GUMBEL STRAIGHT FLOW: DISTILLING AUTOREGRESSIVE MODELS INTO ONE-STEP FLOW MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present Gumbel Straight Flow (GSF), a continuous flow map language model that leverages the noise-data coupling of a pretrained autoregressive language (AR) model. We theoretically demonstrate that the coupling between Gumbel noise and one-hot token sequences induced by an autoregressive model yields non-intersecting linear paths connecting the noise to the sequence representations. To further enhance high-quality few-step path sampling, we use a flow map semigroup objective where the tangent (velocity) condition is guided directly by the AR teacher. Across various benchmarks, including pretraining and downstream tasks, GSF can outperform current few-step language generation baselines.

1 INTRODUCTION

Autoregressive models (AR) are widely used for language generation due to their strong generative performance. However, their sequential generation process is computationally expensive. To address this limitation, parallel-generation language models based on discrete diffusion or continuous flow models have been introduced (Austin et al., 2021; Dieleman et al., 2022; Li et al., 2022; Sahoo et al., 2024; Shi et al., 2024; Hu et al., 2026a; Lee et al., 2026). Nevertheless, discrete diffusion and continuous-flow models often require many sampling steps to control factorization or discretization errors.

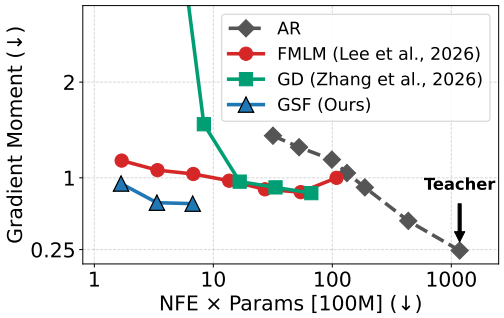


Figure 1: Efficiency-performance trade-offs for GSF, primary distillation baselines (varying step sizes), and AR models (varying model sizes).

To accelerate inference in parallel generative models, a natural direction is distillation, i.e., compressing multi-step iterative generation into a few sampling steps or a single step (Song et al., 2023; Deschenaux and Gulcehre, 2025; Hayakawa et al., 2025; Hoogeboom et al., 2026), often using a discrete diffusion model as a teacher. Given their better empirical performance on text modeling, we explore whether autoregressive (AR) models can be used as more capable teachers. Recent work harnesses AR teachers to supervise parallel student models by constructing deterministic sequence-level noise–data pairs (Draxler et al., 2026; Zhang et al., 2026). However, learning sequence-level noise-to-data mappings is an exceptionally hard supervised problem; for instance, Gumbel-Distill (GD) (Zhang et al., 2026) must still fall back to multi-step refinement over a discrete diffusion backbone.

Table 1: Comparison of GSF against existing language generation paradigms. This paper shows that the AR model's coupling forms a non-intersecting straight path and distills the path into GSF.

Model	Parallel Generation	Deterministic Pairing	Straight Path
Autoregressive Model	✗	✓	✓
Discrete Diffusion Model	✓	✗	✗
Continuous Flow (Diffusion)	✓	✓	✗
Gumbel Straight Flow (GSF)	✓	✓	✓

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

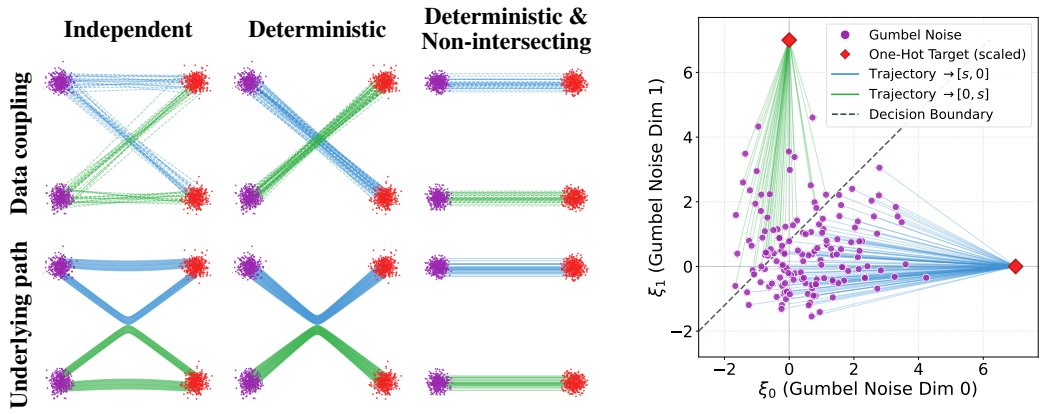


Figure 2: (Left) Three types of noise–data coupling and their corresponding optimally learned flow paths. (Right) The Gumbel coupling induced by an AR model belongs to the non-intersecting case and yields straight underlying paths.

