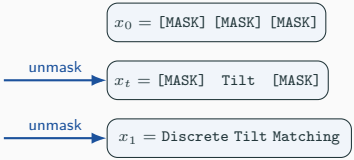
How do we fine-tune masked diffusion LLMs when sequence-level likelihoods are intractable?

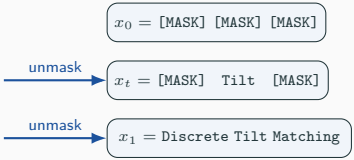
How do we fine-tune masked diffusion LLMs when sequence-level likelihoods are intractable?

Masked diffusion LLMs generate by **iteratively unmasking** tokens, enabling flexible any-order inference.

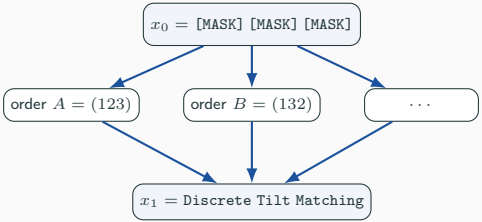


How do we fine-tune masked diffusion LLMs when sequence-level likelihoods are intractable?

Masked diffusion LLMs generate by **iteratively unmasking** tokens, enabling flexible any-order inference.



Standard RL post-training is built around **sequence-level likelihoods**.

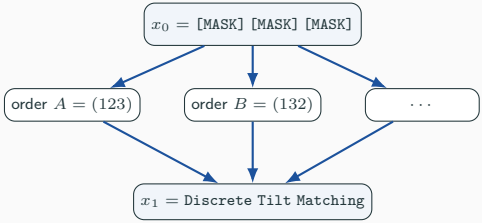


One sequence can be reached through exponentially many unmasking trajectories.

2. Why is Discrete Tilt Matching (DTM) needed

- RL objectives such as GRPO need $\log \rho_\theta(x_1)$ or likelihood ratios.

Standard RL post-training is built around **sequence-level likelihoods**.

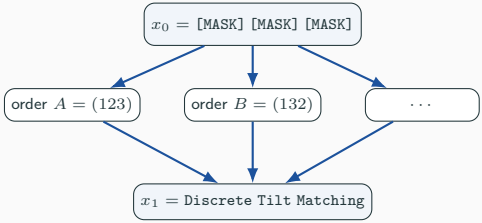


One sequence can be reached through exponentially many unmasking trajectories.

2. Why is Discrete Tilt Matching (DTM) needed

- RL objectives such as GRPO need $\log \rho_\theta(x_1)$ or likelihood ratios.
- For masked diffusion models, $\rho_\theta(x_1)$ sums over all reveal orders.

Standard RL post-training is built around **sequence-level likelihoods**.

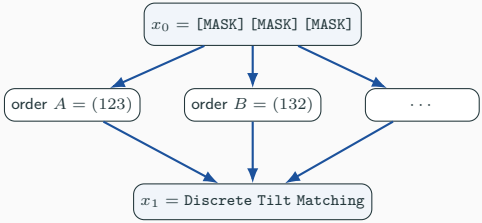


One sequence can be reached through exponentially many unmasking trajectories.

2. Why is Discrete Tilt Matching (DTM) needed

- RL objectives such as GRPO need $\log \rho_\theta(x_1)$ or likelihood ratios.
- For masked diffusion models, $\rho_\theta(x_1)$ sums over all reveal orders.
- Exact likelihood evaluation is impractical; common alternatives use biased surrogates or high-variance estimators.

Standard RL post-training is built around **sequence-level likelihoods**.



One sequence can be reached through exponentially many unmasking trajectories.

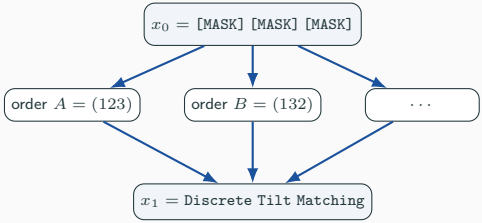
2. Why is Discrete Tilt Matching (DTM) needed

- RL objectives such as GRPO need $\log \rho_\theta(x_1)$ or likelihood ratios.
- For masked diffusion models, $\rho_\theta(x_1)$ sums over all reveal orders.
- Exact likelihood evaluation is impractical; common alternatives use biased surrogates or high-variance estimators.

Gap in prior work

Post-training needs an objective native to dLLMs: **tractable local states** rather than sequence-level marginal likelihoods.

