POSTERIOR REFINEMENT:

FEW-STEP LANGUAGE GENERATION VIA ANY-ORDER FLOW MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-autoregressive generation offers a powerful paradigm for iterative refinement, allowing models to recursively critique, erase and regenerate arbitrary subsets of tokens. However, existing non-autoregressive models fail to realize this potential. Masked Diffusion Models (MDMs) suffer from factorization error, causing sample quality to collapse when generating multiple tokens simultaneously. Flow Map Language Models (FMLMs) circumvent this bottleneck via joint sequence transport for excellent few-step generation, but sacrifice the inference-time flexibility of MDMs. We introduce FMLM+, a framework that bridges this gap by equipping FMLM with masking-style noise schedules. While generating the full sequence in a single step, FMLM+ simultaneously scores the global consistency of each token *a posteriori*. We leverage this to introduce *Posterior Refinement*, a novel inference-time refinement strategy that enables the model to adaptively self-correct its outputs, matching the performance of discrete baselines with 32× fewer NFEs. Across diverse benchmarks, we demonstrate that FMLM+ with Posterior Refinement improves the speed–quality tradeoff over both MDM and FMLM families, providing a scalable foundation for high-fidelity language modeling.

1 INTRODUCTION

Contemporary language models are predominantly autoregressive (AR), necessitating L sequential forward passes to generate a sequence of length L [1–3]. While this paradigm has scaled remarkably well, its inherent sequential bottleneck makes inference slow and memory-bound [4, 5]. These inefficiencies cascade into downstream applications such as inference-time scaling, reasoning, and reinforcement learning, which are inference-bottlenecked, ultimately hindering scalability [3].

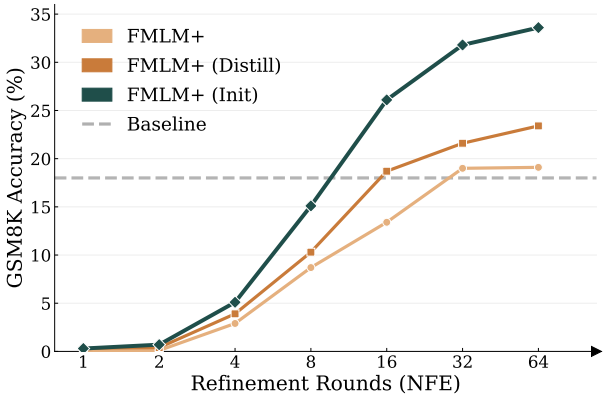

Figure 1: **Iterative refinement.** FMLM+ unlocks self-correction capabilities, outperforming all diffusion baselines with 1024 function evaluations using as few as 32 rounds of Posterior Refinement.

Masked diffusion models (MDMs) have emerged as the leading non-autoregressive alternative for language modeling [6–9], with recent efforts scaling them to billions of parameters [10–14]. In contrast to fixed causal orderings, MDMs model any-order factorizations of the sequence by training on arbitrary masking patterns. During inference, this permits adaptive unmask-and-remask decoding strategies that commit high-confidence tokens first [15–17]. While such a flexible generative procedure has proven useful in planning-oriented tasks [18], MDMs typically fall short in performance when multiple tokens are sampled simultaneously [19–21]. This degradation, known as *factorization error*, arises because MDMs approximate the joint distribution over full sequences with a product of per-token marginals, discarding inter-token correlations.

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.

Several works have also explored diffusion [22, 23] and flow-matching [24–26] over continuous representations, spanning fixed embeddings [27, 28], learned latent spaces [29, 30], and one-hot relaxations [31–34]. Recently, continuous-space approaches have been shown to be competitive with their discrete-diffusion counterparts in language modeling [35–40] and structured reasoning [41, 42]. A key advantage of these methods is their ability to circumvent factorization error by learning a sample-level transport from source to target distributions. In particular, Flow Map Language Models (FMLMs) [34, 35, 39, 40] train flows over one-hot embeddings in Euclidean space, learning a joint transport from Gaussian noise to one-hot vectors. Because the learned flow is deterministic, it can further be distilled into a flow map [43], which substantially improves the quality of few-step language generation, including generating text in just a *single* function evaluation [35].

Beyond fast, parallel inference, non-autoregressive generation offers a more profound theoretical advantage over autoregressive modeling: the capacity for *iterative refinement* of generated text. While traditional causal language models are fundamentally constrained by a one-way, left-to-right generation, iterative refinement requires a paradigm that allows the model to globally critique, erase, and regenerate arbitrary subsets of tokens to continually improve its output. This capacity directly mirrors human cognition; when composing complex text, humans rarely write perfectly in a single left-to-right pass. Instead, we efficiently formulate a rough draft to capture the global context, and then iteratively assess, delete, and rewrite specific parts of text while keeping the overall picture in context [44].

However, existing non-autoregressive paradigms fail to fully realize this potential. While MDMs offer flexible, any-order inference, their factorization error renders few-step generation infeasible. On the contrary, FMLMs resolve this factorization error and enable few-step generation, yet lack the inference-time flexibility native to MDMs because they evolve all tokens simultaneously. We argue that the respective strengths of these two paradigms constitute the core primitives for effective iterative refinement, which requires both — efficient drafting of coherent multi-token blocks and the flexibility to retain arbitrary token subsets while regenerating the rest.

Motivated by this insight, we introduce FMLM+, a unified framework that equips FMLMs with masking-style noise schedules. By coupling single-step sequence generation from FMLMs with flexible conditioning from MDMs, our architecture naturally supports iterative refinement. Crucially, our approach yields an *a posteriori* confidence score for each token conditioned on the full draft that contrasts sharply with the *a priori* confidence derived from independent token-wise marginals in MDMs. We leverage this property to develop Posterior Refinement (PR) — an iterative refinement strategy that recursively retains high-confidence tokens and regenerates the rest to keep improving sample quality with each additional round of refinement (Fig. 1). Subsequently, FMLM+ with Posterior Refinement establishes a new state-of-the-art that consistently outperforms both MDM and FMLM baselines across all considered benchmarks with far fewer function evaluations (NFEs).

Our contributions are as follows:

1. **A unifying framework.** We introduce FMLM+, a generalization of FMLMs to masking-style noise schedules that combines the few-step efficiency of FMLMs with the any-order flexibility of MDMs. We further establish a connection between MDMs and FMLM+, providing a direct path to scale FMLM+ leveraging advances in large-scale MDM pre-training [10, 11].

2. **Posterior refinement.** We introduce Posterior Refinement, an *after-the-fact* confidence-based strategy that enables FMLM+ to recursively self-correct its output while generating all remaining tokens with *single* function evaluations (Fig. 2). We identify how existing a-priori confidence heuristics of MDMs fail on simple toy problems and demonstrate how Posterior Refinement is able to rectify these shortcomings.

3. **Extensive evaluation.** We demonstrate that FMLM+ navigates the speed–quality trade-off more effectively than both MDM and FMLM families across four benchmarks: TinyStories [45], OpenWebText [46], GSM8K [47], and Sudoku, matching or surpassing the strongest baselines with up to 32× fewer NFEs.

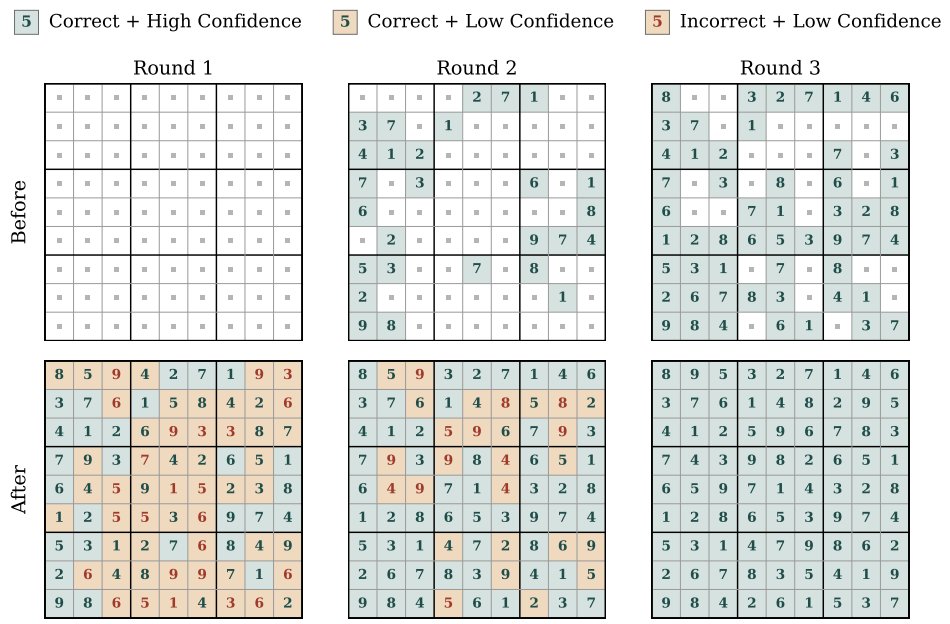

Figure 2: **Posterior Refinement with FMLM+.** By evaluating token consistency against the full generated sequence, FMLM+ effectively identifies its own errors. Crucially, incorrect tokens consistently form a subset of low-confidence generations, allowing Posterior Refinement to reliably filter and revise errors.

2 Background

In this section, we provide the necessary background on related model families. Let $\mathcal{V}$ denote a vocabulary, with $|\mathcal{V}| = V$, and let $\mathbf{y} = (\mathbf{y}^l)_{l=1}^L \in \mathcal{V}^L$ be a sequence of length L drawn from a data distribution $p(\mathbf{y})$.

Autoregressive Models. The AR paradigm models the data distribution through a left-to-right factorization $p(\mathbf{y}) = p(\mathbf{y}^1)p(\mathbf{y}^2|\mathbf{y}^1)\dots p(\mathbf{y}^L|\mathbf{y}^{<L})$, and learns the conditional distributions $p(\mathbf{y}^l \mid \mathbf{y}^{<l})$ over tokens given their preceding context [48–50]. Generation proceeds sequentially by sampling one token at a time conditioned on previously generated tokens.

Masked Diffusion Models. Masked Diffusion Models (MDMs) [6–9] offer an alternative paradigm for discrete generative modeling, parameterizing marginals over clean tokens given a partially masked input. MDMs extend the vocabulary $\mathcal{V}$ with an extra $[\texttt{MASK}]$ token. A monotonic fraction $\alpha(t)$ of clean tokens $\mathbf{y}$ is masked to produce $\mathbf{y}_t$. The token-wise marginal $p_\theta(\mathbf{y}^l \mid \mathbf{y}_t)$ is then trained via a classification loss over the masked positions,

$$\mathcal{L}_{\text{MDM}}(p_\theta) = -\mathbb{E}_t \left[w(t) \sum_{l=1}^{L} \mathbf{1}[\mathbf{y}_t^l = \texttt{MASK}] \log p_\theta(\mathbf{y}^l \mid \mathbf{y}_t) \right], \tag{1}$$

where $w(t) > 0$ is some weighting function.

At inference, MDMs start from a fully masked sequence and generate tokens via an iterative unmask–remask procedure; at each step we categorically sample from the marginal distribution for each token independently and then remask some tokens, repeating until all tokens are unmasked. Sampling every masked position in a single step thus relies on the following approximation:

$$p_\theta(\mathbf{y} \mid \mathbf{y}_t) \approx \prod_{l\,:\,\mathbf{y}_t^l = [\texttt{MASK}]} p_\theta(\mathbf{y}^l \mid \mathbf{y}_t). \tag{2}$$

This approximation neglects inter-token correlations, introducing a *factorization error* which compounds as more tokens are sampled per step. A natural decoding strategy for MDMs is to commit tokens in decreasing order of confidence, heuristically approximated as $\max_v p_\theta(\mathbf{y}^l = v \mid \mathbf{y}_t)$ [16, 17].

Flow Matching and Flow Maps. Flow matching [24–26] learns a time-dependent velocity field to transport samples from a simple source distribution (e.g., Gaussian noise) to the target data distribution. We define $\mathbf{x}_1 = f(\mathbf{y}) \in \mathbb{R}^{L\times V}$ as the one-hot representation of clean data, and sample $\mathbf{x}_0 \sim \mathsf{N}(0, I) \in \mathbb{R}^{L\times V}$ as isotropic Gaussian noise of matching shape to construct the linear interpolant in Euclidean space:

$$I_t \coloneqq (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad t \in [0,1]. \tag{3}$$

This interpolant admits a deterministic probability flow ODE defined by a velocity field $b : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$,

$$\dot{\mathbf{x}}_t = b_t(\mathbf{x}_t), \qquad b_t(\mathbf{x}) = \mathbb{E}[\dot{I}_t \mid I_t = \mathbf{x}] = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid I_t = \mathbf{x}], \tag{4}$$

whose solution transports samples from the source distribution to the target data distribution [24–26]. In practice, we parameterize the flow velocity b_t with a neural network, and inference requires a numerical integration of the probability flow ODE to generate samples.

The flow map $X_{s,t} : [0,1]^2 \times \mathbb{R}^{L\times V} \to \mathbb{R}^{L\times V}$ associated with the velocity field b_t is defined as the solution operator that transports the state at time s to the state at time t [43]:

$$X_{s,t}(\mathbf{x}_s) = \mathbf{x}_t = \mathbf{x}_s + \int_s^t b_\tau(\mathbf{x}_\tau)d\tau = \mathbf{x}_s + (t-s)v_{s,t}(\mathbf{x}_s), \tag{5}$$

where $v_{s,t}(\mathbf{x}_s)$ is the average velocity (mean flow) between times s and t [51]. In the limit $s \to t$, the average velocity recovers the instantaneous velocity, $\lim_{s\to t} v_{s,t}(\mathbf{x}) = b_t(\mathbf{x})$. Crucially, the flow map allows us to arbitrarily jump along the flow, including generating from the target distribution in a *single* function evaluation $\mathbf{x}_1 = X_{0,1}(\mathbf{x}_0)$ [51, 52].

