Masked Diffusion Models are Secretly Learned-Order Autoregressive Models

Learning a distribution over decoding orders

Paper authors: Prateek Garg, Bhavya Kohli, Sunita Sarawagi

arXiv:2511.19152v1 / 24 November 2025

Research discussion, 14 October 2026 / 20–30 minutes

Position-specific clocks reweight autoregressive learning tasks.
The theory is precise; improvements in tabular fidelity are mixed.

Reading presentation, not an author presentation. Versioned paper: arxiv.org/abs/2511.19152v1.

Can the training objective learn a useful order?

The starting point

A standard MDM trains equally over all token orders.

- Tables and graphs lack a canonical sequence.
- Finite models may find some conditional predictions easier.

The proposal

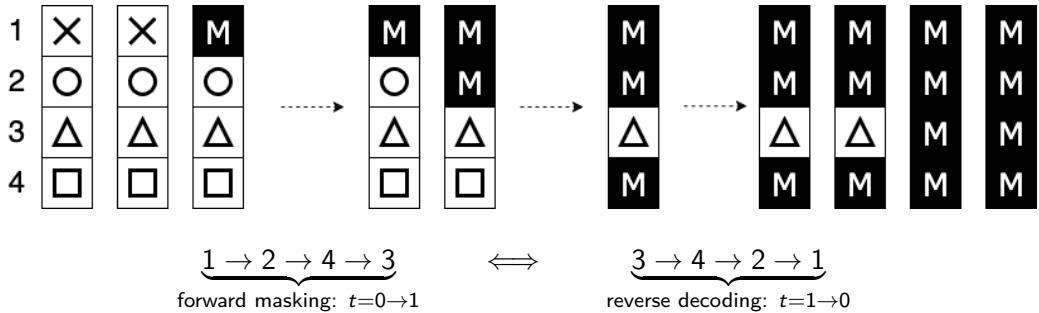
Learn a separate masking schedule for each position.

- Change the distribution of orders during training.
- Sample from that distribution at generation time.

Scope: the order distribution is independent of sample content.

Paper, Sec. 1, PDF pp. 1–2; Appendix A, PDF p. 9.

Reverse generation undoes a masking trajectory



- Each position changes once: token → mask in the forward process.
- The last position masked is the first position decoded.

Original Fig. 1, PDF p. 2. Arrows below the image are an explanatory annotation.

The schedule chooses when; the network chooses what

For one position, α_t is the probability of retaining its value.

$$q(x_t \mid x_s) = \frac{\alpha_t}{\alpha_s}\delta_{x_s} + \left(1 - \frac{\alpha_t}{\alpha_s}\right)\delta_{\mathrm{M}}, \qquad s < t.$$

A masked position in the reverse process

$$p_\theta(x_s \mid x_t = \mathrm{M}) = \underbrace{\frac{1-\alpha_s}{1-\alpha_t}}_{\text{stay masked}}\ \delta_{\mathrm{M}} + \underbrace{\frac{\alpha_s - \alpha_t}{1-\alpha_t}}_{\text{unmask}}\ \mu_\theta(\mathbf{x}_t, t).$$

An unmasked position is copied unchanged. Other visible positions provide context.

Paper, Sec. 2, Eqs. (1)–(2), PDF p. 2; mean parameterization.

Every position has a random transition clock

Let T_ℓ be the noise-time coordinate of position ℓ's transition.

$$\boxed{\mathbb{P}(T_\ell \le t) = 1 - \alpha_{t,\ell}, \qquad f_\ell(t) = -\alpha'_{t,\ell}.}$$

- At forward time t, being masked means $T_\ell \le t$.
- Independent forward masking gives independent transition clocks.
- Sample $U_\ell \sim U(0,1)$, then $T_\ell = \alpha_\ell^{-1}(U_\ell)$.

Time convention used in this presentation

Decode in **decreasing** T_ℓ, since reverse time runs from 1 to 0.

