, 0.084]$	-0.043 $[-0.062, -0.026]$
Free fit	$+0.114$ $[0.098, 0.130]$	$+0.128$ $[0.114, 0.143]$	$+0.127$ $[0.116, 0.138]$	$+0.098$ $[0.089, 0.107]$
Finite-gate projection	$+0.252$ $[0.224, 0.284]$	$+0.192$ $[0.176, 0.211]$	$+0.151$ $[0.140, 0.163]$	$+0.103$ $[0.094, 0.112]$
aMUSEd-256				
Free fit	$+0.244$ $[0.165, 0.328]$	$+0.225$ $[0.160, 0.291]$	$+0.194$ $[0.168, 0.224]$	$+0.113$ $[0.101, 0.126]$
Finite-gate projection	$+0.545$ $[0.429, 0.671]$	$+0.332$ $[0.276, 0.390]$	$+0.213$ $[0.187, 0.241]$	$+0.115$ $[0.103, 0.128]$
(b) Guidance parameters				
MMaDA, free fit	2.00	1.45	1.51	2.21
MMaDA, projection	1.14	1.26	1.51	2.20
aMUSEd, free fit	0.91	1.11	1.45	1.89
aMUSEd, projection	0.75	1.05	1.44	1.91
(c) Power-form / free-fit gain ratio				
MMaDA, original decision*	0.90 $[0.81, 0.99]$	0.80 $[0.76, 0.85]$	0.56 $[0.48, 0.64]$	-0.44 $[-0.64, -0.26]$
aMUSEd	1.58 $[1.28, 2.11]$	1.08 $[0.96, 1.28]$	0.66 $[0.60, 0.73]$	-0.11 $[-0.24, 0.01]$

*Against the free fit of the first readout, on which the criteria were read. Every other entry comes from solving both fits exactly with one solver, which changes no conclusion. Against the exactly solved free fit, the MMaDA gain ratios are 0.95, 0.85, 0.58 and -0.44, and the first readout gave aMUSEd gain ratios of 1.26, 0.88, 0.65 and -0.11. The criteria asked the power form for a gain ratio of at least one half at 1% and 4% and a significant loss to the free fit at 33%, and they asked the finite-gate projection for a gain ratio of at least 0.9 at 1% and 4% and of one half at 33%. All hold on both models. With sixteen times the training draws, the free fit's guidance is 2.04 $[1.72, 2.35]$ at 1% and 1.56 $[1.38, 1.78]$ at 4%. On MMaDA's released schedule, steps that commit at most 12% of their masked tokens make 46% of all commits.

B.3 Controlled checks of the guidance coefficient

These experiments test the guidance relation of Proposition 1 using categorical proposals followed by Gumbel top-1 selection. The main samples are not drawn from the predicted power-form distribution. For Cond at $s = 1$, $w = 3.5$ and $\tau = 0.3$, the predicted coefficient is $w' = 21/26 = 0.807692$. Each synthetic world, or each captured state in the replay, shares one fitted guidance and one inverse temperature across positions.

Synthetic positions. We generate 24 worlds, eight each with independent, low-rank and heteroskedastic logits, using a vocabulary of 256 tokens. Within each world, four nested sets contain 256, 1,024, 4,096 and 16,384 candidate positions. Each draw retains one winner. Each set uses 8,192 training draws and 8,192 independent test draws, with common candidate and Gumbel randomness across set sizes. The fitted coefficients are 0.8400, 0.8163, 0.8081 and 0.8021. The last has a 95% interval of $[0.7958, 0.8085]$. The mean within-world absolute error falls by 0.0195 $[0.0108, 0.0291]$ between the endpoints. Intervals bootstrap worlds within their three generating families. The held-out cost of fixing the theoretical parameters also decreases. Samples from the exact power-form distri-

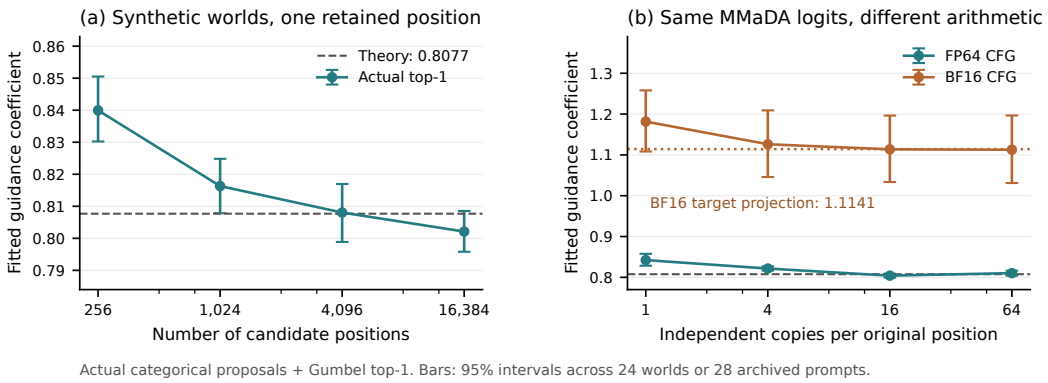


Figure 2: Guidance coefficients under controlled sparse selection. Both panels use actual categorical proposals and Gumbel top-1 with $w = 3.5$, $\tau = 0.3$ and the conditional score. (a) Increasing the number of candidate positions across 24 synthetic worlds. (b) Artificially copying positions from identical captured MMaDA-8B logits under two CFG arithmetic paths, summarized over 28 archived prompts. Reference lines mark the theoretical coefficient and the projection of the BF16 power-form target onto the shared-parameter family. Bars are 95% world- or prompt-bootstrap intervals. Each fit shares guidance and temperature across positions within one world or state. These controlled replays do not generate new image runs.

bution serve only as a separately labelled control. An additional 24-world study varies guidance, ranking noise and the number of retained positions. Its coefficient error, normalized by the input guidance, rises from 0.0154 at one retained position to 0.1059 at 64.

Identical model logits, two arithmetic paths. We use two archived MMaDA-8B states for each of 32 prompts. Four prompts first checked reproducibility, and the curves and intervals use the other 28. These are additional archived prompts for this replay, not newly generated model runs. We copy each position $R \in \{1, 4, 16, 64\}$ times, draw each copy's candidate independently, and retain one winner. Every state uses 65,536 training draws and 16,384 independent test draws. Increasing R increases the competitors, not the retained sample count. We compare the original BF16 CFG arithmetic with FP64 CFG computed from the same captured conditional and unconditional logits. We average the two states within each prompt before a paired-prompt bootstrap.

The FP64 coefficient changes from 0.8422 at $R = 1$ to $0.8104\ [0.8042, 0.8164]$ at $R = 64$, close to the theoretical 0.8077. The BF16 coefficient changes from 1.1817 to $1.1126\ [1.0311, 1.1966]$, close to 1.1141, the projection of the BF16 power-form target. This projection uses the same two-parameter family, weighting each position by its limiting selection mass $\sum_x q_i(x)h_i(x)^{1/\tau}$, where h_i is the conditional-score probability. The endpoint minus this target is $-0.0015\ [-0.0088, 0.0056]$. Both paths therefore approach their respective targets in this controlled replay. Recomputing CFG does not restore information lost in the model's BF16 forward pass.

Table 7: Guidance estimates and estimation bias under different precision. MMaDA-8B conditional scores, $w = 3.5$ and $\tau = 0.3$, with ideal prediction $w' = 0.8077$. Shared fits estimate one temperature and guidance per state, whereas per-position fits estimate them separately. Intervals are 95% prompt bootstraps. The bias check uses direct target samples instead of position selection (Appendices B.2 and B.3).