Standard RL post-training is built around **sequence-level likelihoods**.



One sequence can be reached through exponentially many unmasking trajectories.

3. Discret Tilt Matching

Takeaway

DTM replaces sequence-level likelihood policy optimization with **local unmasking posterior matching** under reward tilting.

3. Discret Tilt Matching

Takeaway

DTM replaces sequence-level likelihood policy optimization with **local unmasking posterior matching** under reward tilting.

- The likelihood $\rho_\theta(x_1)$ is sampled with the local unmasking posterior $\pi_\theta(\,v\,|x_t, i)$, the likelihood of having token v at position i at partially masked state x_t

3. Discret Tilt Matching

Takeaway

DTM replaces sequence-level likelihood policy optimization with **local unmasking posterior matching** under reward tilting.

- The likelihood $\rho_\theta(x_1)$ is sampled with the local unmasking posterior $\pi_\theta(v\,|x_t, i)$, the likelihood of having token v at position i at partially masked state x_t
- Instead of estimating $\rho_\theta(x_1)$, DTM works with $\pi_\theta(v\,|x_t, i)$ solely.

3. Discret Tilt Matching

Takeaway

DTM replaces sequence-level likelihood policy optimization with **local unmasking posterior matching** under reward tilting.

- The likelihood $\rho_\theta(x_1)$ is sampled with the local unmasking posterior $\pi_\theta(v\,|x_t, i)$, the likelihood of having token v at position i at partially masked state x_t
- Instead of estimating $\rho_\theta(x_1)$, DTM works with $\pi_\theta(v\,|x_t, i)$ solely.

Why care?

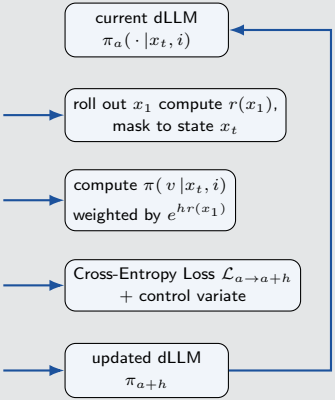
At scale on LLaDA-8B-Instruct, DTM gives **strong gains** on Sudoku and Countdown while staying competitive on math reasoning.

3. Discrete Tilt Matching

Fine-tuning as Tilting

- Want to go from base model $\rho_0(x)$ to $\rho_A(x)$, where $\rho_A(x) = \rho_0(x)e^{hr(x)}$ with reward $r(x)$.

DTM algorithm



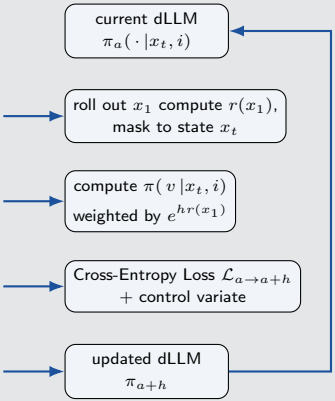
3. Discrete Tilt Matching

Fine-tuning as Tilting

- Want to go from base model $\rho_0(x)$ to $\rho_A(x)$, where $\rho_A(x) = \rho_0(x)e^{hr(x)}$ with reward $r(x)$.
- DTM performs tilts in small steps:

$$\rho_{a+h}(x) \propto \rho_a(x)\, e^{hr(x)}.$$

DTM algorithm



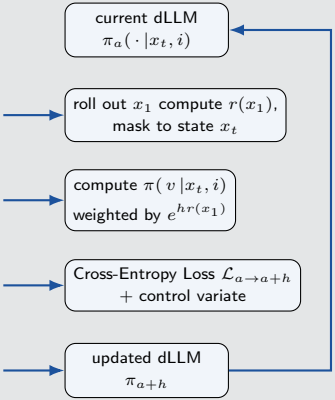
3. Discrete Tilt Matching

Fine-tuning as Tilting

- Want to go from base model $\rho_0(x)$ to $\rho_A(x)$, where $\rho_A(x) = \rho_0(x)e^{hr(x)}$ with reward $r(x)$.
- DTM performs tilts in small steps:
$$\rho_{a+h}(x) \propto \rho_a(x)\, e^{hr(x)}.$$