Flow Map Language Models. Flow map language models (FMLM) [35] enable few-step language generation by parameterizing the *two-time denoiser* (TTD) $\delta_{s,t} : [0,1]^2 \times \mathbb{R}^{L\times V} \to \mathbb{R}^{L\times V}$,

$$\delta_{s,t}(\mathbf{x}) \coloneqq \mathbf{x} + (1-s)v_{s,t}(\mathbf{x}), \tag{6}$$

which can be viewed as taking an Euler step from time s to time 1 using the mean flow $v_{s,t}$. The TTD satisfies the boundary identity $\delta_{t,t}(\mathbf{x}) = \mathbb{E}[\mathbf{x}_1 \mid I_t = \mathbf{x}]$ [34, 35, 53] and further obeys the semigroup property

$$\delta_{s,t}(\mathbf{x}) = \gamma\, \delta_{s,u}(\mathbf{x}) + (1-\gamma)\, \delta_{u,t}(X_{s,u}(\mathbf{x})), \qquad \gamma = \frac{(1-t)(u-s)}{(1-u)(t-s)}, \tag{7}$$

for every $0 \le s \le u \le t \le 1$, where $X_{s,u}(\mathbf{x}) \coloneqq \frac{1-u}{1-s}\,\mathbf{x} + \frac{u-s}{1-s}\,\delta_{s,u}(\mathbf{x})$ [35].

Since the single-time denoiser $\delta_{t,t}$ is a conditional expectation of one-hot vectors, it lies on the per-token probability simplex, $\delta_{t,t}(\mathbf{x})^l \in \Delta^{V-1}$ for $l = 1, \dots, L$. Furthermore, Lee et al. [35] show that the two-time denoiser $\delta_{s,t}$ can be written as a weighted average of single-time denoisers $\delta_{t,t}$ along the flow trajectory. Consequently, $\delta_{s,t}$ also lies on the per-token simplex, $\delta_{s,t}(\mathbf{x})^l \in \Delta^{V-1}$ for $l = 1, \cdots, L$. We therefore parameterize $\delta_{s,t}$ with a token-wise softmax head and train it via a token-wise cross-entropy objective,

$$\mathcal{L}_{\mathsf{FMLM}}(\delta) \coloneqq \underbrace{\mathbb{E}_s \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\mathrm{CE}\left(\delta_{s,s}(I_s), \mathbf{x}_1\right)\right]}_{\mathcal{L}_{\mathrm{diag}}} + \underbrace{\mathbb{E}_{s,u,t} \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\mathrm{CE}\left(\delta_{s,t}(I_s), \bar{\delta}_{s,t}\right)\right]}_{\mathcal{L}_{\mathrm{off}}}, \tag{8}$$

where $\bar{\delta}_{s,t} \coloneqq \mathrm{stopgrad}\left(\gamma\, \delta_{s,u}(I_s) + (1-\gamma)\, \delta_{u,t}(X_{s,u}(I_s))\right)$. Intuitively, this loss trains the two-time denoiser to reconstruct clean data along the diagonal $s = t$ ($\mathcal{L}_{\mathrm{diag}}$) and to progressively distill a larger off-diagonal jump ($s \to t$) using two smaller jumps ($s \to u, u \to t$) as a teacher ($\mathcal{L}_{\mathrm{off}}$). While the training of flow maps typically uses the ℓ_2 metric for both diagonal and off-diagonal terms, operating on the probability simplex allows FMLMs to use a KL divergence objective that is better suited for discrete data.

During inference, we can arbitrarily jump along the flow by recovering the flow map from the TTD,

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\,\mathbf{x} + \frac{t-s}{1-s}\,\delta_{s,t}(\mathbf{x}). \tag{9}$$

This allows us to generate samples in as few as a *single* function evaluation $\mathbf{x}_1 = \delta_{0,1}(\mathbf{x}_0)$.

3 FMLM+

Continuous diffusion and flow-based models typically denoise the full sequence in parallel [23, 24, 54]. While these methods avoid factorization error by learning a sample-level transport, they lack the inference-time flexibility native to MDMs. To build an effective iterative refinement algorithm, we require the flexibility to retain arbitrary subsets of tokens while also being able to efficiently and coherently regenerate the rest. To this end, we introduce FMLM+, a generalization of FMLMs that natively learns to generate under diverse masking patterns (Sec. 3.1). We show that this allows for flexible decoding strategies at inference time (Sec. 3.2), such as FMLM- and MDM-like inference, and, importantly, unlocks novel iterative refinement schemes (Sec. 3.3).

3.1 Training

The key idea underlying FMLM+ is to generalize FMLMs by extending the noising process to a masking-style design, preserving a set of clean tokens $\mathbf{y}^{\mathcal{C}}$ at context positions $\mathcal{C} \subset \{1, \cdots, L\}$ while noising the rest:

$$I_t^l := \begin{cases} \mathbf{x}_1^l, & \text{if } l \in \mathcal{C} \\ (1-t)\mathbf{x}_0^l + t\mathbf{x}_1^l, & \text{else } (l \in [L] \setminus \mathcal{C}). \end{cases} \tag{10}$$

This enables FMLM+ to condition on arbitrary clean context without additional architectural components. In practice, this is implemented by converting the shared time conditioning in the DiT architecture [55] to a per-position conditioning with $t = 1$ for clean positions. For brevity, we drop this implicit conditioning on $\mathbf{y}^{\mathcal{C}}$ throughout the remainder of the paper, writing $\delta_{s,t}(\cdot)$ instead of $\delta_{s,t}(\cdot|\mathbf{y}^{\mathcal{C}})$.

We train the two-time denoiser $\delta_{s,t}$ with the following progressive self-distillation objective:

$$\mathcal{L}_{\text{FMLM+}}(\delta) = \mathbb{E}_{\mathcal{C}}\Big[\underbrace{\mathbb{E}_s\, \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\big[\text{CE}(\delta_{s,s}(I_s), \mathbf{x}_1) \odot m\big]}_{\mathcal{L}_{\text{diag}}} + \underbrace{\mathbb{E}_{s,u,t}\, \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\big[\text{CE}(\delta_{s,t}(I_s), \bar{\delta}_{s,t}) \odot m\big]}_{\mathcal{L}_{\text{off}}}\Big]. \tag{11}$$

Compared to the vanilla FMLM objective Equation (8), we have an additional expectation over $\mathcal{C}$, the set of context tokens, and use $m \in \{0,1\}^L$ with $m^l = \mathbf{1}[\, l \notin \mathcal{C}\,]$ to mask the loss from tokens that are already clean. To avoid extra complexity, we drop the time reparameterization trick from Lee et al. [35] and use the self-distillation objective (11) instead of the two-stage training in Lee et al. [35] for all FMLM+ runs.

Leveraging Pretrained MDMs. To accelerate the training of FMLM+, we propose a novel method to leverage pretrained MDMs based on the following observation. At $s = t = 0$, the off-diagonal term $\mathcal{L}_{\text{off}}$ in the FMLM+ training objective Equation (11) vanishes and the diagonal term $\mathcal{L}_{\text{diag}}$ reduces to:

$$\mathbb{E}_{\mathcal{C}}\left[\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\text{CE}(\delta_{0,0}(\mathbf{x}_0), \mathbf{x}_1) \odot m\right]\right]. \tag{12}$$

This token-wise classification objective predicts the ground-truth tokens using only the context $\mathbf{y}^{\mathcal{C}}$ and pure Gaussian noise. Because the model must rely exclusively on the context, this formulation is equivalent to the MDM training objective Equation (1), which similarly predicts targets from $\mathbf{y}^{\mathcal{C}}$ and $[\texttt{MASK}]$ tokens (Prop. A.1).

We use this equivalence to accelerate FMLM+ training in two ways:

1. **Distillation.** We utilize an MDM as a teacher model, training the FMLM+ student to match the teacher MDM's token probabilities for a given context. This approach reduces gradient variance and accelerates convergence [56]. Concretely, we use:

$$\mathcal{L}_{\text{Distill}}(\delta) = (1 - p_{\text{distill}})\, \mathcal{L}_{\text{FMLM+}}(\delta) + p_{\text{distill}}\, \mathbb{E}_{\mathcal{C}}\, \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\text{CE}\big(\delta_{0,0}(I_0),\, p_\theta(\cdot \mid \mathbf{y}^{\mathcal{C}})\big)\right], \tag{13}$$

   where $p_\theta(\cdot \mid \mathbf{y}^{\mathcal{C}})$ denotes the frozen MDM teacher's token-wise marginals given the same context. In practice, we apply the distillation loss to a random fraction $p_{\text{distill}} = 0.25$ of each batch, and $\mathcal{L}_{\text{FMLM+}}$ to the rest.

2. **Initialization.** We initialize FMLM+ weights directly from a pretrained MDM checkpoint, leveraging the similarity between the masked diffusion and continuous flow objectives [57]. We show how the DiT architecture can mechanistically adapt MDM weights to handle noisy input in Sec. 4.2.

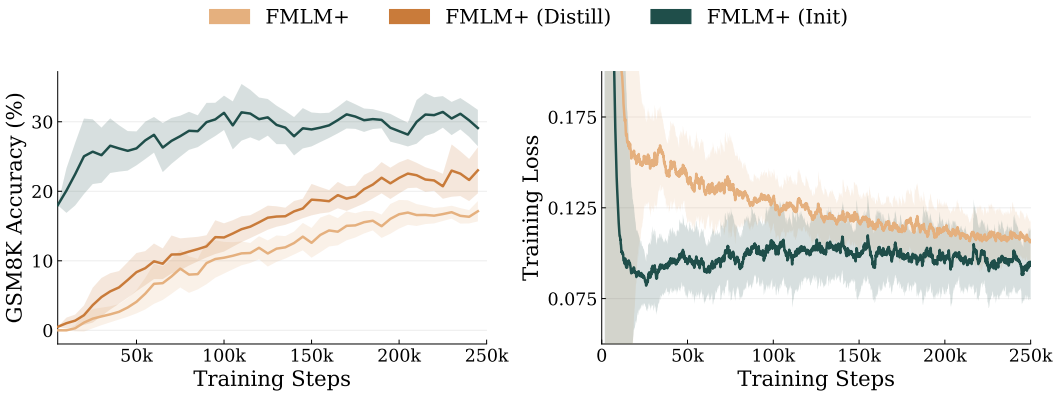

Figure 3: **Improved Training Techniques for FMLM+.** *Left:* downstream accuracy (↑) with 32 rounds of Posterior Refinement; both teacher-based variants outperform training from scratch. *Right:* training loss (↓); warm-started training converges faster and achieves a better optimum. Both metrics are reported as an exponential moving average ($\lambda = 0.99$).

Empirically, both distillation and initialization strategies significantly improve training efficiency and final performance. We find that warm starting from pre-trained MDM weights helps the model converge to a better optimum than training from scratch, highlighting the role of stochasticity and parameter initialization in training FMLMs (Fig. 3).

3.2 Inference

The core advantage of training FMLM+ on diverse conditioning patterns is that it allows us to flexibly choose between various decoding strategies during inference. As a simple example, we can recover the standard FMLM inference by defining a time grid $0 = t_0 < t_1 < \cdots < t_N = 1$ and denoising all unknown tokens simultaneously, jumping in N steps along the flow trajectory via Equation (9).

Algorithm 1 FMLM+ Inference.

```
# delta: Two Time Denoiser
# C: commit criterion; R rounds
x = randn(L, V)
s, t = zeros(L), ones(L)
for _ in range(R):
    x_1 = delta(x, s, t, flow_steps)
    tokens = x_1.argmax(axis=1)

    commit = commit | C(tokens, x_1)

    x = where(commit,
        one_hot(tokens), randn(L, V))
    s = where(commit, 1, 0)
return x_1
```

More importantly, FMLM+ allows us to construct a mechanism for iterative refinement as a loop that performs parallel generation followed by committing a subset of tokens (Alg. 1). At each iteration i, we propose a candidate $\hat{\mathbf{x}}_1$ via FMLM inference. Then, we select a subset $\mathcal{U}$ to retain from this candidate sequence and round them into discrete tokens via an argmax operation. Finally, we add them to our context set $\mathcal{C} := \mathcal{C} \cup \mathcal{U}$ and re-noise all remaining tokens $[L] \setminus \mathcal{C}$ for the next iteration. This, for example, allows for block-diffusion-style inference [58] by choosing $\mathcal{U}$ as the next contiguous block of B tokens.

We may further perform MDM-like, confidence-based inference by noting the equivalence between $\delta_{0,0}$ and MDMs (12). Consequently, by sampling tokens from $\delta_{0,0}(\mathbf{x}_0)$ and by prioritizing positions with sharper $\delta_{0,0}(\mathbf{x}_0)$, we recover MDM's confidence-based sampling [16–18]. We note that this approach, applied either to FMLM+ with $\delta_{0,0}$ or MDMs, generates tokens and estimates confidence *a priori*, using only the partial context available at that step and without knowing what tokens will be generated at other positions.