Proposition 3.2 and Algorithm 1, PDF p. 3; proof, pp. 9–10. Exact endpoints: $\alpha_{0,\ell} = 1$, $\alpha_{1,\ell} = 0$. See backup for the paper's order notation.

Different clocks assign different weights to orders

For decoding order π, define

$$\Omega_{\pi}^{\mathrm{dec}} = \{\mathbf{t} : t_{\pi(1)} > \cdots > t_{\pi(L)}\}.$$

$$P_{\mathrm{dec}}(\pi) = \int_{\Omega_{\pi}^{\mathrm{dec}}} \prod_{\ell=1}^{L} [-\alpha'_{t_\ell,\ell}]\, d\mathbf{t}.$$

One shared schedule

Exchangeable continuous clocks.

$P_{\mathrm{dec}}(\pi) = 1/L!$ for every order.

Position-specific schedules

Relative clock speeds change.

Some orders become more likely.

Paper, Eq. (7), PDF p. 3; written here using descending decoding times.

The loss weights each masked prediction by its rate

Let $M_\ell = \mathbf{1}\{x_t^\ell = \mathrm{M}\}$ and $c_\ell = -\log \mu_\ell^{x_0^\ell}(\mathbf{x}_t; \theta)$.

$$\mathcal{L} = \int_0^1 \sum_{\ell=1}^{L} \underbrace{\frac{-\alpha'_{t,\ell}}{1-\alpha_{t,\ell}}}_{\text{reverse unmasking rate}} \mathbb{E}_q[M_\ell c_\ell]\, dt.$$

- The rate and cross-entropy are both nonnegative.
- A visible token is copied; only masked positions contribute.
- Learning a schedule changes both the masking distribution and the weight.

Paper, Eq. (8), PDF p. 4. The masking indicator makes the carry-over convention explicit.

The event rate turns the loss into a sum over jumps

$$\underbrace{\mathbb{P}(M_\ell = 1)}_{1-\alpha_{t,\ell}} \; \underbrace{\frac{-\alpha'_{t,\ell}}{1-\alpha_{t,\ell}}}_{\text{unmasking rate}} = \underbrace{-\alpha'_{t,\ell}}_{\text{transition-time density}} .$$

1. Integrating the weighted masked loss counts each token's jump once.
2. Continuous independent clocks have no simultaneous jumps almost surely.
3. At $T_{\pi(i)}$, positions $\pi(< i)$ have already been decoded.
4. Group trajectories by their decoding order π.

Explanatory derivation from Eq. (8), PDF p. 4. The paper instead telescopes trajectory terms in Appendix B, pp. 10–11.

Without time input, the objective is weighted AR loss

Define an autoregressive log-likelihood for each order:

$$LL_\pi(x,\theta) = \sum_{i=1}^{L} \log \mu(x_{\pi(i)} \mid x_{\pi(<i)};\theta).$$

Proposition 3.1, with a consistent decoding convention

$$\boxed{\mathcal{L}(x) = -\mathbb{E}_{\pi\sim P_{\mathrm{dec}}}[LL_\pi(x,\theta)].}$$

- One network shares parameters across all orders.
- $x_{\pi(<i)}$ is a full-length input with other positions masked.
- With a time-dependent network, retain $\mathbb{E}_{T_{\pi(i)}|\pi}$ inside the sum.

Paper, Eqs. (5)–(6), PDF p. 3. Continuous time, absorbing masks, independent clocks.

An average AR loss is still a likelihood bound

If we sample an order and then generate tokens, the marginal model is

$$p_{\theta,w}(x) = \sum_{\pi} P_w(\pi) \underbrace{\prod_i \mu(x_{\pi(i)} \mid x_{\pi(<i)}; \theta)}_{p_\theta(x|\pi)}.$$

Jensen's inequality gives

$$\boxed{-\log p_{\theta,w}(x) \leq \mathbb{E}_{\pi\sim P_w}[-\log p_\theta(x \mid \pi)] = \mathcal{L}(x).}$$

- The theorem identifies the *training objective* exactly.
- It does not replace a log of a mixture with an average of logs.

