COMPUTING GUIDANCE SCHEDULES FROM CONDITIONAL INFORMATION IN MASKED GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked generative models apply classifier-free guidance (CFG) repeatedly during parallel decoding, yet existing methods schedule guidance using a predefined function of normalized time. The appropriate time-based shape can vary across checkpoints, token grids, and reveal policies, requiring a new curve-family or exponent search for each setting. We introduce INFOPROGRESS, an information-coordinate view of CFG scheduling. A short Monte Carlo probe of the frozen decoder measures the conditional-to-unconditional log-likelihood ratio carried by tokens committed at each round. Accumulating this evidence produces a checkpoint-specific information clock, from which dose normalization computes all per-round guidance weights in closed form. INFOPROGRESS fixes a shared clock-to-guidance mapping and retains only the ordinary one-dimensional CFG-strength scan. On MaskGIT-256, it improves tuned constant CFG from FID 4.481 to 3.218 at 32 network evaluations and surpasses the checkpoint's reported FID 4.17 at 64 evaluations. Across decoding budgets, token grids, model capacities, and MaskBit codebook widths, it consistently improves constant CFG and remains within 0.28 FID of the best hindsight-selected fixed curve.

1 INTRODUCTION

Masked generative models produce image tokens in parallel and refine the image over a small number of decoding rounds (Chang et al., 2022; Besnier et al., 2025a). Classifier-free guidance (CFG) controls how strongly the condition influences these token predictions. Its schedule determines when that guidance is applied: two schedules can have the same average strength while allocating it to different stages of generation. Common choices include a constant weight and predefined functions of normalized decoding time (Wang et al., 2024; Weber et al., 2024).

Normalized time counts completed rounds. It does not measure how much evidence for the condition has accumulated in the generated tokens. At the same fraction of the decoding budget, different checkpoints, token grids, or token-selection rules may have made different progress. Choosing a guidance curve in time therefore also chooses an assumption about that progress. Searching over curve families and exponents repeats this choice for each setting.

We construct the schedule from the decoder's own predictions. For each token accepted during calibration, we compare its probability with and without the condition. A large positive log-probability ratio indicates that the condition strongly favors that token. Summing these ratios over accepted tokens and averaging across images gives the evidence accumulated at each round. The cumulative profile defines an *information clock*: progress is measured by accumulated conditional evidence rather than elapsed rounds.

INFOPROGRESS maps this clock to guidance weights using a power function with a shared exponent (Figure 1). Normalization separates the shape of the schedule from its average strength. The clock is measured once for each checkpoint, decoding budget, and token-selection policy; subsequent generation uses the resulting fixed schedule. The model weights remain unchanged, and only the global guidance strength is searched after calibration.

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.

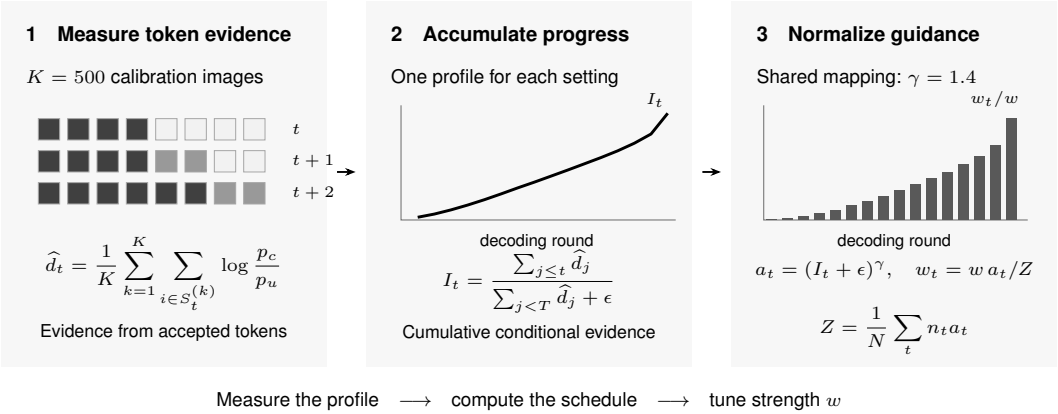

Figure 1: **From conditional evidence to guidance weights.** A 500-image calibration run measures the evidence carried by tokens accepted at each round. Accumulating this evidence gives the information clock I_t. A shared exponent converts I_t into a schedule with mean strength w. Dark, light-gray, and white squares indicate previously visible, newly accepted, and masked tokens. The plotted profile is from MaskGIT-256 at $T = 16$.

The construction recovers polynomial time schedules when measured progress follows a power law, a regime illustrated by our held-out checkpoint.

The contributions are:

- **An information clock for guidance.** We measure decoding progress from the conditional evidence carried by accepted tokens, providing a model-derived coordinate for guidance scheduling.
- **A shared schedule-construction rule.** A Monte Carlo estimate and an average-strength normalization turn this clock into per-round weights. The construction separates schedule shape from strength and recovers polynomial schedules when progress follows a power law.
- **Improved generation across settings.** The resulting schedules reduce FID by 28.2% at 32 NFE on MaskGIT-256 and recover 98.3% of the improvement achieved by a separately tuned polynomial. They remain within 0.28 FID of the best tested fixed schedule across eight settings and improve constant CFG across five MaskBit codebook sizes with one shared exponent.

2 Related Work

Parallel masked generation. MaskGIT introduced iterative masked-token synthesis (Chang et al., 2022). Token-Critic changes how generated tokens are accepted (Lezama et al., 2022), while Halton scheduling improves spatial coverage without retraining (Besnier et al., 2025a). MaskBit represents images with bit tokens (Weber et al., 2024); Meissonic and Show-o extend masked generation to text-conditioned and unified multimodal settings (Bai et al., 2025; Xie et al., 2024). InfoProgress operates within these iterative decoders and controls the guidance applied at each round.

