

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>2</b>
2.1	Discrete Probability Flows and Masked Diffusion Models . . . . .	2
2.2	Kinetic and Conditional Kinetic Energy . . . . .	3
2.3	MDM as Geodesic Curve . . . . .	3
<b>3</b>	<b>Main Results</b>	<b>4</b>
3.1	Equivalence of Energies in MDMs . . . . .	4
3.2	MDMs with the Optimal Mask Schedule Are Energy Minimizers . . . . .	5
3.3	Energy-Inspired Fast Samplers . . . . .	6
<b>4</b>	<b>Experiments</b>	<b>8</b>
<b>5</b>	<b>Related Work</b>	<b>8</b>
<b>6</b>	<b>Conclusion</b>	<b>9</b>
	<b>Appendix A Detailed Notations and Definitions</b>	<b>21</b>
	<b>Appendix B An Intuitive Explanation of Geodesics</b>	<b>21</b>
	<b>Appendix C Proof of Auxillary Lemmas</b>	<b>22</b>
C.1	Derivation of the relationship between $\alpha_t$ and $\sigma_t$ in MDM . . . . .	22
C.2	Proof of the Invariance of Training Loss to the Mask Schedule . . . . .	23
C.3	Derivation of Conditional Rate Matrix of MDM . . . . .	23
C.4	Decomposition of Conditional Kinetic Energy along Sequence Dimension . . . . .	24
	<b>Appendix D Proof of Main Results</b>	<b>25</b>
D.1	Proof of Theorem 3.1 . . . . .	25
D.2	Proof of Theorem 3.2 . . . . .	26
D.3	Proof of Example 3.3 . . . . .	28
D.4	Proof of Lemma 3.5 . . . . .	28
D.5	Proof of Theorem 3.6 . . . . .	29
D.6	Proof of Proposition 3.7 . . . . .	31
	<b>Appendix E Experimental Details</b>	<b>32</b>
E.1	Standard Benchmarks and Evaluation Settings . . . . .	32
E.2	Additional Results and Raw Data . . . . .	32
E.3	Additional Samples . . . . .	32
E.4	Details of Toy Sampling Experiment in Fig. 4 . . . . .	34

## Appendix A Detailed Notations and Definitions

- $n$ : the sequence length.
- $x, z$ : a  $n$ -dimensional vector representing states in a model.
- $\mathcal{D}$ : the vocabulary with size  $|\mathcal{D}| = d$ .
- $x^i, z^i \in \mathcal{D}$ : the  $i$ -th token of data  $x, z$ .
- $m(z)$ : the number of mask tokens in  $z$ .
- $(p_t(z))_{t \in [0,1]}$ : the DPF that connects a simple distribution  $p_0(z)$  and a data distribution  $p_1(z) = q(z)$ .
- $p_{t|1}(z|x_1)$ : the conditional probability flow conditioned on the data.
- $p_{1|t}(x_1|z)$ : the posterior distribution conditioned on time  $t$ .
- $\alpha_t$ : The mask schedule function.
- $Q_t$ : The transition rate matrix of the unmasking process of MDM at time  $t$ .
- $Q_{t|1}$ : The conditional rate matrix of the unmasking process of MDM at time  $t$ .
- $\overleftarrow{Q}_t$ : The transition rate matrix of the masking process of MDM at time  $t$ .
- $\sigma_t$ : The transition rate that uniquely determines  $\overleftarrow{Q}_t$  in MDM settings.
- $\gamma_t$ : The interpolation schedule function of the geodesic curve on the high-dimensional sphere. Also the weight function we choose in three energy functionals for energy minimization.
- $B(a, b)$ : Beta distribution with parameters  $a$  and  $b$ .
- CDF: cumulative distribution function

## Appendix B An Intuitive Explanation of Geodesics

This appendix provides an intuitive introduction to geodesics and exponential maps to clarify the geometric interpretation of MDM in [42]. Readers seeking formal mathematical definitions of these differential geometry concepts may refer to [5] for complete technical specifications.

**Manifolds and Tangent Spaces.** A manifold is a smooth high-dimensional surface, such as a  $D$ -dimensional sphere  $\mathbb{S}^{D-1}$ . In our scenario, the embedding Eq. (8) maps the per-token conditional probability flow Eq. (3) onto  $\mathbb{S}^{D-1}$ , as shown in Sec. 2.3. On every point  $y_0$  on the manifold, there exists a tangent space  $\mathcal{T}_{y_0}$  containing all vectors starting from  $y_0$  and tangent to the manifold.

**Exponential Map and Geodesics.** The exponential map is denoted as  $\exp_{y_0}(v)$ , which maps a tangent vector  $v \in \mathcal{T}_{y_0}$  to a point  $y_1$  on the manifold. Geometrically, this represents moving from  $y_0$  along the "direction" of  $v$  at constant speed until reaching  $y_1$ . This movement follows a geodesic path, which is both the "shortest path" and the "straight path" between two points on a manifold, generalizing straight lines in Euclidean space. For example, great circles are geodesics on spheres.

**Inverse Exponential Map and Parameterized Geodesic Trajectory:** The inverse exponential map is denoted as  $\exp_{y_0}^{-1}(y_1)$ , which maps a manifold point  $y_1$  back to a tangent vector  $v \in \mathcal{T}_{y_0}$ . This vector encodes both direction and distance from  $y_0$  to  $y_1$ . Therefore, given start/end points  $y_0$  and  $y_1$  on the manifold, a geodesic trajectory parameterized by  $\gamma_t$  (strictly increasing with  $\gamma_0 = 0, \gamma_1 = 1$ ) can be expressed as:

$$\exp_{y_0}^{-1}(y_t) = \gamma_t \cdot \exp_{y_0}^{-1}(y_1), \quad t \in [0, 1] \quad (19)$$

This formulation implies:

- $\exp_{y_0}^{-1}(y_1)$  is the tangent vector encoding the direction and scale that generates the geodesic curve from  $y_0$  to  $y_1$ .
- $\gamma_t$  is the interpolation schedule and  $\gamma_t = t$  means constant-speed motion along the geodesic.

Recent theoretical advances [42] reveal that MDM's conditional probability flow in Eq. (3) forms exactly the geodesic curve in spherical geometry (see Fig. 6). The interpolation schedule  $\gamma_t$  is uniquely determined by the mask schedule  $\alpha_t$  as  $\gamma_t = \frac{2}{\pi} \arcsin \sqrt{\alpha_t}$ , which is equivalent with Condition 3.4.

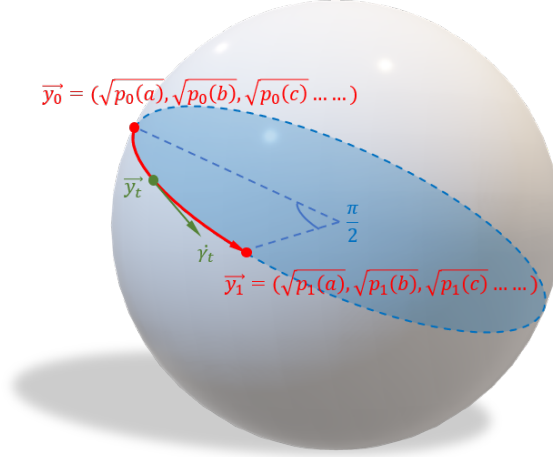


Figure 6: The per-token conditional probability flow in MDM generates exactly the geodesic curve.

## Appendix C Proof of Auxillary Lemmas

In this section, we provide complete proofs for the auxiliary lemmas referenced in the main text for completeness.

### C.1 Derivation of the relationship between $\alpha_t$ and $\sigma_t$ in MDM

**Lemma C.1.** The mask schedule  $\alpha_t$  relates to the rate  $\sigma_t$  via expression

$$\alpha_t = \exp \left( - \int_t^1 \sigma_s ds \right). \quad (20)$$

*Proof.* From the definition of mask schedule in Eq. (3), it suffices to prove

$$\mathbb{P}(x_t^i = x_1^i | x_1) = \exp \left( - \int_t^1 \sigma_s ds \right). \quad (21)$$

Consider infinitesimal time intervals  $(t, t - \Delta t]$  where each token experiences masking probability  $\sigma_t \Delta t + o(\Delta t)$ . The preservation probability therefore satisfies the following product bounds:

$$\prod_{k=0}^{\lfloor (1-t)/\Delta t \rfloor + 1} (1 - \sigma_{(1-k\Delta t)} \Delta t + o(\Delta t)) \leq \mathbb{P}(x_t^i = x_1^i | x_1) \quad (22)$$

$$\leq \prod_{k=0}^{\lfloor (1-t)/\Delta t \rfloor} (1 - \sigma_{(1-k\Delta t)} \Delta t + o(\Delta t)). \quad (23)$$

Analyzing the upper bound through logarithmic transformation, we get

$$\prod_{k=0}^{\lfloor (1-t)/\Delta t \rfloor} (1 - \sigma_{(1-k\Delta t)} \Delta t + o(\Delta t)) = \exp \left( \sum_{k=0}^{\lfloor (1-t)/\Delta t \rfloor} \log(1 - \sigma_{(1-k\Delta t)} \Delta t + o(\Delta t)) \right) \quad (24)$$

$$= \exp \left( \sum_{k=0}^{\lfloor (1-t)/\Delta t \rfloor} (-\sigma_{(1-k\Delta t)} \Delta t + o(\Delta t)) \right) \quad (25)$$

$$\stackrel{(1)}{\rightarrow} \exp \left( - \int_0^{1-t} \sigma_{1-u} du \right) \quad (26)$$

$$\stackrel{(2)}{=} \exp \left( - \int_t^1 \sigma_s ds \right), \quad (27)$$

807 where in (1) follows from Riemann sum convergence as  $\Delta t \rightarrow 0$  and (2) applies the variable  
808 substitution  $s = 1 - u$  to align integration limits.