Samples and CFG arithmetic	Fitted parameters	Estimate / reference
Guidance fitted to selected values		
Sparsest top-1 replay, FP64	Shared across positions	$0.8104\ [0.8042, 0.8164]$
Sparse top-1 selection, BF16	Shared across positions	Approaches target projection 1.11
1% selected, BF16	Separate at each position	$2.04\ [1.72, 2.35]$
Estimation bias under direct sampling		
Direct BF16 power-form samples	Separate at each position	$+0.46$ bias against own projection

Why arithmetic changes the map. At $s = 1$, write $d = z_c - z_u$, $z = z_c + wd$, and let implemented CFG return $\bar{z} = z + \varepsilon$. Let $\eta = \log q_{\text{native}} - \log \text{softmax}(\bar{z})$ and $\xi = \log h_{\text{native}} - \log \text{softmax}(z_c)$ denote additional probability and score rounding. With $\beta = (1+\tau)/\tau$ and $w' = w\tau/(1+\tau)$, the BF16 power-form target obeys

$$\log q_{\text{native}}(x) + \frac{\log h_{\text{native}}(x)}{\tau} = \beta\big[\bar{z}(x) + (w' - w)d(x)\big] - \frac{\varepsilon(x)}{\tau} + \eta(x) + \frac{\xi(x)}{\tau} + C, \quad (14)$$

where C is constant across token values at one position. The token-dependent rounding term need not be absorbed by temperature or guidance, and it remains when the selected fraction tends to zero. This separates arithmetic error from the finite-selection approximation. Earlier work discusses precision-induced positional bias (Hayakawa et al., 2026). The intervention here concerns CFG arithmetic and its effect on the guidance map.

Native fits and estimation bias. The replay fits shared parameters within a state. This differs from fitting every position separately, or sharing guidance within a prompt while fitting each position's temperature. In native BF16 top-1 experiments, the original 32-prompt per-position summary is 1.5047 $[1.2361, 1.7868]$. Sharing guidance within a prompt gives 0.7320 $[0.5957, 0.8716]$. An additional 96 prompts give 0.7061 $[0.6344, 0.7782]$ with the latter estimator. The additional prompts were captured in a different environment. These results do not establish the ideal coefficient for native sampling.

To isolate estimation error, we sample the BF16 power-form target eight times on four fixed prompts, keeping the original per-position sample counts. The per-position estimator exceeds its own numerical projection by 0.4562 $[0.3131, 0.5993]$. The estimator sharing guidance within each prompt has bias -0.0045 $[-0.0100, 0.0010]$ against its own, different projection. These t_7 intervals quantify only Monte Carlo variation on the four fixed prompts. This control identifies an estimation contribution, but does not assign all of the earlier coefficient near 2.0 to rounding or estimation. Numerical inputs and source hashes are provided in `data/coefficient_checks.json`.

C TESTS OF THE TEMPERATURE RULE

Each test first sweeps random positions over temperature, or takes their optimum from earlier development data, and then compares ranking noises with random positions at the optimum and at a hotter temperature (Table 8 and Figure 3). Intervals are corrected for the number of ranking noises compared at one temperature, and not across studies.

C.1 TEMPERATURE SWEEPS

Meissonic-1024. The sampler runs guidance 3 and 64 steps, and earlier development data put the random-position optimum at $s = 1.0$, with FID_{500} of 98.03, 92.51 and 96.77 at $s = 0.8$, 1.0 and 1.2. On 700 new COCO prompts, all eight settings of a prompt share its seed and GPU, and intervals are at level $1 - 0.05/3$. Heating from 1.0 to 1.4 raises random-position FID_{700} from 71.05 to 103.52. At $s = 1.0$ the noises $r = 1$, 3 and 9 are worse than random positions by 26.02, 10.22 and 3.30 FID, and at $s = 1.4$ they are better by 19.72, 33.16 $[30.14, 37.69]$ and 25.38.

LaViDa-O. The sampler generates 1024-pixel images in 32 steps. The prompts are the 241 GenEval prompts that no earlier LaViDa-O study used, with two draws each, and the metric is CLIP-H. Random positions at guidance 4, over eight temperatures from 0.25 to 2.0, reach their best CLIP-H, 0.3810, at the released $s = 0.8$, and no cooler temperature is significantly better at guidance 4 or 1. The rule fixed in advance then set the hot temperature to 1.6, the smallest tested value of at least twice the optimum. There, all five noises $a \in \{0.5, 1.5, 3, 5, 15\}$ beat random positions at level $1 - 0.05/5$, and the best, $a = 1.5$, gains 0.0475 $[0.0414, 0.0537]$ and ties with $a = 3$. With position scores at temperature 1, the map is $s' = s\tau/(s+\tau)$, and setting s' to the optimum gives $a \approx 2.1$ at a step with $\alpha_t = 0.75$, between the two tied values. The release draws its ranking noise in bfloat16, which gives $-\infty$ to 0.2% of eligible positions per step in every setting.

aMUSEd-256. The rule was developed on this model at guidance 4 and 12 steps. The random-position optimum is $s = 0.8$, at 46.51 FID against 52.17 at $s = 1$. At $s = 1.25$ the best noise is

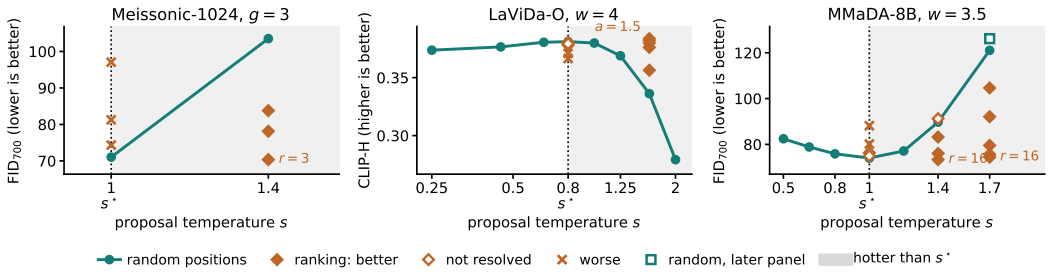


Figure 3: Temperature sweeps on three validation models. Curves show random-position performance. Markers compare each tested ranking strength with random positions at the same temperature, and labels identify the best noise above $s^\star$ (shaded). Ranking helps above $s^\star$ but has no significant gain at it. The MMaDA points at $s = 1.7$ use a later batch of 700 prompts. The separate strength-trend prediction L3 is evaluated most clearly on aMUSEd (Appendix C.1).

Table 8: Tests of the temperature rule and their original decisions. L1 predicts a gain above the temperature optimum, L2 no gain at or below it, and L3 weaker optimal ranking as the temperature approaches the optimum (Section 4). The third column states when each prediction was fixed. Decisions reproduce each study's own criterion rather than a common equivalence test. For an L2 criterion, "supported" means that no tested setting was significantly better than random positions, and the aMUSEd check uses non-inferiority margins fixed in advance. $g^\star$ and $w^\star$ are the random-position optimum along guidance. Tests that cannot fail or that show no significant differences add no evidence for a trend. †: reading made after the results. $*$: execution change made before any outcome was inspected (Appendix C.3).