Guidance scheduling. CFG was introduced for continuous diffusion (Ho & Salimans, 2022). Time-varying guidance weights have since been studied across diffusion models (Wang et al., 2024). MDTv2 and MaskBit use power-cosine schedules (Gao et al., 2023; Weber et al., 2024), and analyses of masked diffusion examine the effect of strong early guidance (Rojas et al., 2026). These methods parameterize guidance by decoding time. InfoProgress instead measures an information clock from accepted tokens and applies a shared mapping in that coordinate.

Dependence in parallel decoding. Parallel decoders predict token marginals within each round. Under exact-marginal parallel sampling, conditional total correlation quantifies the within-round dependence lost by this factorization (Wen et al., 2026). Our evidence statistic uses the marginal

probabilities supplied by the decoder. Appendix A gives its relation to joint evidence through an explicit within-round dependence correction.

3 An Information Clock for Guidance

3.1 Measuring conditional evidence

Let $h_t = x_{R_t}$ denote the tokens visible before round t, and let S_t contain the positions accepted at that round. Accepted tokens remain visible in subsequent rounds. The frozen model provides token probabilities $p_c(x_i \mid h_t)$ with the condition and $p_u(x_i \mid h_t)$ without it. For positive probabilities at the accepted tokens, define the round evidence

$$\widetilde{\Delta}_t = \sum_{i \in S_t} \log \frac{p_c(x_i \mid h_t)}{p_u(x_i \mid h_t)}. \tag{1}$$

Each term is positive when conditioning raises the token's probability and negative when it lowers it. The sum records the evidence carried by the tokens accepted during that round.

Proposition 1 (Accumulated evidence)**.** *For a realized decoding trajectory τ, define the token-probability products $Q_c(\tau) = \prod_t \prod_{i \in S_t} p_c(x_i \mid h_t)$ and Q_u analogously. Then*

$$\sum_{t=0}^{T-1} \widetilde{\Delta}_t = \log \frac{Q_c(\tau)}{Q_u(\tau)}. \tag{2}$$

The accumulated evidence compares the two model branches on the same sequence of accepted tokens. Appendix A gives the proof and the relation between these products and trajectory likelihoods.

3.2 Constructing the guidance schedule

A calibration run fixes the checkpoint, round budget, token-selection policy, and calibration strength. Let q_{cal} denote its trajectory distribution. The expected evidence at round t is

$$d_t^\star = \mathbb{E}_{\tau \sim q_{\mathrm{cal}}} \left[\sum_{i \in S_t(\tau)} \log \frac{p_c(x_i \mid h_t)}{p_u(x_i \mid h_t)} \right]. \tag{3}$$

We estimate this profile from K independent calibration trajectories:

$$\widehat{d}_t = \frac{1}{K} \sum_{k=1}^{K} \sum_{i \in S_t^{(k)}} \log \frac{p_c(x_i^{(k)} \mid h_t^{(k)})}{p_u(x_i^{(k)} \mid h_t^{(k)})}. \tag{4}$$

The inner sum collects the evidence within one image; the outer average combines the images into a single profile. Write $d_t = \widehat{d}_t$ and define

$$I_t = \frac{\sum_{j \le t} d_j}{\sum_{j < T} d_j + \epsilon}, \qquad a_t = (I_t + \epsilon)^\gamma, \qquad w_t = w\, a_t / Z, \tag{5}$$

where $\epsilon > 0$ stabilizes the normalization and the power map. The information clock I_t tracks cumulative evidence, and a_t assigns the relative guidance at that progress level. For $d_t \ge 0$ and $\sum_t d_t > 0$, the clock is nondecreasing; $\gamma > 1$ places more weight on later information progress. The domain conditions for general signed profiles are given in Appendix A.1.

Let $n_t = |S_t|$ be the prescribed number of tokens accepted at round t and $N = \sum_t n_t$. We normalize by

$$Z = \frac{1}{N} \sum_t n_t a_t, \qquad \frac{1}{N} \sum_t n_t w_t = w. \tag{6}$$

Thus w is the average guidance weight per accepted token. Weighting by n_t accounts for the different numbers of tokens accepted across rounds. The profile sets the timing of guidance, while w controls its overall strength. All images generated under a given setting use the same calibrated schedule.

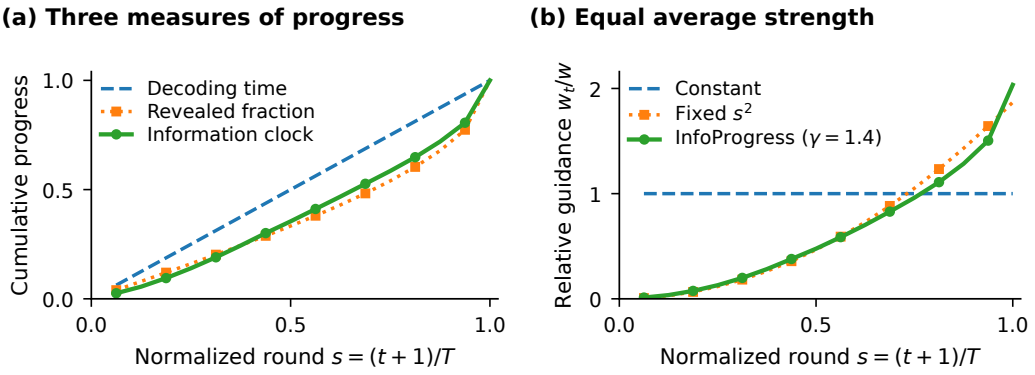

Figure 2: **Decoding time, information progress, and guidance.** MaskGIT-256, $T = 16$, with a 500-image calibration profile. **(a)** Normalized decoding time, the fraction of revealed tokens, and the information clock. The revealed fraction follows the continuous arccosine schedule. **(b)** Constant, quadratic, and INFOPROGRESS schedules at the same token-weighted mean strength w. The vertical axis shows w_t/w; $\epsilon = 0$ for this illustration.

We select $\gamma = 1.4$ once on MaskGIT-256 and reuse it in the primary experiments. Each primary profile uses $K = 500$ calibration images at the corresponding tuned constant-CFG strength. The held-out experiment fixes the calibration strength at $w_{\text{cal}} = 0.7$ in advance. After measuring the profile, we search only w; Appendix B lists the calibration settings.