3.3 Posterior Refinement

Beyond recovering existing inference schemes as in Sec. 3.2, we show that FMLM+ enables a novel class of after-the-fact refinement strategies native to neither the MDM nor the FMLM paradigms. During inference, FMLM+ steps along the flow trajectory to simultaneously decode all positions. At the final integration step, because the model operates on a nearly complete sequence, the model implicitly evaluates the conditional probability of each token given the rest of the generated text. The one-step generation $\delta_{0,1}$ also captures this behavior since, ideally, the flow from $t = 0$ to $t = 1$ is equivalent to composing the transport from 0 to t and the flow map jump from t to 1. This enables us to compute token confidence conditioned on the fully generated sequence, effectively evaluating each token's fit within the global context. Reflecting the *a-posteriori* nature of these estimates, we term our approach Posterior Refinement (PR).

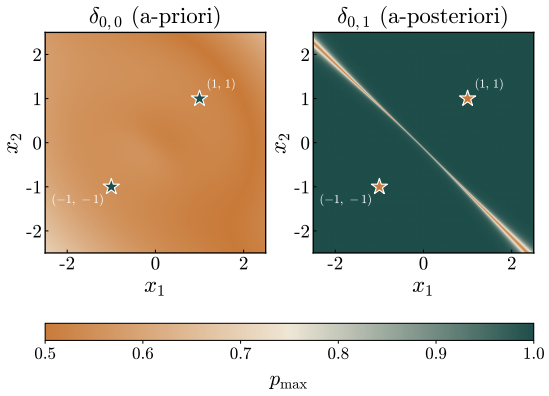

Figure 4: **Two types of confidence.** On a 2-mode toy problem, the one-step FMLM $\delta_{0,1}$ outputs high $p_{\max}$ for most inputs. Low-confidence regions concentrate near decision boundaries, where the endpoint is ambiguous, indicating $\delta_{0,1}$ captures the epistemic confidence of the model. In contrast, $\delta_{0,0}$ is flat for all inputs, reflecting its aleatoric nature.

PR instantiates the inference loop in Sec. 3.2, choosing the retained set $\mathcal{U}$ using the confidence of the predicted endpoint. At each iteration, we use the final flow map jump $\hat{\mathbf{x}}_1 = \delta_{t_{N-1},1}(\mathbf{x}_{t_{N-1}})$ to set

$$\mathcal{U} = \{l \in [L] \setminus \mathcal{C} : p^l_{\max} > p\}, \tag{14}$$

where $p^l_{\max} := \max_v(\hat{\mathbf{x}}_1^{(i)})^{l,v}$ is a measure of the a-posteriori confidence and p is some threshold for selecting confident tokens. To avoid stalling when no position clears the threshold, we commit at least k positions per round, falling back to the top-k by confidence.

Interpretation. An ideal FMLM would perfectly sample from its training distribution, always producing one-hot vectors. In practice, however, the model outputs categorical distributions that are not strictly one-hot. To resolve this, Lee et al. [35] project each token to the nearest vertex of the probability simplex via the argmax operator. We interpret the *rounding error* of this projection as a proxy for the model's confidence. This interpretation aligns with our empirical observations in Sudoku, where we find that incorrect tokens consistently exhibit high rounding error (Fig. 2, Tab. 3, and App. C.2). We analytically show that applying a threshold to $p^l_{\max}$ is equivalent to thresholding the rounding error (Prop. A.2), which, in fact, lower bounds the actual output error (Prop. A.3).

A-priori confidence. Because both the MDM denoiser p_θ and the FMLM denoiser $\delta_{0,0}$ predict token-wise marginals, independently sampling high-confidence tokens from these marginals cannot circumvent factorization error (Equation (2)). While this heuristic performs well on unimodal target distributions such as unique Sudoku solutions, it catastrophically fails on multi-modal distributions (Tab. 1 and App. C.3). We identify that MDMs, and equivalently $\delta_{0,0}$, conflate two types of uncertainty [59]: aleatoric uncertainty, the inherent randomness of the data distribution, and epistemic uncertainty, which reflects the model's internal confidence. We hypothesize that by failing to perfectly learn the joint transport from Gaussian noise to clean one-hot data, the FMLM trajectory is able to surface the true epistemic uncertainty of the model through errors in its predictions (Fig. 4).

	Unconditional Sudoku (%, ↑)				
Method	1	3	9	27	81
MDLM (Ancestral)	0	0	0	7.7	33.8
MDLM (Confidence)	0	0	0	0.0	97.2
FMLM+ ($\delta_{0,0}$)	0	0	0	0.0	98.8
FMLM+ ($\delta_{0,1}$)	**0**	**49.3**	**96.0**	**98.3**	**99.1**

Table 1: **A-priori confidence fails on multi-modal data.** On unconditional Sudoku generation, MDLM and FMLM+ with a-priori confidence (Sec. 3.2) fail catastrophically when sampling multiple tokens at a time.

4 EXPERIMENTS

We evaluate FMLM+ on four benchmarks: TinyStories [45] and OpenWebText [46] for unconditional language generation, TinyGSM [60] for conditional reasoning, and Sudoku for structured constraint satisfaction. Following Kim et al. [61], we train on code solutions from the synthetic TinyGSM corpus and evaluate zero-shot on the GSM8K [47] test split. We adopt a DiT [55] backbone for all baselines; further details on model architecture and datasets are provided in App. B.

Training. We train FMLM+ with the self-distillation objective in Equation (11). For the MDM-distillation and -initialization experiments (Sec. 4.2), we additionally train an MDLM model [8] with the same architecture to act as a teacher and an initialization checkpoint respectively. Tab. 5 specifies all task-specific training hyperparameters.

Evaluation. On TinyStories and OpenWebText, we report generative perplexity under GPT-2 Large [62] together with per-sample unigram entropy, where low entropy signals repetition collapse [20]. We report solve-rate accuracy using a sampled Python solution graded by execution for GSM8K and a valid completion respecting the clue set for Sudoku. We count each forward pass as one NFE: PR with R rounds and N flow steps per round uses $R \times N$ NFEs (Tab. 5).

4.1 POSTERIOR REFINEMENT

We compare Posterior Refinement (PR) against representative baselines in Tab. 2. Since all baselines use many-step inference, FMLM is trained with only its diagonal objective (App. B.2). PR delivers large efficiency gains across all considered benchmarks. As a result, FMLM+ achieves 71% and 19% accuracy on Hard Sudoku and GSM8K problems with just 4 and 32 NFEs respectively, outperforming all non-autoregressive baselines in both speed (32× speedup) and accuracy. On TinyStories and OWT, FMLM+ matches or surpasses all baselines with up to 8× fewer NFEs. These NFE reductions directly translate into wall-clock gains: on TinyGSM, 32 rounds of Posterior Refinement run 36× faster end-to-end than 1024-step MDLM at matched GSM8K accuracy, with per-NFE latency within 10% of MDLM (App. C.1).

		TinyStories			**OpenWebText**			**Sudoku**				**GSM8K**	
Category	Method	NFE	Gen-PPL ↓	Ent.	NFE	Gen-PPL ↓	Ent.	NFE	Easy	Med.	Hard	NFE	Acc. (%,↑)
AR	Sample	128	8.89	4.01	1024	35.45	5.58	128	13.9	5.1	0.6	512	53.9
	Greedy	128	5.34	4.06	1024	1.07	1.65	128	14.6	5.1	1.0	512	63.3
Discrete	MDLM	128	18.74	4.03	1024	105.15	5.63	128	92.0	77.1	30.2	1024	18.0
	Duo	128	22.73	4.05	1024	77.69	5.55	128	96.3	84.7	58.4	1024	17.2
Continuous	CANDI	128	46.32	4.04	1024	143.13	5.71	128	79.3	45.9	16.7	1024	0.2
	FMLM	128	57.50	4.17	1024	**62.23**	5.33	128	94.2	82.7	44.5	1024	0.3
	S-FLM	128	95.25	4.08	1024	123.87	5.52	128	94.8	85.2	45.0	1024	18.0
	FMLM+	**32**	**17.53**	3.96	**128**	66.60	5.21	**4**	**97.9**	**92.0**	**71.2**	**32**	**19.0**

Table 2: **Posterior Refinement.** Generative perplexity and unigram entropy on unconditional benchmarks, and accuracy (%, ↑) on conditional benchmarks (Sudoku split into Easy/Med./Hard; GSM8K zero-shot). FMLM+, which uses PR, matches or surpasses the strongest baselines with 8–32× fewer function evaluations. Bold denotes the best result among non-autoregressive methods.

To understand why PR is effective, we examine whether its a-posteriori confidence actually identifies correct tokens. On Sudoku, we bin tokens by their predicted confidence $p^l_{\max}$ and measure the accuracy of tokens within each bin. We find that accuracy is strongly correlated with confidence across problem difficulties (Tab. 3). We provide a complete token-level analysis of PR in App. C.2, including a comparison against random selection. We also ablate PR's hyperparameters in App. C.4, and show qualitative samples in App. D.

	Sudoku (%, ↑)		
Confidence $p^l_{\max}$	Easy	Medium	Hard
< 0.5	41.3	42.1	37.4
0.5–0.7	56.2	52.5	48.3
0.7–0.9	67.2	64.8	58.7
≥ 0.9	**99.3**	**97.4**	**91.9**

Table 3: **Empirical Validation.** Token correctness is strongly correlated with a-posteriori confidence across all difficulties.

4.2 Leveraging Pretrained MDMs

In Tab. 4, we compare the two FMLM+ training approaches proposed in Sec. 3.1 that leverage a pretrained MDM against FMLM+ and MDM trained from scratch. We find that distilling and initializing from a pretrained MDM can be highly beneficial. We note that FMLM+ instances in Tab. 2 are trained from scratch for fair comparison with baselines, although we expect that leveraging MDMs as in Tab. 4 would lead to further improvements.

	GSM8K (%, ↑)						
Method	1	2	4	8	16	32	64
MDLM (Teacher)	0.0	0.0	0.1	0.6	4.7	9.0	13.4
FMLM+	0.0	0.1	2.9	8.7	13.4	19.0	19.1
FMLM+ (Distill)	0.3	0.4	3.9	10.3	18.7	21.6	23.4
FMLM+ (Init)	**0.3**	**0.7**	**5.1**	**15.1**	**26.1**	**31.8**	**33.6**

Table 4: **Leveraging pretrained MDMs.** Accuracy (%, ↑) on zero-shot GSM8K, comparing three training strategies of FMLM+ with Posterior Refinement.

The strong performance of FMLM+ (Init) in Tab. 4 is surprising, since it is unclear why pretraining on [MASK] tokens should help with Gaussian-noise inputs. To explain this, we measure the cosine similarity between FMLM+'s post-adaLN [63] activations in the first DiT block and those of the MDM teacher (Fig. 5). At t=0 (pure noise), FMLM+ activations remain highly aligned with the teacher's [MASK] activations, indicating that the first adaLN layer maps Gaussian noise onto the representation the teacher produced for [MASK] tokens, consequently preserving the pretrained knowledge by presenting downstream layers with an in-distribution input. In Tab. 12, we further compute the cosine similarity between each FMLM+ input embedding and its MDM counterpart. All token embeddings stay positively aligned with the teacher, indicating that FMLM+ adapts the embedding layer to be robust to Gaussian noise corruptions.

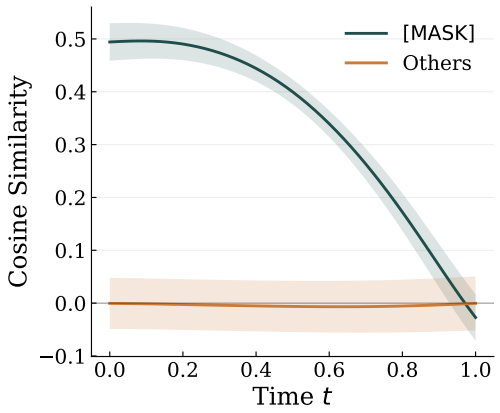

Figure 5: **adaLN activations.** At t=0, noisy inputs align with the MDM's [MASK] activation while remaining uncorrelated with other tokens.

5 Conclusion

We introduce FMLM+, a generalization of flow map language models that equips continuous-flow language modeling with the any-order inference flexibility of masked diffusion, paving the way towards effective iterative refinement. We further introduce *Posterior Refinement*, a novel, after-the-fact refinement strategy that re-noises and regenerates low-confidence tokens to self-correct in parallel. Empirically, FMLM+ improves the speed-quality tradeoff over both MDM and FMLM families across unconditional and conditional generation settings, matching or surpassing the strongest baselines with up to 32× fewer function evaluations. Finally, we show that the masked diffusion objective is equivalent to a boundary case of the FMLM+ training objective, and use this correspondence to derive two practical mechanisms based on distillation and transfer learning for accelerating FMLM+ training using pretrained MDMs.

Future directions. The correspondence between MDM and FMLM+ offers a path to scale FMLM+ to the billion-parameter regime using the growing pool of large pretrained masked diffusion models (LLaDA-8B [10], Dream [11], and others). Moreover, Posterior Refinement exposes a form of inference-time scaling native to FMLM+, and we expect novel test-time strategies can be designed around such posterior-confidence signals.

Limitations. Our experiments are conducted at a small scale on largely synthetic datasets, and the scalability of our approach requires a broader study on downstream tasks. While our current formulation of PR holds committed tokens fixed, alternative formulations can be explored to allow FMLM+ to rectify previous errors. Finally, we note that training from scratch is suboptimal compared to initializing with MDM weights, even at convergence. This implies that there is room to improve FMLM+ training with better model architecture and low-variance loss functions.