Study	Prediction	Fixed	Result under the stated criterion
(a) Temperature sweeps and follow-ups			
Meissonic; s, FID	L1 at s=1.4; L2 at $s^\star$=1	before images	L1 and L2 supported: all four criteria met
LaViDa-O; s, CLIP-H	no cooler s wins; L1 at $2s^\star$	before images	both temperature criteria met
MMaDA-8B; s, FID	L2 at $s^\star$; L1 at s_{hot}; L3	before images	L1 and L2 supported*; L3 cannot fail in this design
MMaDA-8B; equal commits, FID	r=16 loses at s=1, wins at 1.4	before images	both signs supported* on new prompts
MMaDA-8B; frontier, FID/CLIP-B	heated ranking beats random on FID or CLIP	before images	only one metric criterion met; no gain established
MMaDA-8B; s=1.7, FID	L1; best r in the finite-gate set	before images	L1 and the predicted set supported; exact r missed
aMUSEd-256; s, FID	L1 at s=1; L2 at $s^\star$=0.8	development, then new prompts	L1 supported; all four L2 non-inferiority margins met
(b) Guidance sweeps near the temperature optimum			
H-MaskGIT-B; w, FID	best noise non-decreasing	before the law	noise trend supported
H-MaskGIT-S; w, FID	best noise $4.5 \to 18 \to \infty$	before images	predicted noise sequence in 16/16 seeds
Meissonic; g, CLIP	L3; random best above $g^\star$	before images	both guidance criteria met
(c) Guidance used as a proxy for temperature			
aMUSEd-256, s=1; g, FID	random wins from $g^\star$ up	before the law	prediction fails at 4 and 6.5; partial explanation†
LaViDa-O; w, CLIP-H	L1 below $w^\star$; L2, L3	before images	L1 fails; temperature already optimal†
Meissonic, g=9; s, FID	interior r wins at s=1.4	before images	no interior r wins; heating effect within noise†
(d) Inconclusive tests and limits of the rule			
aMUSEd-512; g, CLIP	L3 with a significant change	before images	no significant differences (three stages)
MaskBit; w, FID	wins below $w^\star$, loses above	before the law	gain below the optimum supported; $w^\star$ not reached
H-MaskGIT-S; w, FID	swap of confidence and random next to $w^\star$	before the law	predicted swap fails; outside the temperature rule

$r = 2$, at 47.19 against 77.83 for random positions, and a separate search ranks it above $r = 0.5$ and 1. At $s = 1$ the best is $r = 6$, at 46.18 against 51.87. A check on 2,000 new prompts, in a loop that never re-masks committed tokens, confirms this value. Greedy confidence, $r = 2$, 6 and 16, and random positions reach 40.83, 37.29, 36.49, 37.71 and 41.45 FID, and $r = 6$ beats random positions by 4.97 $[4.20, 5.91]$. At $s = 0.8$, random positions are no worse than tuned confidence within four margins fixed in advance (Appendix E).

C.2 SWEEPS ALONG GUIDANCE

These sweeps vary guidance, which is only a proxy for the distance to the temperature optimum (Section 7).

Near-optimal temperature. H-MaskGIT-B crosses guidance 0.3, 1 and 3 with first-step noise $\tau_0 \in \{0, 4.5, 18, \infty\}$ and 16 seeds of 10,000 images. Its random-position optimum lies near guidance 0.7. Below it, the released $\tau_0 = 4.5$ is best, at 8.995 FID against 15.010 for random positions, and at guidance 1 and 3 random positions are best, at 7.608 and 18.786. On H-MaskGIT-S the optimum lies near 1.0, and the best noise moves from 4.5 to 18 to random positions at guidance 0.3, 0.7 and 1.5, each time winning in all 16 seeds. At twice its sampling temperature, every setting of H-MaskGIT-B is 16 to 40 FID worse than at its default temperature, and ranking only limits the damage, so the rule does not apply far from the optimum. Meissonic-1024, scored by CLIP on 2,688 images, has its optimum at guidance 4.5. Its best noise is 3 at guidance 1.5, 9 at 3 and 4.5, and random positions from 6 to 12. Confidence at noise 3 gains 0.0103 $[0.0015, 0.0196]$ CLIP at guidance 1.5 and loses 0.0069 $[0.0007, 0.0138]$ at 4.5.

Guidance as a proxy. Three sweeps used guidance in place of the distance to the temperature optimum, and all three predictions failed. In aMUSEd-256 at $s = 1$, the random-position optimum lies at guidance 4, where confidence still wins by 3.70 FID. The difference is not significant at 6.5, and random positions lead by 0.77, 3.34 and 5.35 at 8, 10 and 12, below the crossing between 10 and 13 that the temperature reading places. In LaViDa-O, random positions are best at guidance 4 among six levels. At guidance 1, 2, 4 and 8, no noise beats them under a Bonferroni correction over five noises, the weak noises $a = 5$ and 15 tie them within 0.0023, and greedy ranking loses 0.088 $[0.074, 0.103]$ at guidance 1, where it leaves 11% of images nearly blank. In Meissonic-1024 at the released guidance 9, on the prompts and seeds of the guidance-3 sweep, heating from $s = 1$ to 1.4 changes random-position FID by only 0.10 $[-2.48, 1.79]$. No noise beats random positions at $s = 1.4$, and at $s = 1$ the noises $r = 1$, 3 and 9 are worse by 13.86, 6.25 and 5.15. Two of its four shards ran on a host with a different driver, common to all settings of a prompt.

Other sweeps. On MaskBit, random-position FID still falls at guidance 35, the top of the grid, and confidence is better at every level, by 20.3 $[19.8, 20.7]$ FID at 7.1 and by 8.2 at 35. On aMUSEd-512, scored by CLIP-B, the best noise moves from 4.5 to 18 as guidance rises from 2.5 to 40, but no change between adjacent levels exceeds the threshold of 0.001 fixed in advance.

Step count. The rule does not cover the number of steps. In aMUSEd-256 and H-MaskGIT-S, the extra FID that heating costs random positions relative to confidence changes with the step count in opposite directions. Two tests fixed in advance tried to predict this response on H-MaskGIT-S, and neither was supported. A swap of step schedules left the sets of best noises overlapping, and a crossed test of step count and temperature gave -3.52 $[-11.35, 2.64]$ and -0.22 $[-10.68, 10.62]$. A fixed noise multiplier also does not hold the ranking noise fixed across step counts, since the last noise before the final step scales as $1/(T - 1)$.

C.3 MMADA-8B

Model. We use `Gen-Verse/MMaDA-8B-MixCoT` at 512 pixels in bf16, with the released schedule of Table 5, and change only r, the temperature, and whether positions are ranked or random. The README's command sets guidance 3.5 and 15 steps, while the value 1.5 in the demo configuration belongs to its training block, and we recorded this choice before any image. One earlier study in this project used MMaDA-8B-Base, a different checkpoint. It compared a scrambled Halton order with random positions at one temperature, scored only CLIP-B on 250 prompts, had no confidence setting and found no significant difference. No ranking, temperature or FID result existed for any MMaDA checkpoint when the predictions below were written.

The temperature sweep. The 700 COCO captions had never been generated in this project, and all settings of a prompt share its seed and GPU. Random positions ran first over seven temperatures from 0.5 to 1.7 (Figure 3). The rule fixed in advance set $s^\star = 1.0$, the lowest FID_{700} at 74.03, and the hot temperature 1.4, the first hotter value at least 10 FID worse, at 89.72.

Confidence with $r \in \{1, 16, 256, 4096\}$ then ran at both temperatures, with intervals at level $1 - 0.05/4$. This correction covers the four noises at one temperature. It does not adjust for choosing $s^\star$ on the same prompts, which the later studies on further prompts address.

At $s^\star$ the intervals still allow up to 2.03 FID of benefit, and the prediction that the best ranking strength decreases as the proposal cools holds only trivially, because random positions are the weakest setting.

Heated ranking at $s = 1.4$ and $r = 16$ improves on the released sampler by 14.81 $[12.89, 17.77]$ FID and 0.00744 $[0.00528, 0.00959]$ CLIP-B, in a reading made after the results.