Estimation accuracy and cost. Let $Y_t(\tau)$ denote the inner sum in Eq. (3). For independent trajectories with finite variance,

$$\mathbb{E}[\widehat{d}_t] = d_t^\star, \qquad \mathrm{Var}(\widehat{d}_t) = \mathrm{Var}(Y_t)/K. \tag{7}$$

The evidence estimate is unbiased, and its standard error decreases as $K^{-1/2}$. For fixed T, continuity of the normalization and power map gives almost-sure convergence of the clock and relative guidance weights (Appendix A.1). At the same round budget, a 500-image probe requires 1% as many model evaluations as a 50k-image generation run. Calibration-strength and final-strength searches are additional costs.

Relation to time schedules. Write $s = (t+1)/T$ for normalized decoding time. The power map gives a direct connection between an information clock and a polynomial time schedule.

Proposition 2 (Power-law progress)**.** *For $\epsilon = 0$, suppose $I(s) = s^\beta$ on $s \in (0, 1]$, with $\beta > 0$ and $\gamma > 0$. Then $a(s) = s^{\beta\gamma}$. A polynomial schedule s^q, $q > 0$, therefore corresponds to the information clock $I(s) = s^{q/\gamma}$.*

A quadratic schedule approximates INFOPROGRESS when the clock is close to $s^{2/\gamma}$. The held-out experiment illustrates this correspondence. Figure 2 compares decoding time, revealed-token fraction, and measured progress, together with their resulting guidance weights.

4 EXPERIMENTS

Evaluation protocol. We evaluate frozen public MaskGIT/Halton-MaskGIT and MaskBit checkpoints, with additional diagnostics on Meissonic and Show-o. Each direct FID comparison (Heusel et al., 2017) uses the same reference statistics, preprocessing, sample count, and software/hardware environment. Reported "±" values are standard deviations across seeds. We count each conditional or unconditional model forward pass as one network evaluation, so T decoding rounds use $2T$ NFE. This count covers model evaluation during generation. Calibration and the remaining evaluation costs are described in Appendix B.

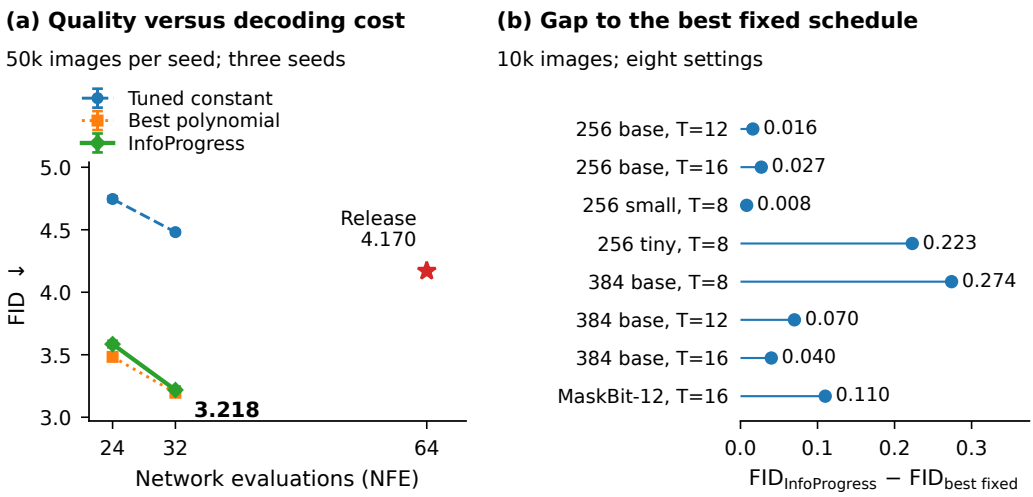

Figure 3: **Generation quality across decoding budgets and model settings.** **(a)** MaskGIT-256, evaluated with 50k images per seed and three seeds (mean ± SD). The star marks the external 64-NFE release reference, FID 4.170 (Besnier et al., 2025b). **(b)** Difference from the best tested fixed schedule across the eight 10k-image settings in Table 2. Replication details are given in Appendix C.3.

Table 1: MaskGIT-256, 50k images per seed, three seeds (mean ± SD). Fixed polynomial schedules tune both the exponent and the strength at each budget; INFOPROGRESS uses the shared exponent $\gamma = 1.4$. The 64-NFE row is an external release reference.

NFE	schedule	FID ↓	improvement over constant
24	tuned constant	4.746±.022	–
	best polynomial	**3.483±.026**	26.6%
	INFOPROGRESS	3.585±.027	24.5%
32	tuned constant	4.481±.015	–
	best polynomial	**3.196±.007**	28.7%
	INFOPROGRESS	3.218±.028	28.2%
64	release reference	4.170	–

Baselines. Every schedule receives an independent strength search. Table 1 also tunes the polynomial exponent at each decoding budget. Table 2 compares the tested constant, linear, quadratic, and power-cosine schedules. The best fixed schedule is the one with the lowest FID among those tested in the corresponding comparison. A strength search is bracketed when both a lower and a higher tested strength give worse FID; values at a search boundary are marked in the tables.

4.1 GENERATION QUALITY AT LOW EVALUATION BUDGETS

At 32 NFE, INFOPROGRESS reduces FID from 4.481 to 3.218, improving tuned constant CFG by 28.2%. The best separately tuned polynomial reaches 3.196, so INFOPROGRESS recovers 98.3% of its improvement over constant guidance. At 24 NFE, INFOPROGRESS reaches 3.585, a 24.5% improvement over constant CFG and 91.9% of the polynomial schedule's improvement (Table 1). These results use 50k images per seed and three seeds. Figure 4 shows class-matched samples.

The external release reference for the 142M-parameter, 16×16-token checkpoint is FID 4.17 at 32 rounds, or 64 NFE (Besnier et al., 2025b). Figure 3(a) and Table 1 distinguish this published value from our evaluations.