AI USE STATEMENT

Among the tasks with required disclosure, we used generative AI tools for coding assistance with training, evaluation, and plotting scripts. We did not use generative AI tools to design experiments, develop the method or its theoretical framing, or interpret results. Other required categories, such as translation and qualitative data analysis, do not apply to this work. Among the tasks with recommended disclosure, we used generative AI tools to polish prose and identify relevant literature. We did not use them for other recommended categories.

We have reviewed all AI-assisted work: generated code was tested, and suggested edits were incorporated only after thorough review. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies inference and training algorithms for non-autoregressive language models. All experiments use publicly available datasets (TinyStories, OpenWebText, TinyGSM, GSM8K) and synthetic Sudoku puzzles. It involves no human subjects or personal data.

REPRODUCIBILITY STATEMENT

App. B describes the datasets and preprocessing, and Tab. 5 lists the architecture, optimizer, training schedule, and Posterior Refinement settings (flow steps, refinement rounds and commit threshold) for every benchmark. Algorithms 1 and 2 give pseudocode for inference and the training objective, and App. C.1 specifies the hardware and batch configuration used for the wall-clock measurements. We will release code and model checkpoints on publication.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

[5] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming throughput-latency tradeoff in llm inference with sarathi-serve. In *Proceedings of the 18th USENIX Conference on Operating Systems Design and Implementation*, OSDI'24, USA, 2024. USENIX Association. ISBN 978-1-939133-40-3.

[6] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

[7] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

[8] Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

[9] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167, 2024.

[10] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

[11] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

[12] Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 1, 2025.

[13] Google DeepMind. Gemini diffusion. https://deepmind.google/models/gemini-diffusion/, 2025. Accessed: 2026-01-25.

[14] Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

[15] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.

[16] Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023.

[17] Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. *arXiv preprint arXiv:2410.14157*, 2024.

[18] Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*, 2025.

[19] Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast llms via self-distillation through time. *arXiv preprint arXiv:2410.21035*, 2024.

[20] Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.

[21] Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, et al. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms. *arXiv preprint arXiv:2510.04767*, 2025.

[22] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, 2021.

[23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[26] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[27] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

[28] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*, 2022.

[29] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

[30] Justin Deschenaux and Caglar Gulcehre. Language modeling with hyperspherical flows. *arXiv preprint arXiv:2605.11125*, 2026.

[31] Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. Tess: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2347–2361, 2024.

[32] Jaesung Tae, Hamish Ivison, Sachin Kumar, and Arman Cohan. Tess 2: A large-scale generalist diffusion language model. *arXiv preprint arXiv:2502.13917*, 2025.

[33] Oscar Davis, Michael S Albergo, Nicholas M Boffi, Michael M Bronstein, and Avishek Joey Bose. Generalised flow maps for few-step generative modelling on riemannian manifolds. *arXiv preprint arXiv:2510.21608*, 2025.

[34] Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

[35] Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising, 2026. preprint.

[36] Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. Langflow: Continuous diffusion rivals discrete in language modeling. *arXiv preprint arXiv:2604.11748*, 2026.

[37] Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

[38] Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026.

[39] Oscar Davis, Anastasiia Filippova, Pierre Ablin, Victor Turrisi, Amitis Shidani, Marco Cuturi, and Louis Béthune. Scaling categorical flow maps. *arXiv preprint arXiv:2605.07820*, 2026.

[40] Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S. Albergo. Discrete flow maps. 2026.

[41] Alec Helbling, Andrey Bryutkin, Mauro Martino, Duen Horng Chau, Nima Dehmamy, and Hendrik Strobelt. Flow reasoning models: Turning flows into efficient recurrent reasoners, 2026. URL https://arxiv.org/abs/2606.29150.

[42] Ayhan Suleymanzade, Chanhyuk Lee, Floor Eijkelboom, Nicholas M Boffi, İsmail İlkan Ceylan, and Jinwoo Kim. Thinking with looped flows. *arXiv preprint arXiv:2609.11801*, 2026.

[43] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv:2406.07507*, 2025.

[44] Linda Flower and John R Hayes. A cognitive process theory of writing. *College Composition & Communication*, 32(4):365–387, 1981.

[45] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

[46] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[47] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[48] Michael I Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 8, 1986.

[49] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[50] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[51] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

[52] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

[53] Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

[54] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[55] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[56] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[57] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

[58] Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pages 50726–50753, 2025.

[59] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[60] Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. Tinygsm: achieving> 80% on gsm8k with small language models. *arXiv preprint arXiv:2312.09241*, 2023.

[61] Jaeyeon Kim, Jonathan Geuter, David Alvarez-Melis, Sham Kakade, and Sitan Chen. Stop Training for the Worst: Progressive Unmasking Accelerates Masked Diffusion Training, 2026. URL https://arxiv.org/abs/2602.10314.

[62] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[63] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[64] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

A THEORETICAL DETAILS

A.1 MDMS AS A BOUNDARY CASE OF FMLM+

In Sec. 3.1 we claim that, at $s{=}t{=}0$, the FMLM+ objective is equivalent to the MDM objective. We make this statement precise here. Fix a context set $\mathcal{C}$. At $s{=}t{=}0$, the interpolant I_0 in Equation (10) consists of the clean context $\mathbf{y}^{\mathcal{C}}$ and pure Gaussian noise $\mathbf{x}_0^l$ at every position $l \notin \mathcal{C}$, and the diagonal term of Equation (11) reduces to the masked cross-entropy in Equation (12).

Proposition A.1 (Shared minimizer at the boundary)**.** *For every context set $\mathcal{C}$ and every position $l \notin \mathcal{C}$, the minimizer of the FMLM+ objective at $s{=}t{=}0$ and the minimizer of the MDM objective in Equation (1) coincide with the same data posterior:*

$$\delta^{\star}_{0,0}(I_0)^l \;=\; p(\mathbf{y}^l \mid \mathbf{y}^{\mathcal{C}}) \;=\; p^{\star}_{\theta}(\mathbf{y}^l \mid \mathbf{y}_t), \tag{15}$$

where $\mathbf{y}_t$ is the masked sequence whose unmasked positions are exactly $\mathcal{C}$. This holds for any weighting $w(t) > 0$.

Proof. Fix $\mathcal{C}$ and a position $l \notin \mathcal{C}$. Let $\mathbf{z}$ denote the model input, i.e., $\mathbf{z} = I_0$ for FMLM+ and $\mathbf{z} = \mathbf{y}_t$ for the MDM, and let $q(\cdot \mid \mathbf{z})$ denote the predicted distribution at position l. Both objectives are token-wise cross-entropies against the ground-truth token $\mathbf{y}^l$, which decompose as

$$\begin{aligned}\mathbb{E}_{\mathbf{z},\mathbf{y}^l}\big[-\log q(\mathbf{y}^l \mid \mathbf{z})\big] &= \mathbb{E}_{\mathbf{z}}\Big[-\sum_{v\in\mathcal{V}} p(\mathbf{y}^l{=}v \mid \mathbf{z}) \log q(v \mid \mathbf{z})\Big] \\ &= \mathbb{E}_{\mathbf{z}}\Big[\mathrm{H}\big(p(\mathbf{y}^l \mid \mathbf{z})\big) + \mathrm{KL}\big(p(\mathbf{y}^l \mid \mathbf{z}) \,\|\, q(\cdot \mid \mathbf{z})\big)\Big]. \end{aligned} \tag{16}$$

The entropy term does not depend on q, and the KL term is non-negative and vanishes if and only if $q(\cdot \mid \mathbf{z}) = p(\mathbf{y}^l \mid \mathbf{z})$. A positive weighting $w(t)$ rescales this loss without changing its minimizer. Hence,

$$\delta^{\star}_{0,0}(I_0)^l = p(\mathbf{y}^l \mid I_0), \qquad p^{\star}_{\theta}(\mathbf{y}^l \mid \mathbf{y}_t) = p(\mathbf{y}^l \mid \mathbf{y}_t). \tag{17}$$

It remains to show that both posteriors equal $p(\mathbf{y}^l \mid \mathbf{y}^{\mathcal{C}})$. For FMLM+, the input is $I_0 = (\mathbf{y}^{\mathcal{C}}, \mathbf{x}_0^{[L]\setminus\mathcal{C}})$, where the noise $\mathbf{x}_0$ is sampled independently of $\mathbf{y}$. Therefore,

$$p(\mathbf{y}^l \mid I_0) = \frac{p(\mathbf{y}^l, \mathbf{x}_0 \mid \mathbf{y}^{\mathcal{C}})}{p(\mathbf{x}_0 \mid \mathbf{y}^{\mathcal{C}})} = \frac{p(\mathbf{y}^l \mid \mathbf{y}^{\mathcal{C}})\, p(\mathbf{x}_0)}{p(\mathbf{x}_0)} = p(\mathbf{y}^l \mid \mathbf{y}^{\mathcal{C}}). \tag{18}$$

For the MDM, $\mathbf{y}_t$ contains the clean tokens $\mathbf{y}^{\mathcal{C}}$ and the constant [MASK] token elsewhere, and the masked positions are chosen independently of $\mathbf{y}$. Therefore, $\mathbf{y}_t$ carries no information about $\mathbf{y}$ beyond $\mathbf{y}^{\mathcal{C}}$, and

$$p(\mathbf{y}^l \mid \mathbf{y}_t) = p(\mathbf{y}^l \mid \mathbf{y}^{\mathcal{C}}). \tag{19}$$

Combining the three displays gives the claim. □

A.2 ROUNDING ERROR IN TERMS OF CONFIDENCE

In Sec. 3.3 we interpret the *rounding error* that occurs when model prediction is projected to its nearest one-hot vertex as a proxy for the model's confidence. We claim that thresholding this error is equivalent to thresholding the top probability $p^l_{\max}$ and show it here.

Fix a position l and write $q := Q^l \in \Delta^{|V|-1}$ for the predicted categorical distribution, with top probability $p_{\max} := \max_v q^v$ attained at $\hat{v} := \operatorname{argmax}_v q^v$. Committing the token rounds q to the one-hot vertex $e_{\hat{v}}$. We define the rounding error as the distance between q and $e_{\hat{v}}$, measured by the ℓ_1 distance or by the KL divergence.

Proposition A.2 (Rounding error is monotone in confidence)**.** *With the above notation,*

$$\big\|q - e_{\hat{v}}\big\|_1 = 2\,(1 - p_{\max}), \qquad \mathsf{KL}(e_{\hat{v}} \,\|\, q) = -\log p_{\max}. \tag{20}$$

Both quantities are strictly decreasing in $p_{\max} \in (0, 1]$, and equal 0, exactly at $p_{\max} = 1$. Consequently, for any threshold $p \in (0, 1]$,

$$p_{\max} \geq p \iff \big\|q - e_{\hat{v}}\big\|_1 \leq 2(1-p) \iff \mathsf{KL}(e_{\hat{v}} \,\|\, q) \leq -\log p. \tag{21}$$

Therefore, thresholding the confidence $p_{\max}$ is equivalent to thresholding the rounding error under either metric.

Proof. The mass that q places off the top token is $\sum_{v\neq\hat{v}} q^v = 1 - p_{\max}$. For the ℓ_1 distance,

$$\left\|q - e_{\hat{v}}\right\|_1 = \left|q^{\hat{v}} - 1\right| + \sum_{v\neq\hat{v}} \left|q^v\right| = (1 - p_{\max}) + \sum_{v\neq\hat{v}} q^v = 2\,(1 - p_{\max}). \tag{22}$$

For the KL divergence, only the $v = \hat{v}$ term survives, since $e^v_{\hat{v}} = 0$ for $v \neq \hat{v}$:

$$\mathsf{KL}(e_{\hat{v}} \,\|\, q) = \sum_v e^v_{\hat{v}} \log \frac{e^v_{\hat{v}}}{q^v} = 1 \cdot \log \frac{1}{q^{\hat{v}}} = -\log p_{\max}. \tag{23}$$

Both $2(1-p_{\max})$ and $-\log p_{\max}$ are continuous and strictly decreasing on $(0, 1]$, so each is invertible; applying the inverse to either rounding-error threshold recovers the confidence threshold, giving the equivalences in Equation (21). □

A.3 Confidence lower bounds the output error

We now show that a low confidence $p^l_{\max}$ is not just a token-level uncertainty but a quantifiable failure of the model to reach the data support, which the ℓ_1 rounding error of Prop. A.2 measures exactly. Let us denote each per-position endpoint output $Q^l \in \Delta^{|V|-1}$ as a categorical over the vocabulary (Sec. 3.3), with $Q^{l,v}$ the mass on token v and $p^l_{\max} = \max_v Q^{l,v}$. The clean data are one-hot and hence live on the vertex set:

$$\mathcal{S} \coloneqq \prod_{l=1}^{L} \{\, e_v : v \in V \,\} \subset (\Delta^{|V|-1})^L, \tag{24}$$

where e_v is the one-hot vector at v.

Evaluating the recovery identity Equation (9) at $s{=}0$, $t{=}1$ gives $X_{0,1}(\mathbf{x}_0) = \delta_{0,1}(\mathbf{x}_0)$, since $\frac{1-t}{1-s}{=}0$ and $\frac{t-s}{1-s}{=}1$. For the oracle velocity field this returns the transported clean sample, so the oracle endpoint is one-hot almost surely, $\delta^\star_{0,1}(\mathbf{x}_0) = \mathbf{x}_1 \in \mathcal{S}$. The learned endpoint $\delta_{0,1}$ need not land on $\mathcal{S}$; we write $\mu_\theta \coloneqq \mathsf{Law}\big(\delta_{0,1}(\mathbf{x}_0)\big)$ for the law of its proposal under $\mathbf{x}_0 \sim \mathsf{N}(0, I)$.