Reader derivation from the negative ELBO, Eq. (3), p. 2, and Eq. (6), p. 3. Assumes no time input and generation one token at a time.

Power schedules turn weights into order preferences

$$\alpha_{t,\ell} = 1 - t^{w_\ell}, \quad w_\ell > 0; \qquad T_\ell \overset{d}{=} U_\ell^{1/w_\ell}.$$

Two-position example

$$\mathbb{P}(1 \text{ decoded first}) = \frac{w_1}{w_1 + w_2}.$$

With $(w_1, w_2) = (1, 3)$:

$$P(1 \to 2) = \tfrac{1}{4}, \quad P(2 \to 1) = \tfrac{3}{4}.$$

- Larger w_ℓ: masked later in forward time; decoded earlier in reverse time.
- Equal weights recover uniform orders.
- Relative weights matter; the order remains stochastic.

Schedule family: Sec. 4, PDF p. 4. Probability calculation and example: reader derivation. General Plackett–Luce formula in backup.

Learning clocks requires gradients through masking

With the power schedule,

$$M_\ell \sim \mathrm{Bernoulli}(t^{w_\ell}), \qquad \frac{-\alpha'_{t,\ell}}{1-\alpha_{t,\ell}} = \frac{w_\ell}{t}.$$

- A schedule parameter affects the loss weight *and* the discrete mask distribution.
- Ordinary differentiation through the weight misses the second effect.
- The paper uses RLOO: REINFORCE with a leave-one-out baseline.

Training and generation

Train on randomly masked rows. Generate by sampling clocks, sorting them, and filling one position at each event.

Paper, Sec. 4, Eq. (8), PDF p. 4; Algorithm 1, p. 3. The RLOO sample count and full optimizer setup are not specified.

The direct ablation compares two 86K-parameter models

Model	Schedule	Role
MDM	Fixed, $\alpha_{t,\ell} = 1 - t$	Direct control
MDM(LS)	Learned, $\alpha_{t,\ell} = 1 - t^{w_\ell}$	Proposed variant
TabDiff	Learned mixed-type diffusion	Larger strong baseline

- Six datasets: Adult, Default, Beijing, Shoppers, Magic, News.
- MDM and MDM(LS): same hyperparameter set; roughly 86K parameters.
- Other baselines: STaSy, CoDi, TabDDPM, TabSyn; scores taken from TabDiff.
- TabDiff is reported as 10M parameters; this is not a matched-budget comparison.

Paper, Sec. 4.1, PDF p. 4; Appendix C.1, p. 11. MDM variants: reported means and standard deviations over 10 runs.

Four metrics separate fidelity from coverage

Metric	What it measures	Direction
Trend	Pairwise column relationships	Higher
Shape	Marginal distributions of individual columns	Higher
α-Precision	Realism of generated samples	Higher
β-Recall	Coverage of the real data distribution	Higher

Best validation loss is a separate, lower-is-better quantity.

High precision can coexist with limited coverage.

Paper, Tables 1–3, PDF pp. 4–5; metric descriptions referred to TabDiff (2410.20626v3), Appendix A.2. Metric α is unrelated to the noise schedule.

Best validation loss improves on four of six datasets

Dataset	MDM	MDM(LS)	LS − MDM
Adult	13.1861	13.1524	−0.0337
Default	19.4247	18.3272	−1.0975
Beijing	16.7610	17.0060	+0.2450
Shoppers	16.4246	16.4604	+0.0358
Magic	12.6114	12.4712	−0.1402
News	49.9699	48.9401	−1.0298

Default and News show the largest absolute reductions.

Beijing and Shoppers become worse.

Paper, Fig. 2a, PDF p. 5. Best validation losses; differences calculated from reported values. Negative difference is better.