Figure 4: **Class-conditional samples at 32 NFE.** MaskGIT-256, $T = 16$, with tuned constant CFG ($w = 0.6$, top) and INFOPROGRESS ($w = 1.6$, bottom). Each column uses the same ImageNet class.

Table 2: Comparison across eight settings with a shared $\gamma = 1.4$. All schedules tune their strength independently and use 10k images. Bold marks the lowest FID in each row; "–" marks an untested schedule. Seed counts are given in Appendix C.3.

setting	constant	linear	quadratic	power-cosine	INFOPROGRESS
256 base, $T = 12$	7.082	–	**5.974**	6.274	5.990
256 base, $T = 16$	6.820	–	**5.639**	5.888	5.666
256 small, $T = 8$	9.246	–	**7.508**	7.705	7.516
256 tiny, $T = 8$	14.189	–	**11.694**	12.176	11.917
384 base, $T = 8$	15.214	**14.798**	15.079	15.855	15.072
384 base, $T = 12$	13.182	**12.919**	12.920	13.281	12.989
384 base, $T = 16$	12.441	**12.001**	12.087	12.254	12.041
MaskBit-12, $T = 16$	5.711	–	**4.825**	5.777	4.935

256 and 384 denote 16×16 and 24×24 MaskGIT token grids. MaskBit seed SDs: constant 0.022; quadratic 0.048; power-cosine 0.043; INFOPROGRESS 0.042.

4.2 A SHARED MAPPING ACROSS MODEL SETTINGS

We keep $\gamma = 1.4$ fixed while varying the decoding budget, token grid, model capacity, and tokenization. Across the eight settings in Table 2, INFOPROGRESS improves tuned constant CFG and stays within 0.008–0.274 FID of the best tested fixed schedule. The FID reductions are 1.730 on MaskGIT-small, 2.272 on MaskGIT-tiny, and 0.776 on MaskBit-12bit. Both reduced-capacity MaskGIT models use $T = 8$.

The best tested fixed schedule changes with the setting: quadratic on the 16×16 MaskGIT and MaskBit grids, and linear on the 24×24 grid. The quadratic schedule also performs well on the larger grid, trailing the linear schedule by 0.001–0.281 FID. INFOPROGRESS obtains similarly competitive schedules by recomputing the evidence profile and retaining the same mapping.

The power-cosine schedule illustrates variation across model families. It improves constant CFG on the 16×16 MaskGIT grid, whereas on MaskBit its FID is 5.777 ± 0.043, compared with 5.711 ± 0.022 for constant CFG. INFOPROGRESS reaches 4.935 ± 0.042 on MaskBit using the exponent selected on MaskGIT.

4.3 GUIDANCE TIMING AND TOKEN-SELECTION CONTROLS

Later guidance gives better FID in the independently tuned comparison in Table 3(a). The late-interval and increasing schedules improve constant CFG, whereas the early-interval and decreasing schedules worsen it. Panel (b) compares constant and quadratic guidance under two token-selection rules. The quadratic schedule improves FID by 17.3% with fixed-position selection and 21.2% with confidence-Gumbel selection. The gain from increasing guidance is present under both rules.

Table 3: Guidance timing and token-selection controls on MaskGIT-256, $T = 16$, 10k images. Each strength search has an interior best value. Positive Δ denotes an FID reduction relative to constant CFG.

(a) Independently tuned time shapes

shape	temporal emphasis	FID ↓	Δ vs. constant
late interval	late only	$5.600 \pm .075$	$+1.220$
ramp-up	increasing	5.680	$+1.140$
power-cosine	increasing	5.912	$+0.908$
constant	uniform	6.820	0
ramp-down	decreasing	7.277	-0.457
early interval	early only	7.519	-0.699

(b) Commit-policy control: constant vs. fixed s^2

commit rule	constant	fixed s^2	relative gain
fixed position	6.820	5.639	17.3%
confidence-Gumbel	9.213	7.261	21.2%

A fixed-strength experiment on MaskBit separates timing from average guidance strength. At $T = 64$ and nominal $w = 7.1$, constant CFG gives 20.656 ± 0.249 and the released power-cosine schedule gives 4.213 ± 0.028. Rescaling the cosine to the same token-weighted average strength as constant CFG gives 4.555 ± 0.027. Matching the average strength changes the scheduled FID by 0.342, compared with the original 16.443-FID gap. The large remaining difference at equal average strength demonstrates the importance of when guidance is applied at this operating point.

4.4 Transfer across codebook sizes

Five MaskBit checkpoints span codebooks from 2^{10} to 2^{18} entries, a 256-fold range in codebook size corresponding to 10–18 bits per token. With the shared $\gamma = 1.4$, InfoProgress improves tuned constant CFG by 9.8–19.0% (Table 4). Its FIDs range from 4.790 to 4.930, compared with 5.434 to 6.083 for constant CFG. All five settings use one seed and the same secondary machine. The two daggered entries continue to improve at the largest tested strength.

Table 4: Transfer across MaskBit codebook sizes, $T = 16$, 10k images, one seed per width. All runs use the same secondary machine. † marks the highest tested strength, where FID is still decreasing.

codebook	tuned constant	InfoProgress	FID gain	relative gain
2^{10}	6.083	$4.926^{\dagger}$	1.157	19.0%
2^{12}	5.716	4.930	0.786	13.8%
2^{14}	5.523	4.790	0.733	13.3%
2^{16}	5.434	4.904	0.530	9.8%
2^{18}	5.812	$4.892^{\dagger}$	0.920	15.8%

4.5 Evaluation on a held-out checkpoint

For H-MaskGIT-S-384, the calibration protocol, $\gamma = 1.4$, strength-search procedure, and screening rule were fixed before evaluation. The probe uses 500 images and constant strength $w_{\mathrm{cal}} = 0.7$. The resulting guidance schedules closely track a quadratic ramp: after scaling each final weight to one, the shape RMSE is 0.044–0.049 across $T = 8, 12, 16$. The corresponding FID differences between InfoProgress and quadratic guidance are $+0.048$, -0.045, and -0.010 (Table 5). The proximity of both shape and FID illustrates the polynomial approximation described in Proposition 2.