809 The lower bound converges identically through analogous arguments. This completes the proof.  $\square$

## 810 C.2 Proof of the Invariance of Training Loss to the Mask Schedule

811 Different equivalence expressions of the training loss of MDM has been proved invariant to the choice  
812 of  $\alpha_t$  in multiple works [12, 38, 36]. We adapt a proof from [36] by examining the negative evidence  
813 lower bound (NELBO) through token-level denoising components:

$$\mathcal{L}_{\text{NELBO}} = \mathbb{E}_{p_1(x_1), p_{t|1}(z|x_1)} \int_0^1 \frac{-\dot{\alpha}_t}{1 - \alpha_t} \log \langle x_\theta(z_t, t), x_1 \rangle dt \quad (28)$$

$$= \mathbb{E}_{p_1(x_1), p_{t|1}(z|x_1)} \int_0^1 \frac{-\dot{\alpha}_t}{1 - \alpha_t} \sum_{i=1}^n \log \langle x_\theta^i(z^{1:n}, t), x_1^i \rangle dt. \quad (29)$$

814 Despite the apparent dependence on  $\alpha_t$  in its parametric form, the loss exhibits fundamental invariance  
815 as formalized below.

816 **Proposition C.2** (Schedule Invariance, Proof Adapted from [36]).  *$\mathcal{L}_{\text{NELBO}}$  is invariant to  $\alpha_t$ 's  
817 functional form, depending only on its boundary values  $\alpha_0 = 0, \alpha_1 = 1$ .*

818 *Proof.* The invariance emerges through variable substitution via the chain rule. Let  $\gamma \equiv \log(1 - \alpha_t)$ ,  
819 then

$$\mathcal{L}_{\text{NELBO}} = \mathbb{E}_{p_1(x_1), p_{t|1}(z|x_1)} \int_{t=0}^{t=1} \frac{-\alpha'_t}{1 - \alpha_t} \log \langle x_\theta(z_t, t), x \rangle dt \quad (30)$$

$$\stackrel{(1)}{=} \mathbb{E}_{p_1(x_1), p_{t|1}(z|x_1)} \int_{t=0}^{t=1} \log \langle x_\theta(z_t, t), x \rangle d[f(t)] \quad (31)$$

$$\stackrel{(2)}{=} \mathbb{E}_{p_1(x_1), p_{t|1}(z|x_1)} \int_{\gamma=0}^{\gamma=-\infty} \log \langle x_\theta(z_{f^{-1}(\gamma)}, f^{-1}(\gamma)), x \rangle d\gamma \quad (32)$$

$$\stackrel{(3)}{=} -\mathbb{E}_{p_1(x_1), p_{t|1}(z|x_1)} \int_{\gamma=-\infty}^{\gamma=0} \log \langle \tilde{x}_\theta(\tilde{z}_\gamma, \gamma), x \rangle d\gamma \quad (33)$$

820 Here (1) applies the substitution  $f(t) = \log(1 - \alpha_t)$ . (2) applies change of variable  $\gamma \equiv f(t)$ . In  
821 (3) we let  $\tilde{z}_\gamma \equiv z_{f^{-1}(\gamma)}$ ,  $\tilde{x}_\theta(\tilde{z}_\gamma, \gamma) \equiv x_\theta(\tilde{z}_\gamma, f^{-1}(\gamma))$ . The final expression contains no explicit  
822 dependence on  $\alpha_t$ 's trajectory between its fixed endpoints, thereby establishing the claimed invariance.  
823  $\square$

## 824 C.3 Derivation of Conditional Rate Matrix of MDM

825 Although explicit sampling through  $Q_{t|1}(z, x|x_1)$  remains unnecessary in MDM, establishing closed-  
826 form representations proves valuable for theoretical characterization.

827 **Lemma C.3** (Conditional Rate of MDM). *The following conditional rate matrix generates MDM's  
828 unmasking process:*

$$Q_t(z, x|x_1) = \begin{cases} \overleftarrow{Q}_t(x, z) \frac{p_{t|1}(x|x_1)}{p_{t|1}(z|x_1)} & p_{1|t}(x_1|z) > 0 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \frac{\sigma_t \alpha_t}{1 - \alpha_t} & z \rightarrow x \Rightarrow x_1 \\ -\sum_{x \neq z} Q_t(z, x|x_1) & x = z \\ 0 & \text{otherwise} \end{cases}. \quad (34)$$

829 Here  $z \rightarrow x$  denotes single-token unmasking transitions defined in Sec.2.1 and  $x \Rightarrow x_1$  denotes that  
830  $x_1$  can be generated from  $x$  through one or several steps of unmasking.

831 Applying Bayes' theorem establishes the posterior relationship:

$$p_{1|t}(x_1|z) = \frac{p_{t|1}(z|x_1)p_1(x_1)}{p_t(z)}, \quad (35)$$

we know that the positivity condition  $p_{1|t}(x_1|z) > 0$  consequently requires  $p_{t|1}(z|x_1) > 0$  and  $z \Rightarrow x_1$ , ensuring the rate matrix's well-posedness.

*Proof.* To verify that  $Q_t(z, x|x_1)$  generates MDM's demasking dynamics, we confirm the consistency condition in Eq. (6) holds. We verify this through the following derivation:

$$\sum_{x_1} Q_t(z, x|x_1) p_{1|t}(x_1|z) = \sum_{x_1} Q_t(z, x|x_1) \frac{p_{t|1}(z|x_1) p_1(x_1)}{p_t(z)} \quad (36)$$

$$= \sum_{x_1} \overleftarrow{Q}_t(x, z) \frac{p_{t|1}(x|x_1)}{p_{t|1}(z|x_1)} \frac{p_{t|1}(z|x_1) p_1(x_1)}{p_t(z)} \quad (37)$$

$$= \sum_{x_1} \overleftarrow{Q}_t(x, z) \frac{p_{t|1}(x|x_1) p_1(x_1)}{p_t(z)} \quad (38)$$

$$= \overleftarrow{Q}_t(x, z) \frac{p_t(x)}{p_t(z)} = Q_t(z, x). \quad (39)$$

We subsequently derive its closed-form of  $Q_t(z, x|x_1)$ . If  $x \not\Rightarrow x_1$ , then  $p_{t|1}(x|x_1) = 0$ , yielding  $Q_t(z, x|x_1) = 0$ . If  $z \rightarrow x \Rightarrow x_1$ , on the other hand, we have

$$\begin{aligned} Q_t(z, x|x_1) &= \overleftarrow{Q}_t(x, z) \frac{p_{t|1}(x^i|x_1^i)}{p_{t|1}(z^i|x_1^i)} \\ &= \sigma_t \frac{\alpha_t}{1 - \alpha_t}, \end{aligned} \quad (40)$$

thus completing the proof.  $\square$

#### C.4 Decomposition of Conditional Kinetic Energy along Sequence Dimension

In this section, we show that under many DPF frameworks such as MDMs, the conditional kinetic energy can be decomposed along sequence dimension as following:

$$\mathcal{E}_c(p_{t|1}, Q_{t|1}; \gamma_t) = \mathbb{E}_{t, p_1(x_1), p_{t|1}(z|x_1)} \sum_{x: x \neq z} \frac{1}{\dot{\gamma}_t p_{t|1}(x|x_1)} Q_{t|1}(z, x|x_1)^2 \quad (41)$$

$$= \mathbb{E}_{t, p_1(x_1)} \sum_{z, x: x \neq z} \frac{p_{t|1}(z|x_1)}{\dot{\gamma}_t p_{t|1}(x|x_1)} Q_{t|1}(z, x|x_1)^2 \quad (42)$$

$$\stackrel{(1)}{=} \mathbb{E}_{t, p_1(x_1)} \sum_{i=1}^n C \sum_{z^i, x^i: x^i \neq z^i} \frac{p_{t|1}(z^i|x_1)}{\dot{\gamma}_t p_{t|1}(x^i|x_1)} Q_{t|1}(z^i, x^i|x_1)^2 \quad (43)$$

$$= \sum_{i=1}^n C \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \sum_{x^i: x^i \neq z^i} \frac{1}{\dot{\gamma}_t p_{t|1}(x^i|x_1)} Q_{t|1}(z^i, x^i|x_1)^2 \quad (44)$$

The pivotal decomposition in (1) leverages two structural properties: (i) The conditional probability flow (Eq. (3)) exhibits token-wise independence, and (ii) The conditional rate matrix (Lemma C.3) nullifies transitions altering multiple tokens simultaneously. These enable reduction of full-sequence transitions to single-token operations, with remaining  $n-1$  tokens contributing constant combinatorial factors. This structural property persists across various DPF implementations including MDM and Discrete Flow Matching [32, 37], validating the conditional kinetic energy as theoretically sound surrogate objective. Notably, standard kinetic energy  $\mathcal{E}$  lacks such decomposition due to  $p_t(z)$ 's dependence on cross-token correlations.