Fidelity changes are mixed; coverage often decreases

LS − MDM	Trend	Shape	α-Precision	β-Recall
Adult	+0.0031	+0.0027	+0.0153	−0.0061
Default	+0.0063	+0.0043	+0.0122	−0.0029
Beijing	−0.0030	−0.0025	+0.0152	−0.0147
Shoppers	+0.0007	+0.0003	−0.0009	−0.0148
Magic	−0.0096	−0.0072	−0.0181	−0.0042
News	+0.0015	+0.0002	+0.0093	+0.0068
Improved means	4/6	4/6	4/6	1/6

Magic: lower validation loss, yet all four fidelity metrics decline.

Differences calculated from Tables 1–3, PDF pp. 4–5; Tables 4–7, pp. 11–12. Positive is better. Absolute score differences, not relative percentages or significance tests.

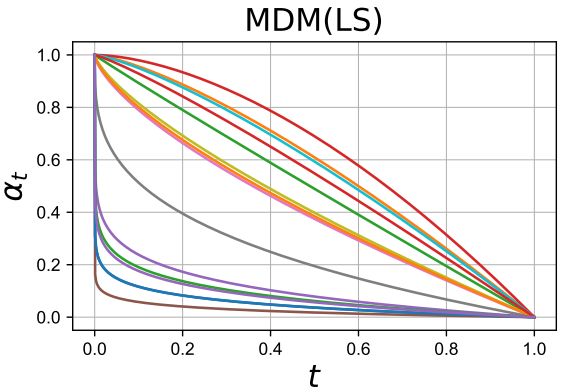
News has the best precision, but limited coverage

News dataset	α-Precision ↑	β-Recall ↑
MDM	0.9779	0.3405
MDM(LS)	**0.9872**	0.3473
TabDiff	0.9736	0.4210
TabSyn	0.9505	**0.4301**

- MDM(LS) attains the highest News precision among all compared methods.
- Its recall remains below both strong baselines shown here.
- TabDiff also leads Trend on all six datasets.

Paper, Tables 1 and 3, PDF pp. 4–5. Selected reported means; uncertainty examples in backup.

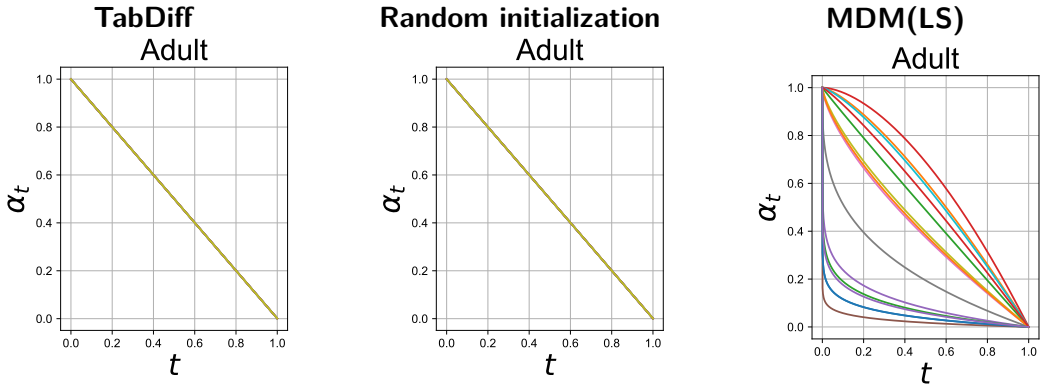
Adult schedules visibly separate across columns



- Horizontal: forward noise time t.
- Vertical: survival probability α_t.
- Different curves imply different clock distributions.
- Higher curves favor earlier reverse decoding.

Original Fig. 2b, PDF p. 5. Colors are not labeled with column names; the plot does not identify semantic or causal order.

The authors' TabDiff reproduction keeps similar clocks



Evidence from this reproduction: random initialization did not separate the TabDiff curves. TabDiff masks categorical columns; MDM(LS) masks all columns.

Adult panels cropped from original Figs. 3–5, PDF p. 12; Appendix C.2, pp. 12–13. A comparison of implementations, not an isolated causal explanation.