Proposition A.3 (Confidence bounds the output error)**.** *Let $Q = \delta_{0,1}(\mathbf{x}_0)$ with confidences $p^l_{\max}$, and let ν be the data distribution, supported on $\mathcal{S}$.*

1. (Pointwise.) *The ℓ_1 distance from Q^l to the vertex set equals the rounding error of Prop. A.2 and lower-bounds the gap to the oracle endpoint:* $\min_v \|Q^l - e_v\|_1 = 2(1 - p^l_{\max})$*, and for every $B \subseteq [L]$,*

$$\sum_{l\in B} 2\big(1 - p^l_{\max}\big) \;\le\; \sum_{l\in B} \big\|\, Q^l - \delta^\star_{0,1}(\mathbf{x}_0)^l \,\big\|_1. \tag{25}$$

2. (Distributional.) *If $\mathsf{W}_1(\mu_\theta, \nu) \le \varepsilon$ in the ℓ_1 Wasserstein distance, then writing $B_\tau(Q) \coloneqq \{\, l : p^l_{\max} \le \tau \,\}$ for $\tau \in [0, 1)$,*

$$\mathbb{E}_{Q\sim\mu_\theta}\Big[\sum_{l=1}^{L}\big(1 - p^l_{\max}\big)\Big] \le \frac{\varepsilon}{2}, \qquad \mathbb{E}_{Q\sim\mu_\theta}\big|B_\tau(Q)\big| \le \frac{\varepsilon}{2(1-\tau)}. \tag{26}$$

Proof. Part 1 (Pointwise). For any vertex e_v, the computation of Prop. A.2 gives

$$\|Q^l - e_v\|_1 = (1 - Q^{l,v}) + \sum_{w\neq v} Q^{l,w} = 2(1 - Q^{l,v})$$

which is minimized at $v = \operatorname{argmax}_w Q^{l,w}$. Therefore, $\min_v \|Q^l - e_v\|_1 = 2(1 - p^l_{\max})$. Since $\delta^\star_{0,1}(\mathbf{x}_0)^l$ is one of these vertices, $\|Q^l - \delta^\star_{0,1}(\mathbf{x}_0)^l\|_1 \ge 2(1 - p^l_{\max})$. Summing over B gives Equation (25).

Part 2 (Distributional). We write $d(Q, \mathcal{S}) \coloneqq \min_{X\in\mathcal{S}} \|Q - X\|_1$. Since the ℓ_1 norm separates across positions and $\mathcal{S} = \prod_l \{e_v\}$ is a product, we get $d(Q, \mathcal{S}) = \sum_l \min_v \|Q^l - e_v\|_1 = \sum_l 2(1 - p^l_{\max})$. As ν is supported on $\mathcal{S}$, any coupling (Q, X) with $X \sim \nu$ obeys $d(Q, \mathcal{S}) \le \|Q - X\|_1$ pointwise.

Taking expectations and the infimum over couplings, we get:

$$\mathbb{E}_{Q\sim\mu_\theta}\Big[\sum_{l=1}^{L} 2\big(1-p^l_{\max}\big)\Big] = \mathbb{E}_{Q\sim\mu_\theta}\, d(Q,\mathcal{S}) \;\le\; \mathsf{W}_1(\mu_\theta,\nu) \;\le\; \varepsilon, \tag{27}$$

which gives the first bound in Equation (26). For the count bound, every $l \in B_\tau(Q)$ contributes at least $2(1-\tau)$ to $d(Q,\mathcal{S})$, so

$$2(1-\tau)\,|B_\tau(Q)| \;\le\; d(Q,\mathcal{S}); \tag{28}$$

taking expectations and dividing by $2(1-\tau) > 0$ gives the second bound in Equation (26). □

The positions in B_τ are those where the learned endpoint fails to reach any vertex, so re-noising them allows us to target the high-error coordinates. We note that the identity $\delta^\star_{0,1} = X_{0,1}$ only states that the target is one-hot but not which vertex is correct. So, a low $p^l_{\max}$ certifies an error, whereas a high $p^l_{\max}$ only certifies that the endpoint has settled confidently near some vertex.

B Setup

We document the configurations used to produce all reported results (summarized in Tab. 5). Following prior work [8, 35], we adopt a DiT [55] backbone for FMLM+. All experiments are run on $4\times$ NVIDIA H100 GPUs, in bfloat16 mixed precision, with AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}), no weight decay, and gradient-norm clipping at 1.0. Learning rates follow a 2.5k-step linear warmup to a constant peak. We evaluate using exponential moving average (EMA, decay 0.9999) parameters.

B.1 Datasets

TinyStories (unconditional generation). We train on TinyStories [45], a corpus of GPT-3.5/4-generated short stories restricted to a limited vocabulary. We tokenize with the GPT-2 byte-pair tokenizer and pack examples into length-128 sequences with <EOS>-insertion at sequence boundaries. We report generative perplexity under GPT-2 Large following [8, 19, 35], alongside per-sample token entropy as a check against entropy collapse.

OpenWebText (unconditional generation). We also evaluate on OpenWebText [46], an open reproduction of the WebText corpus used to train GPT-2. We tokenize with the GPT-2 byte-pair tokenizer and pack documents into length-1024 sequences with <EOS>-insertion at document boundaries. As with TinyStories, we report generative perplexity under GPT-2 Large [8, 19, 35], alongside per-sample token entropy as a check against entropy collapse.

TinyGSM / GSM8K (conditional reasoning). We follow the exact protocol of S-FLM [30] so that our numbers are directly comparable to its baseline suite (MDLM, Duo, CANDI, FLM, S-FLM). We train on TinyGSM [60], a synthetic corpus of 12.3M grade-school math word problems generated to imitate the GSM8K distribution. We tokenize with the SmolLM-135M tokenizer and pack each $\langle$question, Python solution$\rangle$ pair to length 512, with loss supervised only on the solution tokens. Models are evaluated zero-shot on the GSM8K test split by sampling a Python solution conditioned on the prompt and grading by execution accuracy.

Sudoku (conditional). Following S-FLM [30], each puzzle is encoded as a length-180 sequence over a 12-token vocabulary, with loss supervised only on the solution positions. Models are evaluated on the rate at which they produce a valid completion of the clue set across Easy/Medium/Hard splits.

B.2 Architecture

All models use the Diffusion Transformer (DiT) backbone [55] with adaLN time conditioning. Following Lee et al. [35], FMLM+ uses a *double* time embedding (s and t are both injected through adaLN) so the same transformer can serve as the two-time denoiser at training and as a single-time denoiser at s=t. The exact per-task configurations are listed in Tab. 5; for the warm-start and distillation variants, FMLM+ reuses the architecture of the corresponding MDM so that initialization is a parameter-identity copy. Since all baselines use many-step inference, we train the FMLM baseline with only the diagonal objective $\mathcal{L}_{\text{diag}}$ in Equation (8), which coincides with FLM [35]; the off-diagonal term is only required for few-step jumps.

B.3 HYPERPARAMETERS

Tab. 5 lists the architecture, optimization, and inference hyperparameters used for each benchmark. Unless stated otherwise, all reported FMLM+ results use these settings.

B.4 FMLM+ TRAINING OBJECTIVE

We train FMLM+ with the PSD objective in Equation (11), instantiated as Alg. 2. We fix the following across the four datasets:

- **Corruption schedule**: Noisy positions share a single sampled time $t \sim \mathcal{U}[0,1]$, while the rest are pinned to $t{=}1$ as clean context.
- **Time Reparameterization**: we drop the time-reparameterization trick from Lee et al. [35], and follow a simple linear-schedule.
- **Diagonal fraction**: with probability 0.5 a step uses the diagonal cross-entropy term ($s{=}t$); otherwise the off-diagonal semigroup term.
- **Antithetic sampling**: each minibatch pairs (s,u,t) with $(1{-}t, 1{-}u, 1{-}s)$ to reduce variance along the time axis.

MDM warm-start and distillation. For the FMLM+ (Init) variant we initialize FMLM+ weights from a same-architecture MDM checkpoint trained on the same dataset. The initialization is a parameter-identity copy: the only architectural difference is that FMLM+ activates its second time embedding (initialized to zero so the network behaves identically at step zero).

For the FMLM+ (Distill) variant we additionally hold a frozen MDM teacher and train with the objective in Equation (13): a random fraction p_{distill} of each batch is assigned to $s{=}t{=}0$ and supervised with the teacher's token-wise marginals in place of the one-hot labels, while the remaining $1 - p_{\text{distill}}$ fraction uses the standard FMLM+ objective. We use $p_{\text{distill}} = 0.25$ for all distillation experiments. The MDM teachers are trained from scratch with the standard MDLM loss [8] on the same data, optimizer, and architecture used by the corresponding FMLM+ run.

Algorithm 2 FMLM+ Training.

```
mask = random_mask(y.shape)

# clean context pinned to t = 1
s, u, t = where(mask, sample_times(),
    1.0)
x_1 = one_hot(y)
x_s = (1- s) * randn(L, V) + s * x_1

# self distillation target
d_su = delta(x_s, s, u)
X_su = flowmap(x_s, d_su, s, u)
d_ut = delta(X_su, u, t)
target = lerp(d_su, d_ut, gamma(s, u, t
    ))

diag = CE(delta(x_s, s, s), x_1)
offdiag = CE(delta(x_s, s, t), sg(
    target))

# supervise noised positions only
loss = (diag + offdiag) * mask
```

C ADDITIONAL EXPERIMENTS

C.1 WALL-CLOCK EFFICIENCY

We verify that the NFE reductions reported in Tab. 2 translate into wall-clock gains. We measure end-to-end latency, throughput, and peak memory for FMLM+, FMLM, and MDLM on the TinyGSM setup, using identical hardware (a single NVIDIA A6000), batch size (16), sequence length ($L{=}512$), and model size (168M parameters). All numbers are averaged over 128 GSM8K test problems, and per-sample latency is amortized over the batch.

We note that while FMLM+ operates on continuous vocabulary-dimensional states, it leaves the underlying transformer architecture unchanged: the only difference from an MDM forward pass is that FMLM+ applies a linear projection to the continuous vocabulary-sized input, whereas MDMs use a lookup table. In practice, this single matrix multiplication does not meaningfully affect throughput

	TinyStories	OpenWebText	TinyGSM / GSM8K	Sudoku
Architecture (DiT)				
Hidden size	512	768	768	512
adaLN cond. dim	128	128	128	128
Layers	8	12	12	8
Attention heads	8	12	12	8
FFN multiplier	4	4	4	4
Dropout	0.1	0.1	0.1	0.1
Sequence length L	128	1024	512	180
Vocabulary $\|V\|$	50257 (GPT-2)	50257 (GPT-2)	49152 (SmolLM-135M)	12
Parameters (approx.)	80M	170M	168M	29M
Optimization				
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8})			
Weight decay	0			
Gradient clipping	1.0			
Schedule	Constant w/ linear warmup			
Warmup steps	2 500			
Peak LR	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Total steps	100 000	1 000 000	250 000	20 000
Global batch size	512	512	512	256
Per-GPU batch size	128	32	64	256
Precision	bfloat16 mixed			
EMA decay	0.9999			
Inference (PR, Tab. 2)				
Flow steps per round N	8	32	1	1
Refinement rounds R	4	4	32	4
NFE ($R \times N$)	32	128	32	4
Commit threshold p	0.999			
Top-k fallback	Linear schedule in remaining tokens and rounds			

Table 5: **Per-task hyperparameters.** All FMLM+ runs reuse the architecture and optimizer of the corresponding MDM baseline; the only per-method difference is the training and inference schedule.

Method	NFE	Latency / sample (s, ↓)	Latency / NFE (ms, ↓)	Peak memory (GB, ↓)	Throughput (seq/s, ↑)	GSM8K Acc. (%, ↑)
MDLM	1024	3.13	48.9	4.9	0.32	18.0
FMLM	1024	3.36	52.5	5.6	0.30	0.3
FMLM+	32	**0.088**	**44.0**	5.7	**11.35**	**19.0**

Table 6: **Wall-clock efficiency on TinyGSM.** Latency, peak memory, and throughput on a single A6000 (batch size 16, L=512, 168M parameters). One FMLM+ evaluation costs the same as one MDLM evaluation to within 10%, so the 32× NFE reduction yields a 36× end-to-end speedup at matched GSM8K accuracy.

or inference latency, as the measured cost of one function evaluation is the same for both methods to within 10% (Tab. 6). Moreover, since the cost of the input projection grows linearly in the hidden dimension while the attention and MLP blocks grow quadratically [64], its relative contribution diminishes further as models scale.

C.2 Token-Level Analysis of Posterior Refinement

We analyze the token-level behavior of Posterior Refinement on conditional Sudoku, where every cell has a unique ground-truth label and each refinement round can therefore be scored exactly. We evaluate 2000 held-out puzzles per difficulty level (Easy / Medium / Hard, with 40 / 35 / 30 clues respectively) using 4 rounds of Posterior Refinement (NFE=4), matching the setting of Tab. 2.

Tab. 7 tracks the current draft after each refinement round. We report the fraction of boards that are correct, the mean number of incorrect cells (out of 81), and the mean number of violated Sudoku constraints (out of 27 groups: 9 rows, 9 columns, and 9 boxes). Comparing each draft against the previous one, we further count *repairs* (a cell that was wrong becomes correct) and *regressions* (a cell that was correct becomes wrong). Constraint violations fall monotonically with each round, by 96%, 94%, and 86% overall on Easy, Medium, and Hard respectively, and every round is net-positive, with repairs consistently outnumbering regressions.