For MDM's binary mask dynamics ( $x_1^i$  vs. [MASK]), the combinatorial constant becomes  $C = 2^{n-1}$ :

$$\mathcal{E}_c(\alpha_t, \gamma_t) = 2^{n-1} \cdot \sum_{i=1}^n \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \sum_{x^i: x^i \neq z^i} \frac{1}{\dot{\gamma}_t p_{t|1}(x^i|x_1)} Q_{t|1}(z^i, x^i|x_1)^2. \quad (45)$$

851 This decomposition permits notational relaxation where  $\mathcal{E}_c$  analysis considers tokens  $z, x \in \mathcal{D}$   
852 independently of sequence context, a slight abuse of notation adopted in Appendix D.2’s equivalence  
853 proof.

## 854 Appendix D Proof of Main Results

### 855 D.1 Proof of Theorem 3.1

856 **Theorem 3.1** (Kinetic-conditional equivalence in MDMs). *For any weight function  $\gamma_t$  and MDM*  
857 *with mask schedule  $\alpha_t$ , the marginal and conditional kinetic energies are proportional:*

$$\mathcal{E}_k(\alpha_t, \gamma_t) = C_1 \mathcal{E}_c(\alpha_t, \gamma_t), \quad (46)$$

858 where  $C_1$  is a scalar depending only on the sequence length  $n$  and vocabulary size  $d$ . As a result, the  
859 two objectives share the same minimizers:

$$\arg \min_{\alpha_t} \mathcal{E}_k(\alpha_t, \gamma_t) = \arg \min_{\alpha_t} \mathcal{E}_c(\alpha_t, \gamma_t). \quad (47)$$

860 Our argument leverages a foundational decomposition from [44] regarding concrete score representa-  
861 tions:

862 **Lemma D.2.** *for  $z = (z^1, \dots, z^i = [\text{M}], \dots, z^n)$ ,  $x = (z^1, \dots, x^i \neq [\text{M}], \dots, z^n)$ , we have*

$$\frac{p_t(x)}{p_t(z)} = \frac{\alpha_t}{1 - \alpha_t} p_1(x^i | z^{UM}), \quad (48)$$

863 where  $z^{UM}$  is the vector consists of all unmasked tokens of  $z$ .

864 We now prove the main theorem:

865 *Proof.* Let  $m(z)$  quantify the masked positions in  $z$ , assumed w.l.o.g. to occupy initial sequence  
866 positions. The key summation decomposes as:

$$\sum_{x: z \rightarrow x} \frac{p_t(x)}{p_t(z)} = \sum_{i=1}^{m(z)} \sum_{\substack{x: z \rightarrow x \\ x^i \neq z^i = [\text{M}]}} \frac{p_t(x)}{p_t(z)} \quad (49)$$

$$= \sum_{i=1}^{m(z)} \sum_{x^i \neq [\text{M}]} \frac{\alpha_t}{1 - \alpha_t} p_0(x^i | z^{UM}) \quad (50)$$

$$= \sum_{i=1}^{m(z)} \frac{\alpha_t}{1 - \alpha_t} \sum_{x^i \neq [\text{M}]} p_0(x^i | z^{UM}) \quad (51)$$

$$= m(z) \frac{\alpha_t}{1 - \alpha_t}. \quad (52)$$

867 Substituting the rate matrix from Eq. (2), we therefore deduce

$$\mathcal{E}_k = \mathbb{E}_{t, p_t(z)} \sum_{x: z \rightarrow x} \frac{1}{p_t(x) \dot{\gamma}_t} \left( \sigma_t \frac{p_t(x)}{p_t(z)} \right)^2 \quad (53)$$

$$= \mathbb{E}_t \sum_z \sum_{x: z \rightarrow x} \frac{p_t(z)}{p_t(x) \dot{\gamma}_t} \sigma_t^2 \frac{p_t^2(x)}{p_t^2(z)} \quad (54)$$

$$= \mathbb{E}_t \sum_z \sum_{x: z \rightarrow x} \frac{\sigma_t^2 p_t(x)}{\dot{\gamma}_t p_t(z)} \quad (55)$$

$$\stackrel{(1)}{=} \mathbb{E}_t \sum_z m(z) \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t} \quad (56)$$

$$= \left( \sum_z m(z) \right) \mathbb{E}_t \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t} \quad (57)$$

$$\triangleq C_k \cdot \mathbb{E}_t \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t}. \quad (58)$$

868 Here we define  $C_k = \sum_z m(z)$ . On the other hand, the conditional kinetic energy can also be break  
869 down as

$$\mathcal{E}_c = \mathbb{E}_{t, p_1(x_1), p_{t|1}(z|x_1)} \sum_{x: z \rightarrow x} \frac{1}{p_{t|1}(x|x_1) \dot{\gamma}_t} \left( \sigma_t \frac{p_{t|1}(x|x_1)}{p_{t|1}(z|x_1)} \right)^2 \quad (59)$$

$$= \mathbb{E}_{t, p_1(x_1)} \sum_{z, x: z \rightarrow x \Rightarrow x_1} \frac{\sigma_t^2 p_{t|1}(x|x_1)}{\dot{\gamma}_t p_{t|1}(z|x_1)} \quad (60)$$

$$= \mathbb{E}_{t, p_1(x_1)} \sum_{z, x: z \rightarrow x \Rightarrow x_1} \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t} \quad (61)$$

$$= \left( \sum_{x_1} \sum_{z, x: z \rightarrow x \Rightarrow x_1} p_1(x_1) \right) \mathbb{E}_t \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t} \quad (62)$$

$$= n 2^{n-1} \left( \sum_{x_1} p_1(x_1) \right) \mathbb{E}_t \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t} \quad (63)$$

$$= n 2^{n-1} \mathbb{E}_t \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t} \quad (64)$$

$$\triangleq C_c \cdot \mathbb{E}_t \frac{\alpha_t}{1 - \alpha_t} \frac{\sigma_t^2}{\dot{\gamma}_t}. \quad (65)$$

870 This expression of  $\mathcal{E}_c$  is equivalent to the decomposition in Appendix C.4 by plugging in the explicit  
871 form of  $Q_{t|1}$ . The proportionality constant  $C_1 = C_k/C_c$  emerges from comparing both expressions,  
872 with  $C_k, C_c$  depending solely on architectural parameters  $n$  and  $|\mathcal{D}| = d$ . The minimizer equivalence  
873 follows directly from the strict positivity of scaling constants.

874 □

## 875 D.2 Proof of Theorem 3.2

876 **Theorem 3.2** (Conditional-geodesic equivalence in MDMs). *For any weight function  $\gamma_t$  and MDM*  
877 *with mask schedule  $\alpha_t$ , the conditional and geodesic energies are proportional:*

$$\mathcal{E}_c(\alpha_t, \gamma_t) = C_2 \mathcal{E}_g(\alpha_t, \gamma_t), \quad (66)$$

878 where  $C_2$  is a scalar depending only on the sequence length  $n$ . This implies that they share the same  
879 minimizers:

$$\arg \min_{\alpha_t} \mathcal{E}_c(\alpha_t, \gamma_t) = \arg \min_{\alpha_t} \mathcal{E}_g(\alpha_t, \gamma_t). \quad (67)$$

880 *Proof.* The geodesic energy is inherently defined using the token-wise independent conditional flows  
 881 in Eq. (3), therefore it admits straightforward decomposition along sequence dimensions. In MDM  
 882 case where all tokens follows the same mask schedule  $\alpha_t$ , we further have

$$\mathcal{E}_g(p_{t|1}; \gamma_t) = \sum_{i=1}^n \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \frac{4}{\dot{\gamma}_t p_{t|1}(z^i|x_1)} \dot{y}_{t|1}(z^i|x_1)^2 \quad (68)$$

$$= n \cdot \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \frac{4}{\dot{\gamma}_t p_{t|1}(z^i|x_1)} \dot{y}_{t|1}(z^i|x_1)^2 \quad (69)$$

$$\triangleq C_g \cdot \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \frac{4}{\dot{\gamma}_t p_{t|1}(z^i|x_1)} \dot{y}_{t|1}(z^i|x_1)^2 \quad (70)$$