The $T = 8$ result uses three 10k-image runs. The $T = 12$ and $T = 16$ results are single-seed 2k-image screens. Appendix C.4 records the screening and confirmation results.

Table 5: Held-out H-MaskGIT-S-384 results. Difference is $\text{FID}_{\text{INFOPROGRESS}} - \text{FID}_{s^2}$. Shape RMSE compares the guidance schedules after scaling each final weight to one.

budget	fixed s^2	INFOPROGRESS	FID difference	shape RMSE	protocol
$T = 8$	$10.114 \pm .061$	$10.162 \pm .068$	$+0.048$	0.044	10k, 3 seeds
$T = 12$	20.687	20.642	-0.045	0.046	2k, 1 seed
$T = 16$	20.288	20.278	-0.010	0.049	2k, 1 seed

5 DISCUSSION

INFOPROGRESS schedules guidance according to the conditional evidence accumulated by a masked decoder. A short calibration run determines the schedule shape, and normalization leaves one adjustable strength. The same mapping improves constant CFG across model capacities, token grids, decoding budgets, and codebook sizes, while approaching the best fixed schedules selected separately for each setting.

The information clock also gives an interpretation of successful time schedules. Polynomial ramps approximate the construction when evidence accumulates as a power law; the held-out checkpoint exhibits this correspondence. The effect of scheduling varies across models: the Show-o diagnostic gives nearly identical FIDs for constant guidance and a mild ramp (Appendix D). Incorporating within-round dependence or updating the clock during generation would extend the present checkpoint-level calibration.

REPRODUCIBILITY STATEMENT

Appendix B gives the calibration settings, reference statistics, strength-search procedure, and checkpoint identifiers. Appendix C reports additional strength searches and model comparisons. The accompanying machine-readable tables contain the values behind every reported figure and table. Proofs and mathematical assumptions are in Appendix A.

AI USE STATEMENT

Generative AI tools assisted with language editing, literature checks, mathematical consistency checks, and figure-layout code. The sample images were not generated or altered by the editing tools. The authors are responsible for the accuracy and originality of the submitted work.

REFERENCES

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

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

Victor Besnier, Mickael Chen, David Hurych, Eduardo Valle, and Matthieu Cord. Halton-MaskGIT: Official checkpoint release, 2025b. URL https://github.com/valeoai/Halton-MaskGIT. Pre-trained model table, 142M-parameter 16x16 checkpoint; accessed September 24, 2026.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. URL https://arxiv.org/abs/2202.04200.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. URL https://arxiv.org/abs/2303.14389.

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 *NeurIPS*, 2017.

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

José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Token-critic: Refining image synthesis via iterative token sampling. In *ECCV*, 2022.

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. In *ICLR*, 2026. URL https://arxiv.org/abs/2507.08965.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2404.13040.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. *arXiv preprint arXiv:2409.16211*, 2024. URL https://arxiv.org/abs/2409.16211.

Chuling Wen, Weijie Liang, and Jian Lu. Conditional total correlation and the serial depth of adaptive parallel sampling. *arXiv preprint arXiv:2608.25505*, 2026. URL https://arxiv.org/abs/2608.25505.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. URL https://arxiv.org/abs/2408.12528.

A PROOFS AND IDENTITIES

Proof of Proposition 1. For a realized ordered partition $(S_0, \dots, S_{T-1})$ and its histories,

$$\begin{aligned}\log Q_c(\tau) - \log Q_u(\tau) &= \sum_t \sum_{i \in S_t} [\log p_c(x_i \mid h_t) - \log p_u(x_i \mid h_t)] \\ &= \sum_t \widetilde{\Delta}_t.\end{aligned}$$

When the reveal order is fixed independently of generated token values, Q_c and Q_u are the likelihoods of the respective factorized decoders. With sample-dependent acceptance, they score the accepted-token predictions along the realized trajectory. The likelihood of the complete guided sampler also depends on its selection probabilities and sampling normalizers. Under the additional assumptions that the products equal consistent joint laws $p(x \mid c)$ and $p(x)$, Bayes' rule gives

$$\log \frac{p(x \mid c)}{p(x)} = \log \frac{p(c \mid x)}{p(c)}.$$

Within-round dependence. When joint within-round laws are available, define

$$\iota_{c,t} = \log \frac{p_c(x_{S_t} \mid h_t)}{\prod_{i \in S_t} p_c(x_i \mid h_t)},$$

and define $\iota_{u,t}$ analogously. The joint evidence increment is

$$\Delta_t^{\text{joint}} = \widetilde{\Delta}_t + \iota_{c,t} - \iota_{u,t}.$$

The signed quantity $\iota_{c,t}$ measures pointwise dependence within a round. Its expectation under the conditional joint law, with history fixed, is the corresponding conditional total correlation. The same relation holds for the unconditional branch.

KL decomposition. Let q be a joint distribution with marginals q_i, and let r_i be reference marginals. When the terms are finite,

$$\begin{aligned}D_{\mathrm{KL}}\left(q \middle\| \prod_i r_i\right) &= \mathbb{E}_q \log \frac{q(X_S)}{\prod_i q_i(X_i)} + \sum_i \mathbb{E}_{q_i} \log \frac{q_i(X_i)}{r_i(X_i)} \\ &= \mathrm{TC}_q(X_S) + \sum_i D_{\mathrm{KL}}(q_i \| r_i).\end{aligned}$$

For fixed q, changing the reference marginals changes only the second term.

Proof of Proposition 2. For $\epsilon = 0$, substituting $I(s) = s^\beta$ into $a(s) = I(s)^\gamma$ gives $a(s) = s^{\beta\gamma}$. Setting $\beta = q/\gamma$ yields s^q. The average-strength normalization multiplies all weights by a common scalar.