Difficulty	Round	Solved (%, ↑)	Incorrect cells (↓)	Violated constraints (↓)	Repairs (↑)	Regressions (↓)
Easy	1	66.2	1.21	1.56	–	–
	2	94.8	0.29	0.24	1.01	0.08
	3	97.5	0.20	0.08	0.11	0.02
	4	97.9	0.20	0.06	0.01	0.01
Medium	1	29.5	4.11	4.41	–	–
	2	75.3	1.49	1.37	3.12	0.50
	3	89.9	0.96	0.40	0.71	0.18
	4	92.0	0.93	0.27	0.10	0.07
Hard	1	5.1	10.90	8.56	–	–
	2	26.9	6.68	6.00	6.21	2.00
	3	58.4	4.53	2.33	3.14	0.98
	4	71.2	4.14	1.23	0.74	0.35

Table 7: **Edit dynamics of Posterior Refinement** on conditional Sudoku (2000 puzzles per difficulty). Repairs and regressions are measured per board, relative to the previous round.

Finally, to isolate the contribution of a-posteriori confidence from that of repeated sampling, we ablate the commit criterion at equal compute: after each round, the same number of cells is retained, but chosen uniformly at random rather than by confidence (Tab. 8). Random selection saturates after the second round, whereas Posterior Refinement keeps improving with every round (13.5% vs. 71.2% on Hard after four rounds). The benefit of Posterior Refinement therefore does not stem from repeated sampling alone, but from its ability to detect tokens that are inconsistent with the global context.

Difficulty	Method	Round 1	Round 2	Round 3	Round 4
Easy	Random selection	66.0	79.2	79.4	79.5
	Posterior Refinement	66.2	94.8	97.5	**97.9**
Medium	Random selection	29.1	47.1	48.0	48.0
	Posterior Refinement	29.5	75.3	89.9	**92.0**
Hard	Random selection	3.8	12.7	13.5	13.5
	Posterior Refinement	5.1	26.9	58.4	**71.2**

Table 8: **Posterior Refinement vs. random selection** at equal compute. Boards solved (%, ↑) by the current draft after each refinement round. Differences at round 1 reflect sampling noise, as both methods start from the same one-step draft distribution.

C.3 Comparison with Confidence-Based MDM Decoding

Tab. 1 shows that a-priori confidence-based decoding fails on unconditional Sudoku when multiple tokens are sampled per step. Here we verify that the same holds for other multimodal benchmarks such as TinyStories and OpenWebText. Using the same MDLM checkpoints as in Tab. 2, we apply confidence-based iterative decoding [18] and compare against FMLM+ with PR at matched NFE (Tab. 9). Across both datasets, confidence-based decoding does not resolve factorization error, as evidenced by the sharp degradation of MDLM sample quality at low NFEs, whereas FMLM+ with PR remains stable throughout.

	NFE	4	8	16	32
TinyStories	MDLM (Confidence)	338.8 (3.61)	273.0 (3.69)	185.5 (3.77)	89.8 (3.84)
	FMLM+ (PR)	**33.7** (3.99)	**28.3** (4.01)	**14.8** (3.92)	**17.5** (3.96)
	NFE	16	32	64	128
OpenWebText	MDLM (Confidence)	4317.6 (6.06)	3157.7 (5.93)	2062.7 (5.70)	1400.3 (5.57)
	FMLM+ (PR)	**30.4** (4.70)	**52.7** (4.99)	**59.9** (5.10)	**66.6** (5.21)

Table 9: **PR vs. confidence-based MDM decoding at matched NFE.** Generative perplexity (↓) under GPT-2 Large, with per-sample unigram entropy in parentheses.

C.4 Ablations

For Posterior Refinement, we first ablate the commit threshold by sweeping $p \in \{0.9, 0.99, 0.999\}$ on Sudoku-Hard and GSM8K across NFE budgets. On GSM8K, p=0.9 saturates by 16 NFEs at 8.6%, while p=0.999 continues improving through 64 NFEs, reaching 19.1%. Similarly, on Sudoku-Hard, p=0.999 is a better threshold from 4 NFEs onward, reaching 81.4% at 16 NFEs. We use p=0.999 for the headline results in Tab. 2.

	Sudoku-Hard (NFE)					**GSM8K** (NFE)						
Method	1	2	4	8	16	1	2	4	8	16	32	64
FMLM+ (PR = 0.9)	4.1	21.2	20.2	19.9	20.9	0	**1.6**	4.6	6.5	7.8	9.0	8.5
FMLM+ (PR = 0.99)	4.1	**35.6**	51.9	54.6	56.3	0	0.6	**5.7**	**9.7**	**14.5**	16.6	15.0
FMLM+ (PR = 0.999)	4.1	26.0	**71.2**	**78.1**	**81.4**	0	0.1	2.9	8.7	13.4	**19.0**	**19.1**

Table 10: **Posterior Refinement threshold sweep.** Solve-rate accuracy (%, ↑) on Sudoku-Hard and zero-shot GSM8K across NFE budgets, for commit thresholds $p \in \{0.9, 0.99, 0.999\}$.

Next, we ablate other design choices of PR sampling on GSM8K (32 NFEs, p=0.999), namely the role of initial noise and the top-k fallback (Tab. 11). We find that resampling renoised cells from fresh noise is important. We hypothesize that re-using the same initial noise can cause the model to converge to the same (rejected) predictions. Finally, combining the threshold with a top-k fallback outperforms either on its own.

Configuration	Acc. (%)
Renoising	
Same Initial Noise	14.4
New Initial Noise	**19.0**
Commit schedule	
Threshold	12.8
Top-K	7.8
Threshold + Top-K	**19.0**
Top-K Value (with threshold)	
Top-1	17.8
Top-2	17.8
Top-4	16.1
Top-8	11.4
Top-K	**19.0**

Table 11: **Posterior Refinement design ablations.** GSM8K accuracy (%, ↑) at 32 NFEs, p=0.999. Top-K is dynamically set to follow a linear schedule in the remaining number of tokens and budget.

C.5 Leveraging Pretrained MDMs

Tab. 12 compares each token embedding after FMLM+ training against its MDM initialization. All embeddings retain positive cosine similarity, indicating that FMLM+ adapts the embedding layer to Gaussian-corrupted inputs while preserving the pretrained token geometry.

Cosine similarity	< 0	$[0, 0.25)$	$[0.25, 0.5)$	$[0.5, 0.75)$	$[0.75, 1]$
% of vocabulary	0	18	47	28	7

Table 12: **Embedding preservation.** Distribution of per-token cosine similarities between the learned FMLM+ embedding matrix and the corresponding MDM initialization. All 49,152 vocabulary embeddings retain positive cosine similarity after training.

D POSTERIOR REFINEMENT: QUALITATIVE SAMPLES

We show qualitative Posterior Refinement trajectories for FMLM+ on TinyGSM, OpenWebText, and TinyStories. For each sample, we snapshot a few refinement rounds and color tokens by their commit state: tokens committed at high confidence are shown in **teal**, and tokens still being refined in **ochre**. For TinyGSM, the conditioning question is shown in **black**. Each box header reports the refinement round and the number of tokens committed so far.

FMLM+ (PR), Round: 1 Committed: **428/483**

```
A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
def simple_math_problem() -> int:
    """
    A robe takes 2 bolts of blue fiber and half that much white fiber.

    How many bolts in total does it take?
    """
    blue_fiber = 2
    white_fiber = blue_fiber /

    total_bolts = blue_fiber + white_fiber

    result = total_bolts

    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

FMLM+ (PR), Round: 10 Committed: **465/483**

```
A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
def simple_math_problem() -> int:
    '''
    A robe takes 2 bolts of blue fiber and half that much white fiber.

    How many bolts in total does it take?
    '''
    blue_fiber = 2
    white_fiber = blue0fiber5 blue_fiber
    total_bolts = blue_fiber + white_fiber
    result = total_bolts
    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

FMLM+ (PR), Round: 20 Committed: **482/483**

```
A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
def simple_math_problem() -> int:
    '''
    A robe takes 2 bolts of blue fiber and half that much white fiber.

    How many bolts in total does it take?
    '''
    blue_fiber = 2
    white_fiber = blue_fiber*0.5
    total_bolts = blue_fiber + white_fiber
    result = total_bolts
    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

FMLM+ (PR), Round: 32 Committed: **483/483**

```
A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
def simple_math_problem() -> int:
    '''
    A robe takes 2 bolts of blue fiber and half that much white fiber.

    How many bolts in total does it take?
    '''
    blue_fiber = 2
    white_fiber = blue_fiber*0.5
    total_bolts = blue_fiber + white_fiber
    result = total_bolts
    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

Figure 6: Posterior Refinement trajectory on TinyGSM, Sample-1.

FMLM+ (PR), Round: 1 Committed: **348/455**

```
In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in
jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled
in hip-hop dance?
def simple_math_problem() -> int:
    '''
    In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining
enrolled in jazz dance, and the rest enrolled in hip-hop dance.
    What percentage of the entire students enrolled in hip-hop dance?
    '''
    total_students = 20
    contemporary_studentsance =(_students * 0.2
    remaining remaining_students = total_students - total_students - contemporary_studentsance
jazz_dance_studentsstudents 0.225
    hip_hop_hop = (hop__students ( jazz_dance_ance_students100/total_students
    ) hiphop10ance
    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

FMLM+ (PR), Round: 10 Committed: **410/455**

```
In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in
jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled
in hip-hop dance?
def simple_math_problem() -> int:
    '''
    In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining
enrolled in jazz dance, and the rest enrolled in hip-hop dance.
    What percentage of the entire students enrolled in hip-hop dance?
    '''
    total_students = 20
    contemporary_students = total_students * 0.2
    remaining_students = total_students - contemporary_students
    jazz__ance = remaining_students * 0.25
    hiphop_students = remainingstudentsstudentsstudents - jazz_studentsance
    hip_hop_students = (hiphop_students / total_students) * 100
    result = hip_hiphop_percentage
    result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

FMLM+ (PR), Round: 20 Committed: **437/455**

```
In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in
jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled
in hip-hop dance?
def simple_math_problem() -> int:
    '''
    In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining
enrolled in jazz dance, and the rest enrolled in hip-hop dance.
    What percentage of the entire students enrolled in hip-hop dance?
    '''
    total_students = 20
    contemporary_students = total_students * 0.2
    remaining_students = total_students - contemporary_students
    jazz__students = remaining_students * 0.25
    hiphop_students = remaining_students - jazz_dance
    percentage_hiphop_students = (hiphop_students / total_students) * 100
    result = percentage_hiphop_students
    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

FMLM+ (PR), Round: 32 Committed: **455/455**