883 On the other hand, through dimensional decomposition established in Appendix C.4, the conditional  
 884 energy admits:

$$\mathcal{E}_c(p_{t|1}, Q_{t|1}; \gamma_t) = 2^{n-1} \cdot \sum_{i=1}^n \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \sum_{x^i: x^i \neq z^i} \frac{1}{\dot{\gamma}_t p_{t|1}(x^i|x_1)} Q_{t|1}(z^i, x^i|x_1)^2 \quad (71)$$

$$= C_c \cdot \mathbb{E}_{t, p_1(x_1), p_{t|1}(z^i|x_1)} \sum_{x^i: x^i \neq z^i} \frac{1}{\dot{\gamma}_t p_{t|1}(x^i|x_1)} Q_{t|1}(z^i, x^i|x_1)^2, \quad (72)$$

885 Where we define  $C_c = n2^{n-1}$  as in Appendix D.1. Therefore, The equivalence proof reduces to  
 886  $n = 1$  analysis through notational relaxation, treating  $z, x \in \mathcal{D}$  as individual tokens.

887 Leveraging the rate matrix expression from Lemma C.3, we have

$$\mathcal{E}_c(p_{t|1}, Q_{t|1}; \gamma_t) = \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \sum_{z, x: z \rightarrow x \Rightarrow x_1} \frac{p_{t|1}(z|x_1)}{p_{t|1}(x|x_1)} Q_{t|1}(z, x|x_1)^2 \quad (73)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \sum_{z=[M], x=x_1} \frac{p_{t|1}(z|x_1)}{p_{t|1}(x|x_1)} \left( \frac{\sigma_t \alpha_t}{1 - \alpha_t} \right)^2 \quad (74)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \sum_{z=[M], x=x_1} \frac{1 - \alpha_t}{\alpha_t} \left( \frac{\sigma_t \alpha_t}{1 - \alpha_t} \right)^2 \quad (75)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \frac{\sigma_t^2 \alpha_t}{1 - \alpha_t}, \quad (76)$$

888 which is equivalent to the expression in Appendix D.1 by letting  $n = 1$ . Applying the relationship  
 889  $\dot{\alpha}_t = \alpha_t \sigma_t$  deduced from Eq. (4), we get  $\dot{\alpha}_t = \alpha_t \sigma_t$ , therefore we further have

$$\mathcal{E}_c(p_{t|1}, Q_{t|1}; \gamma_t) = \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \frac{\dot{\alpha}_t^2}{\alpha_t (1 - \alpha_t)}. \quad (77)$$

890 On the other hand, applying Eq. (3),  $\mathcal{E}_g$  in  $n = 1$  case can be expressed as

$$\mathcal{E}_g(p_{t|1}; \gamma_t) = \mathbb{E}_{t, p_1(x_1)} \frac{4}{\dot{\gamma}_t} \sum_z \left( \frac{d}{dt} \sqrt{p_{t|1}(z|x_1)} \right)^2 \quad (78)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \sum_z \frac{\dot{p}_t(z|x_1)^2}{p_{t|1}(z|x_1)} \quad (79)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \sum_z \frac{[\dot{\alpha}_t (\delta_{x_1}(z) - \delta_m(z))]^2}{\alpha_t \delta_{x_1}(z) + (1 - \alpha_t) \delta_m(z)} \quad (80)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \left( \sum_{z=x_1} \frac{\dot{\alpha}_t^2}{\alpha_t} + \sum_{z=[M]} \frac{\dot{\alpha}_t^2}{(1 - \alpha_t)} \right) \quad (81)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{1}{\dot{\gamma}_t} \left( \frac{\dot{\alpha}_t^2}{\alpha_t} + \frac{\dot{\alpha}_t^2}{(1 - \alpha_t)} \right). \quad (82)$$



891 Since

$$\frac{\dot{\alpha}_t^2}{\alpha_t(1-\alpha_t)} = \frac{\dot{\alpha}_t^2}{\alpha_t} + \frac{\dot{\alpha}_t^2}{(1-\alpha_t)}, \quad (83)$$

892 the functional equivalence in scalar case is established, extended to  $n$ -dimensions through the  
 893 dimensional scaling factor  $C_2 = C_c/C_g = 2^{n-1}$ . The minimizer equivalence follows from strict  
 894 positivity of scaling relations.  $\square$

### 895 D.3 Proof of Example 3.3

896 **Example 3.3.** When  $n = 1$ , the kinetic, conditional kinetic, and geodesic energies all reduce to:

$$\mathcal{E}(\alpha_t, \gamma_t) = \int_0^1 \frac{1}{\dot{\gamma}_t} \cdot \frac{\dot{\alpha}_t^2}{\alpha_t(1-\alpha_t)} dt. \quad (84)$$

897 *Proof.* In the  $n = 1$  case, we have

$$C_k = \sum_z m(z) = m([M]) = 1; \quad (85)$$

$$C_c = n2^{n-1} = 1; \quad (86)$$

$$C_g = n = 1. \quad (87)$$

898 Therefore, we have  $C_1 = C_2 = 1$  and the three energy functions share the same form Eq. (14).  $\square$

### 899 D.4 Proof of Lemma 3.5

900 **Lemma 3.5.** Under Condition 3.4, the schedule  $\alpha_t^*$  minimizes the geodesic energy.

901 Since the geodesic energy (Definition 2.3) is defined by summing the token-wise conditional proba-  
 902 bility flow, we only need to focus on the one-dimensional case. Therefore, we abuse notation slightly  
 903 by regarding  $z \in \mathcal{D}$ :

$$\mathcal{E}_g(p_{t|1}; \gamma_t) = \mathbb{E}_{t, p_1(x_1), p_{t|1}(z|x_1)} \frac{4}{\dot{\gamma}_t p_{t|1}(z|x_1)} \dot{y}_{t|1}(z|x_1)^2 \quad (88)$$

$$= \mathbb{E}_{t, p_1(x_1)} \sum_z \frac{4}{\dot{\gamma}_t} \dot{y}_{t|1}(z|x_1)^2 \quad (89)$$

$$= \mathbb{E}_{t, p_1(x_1)} \frac{4}{\dot{\gamma}_t} \|y_{t|1}\|^2. \quad (90)$$

904 Here  $y_{t|1}$  denotes the  $d$ -dimensional embedding vector induced by the embedding Eq. (8). Therefore,  
 905 the minimizing problem becomes:

$$\arg \min_{y(t)} \int_0^1 \frac{\|\dot{y}(t)\|^2}{\dot{\gamma}(t)} dt \quad (91)$$

$$\text{s.t. } \|y(t)\| = 1, \forall t. \quad (92)$$

906 For baseline case  $\gamma_t = t$ , we construct the augmented functional with Lagrange multiplier  $\lambda(t)$ :

$$\mathcal{L}[y] = \int_0^1 (\|\dot{y}(t)\|^2 + \lambda(t)(\|y(t)\|^2 - 1)) dt. \quad (93)$$

907 The Euler-Lagrange formalism yields:

$$\frac{\partial \mathcal{L}}{\partial y} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{y}} = 0, \quad (94)$$

908 from which we derive the critical differential relationship:

$$\ddot{y} = \lambda y. \quad (95)$$

909 from  $\|y\|^2 = 1$  we have  $y\ddot{y} + \|\dot{y}\|^2 = 0$ , therefore we have

$$-\|\dot{y}\|^2 = y\ddot{y} = \lambda\|y\|^2 = \lambda. \quad (96)$$

910 Plugging this expression of  $\lambda$  into Eq. (95), we get

$$\ddot{y} = -\|\dot{y}\|^2 y, \quad (97)$$

911 which corresponds to the uniform circular motion with its acceleration pointing towards the center of  
 912 the sphere. Therefore, the route follows the great circle connecting  $y_0$  and  $y_1$ . In MDM case where  
 913  $y_0 = \delta_{[M]}$  and  $y_1$  represents clean data without mask token, we have  $y_0 \cdot y_1 = 0$ . Therefore, the angle  
 914 between  $y_0$  and  $y_1$  is  $\pi/2$  and the curve shares the following simple form:

$$y(t) = y_0 \cos(\frac{\pi}{2}t) + y_1 \sin(\frac{\pi}{2}t). \quad (98)$$

915 Now we generalize to arbitrary  $\gamma_t$  schedules through temporal reparameterization:

$$\arg \min_{y(t)} \int_0^1 \frac{\|\dot{y}(t)\|^2}{\dot{\gamma}(t)} dt = \arg \min_{y(\gamma(t))} \int_0^1 \frac{\|\dot{y}(\gamma(t))\|^2}{\dot{\gamma}(t)} dt \quad (99)$$

$$= \arg \min_{y(\gamma(t))} \int_0^1 \left\| \frac{dy}{d\gamma} \right\|^2 \dot{\gamma}_t dt \quad (100)$$

$$= \arg \min_{y(\gamma)} \int_0^1 \left\| \frac{dy}{d\gamma} \right\|^2 d\gamma \quad (101)$$