The plan first stopped any image whose per-step commit count, which bf16 ties can change, deviated from the schedule by more than 8 tokens. One image reached 10 before any image had been scored or viewed, so the limit was raised to 64 and the sweep was regenerated from the same seeds. A check on the 672 prompts that deviate by at most 8 gives the same conclusions.

Follow-ups on further prompts. The first follow-up enforced the scheduled commit counts exactly on 700 further captions, with exact score ties broken by an independent random key and its two main contrasts at 97.5%. A first attempt stopped after 35 images, never scored or viewed, on an assertion that committed positions are never re-masked, which the released sampler does not guarantee. The assertion was removed and all six settings were regenerated.

The second follow-up reused these prompts and seeds, committed uniformly random positions with the exact counts, and drew each candidate from the finite-gate law at the noise of r=16. Its FID gains over confidence ranking at r=16 are $+0.60$ $[-1.62, 2.66]$ at $s = 1$ and -1.61 $[-3.71, 0.67]$ at $s = 1.4$. These are 95% percentile paired-prompt bootstrap intervals. This gap mixes position preference, dependence between commits and the remaining error of the value law, so it bounds the difference between these samplers rather than a pure position effect.

The third follow-up used 256 further captions at $s = 1.6$, a temperature never generated on this model. Before any image, it fixed each noise's equivalent temperature from the finite-gate law and predicted each contrast from the random-position curve of the same prompts (Table 10), while the equivalent temperatures at $s = 1.0$, 1.4 and 1.7 were already known. The signs of the two primary contrasts agree with the predictions, but the magnitude criterion fails. The secondary $r = 4096$ contrast has the opposite observed sign, with an interval that includes zero.

A refit of $\mathrm{FID}(n) = \mathrm{FID}_\infty + a/n$ on subsets of the stored features, made after the results, lowers the levels by about 36 FID and moves the residuals by at most 0.9 (Chong & Forsyth, 2020).

Hotter proposals and calibrated random positions. On 700 further prompts at $s = 1.7$, confidence with $r \in \{4, 16, 64, 256, 1024\}$ ran against random positions at 1.7 and at 1.0, with intervals at level $1 - 0.05/2$. Before the images, a measurement on sampling states of 64 earlier prompts made the finite-gate law pick $r = 4$ and the power form pick $r = 256$, and a looser prediction, that the best noise lies in $\{4, 16\}$, was then added. The best noise is $r = 16$, so the strict prediction fails and the looser one holds. $r = 4$ is 1.22 $[-1.21, 4.40]$ FID behind, the power-form pick is 17.67 $[13.81, 20.97]$ behind, and $r = 16$ ends -1.46 $[-4.21, 1.32]$ from random positions at 1.0. On 1,400 further prompts, heated ranking at $(s, r) = (1.4, 16)$ is not significantly different from random positions at $s = 1$ on FID or CLIP-B, and its CLIP-L advantage, $+0.0021$ $[0.0005, 0.0037]$, is a secondary metric. The ranking side then received nine settings, $s \in \{1.2, 1.4, 1.6\}$ with $r \in \{8, 16, 32\}$, on the same prompts. Each sampler's best setting was chosen on the first 200 prompts and read on the other 1,200, where ranking at $(1.2, 16)$ has 1.49 $[-0.15, 2.85]$ higher FID than random positions at $s = 0.9$ (95% percentile interval). The criterion fixed in advance, a lower bound above -2, holds. Four of these studies changed their execution before any of their results was read, through a restart after a GPU timeout, shards moved to another machine, a corrected identifier format and an analysis program that skips unexpected keys, and none of these changes touched the prompts, seeds, settings or criteria.

Table 9: Temperature tests and controls on MMaDA-8B. Guidance w=3.5, 15 steps. FID gain is baseline FID minus the FID of the sampler named first in each comparison. The value-only sampler uses random positions and finite-gate values, so its baseline is random positions with proposal values. All other gains compare confidence ranking with random positions. Every study used prompts new to this model. The value-only test shares prompts and seeds with the equal-count test, and the nine-setting comparison reuses the 1,400 prompts. Intervals are paired-prompt bootstraps at each study's own level (Appendix C.3). The random-position sweep selected $s^\star$=1 and s_{hot}=1.4 before any ranking image.

Comparison	Settings and prompts	Results
Confidence vs random: temperature sweep	700 prompts, $s \in \{1, 1.4\}$, four noise multipliers	At s=1, released r=1 gains -14.1 $[-18.4, -11.9]$ FID. No noise improves FID, with at most 2.03 of benefit allowed. At s=1.4, three noises improve FID, and r=16 gains $+16.4$ $[12.0, 19.2]$. The same r=16 gains -6.1 at s=1: the effect reverses as predicted by L1/L2.
Confidence vs random: equal commit counts	700 further prompts, r=16, exact counts at every step	FID gains are -5.1 at s=1 and $+17.9$ at s=1.4, both significant. Matching commit counts preserves the reversal.
Value-only vs proposal values	Same 700 prompts and seeds, random positions, exact counts, finite-gate values at r=16	FID gains are -4.5 at s=1 and $+16.3$ at s=1.4. The differences from confidence ranking include zero at both temperatures.
Confidence vs random: unseen temperature	256 further prompts, s=1.6, finite-gate entropy-matched prediction	At r=16, gains are $+36.6$ FID and $+0.032$ CLIP-B. The predicted signs hold, but the magnitude criterion fails (Table 10).
Confidence vs calibrated random	1,400 further prompts, ranking at $(s, r) = (1.4, 16)$ and random at s=1	Gains: FID $+0.1$ $[-1.8, 1.4]$, CLIP-B $+0.0012$ $[-0.0002, 0.0027]$, CLIP-L $+0.0021$. No significant FID gain, with an upper bound of 1.38. Nine ranking settings on these prompts are compared in Table 2.
Confidence vs calibrated random: hotter proposal	700 further prompts, ranking at s=1.7 and random at s=1	Ranking recovers 51.8 [47.7, 57.3] FID of the heating cost. Its gain over calibrated random is $+1.5$ $[-1.3, 4.2]$, not significant, with an upper bound of 4.21.

Table 10: A test at a temperature never generated on MMaDA-8B. Guidance 3.5, 256 new prompts, exact commit counts, s=1.6. Each FID gain is predicted from the random-position curve on these prompts, evaluated at the entropy-matched temperature s_{eq} under the finite-gate law. These temperatures and the criteria were fixed before any image was generated. Predicted and observed are FID gains over random positions at s=1.6, so positive entries favor ranking, and the residual is observed minus predicted. Observed gains have 97.5% paired-prompt basic bootstrap intervals, and residuals have 95% paired-prompt percentile bootstrap intervals.

Noise	Role	s_{eq}	Predicted FID gain	Observed FID gain	Residual
r=1	secondary	0.83	$+34.25$	$+32.44$ $[24.63, 39.30]$	-1.80 $[-4.92, 3.18]$
r=16	main	0.99	$+33.89$	$+36.65$ $[31.24, 45.46]$	$+2.75$ $[-2.44, 6.86]$
r=256	main	1.18	$+31.70$	$+23.97$ $[18.98, 32.98]$	-7.72 $[-13.74, -2.69]$
r=4096	secondary	1.34	$+18.30$	-1.57 $[-8.78, 6.56]$	-19.88 $[-25.23, -12.65]$

Random-position FID on these prompts: 129.7 at s=0.8, 130.1 at 1, 132.6 at 1.2, 150.7 at 1.4 and 164.0 at 1.6. The criteria required the signs of the two main noises to agree with the prediction, which holds, and each residual interval to lie within ± 5 FID, which fails. The secondary noises were named as such before generation. Every noise but the weakest improves on random positions at s=1.6, and r=16 reaches FID 127.4 against 164.0, with CLIP-B higher by 0.0324. The random-position optimum on these prompts lies at the edge of the grid, at s=0.8.