The contribution is a mechanism for learning orders

1. **Theory:** independent position clocks induce a distribution over orders; without time input, the loss is exactly its weighted AR objective.
2. **Evidence:** schedules separate and validation loss improves on $4/6$ datasets; downstream fidelity is mixed, with recall improving on $1/6$.
3. **Next question:** can better variance control turn flexible orders into more reliable gains in fidelity and coverage?

State-independent orders, a restricted schedule family, and limited implementation details define the present scope.

Paper, Secs. 3–5, PDF pp. 3–5; Appendix C, pp. 11–13. Increased training variance is the authors' hypothesis, not a measured causal result.

Backup: power schedules imply Plackett–Luce orders

A reader derivation for the paper's family $\alpha_{t,\ell} = 1 - t^{w_\ell}$:

$$F_\ell(t) = t^{w_\ell}, \quad Z_\ell = -\log T_\ell, \quad \mathbb{P}(Z_\ell > z) = e^{-w_\ell z}.$$

Thus Z_ℓ are independent exponential clocks with rates w_ℓ.

$$\boxed{P_w(\pi) = \prod_{i=1}^{L} \frac{w_{\pi(i)}}{\sum_{j=i}^{L} w_{\pi(j)}}.}$$

- Decreasing T is increasing Z; exponential races select by relative rate.
- Scaling all w_ℓ by the same positive constant leaves orders unchanged.
- This family cannot assign arbitrary probabilities to all $L!$ permutations.

Derived from Proposition 3.2 and the schedule family in Sec. 4, PDF pp. 3–4. The closed-form Plackett–Luce interpretation is not stated in the paper.

Backup: keep masking and decoding directions consistent

Convention	Time ordering	Visible context at step i
Forward masking ρ	$T_{\rho(1)} < \cdots < T_{\rho(L)}$	$\rho(> i)$
Reverse decoding π	$T_{\pi(1)} > \cdots > T_{\pi(L)}$	$\pi(< i)$

- The paper's Eq. (7) uses increasing times, while Eqs. (5)–(6) condition on a decoded prefix.
- Appendix B, p. 11, explicitly describes the increasing-time prefix as masked.
- This presentation uses decreasing times for decoding throughout.

Paper, PDF pp. 3 and 11. Appendix B also has inconsistent transition-table entries on p. 10 and missing minus signs in Eqs. (23)–(24); the event-rate derivation avoids these issues.

Backup: RLOO retains the discrete-distribution gradient

For the loss $g_w(M)$ at a fixed row and time:

$$\nabla_w \mathbb{E}_{M\sim q_w} g_w(M) = \mathbb{E}\big[\nabla_w g_w(M) + g_w(M)\nabla_w \log q_w(M)\big].$$

For $K > 1$ independent samples, set

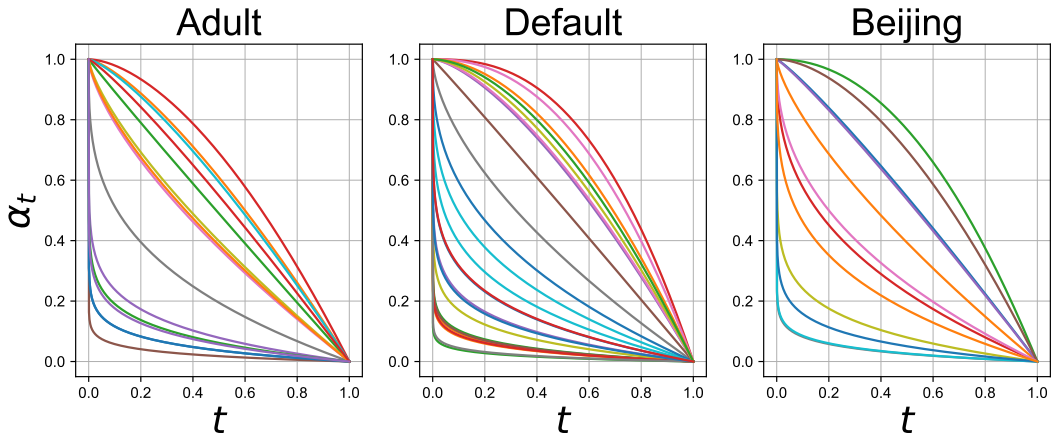
$$b_{-k} = \frac{1}{K-1}\sum_{j\neq k} g_w(M_j).$$

$$\widehat{\nabla_w \mathcal{L}} = \frac{1}{K}\sum_k \Big[\nabla_w g_w(M_k) + (g_w(M_k) - b_{-k})\nabla_w \log q_w(M_k)\Big].$$