916 Therefore in ordinary  $\gamma_t$  cases, the optimized route is the geodesic curve rescheduled by interpolation  
 917 schedule  $\gamma_t$ :

$$y(t) = y_0 \cos(\frac{\pi}{2}\gamma_t) + y_1 \sin(\frac{\pi}{2}\gamma_t). \quad (102)$$

918 By squaring both sides of Eq. (102), we further recover MDM's conditional probability flow:

$$p_{t|1}(t) \stackrel{(1)}{=} p_0 \cos^2(\frac{\pi}{2}\gamma_t) + p_1 \sin^2(\frac{\pi}{2}\gamma_t) \quad (103)$$

$$= \alpha_t^* p_1 + (1 - \alpha_t^*) p_0, \quad (104)$$

919 where (1) leverages orthogonality  $p_0 \cdot p_1 = 0$ . Therefore, we proved that MDM with schedule  
 920 schedule  $\alpha_t^*$  generates the minimal-length curve as well as minimal-energy conditional probability  
 921 path, validating and generalizing the conclusion in [42] from a energy perspective.

## 922 D.5 Proof of Theorem 3.6

923 **Theorem 3.6** (Kinetic energy minimization). *Under Condition 3.4, the MDM schedule  $\alpha_t^*$  simultane-*  
 924 *ously minimizes all three energy functionals.*

925 *Proof.* Theorem 3.1, Theorem 3.2, and Lemma 3.5 collectively establish that  $\alpha_t^*$  optimizes the three  
 926 functionals over all mask schedules  $\alpha_t$ . This conclusion forms the theoretical foundation for our  
 927 data-driven schedule tuning framework presented in Section 3.3.

928 However,  $\mathcal{E}_k$  and  $\mathcal{E}_c$  are defined on both probability flows and rate matrices. Therefore, it remains to  
 929 be further proved that when the probability flow uniquely induced by  $\alpha_t$  is fixed, the conditional rate  
 930 matrix, which is also determined by  $\alpha_t$ , still minimizes the energy functionals, i.e.

$$Q_{t|1}(\alpha_t^*) \in \arg \min_{Q_{t|1}} \mathcal{E}_c(p_{t|1}, Q_{t|1}; \gamma_t), \quad (105)$$

931 where  $Q_{t|1}(\alpha_t^*)$  refer to conditional rate matrix in MDM case derived in Appendix C.3 under  
 932 Condition 3.4. We adapt the following key result from prior analysis [37] of the  $\gamma_t = t$  case:

933 **Lemma D.7.** *The following conditional rate matrix minimize the conditional kinetic energy  $\mathcal{E}_c$  when*  
 934 *the probability flow  $p_t(x)$  and weight function  $\gamma_t = t$  is fixed, i.e.*

$$\arg \min_{Q_{t|1}} \mathcal{E}_c(p_{t|1}, Q_{t|1}; \gamma_t = t) = Q_{t|1}^*(z, x|x_1) = \frac{\dot{\alpha}_t}{1 - \alpha_t} (\delta_{x_1}(x) - \delta_z(x)), \quad (106)$$

935 *where the arg min is taken over any possible  $Q_{t|1}$  that generates the fixed probability flow.*

936 Here  $\alpha_t$  is defined using the conditional probability flow under Discrete Flow Matching (DFM)  
 937 settings, resembling the MDM case (see Eq. (3)) by conditioning on both ends ( $t = 0$  and  $t = 1$ )  
 938 of the flow. In the case when  $p_0(z) = \delta_{[M]}(z)$ , it coincides with the mask schedule defined in our  
 939 work. The notation is also slightly abused by regarding  $z, x \in \mathcal{D}$  justified by the decomposition of  
 940 conditional kinetic energy in Appendix C.4.

941 We now demonstrate a non-trivial result: MDM under Condition 3.4 inherently achieves optimal  
 942 conditional rate matrices for *arbitrary*  $\gamma_t$ , despite structural constraints.

943 First, consider  $\gamma_t = t$  where Appendix C.3 yields the conditional rate matrix. We demonstrate that  
 944 MDM achieves the optimal velocity specified by the RHS of Eq. (D.7). For  $n = 1$ , the conditional  
 945 rate matrix simplifies to:

$$Q_{t|1}(z, x|x_1) = \begin{cases} \frac{\alpha_t \sigma_t}{1 - \alpha_t} & z = [M], x = x_1 \\ -\frac{\alpha_t \sigma_t}{1 - \alpha_t} & z = [M], x = z \\ 0 & \text{otherwise} \end{cases} \stackrel{(1)}{=} \begin{cases} \frac{\dot{\alpha}_t}{1 - \alpha_t} & z = [M], x = x_1 \\ -\frac{\dot{\alpha}_t}{1 - \alpha_t} & z = [M], x = z \\ 0 & \text{otherwise} \end{cases}, \quad (107)$$

946 where (1) follows from the identity  $\dot{\alpha}_t = \alpha_t \sigma_t$  established in Eq. (4), thus completing the  $\gamma_t = t$   
 947 case.

948 For general  $\gamma_t$ , we reformulate the conditional kinetic energy using Definition 2.2:

$$\mathcal{E}_c(Q_{t|1}; \gamma_t) = \mathbb{E}_t \frac{1}{\gamma_t} \sum_{z, x, x_1 \in A} f(z, x, x_1) Q_{t|1}(z, x|x_1) (t)^2, \quad (108)$$

949 where  $f$  and  $A$  are independent of  $Q$ . We first establish the following key lemma:

950 **Lemma D.8.** *If*

$$Q^*(t) \in \arg \min_{Q_{t|1}} \mathbb{E}_t \sum_{z, x, x_1 \in A} f(z, x, x_1) Q_{t|1}(t)^2, \quad (109)$$

951 *then we have*

$$Q^*(\gamma_t) \dot{\gamma}_t \in \arg \min_{Q_{t|1}} \mathbb{E}_t \frac{1}{\dot{\gamma}_t} \sum_{z, x, x_1 \in A} f(z, x, x_1) Q_{t|1}(t)^2, \quad (110)$$

952 *where the arg min is taken over any possible  $Q_{t|1}$  that generates the fixed probability flow.*

953 Proof of the lemma follows via the substitution  $\dot{W}_t = Q_{t|1}(t)$ . Let

$$W^*(t) \in \arg \min_{W(t)} \mathbb{E}_t \sum_{z, x, x_1 \in A} f(z, x, x_1) \dot{W}(t)^2, \quad (111)$$

954 then  $\dot{W}^*(t) = Q^*(t)$ . Therefore, we have

$$\arg \min_{Q_{t|1}} \mathbb{E}_t \frac{1}{\dot{\gamma}_t} \sum_{z,x,x_1 \in A} f(z,x,x_1) Q_{t|1}(t)^2 \quad (112)$$

$$= \frac{d}{dt} \left[ \arg \min_{W_t} \mathbb{E}_t \frac{1}{\dot{\gamma}_t} \sum_{z,x,x_1 \in A} f(z,x,x_1) \dot{W}_t^2 \right] \quad (113)$$

$$= \frac{d}{dt} \left[ \arg \min_{W_{\gamma(t)}} \mathbb{E}_t \frac{1}{\dot{\gamma}_t} \sum_{z,x,x_1 \in A} f(z,x,x_1) \left( \frac{dW_{\gamma(t)}}{dt} \right)^2 \right] \quad (114)$$

$$= \frac{d}{dt} \left[ \arg \min_{W_{\gamma(t)}} \mathbb{E}_t \dot{\gamma}_t \sum_{z,x,x_1 \in A} f(z,x,x_1) \left( \frac{dW_{\gamma(t)}/dt}{d\gamma(t)/dt} \right)^2 \right] \quad (115)$$

$$= \frac{d}{dt} \left[ \arg \min_{W_\gamma} \mathbb{E}_\gamma \sum_{z,x,x_1 \in A} f(z,x,x_1) \left( \frac{dW_\gamma}{d\gamma} \right)^2 \right] \quad (116)$$

$$\ni \frac{d}{dt} [W^*(\gamma)] = Q^*(\gamma_t) \dot{\gamma}_t, \quad (117)$$

955 proving this lemma. Given MDM's optimality under  $\gamma_t = t$ , i.e.