D THE GUIDANCE SETTING OVER WHOLE RUNS

Common design. The studies behind Table 4 share one aMUSEd-256 loop that never re-masks committed tokens, with $T = 12$, base temperature $s = 1$ and the release's noise $\tau_t = 2r(1-t/(T-1))$ at $r = 6$. Every sampler follows the release's cosine schedule, which commits 3 of 256 positions at the first step, 22 of 203 at step 5 and 33 of 66 at step 10. The realized integer counts differ from this schedule by at most three tokens per step, and the last step commits every remaining position from the original proposal. The prompts are 2,000 COCO captions from an earlier study. Each study draws two new images per prompt, so it repeats the sampling on a known prompt set rather than on new prompts.

Under the power-form mapping of Section 5.2, RAND+T uses $s'_t = s\tau_t/(1+\tau_t)$ and RAND+T+G also uses $g'_t = 1 + (g-1)\tau_t/(1+\tau_t)$. CONF+G instead ranks by confidence under the mapped guidance,

so by Propositions 3 and 1 it shares its local value law with COND and RAND+T+G under the power-form approximation. ANTI is a further scoring intervention with score $2 \log q_{0,s}(X) - \log q_{w,s}(X)$. Multiplying the proposal by this score's exponential weight gives the same s' and guidance $w' = w(\tau - 1)/(1 + \tau)$ under that approximation.

Five samplers at guidance 4 and 10. At guidance 4, the four samplers that change the values lie between 36.66 and 36.89 FID, and all beat RAND at 42.65. At guidance 10, CONF reaches 54.77, COND 49.98, CONF+G 47.45, RAND+T+G 46.59 and RAND 53.08. The study required COND and CONF+G to be equivalent within ± 1 FID and ± 0.001 CLIP-B, and CONF+G and RAND+T+G within ± 1 FID, at both guidance levels. Both equivalences fail at guidance 10, where COND trails CONF+G by 2.53 $[1.66, 3.45]$ FID and CONF+G trails RAND+T+G by 0.86 $[0.03, 1.79]$. Matching the power-form value laws therefore does not make whole runs equivalent, in line with the position differences of Appendix B.1. The third prediction, that COND beats CONF at guidance 10, holds by 4.78 $[3.90, 5.78]$ FID.

Changing only the guidance. To separate guidance effects from temperature effects, we compared ten samplers at guidance 10, with 4,000 images each and intervals at level $1 - 0.05/2$. Random positions at fixed guidance form a family of six samplers, RAND, RAND+T and constant temperatures $s \in \{0.45, 0.6, 0.75, 0.9\}$, whose best member, $s = 0.9$ at 52.74 FID, is the comparator. As required, RAND+T+G and COND both beat it, by 6.30 $[5.38, 7.39]$ and 2.83 $[1.82, 4.06]$ FID. RAND+T+G also beats RAND+T, which shares its temperature schedule, by 7.02 $[6.12, 8.22]$, and it raises CLIP-B by 0.0032 $[0.0025, 0.0040]$ over the comparator. Its within-prompt LPIPS is lower, 0.584 against 0.592, so the gain costs some variation between images. The temperature grid has a spacing of 0.15 and does not bound an untested optimum between its points. ANTI beats COND by 1.56 $[0.60, 2.60]$ FID, in line with the lower guidance that Proposition 1 assigns to COND and the expression above assigns to ANTI. Three separate generations of the guidance-10 values of CONF, COND and RAND+T+G agree within 0.6 FID.

Across guidance. A third study ran RAND at guidance 1.25 to 4, CONF and RAND+T at 1.25, 2 and 3, and COND and RAND+T+G at 2 and 3, with 56,000 images and intervals at level $1 - 0.05/2$. RAND is best at guidance 4, so 2 and 3 lie below the guidance optimum. At guidance 1.25, RAND reaches 103.16 FID, CONF 65.19 and RAND+T 63.45. The study predicted that below this optimum confidence marks the reliable positions, so that CONF beats RAND+T. It does not, since the FID gain of CONF over RAND+T is -0.93 $[-2.08, 0.19]$ at guidance 2, -0.35 $[-1.16, 0.56]$ at 3 and -1.74 $[-3.36, 0.25]$ at 1.25. Given the mapped temperature, random positions are thus never significantly worse than confidence at guidance 1.25, 2, 3 or 10. The FID gain of COND over CONF is -1.62 $[-2.61, -0.64]$ at guidance 2 and -0.80 $[-1.57, 0.02]$ at 3.

Finite-gate values at random positions. We also repeated the random-position comparisons with values drawn from the finite-gate law instead of the power form, at guidance 10. The FID changes are not significant: $+0.02$ $[-0.77, 0.84]$ for the confidence score and -0.15 $[-0.91, 0.50]$ for the conditional score. The finite-gate samplers differ from the ranked ones as the power-form samplers did, by -1.06 $[-2.26, 0.01]$ FID against CONF, which is not significant, and by -4.04 $[-5.14, -3.09]$ against COND. These comparisons cannot separate position preferences, dependence between commits and the remaining error of the value law.

Plain guidance tuning. A lower constant guidance is the natural alternative to the mapped schedule. A development comparison on 512 prompts keeps the temperature schedule, commit counts, seeds and device of RAND+T+G and uses constant guidance at every step, including the last (Table 11). Constant guidance 4 gives a point FID 8.6 below RAND+T+G with higher CLIP-B. This comparison supports ordinary guidance tuning as a strong alternative to the mapped schedule, which serves here as a mechanism control.

Two further models. On Meissonic-1024, 1,024 new COCO prompts with two seeds each ran at 64 steps, $s = 1$ and $r = 3$, at guidance 1.5 and 9 (Table 12). The four main FID gains were predicted to be negative at 1.5 and positive at 9, and all four had to be significant. All four point estimates have the predicted sign, but two intervals include zero, so the prediction is not met. COND beats CONF by

2.45 [1.25, 3.82] FID at guidance 9, and both differences between guidance levels, fixed in advance as secondary contrasts, are significant.

On MMaDA-8B, 412 new prompts with two seeds each ran with exact commit counts at $s = 1$, $r = 16$ and $w \in \{1, 3.5, 7\}$. This test was fixed after the Meissonic test and after a reading of its images, which had shown image contrast, the standard deviation of pixel brightness, rising with guidance and falling under conditional-score ranking. It required the FID gain of COND over CONF to rise from $w = 1$ to $w = 7$, and the COND minus CONF image-contrast difference to fall, at 97.5%.

The contrast criterion is met. The FID gain rises from -1.94 to $+0.06$, but the interval for this change includes zero, so the FID criterion is not met. Confidence's own FID still improves at $w = 7$, so no tested setting lies above this model's FID guidance optimum, where the aMUSEd studies found the gain of the conditional score.

Table 11: Constant guidance against the mapped guidance schedule (development). aMUSEd-256 without re-masking, 512 development prompts with two draws each. Every sampler keeps the temperature schedule, commit counts, seeds and device of RAND+T+G, and uses constant guidance at every step, including the last. FID gain is the FID of RAND+T+G minus that of the constant-guidance sampler. Other differences are constant guidance minus RAND+T+G, so positive FID gain and CLIP differences favor constant guidance. LPIPS measures variation between images of the same prompt. FID_{1024} is a point estimate without an interval. The other intervals are 95% paired-prompt percentile intervals, not adjusted for the four guidance values.