Use stopped-gradient coefficients in the score term; compute the explicit derivative separately.

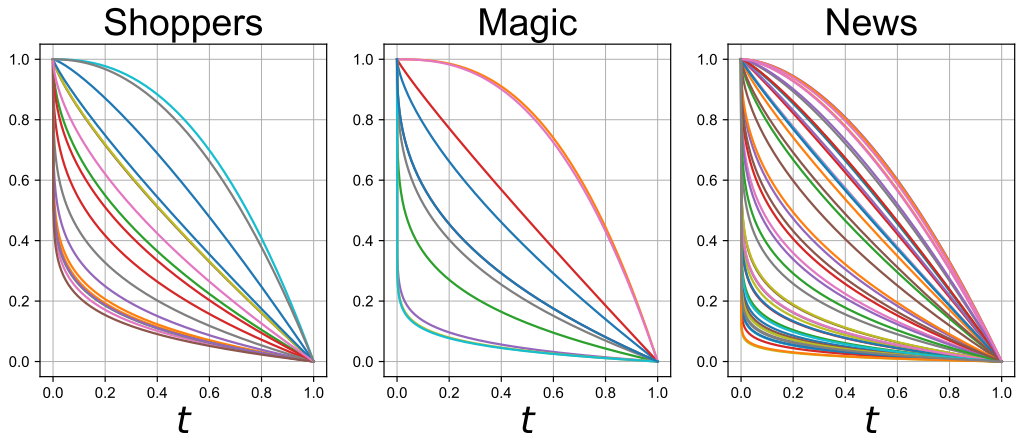
Explanatory score-function identity. Paper, Sec. 4, PDF p. 4, cites Kool et al. (2019) for RLOO but does not specify K or this full estimator.

Backup: learned schedules on Adult, Default, Beijing



Original Fig. 5, PDF p. 12, first three panels; values and axes unchanged. Horizontal: forward time t. Vertical: survival probability α_t.

Backup: learned schedules on Shoppers, Magic, News



Original Fig. 5, PDF p. 12, last three panels; values and axes unchanged. Horizontal: forward time t. Vertical: survival probability α_t.

Backup: reported uncertainty supports a nuanced reading

Dataset	Metric	MDM	MDM(LS)
Default	Trend	0.9642 ± 0.0016	0.9705 ± 0.0028
Magic	Trend	0.9724 ± 0.0026	0.9628 ± 0.0070
News	α-Precision	0.9779 ± 0.0016	0.9872 ± 0.0012
Beijing	β-Recall	0.5309 ± 0.0031	0.5162 ± 0.0027

- These are the paper's means and standard deviations over 10 runs.
- Improvements and regressions occur across distinct quality criteria.
- A best validation loss is a different reporting statistic.

Paper, Appendix C.1, Tables 4, 6, 7, PDF pp. 11–12; Fig. 2a, p. 5.

References and discussion prompts

Main paper
Prateek Garg, Bhavya Kohli, Sunita Sarawagi.
Masked Diffusion Models are Secretly Learned-Order Autoregressive Models.
arXiv:2511.19152v1, 2025. EurIPS PriGM Workshop.

Metric background
Juntong Shi et al. *TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation.*
arXiv:2410.20626v3, Appendix A.2.

- How much flexibility is lost by restricting orders to a global clock family?
- Which variance measurements would explain the loss–fidelity mismatch?
- Would fixed nonuniform schedules separate learning from order selection?