$$Q^*(t) = Q_{t|1}(\alpha_t^*) \Big|_{\gamma_t=t} = \frac{\dot{\alpha}_t}{1 - \alpha_t} \Big|_{\gamma_t=t} = \frac{\pi \sin(\frac{\pi}{2}t) \cos(\frac{\pi}{2}t)}{\cos^2(\frac{\pi}{2}t)} = \pi \tan(\frac{\pi}{2}t), \quad (118)$$

956 it remains to prove that the conditional rate matrix in MDM in general  $\gamma_t$  cases satisfies the LHS of  
957 Eq. (110). Since we have

$$Q_{t|1}(\alpha_t^*) = \frac{\dot{\alpha}_t}{1 - \alpha_t} \Big|_{\gamma_t} = \frac{\pi \sin(\frac{\pi}{2}\gamma_t) \cos(\frac{\pi}{2}\gamma_t) \dot{\gamma}_t}{\cos^2(\frac{\pi}{2}\gamma_t)} = \pi \tan(\frac{\pi}{2}\gamma_t) \dot{\gamma}_t, \quad (119)$$

958 thus  $Q_{t|1}(\alpha_t^*) = Q^*(\gamma_t) \dot{\gamma}_t$  and MDM's intrinsic optimization across arbitrary  $\gamma_t$  is obtained. Remark-  
959 ably, this result transcends geodesic energy  $\mathcal{E}_g$  (defined solely through  $p_{t|1}$ ), demonstrating MDM's  
960 dual optimization of both probability flows and sampling matrices despite structural constraints.  $\square$

## 961 D.6 Proof of Proposition 3.7

962 **Proposition 3.7.** *Linear and squared cosine schedules correspond to specific beta parameterizations:*

$$\alpha_t = t \quad \Leftrightarrow \quad \gamma_t = \text{CDF}_{\mathcal{B}(0.5,0.5)}(t), \quad (120)$$

$$\alpha_t = \sin^2\left(\frac{\pi}{2}t\right) \quad \Leftrightarrow \quad \gamma_t = t = \text{CDF}_{\mathcal{B}(1,1)}(t). \quad (121)$$

963 *Proof.* The probability density function (PDF) of the Beta distribution  $\mathcal{B}(a,b)$  is defined as:

$$p(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \quad (122)$$

964 where  $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$  is the normalizing constant.

965 For the case  $a = b = 0.5$ :

$$B(0.5, 0.5) = \frac{\Gamma(0.5)\Gamma(0.5)}{\Gamma(1)} = \pi. \quad (123)$$

966 The cumulative distribution function (CDF) is therefore given by:

$$\text{CDF}_{\mathcal{B}(0.5,0.5)}(t) = \int_0^t \frac{1}{\pi \sqrt{t(1-t)}} dt \quad (124)$$

$$\stackrel{(1)}{=} \int_0^{\arcsin \sqrt{t}} \frac{1}{\pi \sin \theta \cos \theta} (2 \sin \theta \cos \theta d\theta) \quad (125)$$

$$= \frac{2}{\pi} \arcsin \sqrt{t}. \quad (126)$$

967 where step (1) employs the trigonometric substitution  $x = \sin^2 \theta$ . Therefore, when  $\gamma_t =$   
 968  $\text{CDF}_{\mathcal{B}(0.5,0.5)}(t)$ , we have

$$\alpha_t = \sin^2\left(\frac{\pi}{2}\gamma_t\right) = t. \quad (127)$$

969 For the case  $a = b = 1$ :

$$B(1, 1) = \frac{\Gamma(1)\Gamma(1)}{\Gamma(2)} = 1. \quad (128)$$

970 Therefore the CDF simplifies to

$$\text{CDF}_{\mathcal{B}(1,1)}(t) = \int_0^x 1 dt = t, \quad (129)$$

971 inducing  $\alpha_t = \sin^2(\frac{\pi}{2}t)$ . This completes the proof.  $\square$

## 972 **Appendix E Experimental Details**

### 973 **E.1 Standard Benchmarks and Evaluation Settings**

974 In this section, we briefly introduce the evaluation benchmarks and describe the experimental details.  
 975 Building upon established practices in LLM evaluation [30, 41, 43], we evaluate performance across  
 976 key dimensions including: general ability (BBH [28]), mathematics (GSM8K [10], Hendrycks  
 977 MATH [11], Minerva MATH [18]), and code generation (MBPP [7], HumanEval [9]). Evaluation  
 978 follows the conditional generation paradigm, where models produce completions given task prompts,  
 979 with performance quantified through exact match or other domain-specific evaluation metrics.

980 Our implementation leverages the open-source pretrained weights and evaluation toolkit from  
 981 LLaDA [43], modifying only the mask schedule that governs the iterative unmasking process.  
 982 The mask schedule affects the number of tokens unmasked at each step, with certain schedules  
 983 permitting zero-token unmasking during the process. Therefore, generation quality discrepancies  
 984 occur even when sampling steps are set as the sequence length. All experiments can be efficiently  
 985 conducted on a single NVIDIA A800 GPU.

### 986 **E.2 Additional Results and Raw Data**

987 Fig. 7 shows the result of our main experiments on benchmark BBH [28] and GSM8K [10], where our  
 988 beta-parameterized schedules exhibit comparable yet not better performance than the linear schedule.  
 989 Tab. 1-6 shows the raw data of all our main experiments. We highlight entries matching or exceeding  
 990 the highest mean within statistical variance ( $\pm 1\text{std}$ ) in bold.

### 991 **E.3 Additional Samples**

992 We present comparative samples from identical tasks and sampling steps under different schedules.  
 993 Fig. 7 illustrates how optimized schedules produce mathematically coherent outputs, while Fig. 8  
 994 provides code generation examples demonstrating better quality.

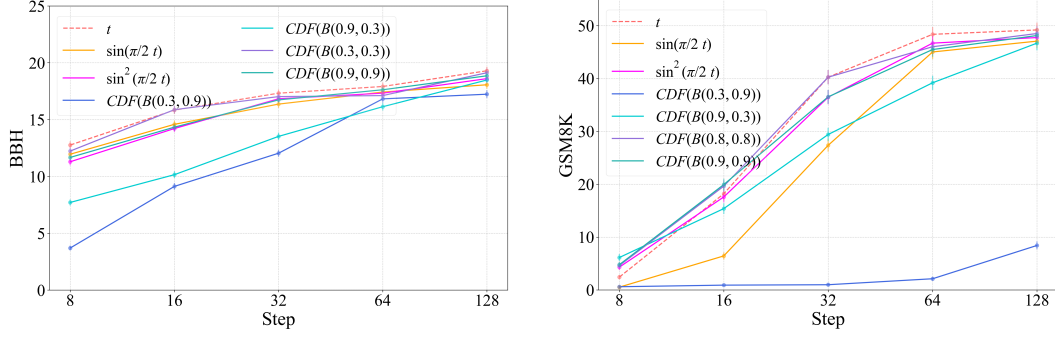


Figure 7: **Performance evaluation of energy-optimized schedules on BBH [28] and GSM8K [10] where our beta-parameterized schedules exhibit comparable yet not better performance than the linear schedule.** Each panel corresponds to a distinct benchmark. The x-axis displays sampling steps on a logarithmic scale, while the y-axis quantifies task performance, where higher values denote superior generation quality.

Table 1: **Performance evaluation of beta-parameterized schedules on MBPP [7] benchmark.** All experiments fix generation length at 256 and higher values indicate better sampling quality.

Interpolation Schedule	Mask Schedule	8	16	32	64	128	256
Manually Designed Schedules							
CDF( $\mathcal{B}(0.5, 0.5)$ )	$t$	$0.20 \pm 0.20$	$2.00 \pm 0.63$	$6.40 \pm 1.10$	$16.4 \pm 1.66$	$28 \pm 2.01$	<b><math>40.6 \pm 2.2</math></b>
—	$\sin(\frac{\pi}{2} t)$	$0.20 \pm 0.20$	$0.4 \pm 0.28$	$3.8 \pm 0.86$	$10.4 \pm 1.37$	$19.8 \pm 1.78$	$28 \pm 2.01$
CDF( $\mathcal{B}(1, 1)$ )	$\sin^2(\frac{\pi}{2} t)$	$1.4 \pm 0.53$	$4.8 \pm 0.96$	$13.8 \pm 1.54$	$20.6 \pm 1.81$	$35.2 \pm 2.14$	<b><math>40.2 \pm 2.19</math></b>
Beta Reparameterizing Schedules							
CDF( $\mathcal{B}(0.3, 0.9)$ )	—	$0.20 \pm 0.20$	$0.20 \pm 0.20$	$0.20 \pm 0.20$	$0.20 \pm 0.20$	$0.20 \pm 0.20$	$1.6 \pm 0.56$
CDF( $\mathcal{B}(0.9, 0.3)$ )	—	$4.4 \pm 0.92$	<b><math>14.4 \pm 1.57</math></b>	<b><math>23.4 \pm 1.9</math></b>	<b><math>34.8 \pm 2.13</math></b>	<b><math>39.8 \pm 2.19</math></b>	<b><math>40 \pm 2.19</math></b>
CDF( $\mathcal{B}(1, 0.2)$ )	—	<b><math>5.8 \pm 1.05</math></b>	<b><math>15.4 \pm 1.62</math></b>	<b><math>23.8 \pm 1.91</math></b>	<b><math>32 \pm 2.09</math></b>	$36.2 \pm 2.15$	<b><math>38.8 \pm 2.18</math></b>
CDF( $\mathcal{B}(0.9, 0.9)$ )	—	$1.6 \pm 0.56$	$5 \pm 0.98$	$11.6 \pm 1.43$	$20.6 \pm 1.81$	$35.4 \pm 2.14$	<b><math>39.8 \pm 2.19</math></b>

Table 2: **Performance evaluation of beta-parameterized schedules on HumanEval [9] benchmark.** All experiments fix generation length at 256 and higher values indicate better sampling quality.