A.1 DOMAIN AND ESTIMATION OF THE INFORMATION CLOCK

For a signed evidence profile, Eq. (5) is defined on

$$\sum_t d_t + \epsilon \neq 0, \qquad I_t + \epsilon > 0 \quad \text{for every } t, \qquad Z > 0.$$

The positivity of the power-map input allows a noninteger γ. Nonnegative round evidence with positive total evidence and $\epsilon > 0$ ensures these conditions and gives a nondecreasing clock. Signed profiles can also satisfy the domain conditions while producing a nonmonotone clock.

Write $Y_t(\tau) = \sum_{i \in S_t(\tau)} \log[p_c(x_i \mid h_t)/p_u(x_i \mid h_t)]$. With the calibration protocol fixed, Eq. (4) is the sample mean of K independent copies of Y_t. Finite variance gives $\mathbb{E}[\widehat{d}_t] = d_t^\star$ and $\mathrm{Var}(\widehat{d}_t) = \mathrm{Var}(Y_t)/K$. If $\mathbb{E}|Y_t| < \infty$, the strong law yields $\widehat{d}_t \to d_t^\star$ almost surely for each round. The vector of estimates converges jointly because T is fixed and finite.

Let $D^\star = \sum_t d_t^\star$. At a population profile satisfying $D^\star + \epsilon \neq 0$, $I_t^\star + \epsilon > 0$ for every t, and $Z^\star > 0$, the normalization and power map are continuous. The continuous mapping theorem therefore gives almost-sure convergence of the estimated clock and relative weights. This nonlinear normalization can introduce finite-sample bias into the clock, even though each $\widehat{d}_t$ is unbiased.

B EXPERIMENTAL PROTOCOL

Strength search. Each schedule has its own strength search. An interior best value is bracketed by a lower and a higher tested strength with worse FID. Search-boundary values are marked in the tables. For a fixed shape, normalization rescales the strength parameter; its best attainable FID over the full strength range is unchanged. Fixed-strength controls and independently tuned comparisons are reported separately.

FID references and sample counts. The reference file is `ImageNet_256_train_stats.pt` for ImageNet-256 and the resolution-matched `stats_384.pt` for the 24×24 token grid. Show-o and Meissonic use COCO references with 256 and 1024 preprocessing, respectively. Direct comparisons match the reference, preprocessing, and sample count. Sample count substantially affects the reported FID: one configuration gives 50.206 with 2k images and 36.439 with 10k. In the Show-o control, the seed standard deviation decreases from 0.445 at 2k images to 0.035–0.102 at 10k images.

Hardware and software controls. Matched configurations evaluated on two machines differ by 0.05–0.17 FID, with the direction changing across strengths. The machines use different PyTorch/CUDA versions. For comparison, the standard deviation of the tuned MaskBit constant baseline on one machine is 0.022. We therefore make each method comparison within a single hardware and software environment. The codebook-size experiment is a separate comparison run entirely on the secondary machine.

Evidence calibration. At each round, we record newly accepted positions and their conditional and unconditional log probabilities. We sum token evidence within each image and average over calibration images, as in Eq. (4). Primary experiments use $K = 500$ images at the strength selected by the constant-CFG FID search. The held-out protocol fixes $w_{\mathrm{cal}} = 0.7$ before evaluation. The resulting profile and shared $\gamma = 1.4$ determine the relative weights; the subsequent search varies only the global strength. Table 6 lists the calibration runs. The probe reuses the conditional and unconditional forward passes of CFG with frozen weights.

Table 6: Calibration settings. Each profile uses a constant-guidance run with the listed image count, seed, and strength. The primary experiments share $\gamma = 1.4$; Meissonic uses the exploratory exponent searches reported in Appendix C.6.

checkpoint / grid	T	probe images	seed	probe strength w
MaskGIT-256 base, 16×16	16	500	0	0.6
MaskGIT-256 base, 16×16	12	500	0	0.6
MaskGIT-256 base, 16×16	8	500	0	0.7
MaskGIT-256, 24×24 grid	8/12/16	500	0	per-budget tuned constant
MaskGIT-small	8	500	0	0.8
MaskGIT-tiny	8	500	0	0.8–1.2
MaskBit-12bit, 16×16	16	500	0	0.6
Meissonic (exploratory grid)	8	200	0	2.0

Checkpoints. Table 7 lists the public models, weight files, and recorded truncated SHA-256 identifiers. All models are used without fine-tuning.

Sampling time. Table 8 reports MaskGIT-256 sampling time on one RTX 3090 at batch size 50. The code path and machine are fixed across round budgets. Time is approximately linear in NFE: a fit to the $T = 8$ and $T = 16$ rows predicts 0.1125 seconds per image at $T = 12$, compared with the measured, rounded value of 0.113. The same fit gives 0.270 seconds per image at 64 NFE, an extrapolated time ratio of $0.270/0.144 \approx 1.9$ relative to 32 NFE.

Table 7: Public checkpoints and truncated SHA-256 identifiers of the weight files used in the experiments.

model	source	file / SHA-256 (truncated)
MaskGIT-256 base	Halton-MaskGIT release (Besnier et al., 2025a)	`ImageNet_256_base.pth` / `63c0f1c4...ba0b778`
MaskGIT-256 small	same release	`ImageNet_256_small.pth` / `52b8c7d6...4cdd8ba8`
MaskGIT-256 tiny	same release	`ImageNet_256_tiny.pth` / `50e149d6...3fc78408`
VQ tokenizer	LlamaGen VQ-16, downsample 16, class-to-image	`vq_ds16_c2i.pt` / `109aa8af...e5e0499f7d`
MaskBit tokenizer/generator, 10/12/14/16/18-bit	MaskBit release (Weber et al., 2024)	e.g. 12-bit tokenizer `3dde0244...53182170`, generator `16bd1383...5a6c3` (12-bit identifiers shown)
Meissonic transformer	Meissonic release (Bai et al., 2025)	transformer weights / `177f4ade...28280bd51`
Meissonic VQ-VAE	same release	VQ-VAE weights / `62ac839c...4b9ae0a8`
Show-o	Show-o release (Xie et al., 2024), Phi-1.5 backbone	model weights / `67ca6506...cad7eaea4`
Inception-v3 (FID/IS)	standard TF-ported weights	standard 2015-12-05 weights

Table 8: MaskGIT-256 sampling time on one RTX 3090, batch size 50. The first three rows are measured; the 64-NFE row is extrapolated.