Constant guidance	FID_{1024} gain	ΔCLIP-B	ΔCLIP-L	ΔLPIPS
g=3	+7.90	−0.0010 [−0.0027, 0.0007]	−0.0036 [−0.0057, −0.0016]	+0.0084 [0.0025, 0.0144]
g=4	+8.62	+0.0019 [0.0003, 0.0034]	+0.0015 [−0.0004, 0.0035]	−0.0039 [−0.0094, 0.0017]
g=6	+5.34	+0.0027 [0.0013, 0.0042]	+0.0029 [0.0011, 0.0047]	−0.0079 [−0.0125, −0.0033]
g=8	+1.79	0.0000 [−0.0014, 0.0014]	+0.0006 [−0.0010, 0.0023]	−0.0017 [−0.0057, 0.0023]

The selection rule fixed in advance picks constant g=4, both on FID and under the alignment constraint. Returning only the last step to g=10 worsens FID by 2.2 to 5.3 for every tested guidance.

Table 12: Score interventions across guidance levels on Meissonic-1024 and MMaDA-8B. For each comparison A vs B, FID gain is FID(B) minus FID(A), while other differences are A minus B. Positive FID gain favors A. The predicted signs refer to FID gain. Interactions are the gain or difference at high guidance minus that at low guidance. Intervals are 95% paired-prompt bootstrap intervals that keep both seeds of a prompt together, except the MMaDA interaction row, which uses 97.5%. Meissonic: 1,024 new prompts, two seeds each, 64 steps, s=1, r=3. MMaDA: 412 new prompts, two seeds each, 15 steps, s=1, r=16, exact commit counts. Image contrast is the standard deviation of pixel brightness.

Meissonic-1024

Comparison	Guidance	Predicted gain sign	FID_{2048} gain	ΔCLIP-B
COND vs CONF	1.5	−	−0.72 [−1.86, 0.34]	−0.0022 [−0.0032, −0.0013]
RAND+T+G vs RAND+T	1.5	−	−1.90 [−2.98, −0.80]	−0.0014 [−0.0023, −0.0005]
COND vs CONF	9	+	+2.45 [1.25, 3.82]	+0.0004 [−0.0004, 0.0013]
RAND+T+G vs RAND+T	9	+	+0.95 [−0.29, 2.05]	−0.0002 [−0.0010, 0.0006]
Secondary interactions: guidance 9 minus 1.5				
COND vs CONF	9 minus 1.5	+	+3.17 [1.62, 4.98]	
RAND+T+G vs RAND+T	9 minus 1.5	+	+2.86 [1.19, 4.35]	

MMaDA-8B, COND vs CONF

Guidance	FID of CONF	FID_{824} gain	Δimage contrast	ΔCLIP-B
1	80.36	−1.94 [−3.80, 0.63]	−0.0024 [−0.0046, −0.0002]	−0.0068 [−0.0085, −0.0050]
3.5	78.97	−1.01 [−3.32, 0.95]	−0.0042 [−0.0065, −0.0020]	−0.0023 [−0.0038, −0.0007]
7	77.94	+0.06 [−2.06, 1.61]	−0.0071 [−0.0093, −0.0048]	−0.0003 [−0.0017, 0.0010]
Interaction: guidance 7 minus 1				
7 minus 1		+2.00 [−1.87, 4.60]	−0.0047 [−0.0082, −0.0011]	

Meissonic: the prediction required all four main contrasts to be significant with the predicted sign. Two are not, so the prediction is not met, while both differences between guidance levels, fixed in advance as secondary contrasts, are significant. MMaDA: the prediction required a significant increase in FID gain and a significant decrease in image contrast from w=1 to w=7. Only image contrast meets this criterion. The FID gain increases, but its interaction interval includes zero.

E CALIBRATED COMPARISONS

E.1 TUNING AND EVALUATION OF CALIBRATED COMPARISONS

Six comparisons, including the earlier aMUSEd evaluation, fixed both samplers before evaluation on prompts excluded from tuning (Tables 13 and 14). None gives confidence ranking a significant FID lead, and the largest point estimate is 0.13.

The two MMaDA evaluations overlap: the expanded search uses 200 of the original 1,400 prompts for tuning and the other 1,200 for evaluation. Earlier confidence settings on aMUSEd and Meissonic were retuned after losing while the corresponding random-position settings stayed fixed, giving ranking more tuning effort.

For Table 2, the aMUSEd interval uses Student-t over eight blocks, while the other intervals use paired-prompt bootstrap. Each prompt generates two images for aMUSEd and for Meissonic at $g = 6$, and one image otherwise.

Table 14: Tuning and evaluation data for Table 13. Model, guidance, step count and evaluation size identify the matching comparison. Settings are (g, s, r) or (s, r). Tuning prompts select the settings, and evaluation prompts provide the reported gains. †: both settings were fixed before the evaluation, on prompts that took no part in the choice. Rows without † use the same prompts for tuning and evaluation.

Comparison	Settings tried or fixed	Tuning prompts	Evaluation prompts
aMUSEd-256, g=4, T=12, new prompts†	Both samplers: $g \in \{2, 4, 7, 10, 13\} \times s \in \{0.4, \ldots, 1.6\}$ at r=1. Further ranking search: $s \in \{1, 1.25\} \times r \in \{2, \ldots, 16\}$.	1,000 search prompts	the next 2,000 prompts of the pool
independent evaluation†	Confidence $(4, 1, 6)$ and random $(4, 0.8, 1)$.	same 1,000 search prompts	81,696 images across eight blocks, metrics and both samplers, on captions used in earlier benchmarks
aMUSEd-256, g=4 and 10, T=48	Both samplers: $s \in \{0.8, 1, 1.25, 1.6\}$ at r=1.	1,000 development prompts	the same prompts, and at g=4 both optima lie at the grid edges
Meissonic-1024, g=3, 700 prompts	Confidence: $s \in \{1, 1.4\} \times r \in \{1, 3, 9\}$. Random: $s \in \{1, 1.4\}$, with $s^\star$=1 from earlier development.	the same 700 prompts	the same 700 prompts
700 held-out prompts†	Confidence $(g, s, r) = (3, 1.4, 3)$ and random $(3, 1, 1)$.	preceding 700-prompt Meissonic sweep	700 prompts new to Meissonic, from caption rows used by the MMaDA sweep
Meissonic-1024, g=6, independent evaluation†	Confidence: six settings, then $s \in \{1, 1.2\} \times r \in \{2, 4, 8\}$, and a check of r=16. Random: fixed at $(6, 1, 1)$.	development prompts	5,000 new COCO prompts and the 553 GenEval prompts
MMaDA-8B, w=3.5, T=15, 700 prompts	Confidence: $r \in \{1, 16, 256, 4096\}$ at $s \in \{1, 1.4\}$. Random: seven temperatures.	the same 700 prompts, read after the results	the same 700 prompts
1,400 prompts†	Confidence: fixed at $(s, r) = (1.4, 16)$. Random: $s \in \{0.8, 0.9, 1.0\}$.	the 700-prompt sweep	1,400 new prompts
nine ranking settings†	Confidence: $s \in \{1.2, 1.4, 1.6\} \times r \in \{8, 16, 32\}$. Random: $s \in \{0.8, 0.9, 1.0\}$.	the first 200 of the 1,400 prompts	the other 1,200, shared with the row above
s=1.7, 700 prompts	Confidence: $r \in \{4, 16, 64, 256, 1024\}$, a set fixed in advance. Random: s=1.7 and 1.0.	the same 700 prompts	the same 700 prompts

E.2 INDEPENDENT EVALUATIONS AND THE RETUNING OF CONFIDENCE

For the independent aMUSEd-256 evaluation, the two settings were fixed before 81,696 images in eight paired blocks, each run on one GPU with two images for 2,000 COCO captions and for the 553 GenEval prompts per setting (Table 15). The criteria fixed in advance bound the random-minus-confidence differences on one side, with an FID upper bound below $+0.25$ and lower bounds above -0.001 CLIP-B, -0.01 GenEval and -0.01 diversity, and all four hold.