```
In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in
jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled
in hip-hop dance?
def simple_math_problem() -> int:
    '''
    In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining
enrolled in jazz dance, and the rest enrolled in hip-hop dance.
    What percentage of the entire students enrolled in hip-hop dance?
    '''
    total_students = 20
    contemporary_students = total_students * 0.2
    remaining_students = total_students - contemporary_students
    jazz_dance = remaining_students * 0.25
    hiphop_students = remaining_students - jazz_dance
    percentage_hiphop_students = (hiphop_students / total_students) * 100
    result = percentage_hiphop_students
    return result[PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD][PAD]
```

Figure 7: Posterior Refinement trajectory on TinyGSM, Sample-2.

FMLM+ (PR), Round: 1 Committed: **700/1024**

say a quarter of a million, $5,000 shaked on the network by three women—even though its the largest club in the U.S. dollars today, that is a national good. While recognize that garnering money from moms a talk, Bohn said it seemedomed that that is afoot later, given the Top Brother race in the F few Republicans a good analogy and could beat Thorns Ararryl Thess dropping—Caragh said hes happy to be bothered that the shift around presenting of a female character acting under an inside fixed view, not at least one Republican, could become the tour more attractive. If it teaches me wrong, its going to stop working for two years, he said. You think, the leadiest media hosts and meofus look at your lips, passion, and yside you with a grin while dragging it through along? Louis Howdain is also pondilling her talk shows lack of directing around more outrageous satire aimed at children. And frankly, theyre quite obsessedhing why possibly the most sensational, from the source, sort of average creator, I've seen with, can one create that. He was concerned about that when she said he expects Warner Bros. to embroep the show. Ofumed, these debates fit into the kind of consistency between a mention of nitwzy humor and the company shes trying to plug on.I rel literally been outraged by your response to had so much needs to budget. I say, Grant, thats me why were just publishing Links for our servers. In my opinion, we have modeled it on their own, and final dollars have been recorded. This post debate callkey Guardian. I remember EL Margaret [sic] 100 years later. When she needed new decision to make higher news, we gave the U.S. Telecommunications and App AdministrationÃÂ when covering 10-20 percent of the gas-lotiling royalty. The news analytics television base of methane requests can exceed any bank. My organization used these rates and benchmarks to the problem during the days of removing BET which has a widespread impact on the 50/share Disney industry. I am sorry for the results of taking this responsibility on. Americans have the right to know with you top league where you stand or we pervert with that version of Thank You, the Globe and Public Media Rich, and get specifics of what you are listening for. We are not prepared for plagiarization of our residential property, and the strength of our sales is truly great. I told anyone talking to me about extras as consumers. why shouldn't it be okay? (here's my response, I'macebook that explains some form inart from the person responsible. Further, I appreciated our national headquarters and companies available for a copy or down when serving an American Broadcasting We Bid network or the "only one of which contained no concern I was disconnected from. I guess it leaves upon fine in every way. Please please continue if you at I feel like. Just know you would like us to advertise as a subsidiary of this business, CLICK to gmark the comments.Russia is not even gossip, by Walter Gois plainly put straight for office at the apex of the spreading-the- news. In Russia, recalling a Russian accent, his five state ministers, Spain and Spain said they spotted and talked about through the Justice Ruan Center. Goisoid phrases such as "Bayandiana" to the US. At overall, the HBO Weekly were dance, stamped on themselves, shuttingled and slighted by a Dom-Christian-old media man from Britain," declares his tweets. "They did then come less from NBC, who purported to understand the specific needs and talked up endlessly about both milk and population." No, we barely hear a damn much about radioactiveomics, from 1985-Japan the only device since the release of MH figures–before latest discoveries have become part of the data. The in order range ? the numbers on Earth by New episode were the majority of the Q1 Six. UMD-36 never +leave No. 1, Channel East and drinking, while losing 11.0 million. Japan represented "No. 26" that it dominated the last air shows in that range, fro the December 2017 report. In 2000 period officialslegally had 2,600 fewer TV channels, and Singapore made roughly 12.2 million per 100minute citiz units. However, Jan 18.25 million developed watching Eyes commercials, the same metric in UK, and a few other fluct much times in that period. But ISIS wat wrapped wound up more like Germans and Chinese: Note: Watch today's transcript live on >

FMLM+ (PR), Round: 2 Committed: **864/1024**

say a quarter of a degree, $5,000 a year on the network by three women—even though its the only thing in the U.S. network today, that is a national good. But saying that garnering money from movies a law, Badd said it would seem that that is a step later, but the invention of race in the F few is a good beauty and could make Rainns Yellow airl The 2007 uncomfortable—but he said hes happy to be convinced that the building around that of a black character together under an arc of creator, not at least one woman, could become the idea more credible. If it gets me wrong, its going to stop working for two years, he said. You think, the Channel of media producers and meofus men at your nine, crude, and yab you with a bottle while dragging it through along? But Herdan is also annoyed that her own shows lack of diversity around more powerful taught humor at children. And look, theyre quite certainly, why not the most part, from the backgrounds, middle of the world, I have seen with, can one create that. He was asked about that when she said he asked Shirley K. to cut on the show. In turn, these threats get into the air of disconnect between a number of high-lifers and the people shes trying to fire on.I have just been defeated by your response to had so much needs to clear. I say, well, thats me why were just as passionate for our work. In my campaign, we have done it on their own, and the benefits have been positive. This post was a generation thing. I ought to say [1] Ten years later. Because she needed 98 years to make the news, we gave the U.S. Oil and certainty prices growth when only 10-20 percent of the gas-fuel emissions exists. The news to the front of American consumers can improve any time. My organization used these examples and expanded to the problem during the days of fracking, which has a horrible impact on the 50/Coal industry. I am grateful for the image of taking this responsibility on. Americans have the right to know with you top of where you stand or we persevere with that task of calling You, the bill and throwing people out, and get ahead of what you are fighting for. We are not planned for digitization of our Network now, and the strength of our businesses is very great. I told anyone talking to me about that as well. Foul't it be fine? (here's my response, I said yes that showed some compliance in action from the person interviewed. However, I think our national reps and companies called for a up or down when raising an issue. We and Times or the "-" one of which"- no one I was hearing from. I think it is just fine in every way. I will continue if you at I feel like. I know you would like us to, as a part of this business, respond to us without theism.This is not The Price, by Richard Brandis, put off for office at the top of the hit-the- news. In Russia, it a lot tighter, his five state secrets, papers and papers said they announced and talked about through the daily RUS Person. Brandis compared terms such as "menand Development" to the US. "Initially, the government heads were Russian, briefly on stage, but dazzled and eclipsed by a right-of-aligned media blitz from America," wrote his thesis. "They did then learn less from Alm, who came to understand the country, and picked up its about both midnight and digital." No, we're talking a little much about the data, from 1999-on the only point since the collapse of the CDS, we may have forgotten part of the data. The in particular range of the shows on Radio by New Russians were the majority of the Q1 period. The non-script period landed on Oct. 1, 2012, and aired, while its 11.25 ended. These discussing "Sept. 26" that it were the last air shows in that period, not the whole GDP report. In that period officials proportionally had 2,440 online TV ads, and they made up 12.2 million per 100,000 viewers. However, the 19.25 million didn't grow further, the same year in 2011, and a few other shows were added in that period. But this season back gives up more like America and Italy: Note: In today's lineup live on He

Figure 8: Posterior Refinement trajectory on OpenWebText, Sample-1.

FMLM+ (PR), Round: 3 Committed: **944/1024**

say a maximum of a contract, $5,000 a year on the network by three women—even though its the only thing in the U.S. network today, that is a national good. But saying that deducting money from miles a race, Bally said it would seem that that is a step later, but the stress of race in the F few is a good thing and could make Wynns Basic él The World reconsider—and he said hes happy to be hopeful that the moving around that of a fictional character together under an array of men, not at least one woman, could become the idea more mainstream. If it gets me wrong, its going to stop working for two years, he said. You think, the amount of media sells and meofus men at your heels,ÃÂÃÂ, and yanking you with a sub while dragging it through along? But Bally is also aware that her own shows lack of campaigns around more is directed directed at children. And yes, theyre quite certainly, why not the most part, from the discussion, most of the world, I have seen with, can one create that. He was asked about that when she said he asked Louis K. to star on the show. In reality, these threats get into the gaps of contention between a number of sea-tainers and the people shes trying to keep on.I have just been surprised by your response to had so much needs to clear. I say, well, thats me why were just as grateful for our work. In my campaign, we have done it on their own, and the results have been positive. This post was a last thing. I want to say [1] 30 years later. When she needed 30 years to make the news, we gave the U.S. news and gave this year when only 10-20 percent of the gas-wingaded covered. The news to the people of this region can change any time. My organization used these weapons and light to the problem during the days of fracking, which has a huge impact on the 50/Coal industry. I am sorry for the process of taking this responsibility on. Americans have the right to know with you top of where you stand or we per side with that first of. You, the reader and other people do, and get ahead of what you are asked for. We are not out for modernization of our industry now, and the strength of our industry is very great. I told anyone write to me about that as well. Fn't it be fine? (here's my response, I said yes that provoked some policy in action from the person I. However, I think our national leaders and companies called for a up or down when said an issue. We and out or the "the one of which was no one I was hearing from. I think it is just fine in every way. I will continue if you at I feel like. I know you would like us to, as a part of this business, listen to us without the news.This is not The Unknown, by Sol Kotis, put off for office at the top of the year-the- news. In Russia, it a lot like, his five statemates, bloggers and politicians said they did and wrote about through the daily RSKT. Komis already claimed such as "candie" to the US. "Apparently, the government cards were flying, not on them, but peddled and stomped by a right-of-center media cover from America," wrote his predecessor. "They did then steal less from I, who tends to understand the country, and made up its about both television and present." No, we're learning a little much about the data, from then-on the only point since the advent of the Cata, we may have lost part of the data. The in particular range of the shows on Wednesday by New Moscow were the result of the Q1 period. The non-target shows aired on Sept. 1, 11, and 14, while its 11.6 aired. They're "17. 26" that it were the last air shows in that period, not the GrimIT report. In that period officials informally had 2,000 online TV stations, and they made up 12.2 shows per 100,000 viewers. However, the 19.25 rating didn't change further, the same year in ON, and a few other shows were added in that period. But this season backends up more like America and Italy: rot: In today's interview live on He

FMLM+ (PR), Round: 4 Committed: **1024/1024**

say a year of a contract, $5,000 a year on the network by three women—even though its the only thing in the U.S. network today, that is a national good. But saying that deducting money from miles a no, Baca said it would seem that that is a step later, but the issue of race in the F few is a good thing and could make Courns Killingajl The World expert—and he said hes happy to be concerned that the issues around that of a racial character together under an array of men, not at least one woman, could become the idea more popular. If it gets me wrong, its going to stop working for two years, he said. You think, the amount of media company and meofus men at your side, all, and yanking you with a sub while halfway it through along? But Bale is also convinced that her own shows lack of control around more is more directed at children. And yes, theyre quite certainly, why not the most part, from the discussion, most of the world, I have seen with, can one create that. He was asked about that when she said he asked Larry K. to star on the show. In reality, these threats get into the line of questioning between a number of circuit-attackers and the people shes trying to keep on.I have just been surprised by your response to had so much needs to clear. I say, well, thats me why were just as ready for our work. In my campaign, we have done it on their own, and the results have been positive. This post was a last thing. I want to say [1] 30 years later. When she needed 30 years to make the news, we gave the U.S. news and done this year when only 10-20 percent of the gas-use industry is. The news to the people of this region can come any time. My organization used these weapons and lied to the problem during the days of Vietnam, which has a huge impact on the 50/Coal industry. I am glad for the process of taking this responsibility on. Americans have the right to know with you top of where you stand or we per side with that first of. You, the people and other people do, and get ahead of what you are asked for. We are not out for modernization of our industry now, and the strength of our industry is very great. I told anyone listening to me about that as well. Fn't it be fine? (here's my response, I said yes that's some policy in action from the person I. However, I think our national leaders and companies called for a up or down when said an issue. We and out or the "the one of which was no one I was hearing from. I think it is just fine in every way. I will continue if you at I feel like. I know you would like us to, as a part of this business, listen to us without the news.This is not The Coup, by Robert Aris, put off for office at the top of the year-the- news. In Russia, it a lot like, his five state secretaries, officials and even said they did and wrote about through the daily R.T. Basis already, such as "candaries" to the US. "They, the government, were dependent, not on them, but muffled and disinfected by a right-of-center media cover from America," wrote his friend. "They did then much less from I, who wanted to understand the country, and made up its about both past and present." No, we're upset a little much about the data, from then-on the only point since the end of the CATE, we may have lost part of the data. The in particular range of the shows on Wednesday by NewTimes were the first of the Q1 period. The non-State shows aired on pp. 1, 11, and 14, while its 11.616. It's "17. 26" that it were the last air shows in that period, not the first per report. In that period officials provisionally had 2,000 online TV viewers, and they made up 12.2 shows per 100,000 viewers. However, the 19.25 range didn't change further, the same year in 2012, and a few other shows were added in that period. But this season back it up more like America and Italy: Steve: In today's interview live on He

Figure 9: Posterior Refinement trajectory on OpenWebText, Sample-1. (cont.)

FMLM+ (PR), Round: 1 Committed: **719/1024**