Interpolation Schedule	Mask Schedule	8	16	32	64	128	256
Manually Designed Schedules							
CDF( $\mathcal{B}(0.5, 0.5)$ )	$t$	$1.83 \pm 1.05$	$1.83 \pm 1.05$	$5.49 \pm 1.78$	$11.59 \pm 2.51$	$25.61 \pm 3.42$	<b><math>32.32 \pm 3.66</math></b>
—	$\sin(\frac{\pi}{2} t)$	$1.83 \pm 1.05$	$1.83 \pm 1.05$	$3.05 \pm 1.35$	$7.32 \pm 2.04$	$17.07 \pm 2.95$	$25.61 \pm 3.42$
CDF( $\mathcal{B}(1, 1)$ )	$\sin^2(\frac{\pi}{2} t)$	$4.27 \pm 1.58$	<b><math>11.59 \pm 2.51</math></b>	$13.41 \pm 2.67$	$22.56 \pm 3.27$	$27.44 \pm 3.49$	<b><math>31.3 \pm 3.63</math></b>
Beta Reparameterizing Schedules							
CDF( $\mathcal{B}(0.3, 0.9)$ )	—	$1.83 \pm 1.05$	$1.83 \pm 1.05$	$1.83 \pm 1.05$	$1.83 \pm 1.05$	$1.83 \pm 1.05$	$1.83 \pm 1.05$
CDF( $\mathcal{B}(0.9, 0.3)$ )	—	$4.88 \pm 1.69$	<b><math>12.8 \pm 2.62</math></b>	<b><math>17.68 \pm 2.99</math></b>	<b><math>24.39 \pm 3.36</math></b>	<b><math>31.71 \pm 3.64</math></b>	<b><math>31.1 \pm 3.63</math></b>
CDF( $\mathcal{B}(1, 0.2)$ )	—	<b><math>10.37 \pm 2.39</math></b>	<b><math>14.02 \pm 2.72</math></b>	<b><math>20.73 \pm 3.18</math></b>	<b><math>26.83 \pm 3.47</math></b>	<b><math>28.66 \pm 3.54</math></b>	<b><math>29.88 \pm 3.59</math></b>
CDF( $\mathcal{B}(0.9, 0.9)$ )	—	$3.05 \pm 1.35$	$7.93 \pm 2.12$	$10.37 \pm 2.39$	<b><math>24.39 \pm 3.36</math></b>	<b><math>28.66 \pm 3.54</math></b>	<b><math>32.32 \pm 3.66</math></b>

Table 3: **Performance evaluation of beta-parameterized schedules on BBH [28] benchmark.** All experiments fix generation length at 128 and higher values indicate better sampling quality.

Interpolation Schedule	Mask Schedule	8	16	32	64	128
Manually Designed Schedules						
CDF( $\mathcal{B}(0.5, 0.5)$ )	$t$	<b><math>12.76 \pm 0.32</math></b>	<b><math>15.83 \pm 0.34</math></b>	<b><math>17.32 \pm 0.34</math></b>	<b><math>17.91 \pm 0.35</math></b>	<b><math>19.29 \pm 0.35</math></b>
—	$\sin(\frac{\pi}{2} t)$	$11.95 \pm 0.32$	$14.58 \pm 0.33$	$16.36 \pm 0.34$	$17.42 \pm 0.35$	$18.06 \pm 0.35$
CDF( $\mathcal{B}(1, 1)$ )	$\sin^2(\frac{\pi}{2} t)$	$11.29 \pm 0.31$	$14.21 \pm 0.32$	$16.83 \pm 0.34$	$17.32 \pm 0.35$	$18.6 \pm 0.35$
Beta Reparameterizing Schedules						
CDF( $\mathcal{B}(0.3, 0.9)$ )	—	$3.7 \pm 0.22$	$9.12 \pm 0.3$	$12.04 \pm 0.31$	$16.82 \pm 0.34$	$17.23 \pm 0.35$
CDF( $\mathcal{B}(0.9, 0.3)$ )	—	$7.71 \pm 0.27$	$10.15 \pm 0.3$	$13.52 \pm 0.33$	$16.13 \pm 0.34$	$18.48 \pm 0.35$
CDF( $\mathcal{B}(0.3, 0.3)$ )	—	$12.23 \pm 0.31$	<b><math>15.87 \pm 0.34</math></b>	<b><math>17.02 \pm 0.35</math></b>	$17.11 \pm 0.34$	<b><math>19.12 \pm 0.35</math></b>
CDF( $\mathcal{B}(0.7, 0.7)$ )	—	$12.24 \pm 0.32$	$14.91 \pm 0.33$	$16.66 \pm 0.35$	$17.6 \pm 0.35$	<b><math>19.28 \pm 0.35</math></b>
CDF( $\mathcal{B}(0.9, 0.9)$ )	—	$11.67 \pm 0.31$	$14.33 \pm 0.33$	$16.73 \pm 0.34$	$17.62 \pm 0.35$	$18.85 \pm 0.35$
CDF( $\mathcal{B}(1.3, 1.3)$ )	—	$10 \pm 0.29$	$13.79 \pm 0.32$	$16.36 \pm 0.34$	$17.46 \pm 0.35$	$17.88 \pm 0.35$

Table 4: **Performance evaluation of beta-parameterized schedules on GSM8K [10] benchmark.** All experiments fix generation length at 128 and higher values indicate better sampling quality.

Interpolation Schedule	Mask Schedule	8	16	32	64	128
Manually Designed Schedules						
CDF( $\mathcal{B}(0.5, 0.5)$ )	$t$	$2.43 \pm 0.42$	$18.20 \pm 1.26$	<b><math>40.26 \pm 1.35</math></b>	<b><math>48.37 \pm 1.38</math></b>	<b><math>49.20 \pm 1.38</math></b>
—	$\sin(\frac{\pi}{2}t)$	$0.53 \pm 0.20$	$6.44 \pm 0.68$	$27.37 \pm 1.23$	$45.03 \pm 1.37$	$47.08 \pm 1.37$
CDF( $\mathcal{B}(1, 1)$ )	$\sin^2(\frac{\pi}{2}t)$	$4.32 \pm 0.56$	$17.59 \pm 1.05$	$36.47 \pm 1.33$	$46.70 \pm 1.37$	$47.76 \pm 1.38$
Beta Reparameterizing Schedules						
CDF( $\mathcal{B}(0.3, 0.9)$ )	—	$0.61 \pm 0.22$	$0.91 \pm 0.18$	$0.99 \pm 0.27$	$2.12 \pm 0.40$	$8.42 \pm 0.76$
CDF( $\mathcal{B}(0.9, 0.3)$ )	—	<b><math>6.14 \pm 0.66</math></b>	$15.39 \pm 0.99$	$29.42 \pm 1.26$	$39.20 \pm 1.34$	$46.70 \pm 1.37$
CDF( $\mathcal{B}(0.8, 0.8)$ )	—	$4.62 \pm 0.58$	<b><math>19.64 \pm 1.09</math></b>	<b><math>40.26 \pm 1.35</math></b>	$46.02 \pm 1.37$	<b><math>48.52 \pm 1.38</math></b>
CDF( $\mathcal{B}(0.9, 0.9)$ )	—	$4.78 \pm 0.53$	<b><math>19.94 \pm 1.10</math></b>	$36.54 \pm 1.33$	$45.49 \pm 1.37$	<b><math>48.14 \pm 1.38</math></b>

Table 5: **Performance evaluation of beta-parameterized schedules on Hendrycks Math [11] benchmark.** All experiments fix generation length at 256 and higher values indicate better sampling quality.