In Meissonic-1024 the two settings were fixed before 10,000 COCO and 4,424 GenEval images per setting on new prompts. No GenEval margin was fixed there, and its GenEval interval would not pass a margin of one point, so we claim no GenEval equivalence for Meissonic.

Calibration also works in the other direction. Earlier fixed pairs lost to random positions, by 1.06 FID in aMUSEd with confidence at $(g, s, r) = (4, 1.25, 2)$ and by 7.71 in Meissonic with confidence at $(6, 1.2, 2)$. After confidence was retuned, to $r = 6$ in aMUSEd and $r = 8$ in Meissonic, those leads fell to 0.20 and 0.83, while the random-position settings stayed as they were. The Meissonic

retuning stopped at the edge of its tested noises, because a check at $r = 16$ lowered both CLIP scores significantly and raised FID_{500} by 4.3, so $r = 8$ is not shown to be optimal.

In the native aMUSEd loop, committed tokens keep a large finite score, and the tuned confidence at $(4, 1, 6)$ re-masks 36 tokens per image on average, which the independent evaluation includes. In a loop that never re-masks, $r = 6$ still beats greedy confidence and random positions (Appendix C.1).

The Meissonic pair chosen on 700 prompts, confidence at $(s, r) = (1.4, 3)$ and random positions at $s = 1$, was regenerated on 700 further prompts that played no part in any Meissonic choice, taken from caption rows that the MMaDA sweep had used. The criterion fixed in advance, that any confidence lead stay within 2 FID, holds. The FID gain of confidence, $+0.69$ on the choosing prompts, becomes -0.84 on the new prompts, while confidence keeps a small CLIP-L edge (Table 15).

Table 15: Calibrated comparisons on prompts excluded from tuning. Every entry is the gain of confidence ranking over random positions, as in Table 2: random-position FID minus confidence FID, and confidence minus random positions for GenEval, CLIP and diversity. Positive FID, GenEval and CLIP entries favor confidence ranking. Positive diversity entries indicate more within-prompt variation. GenEval differences are in percentage points. Intervals are 95% paired-prompt bootstraps unless marked. *Medians of 32 paired regenerations with order-statistic intervals at 97.99%, and 97.87% for diversity. †Level $1 - 0.05/2$. The earlier Meissonic FID has no interval. Dashes mark metrics that were not computed.

Comparison	FID	GenEval (pp)	CLIP-B ($\times 10^{-3}$)	CLIP-L ($\times 10^{-3}$)	Diversity ($\times 10^{-3}$)
Final calibrated pairs					
aMUSEd, tuned pair, eight blocks, Student-t_7	-0.20 $[-0.53, 0.13]$	$+0.06$ $[-0.84, 0.96]$	-0.57 $[-0.85, -0.29]$	-0.94 $[-1.35, -0.53]$	$+1.75$ $[0.37, 3.13]$
Meissonic, tuned pair, g=6	-0.83 $[-1.22, -0.46]$	$+0.19$ $[-0.92, 1.29]$	$+0.20$ $[-0.14, 0.55]$	$+0.56$ $[0.14, 0.99]$	$+0.29$ $[-1.25, 1.81]$
Meissonic, g=3, 700 held-out prompts	-0.84 $[-2.43, 0.89]$	—	$+1.38$ $[-0.01, 2.76]$	$+1.78$ $[0.05, 3.50]$	—
MMaDA, heated ranking against random at s=1, 1,400 prompts†	$+0.13$ $[-1.75, 1.38]$	—	$+1.2$ $[-0.2, 2.7]$	$+2.1$ $[0.5, 3.7]$	—
MMaDA, nine ranking settings, 1,200 prompts	-1.49 $[-2.85, 0.15]$	—	$+0.5$ $[-0.8, 1.8]$	—	—
Earlier pairs, before ranking-noise retuning					
aMUSEd, earlier pair with r=2, 32 regenerations*	-1.060 $[-1.303, -0.919]$	-1.339 $[-1.600, -0.916]$	-1.81 $[-1.96, -1.55]$	—	$+7.45$ $[7.24, 7.82]$
Meissonic, earlier pair with r=2, g=6	-7.713	-1.237 $[-3.447, 0.949]$	-1.065 $[-1.861, -0.275]$	-0.658 $[-1.620, 0.304]$	-5.899 $[-9.461, -2.255]$

E.3 Against the released Meissonic configuration

The Meissonic sweeps at guidance 9 and 3 share prompts, seeds and positions, so the released sampler, confidence at guidance 9 with $s = 1$ and $r = 1$, can be compared prompt by prompt with random positions. It reaches 94.98 FID_{700}, against 81.11 for random positions at the same settings and 71.05 at guidance 3. The ranking step, 13.86 $[11.58, 17.40]$ FID, is the contrast fixed in advance for that sweep, at level $1 - 0.05/3$. The guidance step, 10.07 $[8.41, 12.78]$, and the total, 23.93 $[22.61, 27.55]$, were read after the results at 95%. CLIP-B rises by 0.00267 $[0.00119, 0.00414]$ in the ranking step, by 0.00064 $[-0.00093, 0.00221]$ in the guidance step and by 0.00331 $[0.00163, 0.00499]$ in total.

Because the combined change lowers guidance, FID and CLIP could hide a loss of compositional accuracy. We therefore specified a GenEval study before generating its images (Table 16). We checked whether the combined change caused a decline of at least 2 points whose interval excluded zero, and whether the ranking step's interval excluded a 2-point decline or established a significant decline.

The estimated decline exceeds 2 points and is significant: 3.08 $[0.49, 5.74]$ points. The ranking-step change is $+0.43$ $[-2.34, 2.87]$, whose interval meets neither condition, while the guidance-step change is -3.51 $[-5.87, -1.29]$. Two host restarts interrupted the generation, and the affected prompts were regenerated from their planned seeds before any image was scored.

The evaluator's detections separate the two steps in a reading made after the results. The ranking step leaves the GenEval outcome unchanged on 410 of 553 prompts, raising it on 73 and lowering it on

70. Yet random positions yield 0.61 $[0.46, 0.77]$ more objects per image, larger prompted objects and more object classes that the prompt does not name. The guidance step removes prompted objects instead, since images in which a prompted object is missing rise by 3.2 $[0.5, 5.7]$ percentage points and prompts that lose their GenEval pass outnumber those that gain one, 76 to 51.

F Positions chosen without reading token values

Unlike confidence ranking, the position orders tested here never read candidate values. We compare each with random positions, with both samplers drawing every token from the same proposal at the release's temperature, so any difference comes from the positions. The orders did not receive their own temperature, so these are same-temperature comparisons as defined in Section 2. The Halton order is rotated independently for each sample. FID gain is the FID of random positions minus that of the ordered sampler; CLIP gain uses the reverse difference. Positive gains favor the ordered sampler (Table 17).