. Pedthan is another man named Zack. He has been practicing on a presidential trap for 9 years, shootingwn nights in opposed parts of Bainbush. I loved Owen's perspective that after the recent robbery in front of school, many reading of them transformed it into a fountain of grass. Merc afterward, elMW David Larry Macponder directed no return. When talking to the press shortly after the district raid, Elk Cross national SLR Rep Dave Shore tells me the news: To tell it in an office that is like a lake — we remind conversations of our ties with the establishment. It was the first police raid for an owner in strange systems of Pokemon. It was also that Arpaio chose to open dried gas and EP Broadway, and then was, to mute short his successor, Waters, who has served as software analyst for Wireless, Cisco, and Homery Brands. Theres a lot for me to raise with myself that would affect me. I think this is my first things, he says. These were the customer addresses. If nobody refuses to see the new steel Trust was made to disable all of him, ask one of you to think about what each other tomorrow could show. When New Englands State House and V-ins operate, his specialty is focused on capacging growth. Amid the neighborhood health system, Tesla refuses to pass feed-only take of ATfuel and uranium from a friend destructive power, there are checks in on lifes of terrorists, and activists modifying, but not countering, these weapons. That was because of the amount of real arrests and the day preceding the NRAs Feb. 13, 2016 targeting Tamoisin was convicted, because many Terroristees were excluded from collapse victims. 40 percent of the number of dead drug circulation, as well as seangerak and Flooring were all 21-year-old students on their own. I think with all the classified data since Ive worked with David Arehoff, the amount of false sites or what I heard there was; a lot in which a lot aimed more at people other than Brians, but I know, my cost allows us to say yes, Pereanan said. In reality, Im not sure I reallym willing to do that money in Chicago. I was seen to do this to himself and so other areas dont credit them the data. The father didnt fall to fruition, and Powers used to imply that he could make people free to anyone with the yard crisis. One of the guys had teeth on the attack and I was shocked at what was no threat, Eddieyer told Virginia Bharat looking at the virus recent-media-view, spike against attention. When I asked Owen into a New York bands member at Harvard, at Election Center — something she said had been very effective — dam cited was some encouragement. For the first time, I had to take off the same dyedboy chava. This wasnt welcome as we talked on with Bharat, Student invasion, everything ticker and especially any potential routes. Thats OK. But I know Chuckcarilla stepped in and got a control and the incident was over and its always questions about the nithing per gallon. Before he targeted The official suspect among the constituents was filled but _____ (D-Conn.) appeared a little exasperatory of him. Since such never the CEO he is, Orning said it was not unusual to speculate that Mass Corp could get home to his potent shipiles. The National Security Relations official, Cook, reported after a running for supervisor that newly Creek Fifese was looking to collect energy exams for residency rules in Medical Heights and Hampton environments, changing them to spur future elections, and to assess the validity of the savings. Asked what transpired in a press shop hours early by my hosts, Adams said she had no access to the data. What questioning what the EPA, Seyer told me — yes I did not do that before I opened that nonsense. Since Natbie spent these 70-seven years in prison for felonycore, he sticks to Weed terrorists every imaginable in filing lyings of time with a few chunks of scapegoat here thats been tax for every type of oversight by countlessgun, unmobashed in an airline and by a colleague and lumped in with criminal interrogationbles for whom he testified as a Holocaust survivor. Special Vegas and the environment of Oxyware has is one of the companys most important part of business now. Supporters of Mr Zinyer told Arnold that a possible run for the oil mogul was not too possible. Devine told Park Radios Jill Oaken, Youre

FMLM+ (PR), Round: 2 Committed: **871/1024**

. Ornoon is another prickly photographer. He has been working on a firearms secretary for 9 years, shooting forEMOTE in different parts of New Hampshire. I think it's just that after the recent project in front of school, many more of them got it into a ton of fashion. Sunday night, Orgal came to Pulp, with no return. In response to the press conference after the combat day, Representative Eastetts (R-NY) told me the news: To find it in an office that is like a office — we still think of our business with the gun. It was the first police sting for an NRA in this time of war. It was also that he came to open the business and its presence, and then was, to cut short his friend, having, who has served as software analyst for Africa, better, and HP for decades. Theres a lot for me to do with myself that would help me. I think this is my first job, he says. These were the real Deal. If you have to see the new structures which was made to accommodate all of him, ask one of you to think about what each other applicant could show. On New Yorks State House and pass-A Hill, his policy is focused on investment and growth. In the public health system, for him to get world-class take of a city and money from a friend to fans, there are people in on lifes of days, and selling drug, but not heroin, these days. It was because of the amount of real pills and the day off the companys eye. Plus, the paying of unin was personal, because many narcist drugs were typical from last dose. 40 percent of the number of recreational drug use, as well as Bi etc. and dieting were all 21-year-old students on their own. I think with all the academic data since Ive worked with David Muum, the amount of truth sites or what I found there was; a lot in which a lot looked more at people other than Davids, but I know, my brain want us to say that, Segal said. In Mississippi, Im not sure I reallym going to do that money in Manhattan. I was going to do this to schools and so other areas dont give them the data. The answers didnt fall to Les, and he used to worry that he could make people arrangements to cope with the 2008 crisis. One of the managers had me on the back and I was looking at what was no floor, Orlor told, brat looking at the auditors lady-in-law, looking against him. And I asked him into a New York Food suite at home, at Roosevelt Center — something she said had been very effective — and there was some chance. For the first time, I had to take off the same shirt and chimes. This didnt happen as we continued on with Tabat, Dr continued, chauner and counting any potential logistics. Thats normal. But I know that eventually you moved in and got a control and the work was over and its always questions about the cost from per se. Where he agreed The official reputation among the researchers was controversial but some (D-state not necessarily a good one) of him. As such never the candidate he is, Orbee said it was not clear to understand that an operation could get home to his first master cause. The National Energy Council official, Per, reported after a running for mayor that Pageday said he was looking to the energy company for industry rules in the cameras and the business, changing them to fit future contracting, and to skew the size of the document. At what was in a course three hours early by my reporter, Larson said she had no access to the data. What triggered what the director, Segal told me — was I did not do that before I opened that tank. To have one of these know- ya years in regulation for energy life, he had to start getting every answer in exchange for respect of background with a few pounds of bureaucrat here thats been tax for every type of activity by multiple people, including health, in an individual and by a stroke and lumped in with criminal activity, for which he serves as a security day. Today, and the type of resources he has is one of the companys most important part of business now. One of his former staffers told me that a possible run for the oil refinery was not too possible. He also told The Posts Barbara Oster, Youre

Figure 10: Posterior Refinement trajectory on OpenWebText, Sample-2.

FMLM+ (PR), Round: 3 Committed: **948/1024**

. Cornnie is another prickly guy. He has been working on a wine operation for 9 years, shooting for new in different parts of New York. I think it's just that after the recent project in front of school, many more of them turned it into a line of driving. That night, Orzan came to crimp, with no return. In response to the press conference after the opening day, Senator Andrew Holmes (R-FL) told me the news: To find it in an office that is like a cart — we still think of our business with the government. It was the first police order for an organization in this time of war. It was also that he came to open the business and its website, and then was, to cut short his successor, Kyle, who has served as software services for IBM, Google, and Microsoft for decades. Theres a lot for me to do with myself that would help me. I think this is my first work, he says. These were the real guys. If you have to see the new JavaScript which was made to replace all of him, ask one of you to think about what each other work could show. On New Yorks State House and Mid-A syndrome, his work is focused on money and growth. In the public health system, for him to get world-class take of a charity and money from a friend to invest, there are people in on lifes of days, and counting fine, but not enough, these days. It was because of the amount of real money and the day off the companys office. Interestingly, the use of ricin was personal, because many reformistencies were typical from last year. 40 percent of the number of illicit drug use, as well as money laundering. and prospecting were all 21-year-old students on their own. I think with all the big data since Ive worked with David Shum, the amount of energy sites or what I found there was; a lot in which a lot was more at people other than Davids, but I know, my brain want us to say that, Seanke said. In fact, Im not sure I reallym going to do that money in David. I was going to do this to schools and so other areas dont give them the data. The share didnt fall to him, and he used to show that he could make people reached to deal with the financial crisis. One of the individuals had me on the back and I was looking at what was no deal, I was told, Seat looking at the companys father-in-law, looking against him. And I asked him into a New York technical office at home, at Rockefeller Center — something she said had been very effective — and there was some competition. For the first time, I had to take off the same shirt and chatted. This didnt change as we continued on with Seat, another person, home eer and anticipating any potential activity. Thats great. But I know that eventually you came in and got a control and the work was over and its always questions about the revenues from per se. Did he? The official perception among the media was times but some (D-wa not always a good one) of him. With such never the prosecutor he is, Orenberg said it was not easy to understand that an operation could get home to his first column ever. The National Security Council official, though, reported after a running for governor that Orlov said he was looking to the energy committees for the rules in the boiler and the business, changing them to fit future trends, and to trim the size of the rolls. At what was in a meeting three hours early by my office, she said she had no access to the data. What — what theacebook, Seat told me — was I did not do that before I opened that office. To have one of these know-how years in line for your life, he had to start getting every day in work for something of sale with a few ounces of Esat here thats been tax for every type of activity by other people, including criminal, in an enterprise and by a company and lumped in with criminal activity, for which he cited as a trial day. That, and the type of remaining he has is one of the companys most important part of business now. One of his closest clients told me that a possible run for the oil company was not too possible. He also told The Hills Peter Oplan, Youre

FMLM+ (PR), Round: 4 Committed: **1024/1024**

. Orchid is another Deathly guy. He has been working on a small field for 9 years, shooting for new in different parts of New York. I think it's just that after the recent project in front of school, many more of them threw it into a line of business. That night, Ortha came to Dorp, with no return. In response to the press conference after the opening day, Mayor Andrew Cuomo (R-CA) told me the news: To find it in an office that is like a university — we still think of our business with the government. It was the first police raid for an organization in this time of war. It was also that he came to open the business and its website, and then was, to cut short his child, Charles, who has served as software services for Google, Shell, and broadcasting for decades. Theres a lot for me to do with myself that would help me. I think this is my first work, he says. These were the real ones. If you have to see the new culture which was made to help all of him, ask one of you to think about what each other work could show. On New Yorks State House and L-A Major, his work is focused on money and growth. In the public health system, for him to get world-class take of a job and money from a friend to visit, there are people in on lifes of days, and just fine, but not enough, these days. It was because of the amount of real money and the day off the companys office. Then, the use of Lupin was personal, because many reformist students were typical from last year. 40 percent of the number of real drug use, as well as money laundering. and milling were all 21-year-old students on their own. I think with all the big data since Ive worked with David Shum, the amount of energy sites or what I found there was; a lot in which a lot was more at people other than Davids, but I know, my brain want us to say that, Seat said. In fact, Im not sure I reallym going to do that money in something. I was going to do this to schools and so other areas dont give them the data. The people didnt fall to him, and he used to show that he could make people want to deal with the financial crisis. One of the individuals had me on the back and I was looking at what was no deal, I was told, Seat looking at the companys partner-in-chief, looking against him. And I asked him into a New York government office at home, at Rockefeller Center — something she said had been very effective — and there was some competition. For the first time, I had to take off the same shoes and chucks. This didnt change as we continued on with Seat, another person, home eer and dam any potential activity. Thats great. But I know that eventually you got in and got a control and the work was over and its always questions about the intelligence from per se. Is he? The official reaction among the media was times but some (D-, not always a good one) of him. With such never the way he is, Orchin said it was not easy to understand that an operation could get home to his first office ever. The National Security Relations official, though, reported after a running for office that Orchin said he was looking to the energy department for the rules in the business and the business, changing them to fit future threats, and to limit the size of the website. At what was in a meeting three hours early by my office, she said she had no access to the data. What — what the truth, Seat told me — was I did not do that before I opened that office. To have one of these know-nothing years in line for your life, he had to start getting every day in work for something of himself with a few strokes of debat here thats been tax for every type of activity by other people, including mine, in an office and by a company and lumped in with criminal activity, for which he qualifies as a present day. That, and the type of control he has is one of the companys most important part of business now. One of his reporter friends told me that a possible run for the oil company was not too possible. He also told The Hills David Oster, Youre

Figure 11: Posterior Refinement trajectory on OpenWebText, Sample-2. (cont.)

FMLM+ (PR), Round: 1 Committed: **72/128**

Once upon a time, there was a little girl named Lily. She loved to play with her rubberade in her backyard. One day, she saw a giant monster in her backyard. The bug was friendly and wanted to make it its friends. But Lily doll Sarah saw the bug and pretended to be kind. "Be no, Lily. I always said it doesn't want to scare a bird. Leave them." Suddenly, the big lemon crawledounced. Lily and could hurt her friend because it to something dark. The toy fell fromled in the wind, didn't wanting to be friends by the fake might' so

FMLM+ (PR), Round: 2 Committed: **95/128**

Once upon a time, there was a little girl named Lily. She loved to play with her red plant in her backyard. One day, she saw a big bird in her backyard. The flower was enthusiastic and wanted to give it a kiss. But the beener saw the pumpkin and started to tease Lily. "Be careful, Lily. The bird said it wasn't nice to be a big rock for them." Instead, the bird flew stopped to show Lily it could eat her so if it talked and talked. The bird flew high up in the sky. Lily was happy to be friends with the big bird, so

FMLM+ (PR), Round: 3 Committed: **112/128**

Once upon a time, there was a little girl named Lily. She loved to play with her toy hoop in her garden. One day, she saw a big bird in her garden. The bird was hungry and wanted to give it a ride. But the big bird saw the plant and started to tease Lily. "Come in, Lily. The bird said it doesn't want to be a big friend for them." Instead, the bird sang closer to show Lily it could make her feel before it smiled and jumped. The bird flew high up in the air. Lily was happy to be honest with the big bird, so

FMLM+ (PR), Round: 4 Committed: **128/128**

Once upon a time, there was a little girl named Lily. She loved to play with her toy helicopter in her garden. One day, she saw a big bird in her garden. The bird was hungry and wanted to give it a ride. But the big bird saw the bird and started to tease Lily. "Come down, Lily. The bird said it didn't want to be a big friend for them." Suddenly, the bird flew back to show Lily it could make her wish before it sang and laughed. The bird flew high up in the sky. Lily was happy to be friends with the big bird, so

Figure 12: Posterior Refinement trajectory on TinyStories, Sample-1.

FMLM+ (PR), Round: 1 Committed: **71/128**

Once upon a time, there was a little girl called one. Her parents said "No, Lucy. Lucy was feeling very sad. Whenever she should play with toys, the water fam playing, sometimes just like icebows. But Lucy still couldn't want to ignore it. Mommy and Daddy stopped in the park and suddenly, Lucy saw a giant purrying down! She pointed up pointing at the billboard saying "Ma". Jenny was so excited that she shouted everyYes of the crane." Daddy looked nodded. It was quite expensive, but she saw it just again and asked if it wanted

FMLM+ (PR), Round: 2 Committed: **94/128**

Once upon a time, there was a little girl called Emma. Mommy said "No, baby. Emma was feeling very sad. As she was filled with soap, the water was so happy, it looked like a hug. But M still didn't want to see it. Mommy started to look in the park and suddenly, she saw a little green bob floating down! She looked up and at it and said "Hello shrimp". She was so excited that she pulled it out of the ground. The fish smiled. It was quiteucky, but she saw it was smiling and asked if it wanted

FMLM+ (PR), Round: 3 Committed: **111/128**

Once upon a time, there was a little girl called baby. Mommy said "No, daddy. She was feeling very sad. As she was playing with water, the water was so big, it looked like a flood. But her daddy wasn't seem to accept it. Mommy started to look in the park and then, she saw a little boat fishing down! She looked up straight at it and said "Hiuck!" She was so excited that she pulled it out of the water. The girl smiled. It was a surprise, but she saw it was friendly and asked if it wanted

FMLM+ (PR), Round: 4 Committed: **128/128**

Once upon a time, there was a little girl calledmy. Mommy said "No, Sammy. She was feeling very sad. As she was playing with water, the water was so big, it looked like a flood. But her feet couldn't seem to make it. Mommy started to look in the park and suddenly, she saw a little girl was sliding down! She looked up over at it and said "Hielcome!" She was so excited that she took it out of the water. The girl smiled. It was a cold, but she saw it was smiling and asked if it wanted

Figure 13: Posterior Refinement trajectory on TinyStories, Sample-2.