T	NFE	s / image	images / s
8	16	0.081	12.3
12	24	0.113	8.9
16	32	0.144	6.9
32	64	0.270 (extrapolated)	3.7 (extrapolated)

C Additional Quantitative Results

C.1 MaskGIT-256 strength searches

At $T = 8$, the constant-guidance search gives

w	0.45	0.6	0.7	1.0
FID	10.252	8.144	7.656	8.198

The best tested strength is the interior value $w = 0.7$. At $T = 12$ and $T = 16$, constant CFG attains FIDs 7.082 and 6.820, respectively, at $w = 0.6$.

The quadratic schedule improves both token-selection policies at $T = 16$ with 10k images. Fixed-position selection improves from 6.820 (constant, $w = 0.6$) to 5.639 (s^2, $w = 1.6$), a 17.3% reduction. Confidence-Gumbel selection improves from 9.213 (constant, $w = 0.4$) to 7.261 (s^2, $w = 1.0$), a 21.2% reduction. All four searches are bracketed.

C.2 Time-schedule controls

With strength tuned independently for each schedule on MaskGIT-256 base at $T = 16$, the FIDs are 5.600 for a late interval (three-seed SD 0.075), 5.680 for an increasing ramp, 5.912 for power-cosine, 6.820 for constant, 7.277 for a decreasing ramp, and 7.519 for an early interval. The code uses remaining time $x = 1 - s$; schedule names in the paper refer to increasing decoding progress s. At the shared nominal strength $w = 0.6$, constant guidance is best and the late interval is worst; at $w = 2.4$, their ordering reverses. Independent strength tuning separates this scaling effect from schedule timing.

C.3 Comparison with fixed schedules

Table 2 compares the fixed schedules evaluated on the primary machine. Each schedule tunes its strength. Linear schedules are evaluated on the 24×24 grid; quadratic and power-cosine searches are bracketed in all MaskGIT rows. MaskBit results are three-seed means, as are the InfoProgress results for the small and tiny models. The remaining fixed-schedule entries are generally single-seed 10k-image searches.

Quadratic guidance has the lowest tested FID on the 16×16 and MaskBit grids. Linear guidance has the lowest FID on the 24×24 grid, where the quadratic gap is 0.001–0.281. The InfoProgress gap to the lowest row value is 0.008–0.274 across all eight settings. A single exponent and a newly measured profile per setting give these results.

C.4 Held-out checkpoint protocol

H-MaskGIT-S at 384×384 resolution is the held-out checkpoint. The probe size $K = 500$, calibration strength $w_{\text{cal}} = 0.7$, $\gamma = 1.4$, normalization, and automatic strength-bracketing procedure were fixed before its FID evaluation. Both methods use the same strength-search procedure. The screening protocol specifies a 0.15–0.20 FID improvement threshold for moving from a single-seed 2k-image screen to three 10k-image runs. Table 5 includes all three tested budgets.

At $T = 8$, the initial screen favors InfoProgress by 0.285 FID and triggers the larger evaluation. The three-seed 10k-image means are 10.162 for InfoProgress and 10.114 for the quadratic schedule, a difference of $+0.048$. The $T = 12$ and $T = 16$ screen differences are -0.045 and -0.010, below the expansion threshold. After each schedule's final weight is scaled to one, its RMSE from s^2 is 0.044, 0.046, and 0.049 for $T = 8, 12, 16$. The measured schedules and fixed quadratic ramps are numerically close in both shape and FID in this experiment.

Table 9: MaskBit-12bit, $T = 16$, 10k images, three seeds per schedule. Each strength search has an interior best value.

shape	best w	FID (mean±SD)
s^2 ramp	2.5	$\mathbf{4.825 \pm 0.048}$
INFOPROGRESS	1.0	4.935 ± 0.042
constant	0.6	5.711 ± 0.022
released power-cosine	6.0	5.777 ± 0.043

C.5 MASKBIT STRENGTH SEARCHES AND AVERAGE-STRENGTH CONTROL

The strength-search values used to bracket the scheduled optima are

	tested w			FID		
s^2	1.6	2.5	3.0	4.917	4.873	4.994
INFOPROGRESS	0.6	1.0	1.6	6.084	4.983	5.050
power-cosine	4	6	8	5.920	5.800	5.840

The constant-guidance grid is

w	0.2	0.4	0.6	0.8	1	2
FID	7.557	5.720	5.711	6.530	7.845	13.020
w	3	4.5	6	7.1	9	12
FID	16.106	18.616	19.492	19.548	19.700	19.550

A single evaluation using the MaskGIT evidence profile on MaskBit gives FID 4.922 at $w = 1.6$.

Using the released power-cosine configuration at $T = 64$, $w = 7.1$ gives FID 4.174 and inception score 343.7 with 10k images. In the three-seed fixed-strength control at nominal $w = 7.1$, constant CFG gives 20.656 ± 0.249, power-cosine gives 4.213 ± 0.028, and rescaling power-cosine to the constant schedule's token-weighted mean strength gives 4.555 ± 0.027. Equalizing average strength changes the scheduled FID by 0.342, about 2% of the original 16.443-FID gap. At this fixed operating point, most of the difference remains after average strength is matched.

C.6 EXPLORATORY MEISSONIC EXPONENT SEARCHES

An initial 500-image search uses ImageNet class-name prompts and `stats_meis.pt`. The observed values are

schedule	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 1.0$	$\gamma = 1.4$	constant
w	2	3	3	4	4	2
FID	109.299	110.089	109.572	112.612	116.815	114.286

The strength searches in this protocol have open boundaries.