H-MaskGIT. The released class-conditional checkpoint (Besnier et al., 2025) generates two development groups of ten ImageNet classes (Deng et al., 2009), with 2,500 images and three seeds per cell. At 4 steps, as guidance rises from w=0.2 to 3.0 over eight levels, the FID gain falls from $+21.62$ to -8.51 on the first group and from $+9.70$ to -1.01 on the second. At 32 steps, as guidance rises from 0.1 to 0.5, it falls from $+9.19$ to -1.62 on the first group. All six tested series move toward random positions as guidance rises, and each changes sign. Other metrics do not follow FID, since at 4 steps precision favors Halton at every guidance and at 32 steps inception score favors random positions at every guidance. The sign of some contrasts also depends on whether the FID reference is restricted to the generated classes, so we treat these cells as development evidence. On 1,000 classes with 5,000 images per cell, Halton beats random positions by 6.37 $[6.05, 6.68]$ FID at 4 steps. At 32 steps, applying Halton only during the early steps costs $+0.33$ $[0.13, 0.52]$ FID, and applying it only during the late steps costs $+0.09$ $[-0.17, 0.35]$.

LPD-L-256. The released class-conditional model (Zhang et al., 2026) runs 20 steps with its linear guidance ramp, three seeds and 50,000 images per cell. At cfg 1.5, 2.5, 3.5, 4.8 and 6.0, the FID gains of Halton over random positions are $+1.74$, $+0.90$, $+0.32$, -0.07 and -0.25. Those of LPD's own order are -0.24, -0.02, $+0.02$, $+0.18$ and $+0.38$. The Halton gain crosses zero between cfg 3.5 and 4.8, and LPD's released guidance of 4.8 lies just past that crossing.

aMUSEd-256. We use 2,500 COCO captions that no earlier study had used, three seeds per prompt, and paired-prompt bootstraps with 2,000 resamples. At 4 steps and guidance g=4, 5, 6, 7, 8 and 12, the FID gains are $+3.07$ $[2.00, 4.07]$, $+1.36$ $[0.28, 2.35]$, -0.15 $[-1.28, 0.94]$, -2.29 $[-3.49, -1.24]$, -4.39 $[-5.56, -3.37]$ and -10.33 $[-11.63, -8.97]$, so the crossing lies between 5 and 7, below the released guidance of 10. At 12 steps and g=4, 6, 8 and 12, the gains are $+2.10$ $[1.31, 2.82]$, $+0.44$ $[-0.24, 1.25]$, -1.48 $[-2.28, -0.64]$ and -3.28 $[-4.28, -2.17]$. At 24 steps they are $+1.14$ $[0.46, 1.82]$ at g=4 and $+0.23$ $[-0.63, 1.20]$ at g=12.

Meissonic-1024. Sixty prompts from one template with periodic content, with three seeds each, are scored by CLIP. At 16 steps and guidance 3, 9, 15 and 21, Halton gains 0.0107, 0.0110, 0.0148 and 0.0132 CLIP over random positions, and at 32 steps the gains are 0.0070, 0.0050, 0.0069 and 0.0078. Across guidance the gain moves by less than its seed-to-seed spread of about 0.012, so this small study detects no clear guidance dependence. The guidance-9 cells were generated on a different machine.

Table 13: Complete comparisons before and after temperature calibration. Gain is random-position FID minus confidence-ranking FID (positive favors ranking). Models are aMUSEd-256, Meissonic-1024 and MMaDA-8B. Guidance is w for MMaDA and g otherwise. T denotes steps. Confidence ranking uses the same r in both comparisons. "Own temperatures" uses each sampler's best tested temperature. Upper-block prompts are excluded from tuning. For *, nine (s, r) settings are compared on 200 prompts and both samplers evaluated on the other 1,200. Lower-block settings are selected on the reported prompts. Each setting generates two images per prompt for aMUSEd and Meissonic at $g = 6$, and one otherwise. Paired-prompt bootstrap intervals are 95%, except †97.5%, ‡98.75%, §98.33%, and the second aMUSEd evaluation uses Student-t over eight blocks (Appendix E.1). Dashes denote unreported comparisons, and values without intervals are point estimates.

		Same temperature		Own temperatures		
Model (guidance, T)	Prompts	s	Gain	Conf. (s, r)	Rand. s	Gain
Settings fixed before the evaluation						
aMUSEd $(4, 12)$	2,000 new	1	$+4.81$	$(1, 6)$	0.8	-0.49
aMUSEd $(4, 12)$	$8 \times 2{,}000$	—	—	$(1, 6)$	0.8	-0.20 $[-0.53, 0.13]$
Meissonic $(3, 64)$	700 new	—	—	$(1.4, 3)$	1	-0.84 $[-2.43, 0.89]$
Meissonic $(6, 64)$	5,000 new	—	—	$(1.2, 8)$	1	-0.83 $[-1.22, -0.46]$
MMaDA† $(3.5, 15)$	1,400 new	—	—	$(1.4, 16)$	1	$+0.13$ $[-1.75, 1.38]$
MMaDA* $(3.5, 15)$	1,200 new	—	—	$(1.2, 16)$	0.9	-1.49 $[-2.85, 0.15]$
Settings chosen on the reported prompts						
aMUSEd $(4, 48)$	1,000 dev.	1.6	$+75.48$	$(1.6, 1)$	0.8	$+0.66$
aMUSEd $(10, 48)$	1,000 dev.	1	-8.28	$(1.6, 1)$	0.8	-5.91
Meissonic§ $(3, 64)$	700	1.4	$+33.16$ $[30.14, 37.69]$	$(1.4, 3)$	1	$+0.69$
MMaDA‡ $(3.5, 15)$	700	1.4	$+16.37$ $[12.02, 19.24]$	$(1.4, 16)$	1	$+0.69$
MMaDA† $(3.5, 15)$	700	1.7	$+51.77$ $[47.74, 57.35]$	$(1.7, 16)$	1	$+1.46$ $[-1.32, 4.21]$

Table 16: GenEval for the released Meissonic configuration and random positions. Overall score over six tasks, in percent, on the same 553 prompts and seeds with two images per prompt and setting. The released configuration is confidence with r=1, s=1 and g=9, and every random-position configuration uses s=1. Differences are in percentage points (pp). Ranking step: random positions at g=9 minus the released configuration. Guidance step: random positions at g=3 minus g=9. 95% paired-prompt bootstrap intervals stratified by task.

	Random positions (%)			Change (pp)	
Released (%)	g=9	g=6	g=3	Ranking step	Guidance step
49.33	49.76	50.29	46.25	$+0.43$ $[-2.34, 2.87]$	-3.51 $[-5.87, -1.29]$

No single task changes significantly in either step, and the largest losses of the combined change are on two-object, color and counting prompts.

Table 17: Position-order effects across models and guidance settings. Halton order versus random positions at the lowest and highest tested guidance, in each release's own guidance parameter. Both samplers draw every token from the same proposal. FID gain is random-position FID minus ordered-sampling FID. CLIP gain is ordered sampling minus random positions, so positive values favor the ordered sampler for both metrics. Intervals are paired-prompt bootstraps at 95%. The H-MaskGIT cells are development runs of 2,500 images with three seeds and have no interval, although the change far exceeds their seed-to-seed spread of about 1.3 FID. The last row compares LPD's own order with random positions. Appendix F lists every intermediate guidance level.

Model	Gain metric	Steps	Low guidance	High guidance
H-MaskGIT, ten classes	FID	4	w=0.2: $+21.62$	w=3.0: -8.51
LPD-L-256	FID	20	cfg 1.5: $+1.74$	cfg 6.0: -0.25
aMUSEd-256	FID	4	g=4: $+3.07$ $[2.00, 4.07]$	g=12: -10.33 $[-11.63, -8.97]$
aMUSEd-256	FID	12	g=4: $+2.10$ $[1.31, 2.82]$	g=12: -3.28 $[-4.28, -2.17]$
Meissonic-1024	CLIP	16	g=3: $+0.0107$	g=21: $+0.0132$
LPD-L-256, LPD's own order	FID	20	cfg 1.5: -0.24	cfg 6.0: $+0.38$