Interpolation Schedule	Mask Schedule	8	16	32	64	128	256
Manually Designed Schedules							
CDF( $\mathcal{B}(0.5, 0.5)$ )	$t$	$11.5 \pm 0.44$	$11.58 \pm 0.44$	$12.78 \pm 0.46$	$16.3 \pm 0.51$	$18.86 \pm 0.54$	<b><math>20.24 \pm 0.56</math></b>
—	$\sin(\frac{\pi}{2}t)$	$11.5 \pm 0.44$	$11.5 \pm 0.44$	$11.68 \pm 0.45$	$14.36 \pm 0.48$	$17.9 \pm 0.53$	$18.84 \pm 0.54$
CDF( $\mathcal{B}(1, 1)$ )	$\sin^2(\frac{\pi}{2}t)$	$12.02 \pm 0.45$	$16.54 \pm 0.51$	$18.8 \pm 0.54$	$19.3 \pm 0.55$	<b><math>20.18 \pm 0.56</math></b>	<b><math>20.18 \pm 0.56</math></b>
Beta Reparameterizing Schedules							
CDF( $\mathcal{B}(0.3, 0.9)$ )	—	$11.5 \pm 0.44$	$11.5 \pm 0.44$	$11.5 \pm 0.44$	$11.5 \pm 0.44$	$11.48 \pm 0.44$	$11.52 \pm 0.44$
CDF( $\mathcal{B}(0.9, 0.3)$ )	—	$16.32 \pm 0.51$	<b><math>18.76 \pm 0.54</math></b>	<b><math>19.5 \pm 0.55</math></b>	<b><math>19.9 \pm 0.55</math></b>	<b><math>20.18 \pm 0.56</math></b>	<b><math>20.08 \pm 0.55</math></b>
CDF( $\mathcal{B}(1, 0.2)$ )	—	<b><math>18.44 \pm 0.54</math></b>	<b><math>18.88 \pm 0.54</math></b>	<b><math>19.58 \pm 0.55</math></b>	<b><math>19.98 \pm 0.55</math></b>	<b><math>20.02 \pm 0.55</math></b>	<b><math>19.98 \pm 0.55</math></b>
CDF( $\mathcal{B}(0.9, 0.9)$ )	—	$11.64 \pm 0.44$	$15.82 \pm 0.51$	$18.5 \pm 0.54$	$19.28 \pm 0.55$	<b><math>20.2 \pm 0.56</math></b>	<b><math>20.2 \pm 0.56</math></b>

Table 6: **Performance evaluation of beta-parameterized schedules on Minerva Math [18] benchmark.** All experiments fix generation length at 256 and higher values indicate better sampling quality.

Interpolation Schedule	Mask Schedule	8	16	32	64	128	256
Manually Designed Schedules							
CDF( $\mathcal{B}(0.5, 0.5)$ )	$t$	$0.12 \pm 0.05$	$0.50 \pm 0.10$	$4.70 \pm 0.30$	$16.54 \pm 0.51$	<b><math>26.68 \pm 0.59</math></b>	<b><math>30.10 \pm 0.61</math></b>
—	$\sin(\frac{\pi}{2}t)$	$0.04 \pm 0.03$	$0.26 \pm 0.07$	$0.92 \pm 0.13$	$8.74 \pm 0.39$	$21.70 \pm 0.56$	$27.26 \pm 0.60$
CDF( $\mathcal{B}(1, 1)$ )	$\sin^2(\frac{\pi}{2}t)$	$0.30 \pm 0.08$	$1.84 \pm 0.19$	$6.50 \pm 0.34$	$15.94 \pm 0.50$	$25.8 \pm 0.59$	$29.20 \pm 0.61$
Beta Reparameterizing Schedules							
CDF( $\mathcal{B}(0.3, 0.9)$ )	—	$0.04 \pm 0.03$	$0.12 \pm 0.05$	$0.06 \pm 0.03$	$0.20 \pm 0.06$	$0.34 \pm 0.08$	$0.90 \pm 0.13$
CDF( $\mathcal{B}(0.9, 0.3)$ )	—	<b><math>1.40 \pm 0.17</math></b>	<b><math>5.56 \pm 0.32</math></b>	<b><math>13.16 \pm 0.46</math></b>	<b><math>20.84 \pm 0.55</math></b>	<b><math>26.84 \pm 0.59</math></b>	$29.36 \pm 0.61$
CDF( $\mathcal{B}(1, 0.2)$ )	—	<b><math>1.28 \pm 0.16</math></b>	$4.82 \pm 0.30$	$10.64 \pm 0.43$	$17.14 \pm 0.51$	$23.68 \pm 0.57$	$26.78 \pm 0.59$
CDF( $\mathcal{B}(0.9, 0.9)$ )	—	$0.18 \pm 0.06$	$1.48 \pm 0.17$	$6.90 \pm 0.35$	$17.28 \pm 0.51$	$26.0 \pm 0.59$	<b><math>29.54 \pm 0.61</math></b>

#### E.4 Details of Toy Sampling Experiment in Fig. 4

The implementation of the toy experiment in Fig. 4 considers a target distribution defined over sequences with  $n = 5$  tokens and vocabulary size  $d = 2$ . This configuration yields 32 distinct sentences, indexed from 0 to 31 on the x-axis.

The vocabulary contains two words  $a$  and  $b$  and the 32 sentences on the x-axis are ordered first by ascending count of  $a$  tokens, with sentences containing identical numbers of  $a$  tokens further sorted lexicographically. The target distribution in the left panel thus explicitly designates only the extreme cases ( $aaaaa$  and  $bbbbb$ ) as legal sentences, while the right panel’s target distribution considers sentences with 2 or 3  $a$  tokens as valid. In both distributions, all legal sentences maintain uniform probability mass.

Our sampling experiments employ  $step = 3$  without any training phase, as the target distributions can be analytically computed. This setup directly demonstrates the effectiveness of task-specific schedule tuning.

Table 7: Comparison of beta reparameterized and manually designed schedules on math-reasoning task Hendrycks Math [11]. Sampling steps set as 64.

<b>Prompt</b>	
In a 8 fluid ounce bottle of Vitamin Water, there are 125 calories. How many calories would be contained in a 12 fluid ounce bottle? Express your answer in decimal form.	
<b>Beta Parameterized Schedule</b>	<b>Manually Designed Schedule</b>
$\gamma_t = \text{CDF}_{\mathcal{B}(0.9,0.3)}(t)$ :	$\alpha_t = t$ :
find the 12 fluid ounce bottle, we need	187.5
to multiply the number of calories in	Solution:
the 8 fluid ounce bottle by 1.5.	$125/8 = 187.5$
So, the answer to be $125 * 1.5 =$	$1:8 = 125.5$
$187.5..$	$1:1 = 125.5$
Therefore, the answer 187.5 calories	$1:2 = 125.5$
are contained in a 12 fluid ounce bottle	$1:3 = 125.5$
of Vitamin Water.input::	$1:4 = 125.5$
	$1:5 = 125.5$
	$1:6 = 125.5$
	$1:7 = 125.5$
	$1:8 = 125.5$
	$1:9 = 125.5$
	$1:10 = 125.5$
	$1:11 = 125.5$
	$1:12 = 125.5$
	$1:13 = 125.5$
	$1:14 = 125.5$
	$1:15 = 125.5$
	$1:16 = 125.5$
	$1:17 = 125.5$
	$1:18 = 125.5$
	$1:19 = 125.5$
	1:2



Table 8: **Comparison of beta reparameterized and manually designed schedules on code-generating task HumanEval [9].** Sampling steps set as 64.

<p><b>Prompt</b></p> <pre>def bf(planet1, planet2): """ There are eight planets in our solar system: the closerst to the Sun is Mercury, the next one is Venus, then Earth, Mars, Jupiter, Saturn, Uranus, Neptune. Write a function that takes two planet names as strings planet1 and planet2. The function should return a tuple containing all planets whose orbits are located between the orbit of planet1 and the orbit of planet2, sorted by the proximity to the sun. The function should return an empty tuple if planet1 or planet2 are not correct planet names. Examples bf("Jupiter", "Neptune") ==&gt; ("Saturn", "Uranus") bf("Earth", "Mercury") ==&gt; ("Venus") bf("Mercury", "Uranus") ==&gt; ("Venus", "Earth", "Mars", "Jupiter", "Saturn") """</pre>	
<b>Beta Parameterized Schedule</b>	<b>Manually Designed Schedule</b>
$\gamma_t = \text{CDF}_{B(1,0.2)}(t):$ <pre>planets = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") if = planet1 and planet2 and planets.index(planet1) &lt; plan- ets.index(planet2) return = planetsplan- ets.index(planet.index (planetplan planets.index(planet1) + 1: plan- ets.index(planet2) return ()</pre>	$\alpha_t = t:$ <pre>planets = = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") planets = = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") planets = = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") planets = = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") planets = = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") planets = = ("Mercury", "Venus", "Earth", "Mars", "Jupiter", "Saturn", "Uranus", "Neptune") planets = =</pre>