A separate 2k-image, single-seed search uses COCO prompts, canonical preprocessing, and `stats_coco30k_pre1024.pt`. Constant CFG at $w = 2$ gives FID 52.505. The search grid is $\gamma \in \{0.15, 0.2, 0.3, 0.4, 0.5\}$ and $w \in \{1.5, 2.0, 2.5\}$. The $\gamma = 0.15$ curve has a bracketed best value of 51.159 at $w = 2$. The $\gamma = 0.2, 0.3, 0.5$ curves continue to improve at the upper strength boundary; the lowest observed FID is 50.098 at $\gamma = 0.5, w = 2.5$. These two exploratory searches use separate prompt and reference protocols.

C.7 CODEBOOK-SIZE COMPARISON ON THE SECONDARY MACHINE

The five MaskBit checkpoints use codebooks of 2^{10} to 2^{18} entries with a common architecture, dataset, and training setup. We evaluate all five on the secondary machine at $T = 16$, with 10k images and one seed per width. Table 10 reports this within-machine comparison.

The best tested constant strength is $w = 0.4$ for all five widths. Its FID ranges from 5.434 at 16 bits to 6.083 at 10 bits, a spread of 0.649. With the shared $\gamma = 1.4$, INFOPROGRESS gives FIDs

Table 10: MaskBit codebook comparison on the secondary machine, $T = 16$, 10k images, one seed per setting. Constant CFG uses $w = 0.4$; quadratic schedules use $w = 2.24$. † marks a search boundary and ‡ a single tested strength. "–" denotes an untested schedule.

codebook	constant	INFOPROGRESS	quadratic	power-cosine
2^{10}	6.083	$4.926^{\dagger}$	4.8297	$5.8794^{\ddagger}$
2^{12}	5.716	4.930	4.8638	$5.7782^{\ddagger}$
2^{14}	5.523	4.790	4.7177	–
2^{16}	5.434	4.904	$4.7638^{\dagger}$	–
2^{18}	5.812	$4.892^{\dagger}$	$5.1254^{\dagger}$	–

4.790–4.930 and improves constant guidance by 9.8–19.0%. The 10- and 18-bit INFOPROGRESS searches remain open at the upper strength boundary. The 12-bit FID reduction is 0.786 on this machine; the independently measured primary-machine reduction is $5.711 - 4.935 = 0.776$.

The quadratic searches are bracketed at 10, 12, and 14 bits, where their FIDs are 0.066–0.096 below INFOPROGRESS. They remain open at 16 and 18 bits. Power-cosine has one evaluated strength at each of 10 and 12 bits.

D SHOW-O DIAGNOSTIC

With aligned COCO preprocessing, constant guidance at $w = 12$ gives seed FIDs 29.970, 29.855, and 30.059 (mean 29.961, SD 0.102). The mildest tested ramp, $q = 0.25$ at $w = 12$, gives 29.919, 29.973, and 29.986 (mean 29.959, SD 0.036). From these rounded seed values, the paired improvement is 0.002 FID, with SD 0.104 and standard error 0.060. The mean difference is small relative to the seed variation.

At $w = 12$, top-1 token probabilities range from 0.12 to 0.39 and entropies from 2.4 to 4.6 nats, compared with the codebook maximum $\log 8192 = 9.01$. Under the same evaluation transform, real COCO images have FID 4.435 and their MAGVIT-v2 reconstructions have FID 6.952. Generation FID remains near 29.96. Precision and recall are also similar between the guidance schedules: 0.483/0.433 for constant guidance and 0.482/0.428 for the ramp.

E ADDITIONAL MECHANISTIC TESTS

Forward-KL projection diagnostic. The reference decoder uses $(T_0 = 32, w_0 = 0.6)$; the target uses $T = 16$ on MaskGIT-256 base with fixed-position token selection. We replay 128 reference samples through the target partition. At each round, bisection solves $\mathbb{E}_{r_t^w}[\ell_t] = \widehat{\mathbb{E}}_q[\ell_t]$, where $\ell_t = \log p_c - \log p_u$, r_t^w is the guided distribution, and the right-hand side is estimated from the replayed samples. Negative roots are allowed. With the target reset to $T = T_0$, the estimated mean strength is 0.561, compared with the reference value 0.6. The schedules estimated from $K = 64$ and $K = 128$ samples have correlation 0.972. Isotonic regression gives the monotone variant.

E.1 INTERVAL AND LOCAL-ALLOCATION OBSERVATIONS

Exploratory interval interventions on the 16×16 and 24×24 grids give a reduction in early-guidance benefit by factors of 12 and 24, respectively, compared with factors of 1.6 and 2.1 for late guidance. The class-identity diagnostic is nearly saturated after 60% of the trajectory. In a local allocation test, each of 16 matched perturbations around the tuned increasing schedule changes FID by at most 0.08 in magnitude. The main-text timing and token-selection controls are reported in Table 3.

Table 11: Additional diagnostics of guidance timing and strength selection.

question	observation
Sensitivity to token selection	Quadratic-guidance gains are 17.3% with fixed-position selection and 21.2% with confidence-Gumbel selection; both strength searches are bracketed.
Effect of the Show-o FID reference	Aligned preprocessing changes FID values; the best tested strength remains at $w \geq 12$.
Token certainty under strong Show-o guidance	Top-1 probabilities are 0.12–0.39 and entropies are 2.4–4.6 nats at $w = 12$.
Show-o reconstruction and generation quality	Reconstruction FID is 6.952; generation FID is about 29.97.
Precision as a predictor of scheduling gain	The correlation between precision headroom and gain is $r = 0.022$ across six configurations.
Guidance strength as a predictor of gain	Show-o has the largest best-tested strength w^* and nearly identical constant and ramp FIDs.
Token confidence versus remasking diagnostics	Overlap is 0.246 and Spearman correlation is -0.011 against exact remasking probes.
Strength selected from total evidence	The derived strengths give higher FID than tuned constant CFG at $T = 8, 12, 16$.
Strength selected by per-round marginal KL	Held-out cross-entropy improves while FID increases by 12–14.