- Initialize $\pi_\theta = \pi_a(\,v\,|x_t, i)$

DTM algorithm



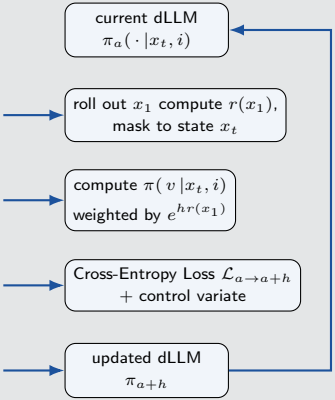
3. Discrete Tilt Matching

Fine-tuning as Tilting

- Want to go from base model $\rho_0(x)$ to $\rho_A(x)$, where $\rho_A(x) = \rho_0(x)e^{hr(x)}$ with reward $r(x)$.
- DTM performs tilts in small steps:
$$\rho_{a+h}(x) \propto \rho_a(x)\, e^{hr(x)}.$$

- Initialize $\pi_\theta = \pi_a(\,v\,|x_t, i)$
- Train π_θ to match the next local unmasking posterior $\pi_{a+h}(\,v\,|x_t, i)$ **exactly** with CE loss and control variate.

DTM algorithm



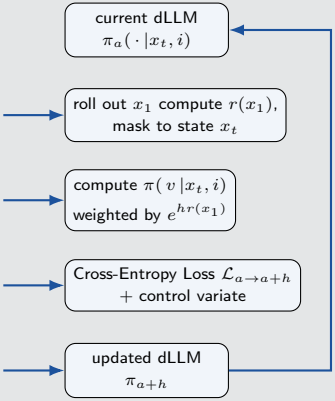
3. Discrete Tilt Matching

Fine-tuning as Tilting

- Want to go from base model $\rho_0(x)$ to $\rho_A(x)$, where $\rho_A(x) = \rho_0(x)e^{hr(x)}$ with reward $r(x)$.
- DTM performs tilts in small steps:
$$\rho_{a+h}(x) \propto \rho_a(x)\, e^{hr(x)}.$$

- Initialize $\pi_\theta = \pi_a(\,v\,|x_t, i)$
- Train π_θ to match the next local unmasking posterior $\pi_{a+h}(\,v\,|x_t, i)$ **exactly** with CE loss and control variate.
- $$\pi_{a+h}(v \mid x_t, i) = \frac{\mathbb{E}_{\pi_a}[e^{hr(x_1)}\mathbf{1}\{x_1^i = v\} \mid x_t]}{\mathbb{E}_{\pi_a}[e^{hr(x_1)} \mid x_t]}$$
which can then be converted into a tractable loss function after rearranging

DTM algorithm



4. Key Results

Benchmark	Base LLaDA	Best prior	DTM	Takeaway
MATH500	34.6	**41.8** (SPG)	40.2	competitive
Countdown	16.8	70.7 (SPG)	**81.3**	best by +10.6
Sudoku	27.7	94.0 (SPG)	**99.4**	nearly solved
GSM8K	79.8	**86.1** (SPG)	83.2	competitive

Accuracy is the best of generation length 256 and 512, following the paper's evaluation summary.

4. Key Results

Benchmark	Base LLaDA	Best prior	DTM	Takeaway
MATH500	34.6	**41.8** (SPG)	40.2	competitive
Countdown	16.8	70.7 (SPG)	**81.3**	best by +10.6
Sudoku	27.7	94.0 (SPG)	**99.4**	nearly solved
GSM8K	79.8	**86.1** (SPG)	83.2	competitive

Empirical message

DTM improves the base model LLaDA on all four tasks and is strongest on structured planning benchmarks.

Accuracy is the best of generation length 256 and 512, following the paper's evaluation summary.

4. Key Results

Benchmark	Base LLaDA	Best prior	DTM	Takeaway
MATH500	34.6	**41.8** (SPG)	40.2	competitive
Countdown	16.8	70.7 (SPG)	**81.3**	best by +10.6
Sudoku	27.7	94.0 (SPG)	**99.4**	nearly solved
GSM8K	79.8	**86.1** (SPG)	83.2	competitive

Empirical message

DTM improves the base model LLaDA on all four tasks and is strongest on structured planning benchmarks.

Convergence guarrantee

KL-divergence of ρ_θ with the ground-truth ρ_{a+h} is bounded by the **regret**

$$\mathrm{KL}(\rho_{a+h} \| \rho_\theta) \leq \frac{1}{Z}(\mathcal{L}_{a \to a+h}(\theta) - \mathcal{L}_{a \to a+h}(\theta^*))$$

Accuracy is the best of generation length 256 and 512, following the paper's evaluation summary.