In this paper, we instead define a flow path in one-hot continuous space directly over the AR model's Gumbel-max sampling. This defines a coupling between the distribution of Gumbel sequences $\mathbf{x}_0$ and corresponding one-hot token vectors $\mathbf{x}_1$ such that, for endpoints distributed according to this coupling, the linear interpolants $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ do not intersect (Figure 2). Gumbel Straight Flow (GSF) then learns these straight paths provided by an AR teacher while enabling parallel generation. Straight paths are desirable: minimizing trajectory curvature allows numerical ODE solvers to generate samples in a few steps without discretization issues (Liu et al., 2023). We further enhance few-step performance using a flow map parameterization (FMLM) (Lee et al., 2026) where the straight tangent (velocity) supervision is given by the AR teacher. As summarized in Table 1, GSF is (to our knowledge) the first parallel-generation model to combine deterministic pairing with the straight, non-intersecting trajectory structure of autoregressive sampling.

Empirical results show that one-step GSF achieves a Gradient Moment (GM)[1] (Hoogeboom et al., 2026) score of 0.94, establishing a new efficiency-performance frontier compared to GD and FMLM as shown in Figure 1. Furthermore, GSF sits on a new, better Pareto frontier than that of even AR models in terms of the efficiency-compute trade-off. The four-step GSF achieves Gen. PPL = 99.7 and Entropy = 5.32, placing it in the frontier region (Figure 6) for fidelity-diversity trade-offs under the identical sampling budget.

2 Preliminaries

2.1 Autoregressive Models

Let $\mathcal{V}$ denote a finite vocabulary and L the sequence length. At each position $i \in \{1, \dots, L\}$, conditioned on the preceding tokens $\mathbf{y}_{<i} = (y_1, \dots, y_{i-1})$, an autoregressive (AR) language model produces a logit vector $\boldsymbol{\ell}_i(\mathbf{y}_{<i}) = (\ell_{i,1}(\mathbf{y}_{<i}), \dots, \ell_{i,|\mathcal{V}|}(\mathbf{y}_{<i})) \in \mathbb{R}^{|\mathcal{V}|}$ and computes the categorical probability of token $j \in \mathcal{V}$ as:

$$p_{\text{AR}}(y_i = j \mid \mathbf{y}_{<i}) = \frac{\exp(\ell_{i,j}(\mathbf{y}_{<i}))}{\sum_{k \in \mathcal{V}} \exp(\ell_{i,k}(\mathbf{y}_{<i}))}. \tag{1}$$

The joint probability of generating a sequence of L tokens $\mathbf{y}_{1:L} = (y_1, \dots, y_L)$ is given by the chain rule: $p_{\text{AR}}(\mathbf{y}_{1:L}) = \prod_{i=1}^{L} p_{\text{AR}}(y_i \mid \mathbf{y}_{<i})$.

2.2 Gumbel-Max Categorical Sampling

Sampling a token sequence from the autoregressive distribution in Eq. (1) can be performed ancestrally via the Gumbel-max trick (Gumbel, 1954). Let $\boldsymbol{\xi}_i = (\xi_{i,1}, \dots, \xi_{i,|\mathcal{V}|}) \in \mathbb{R}^{|\mathcal{V}|}$ denote

[1] GM tends to be more robust against mode collapse than Gen. PPL (Franca and Tong, 2026).

independent standard Gumbel noise vectors with $\xi_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{G}(0,1)$ across all positions $i \in \{1,\dots,L\}$ and vocabulary coordinates $j \in \mathcal{V}$. At each position i, the next token is generated deterministically given the noise and the previous history:

$$y_i = \arg\max_{j \in \{1,\dots,|\mathcal{V}|\}} \big(\ell_{i,j}(\mathbf{y}_{<i}) + \xi_{i,j}\big). \tag{2}$$

By the Gumbel-max trick, the resulting token y_i is distributed according to $p_{\text{AR}}(\cdot \mid \mathbf{y}_{<i})$ from Eq. (1), ensuring that the full sampled sequence satisfies $\mathbf{y}_{1:L} \sim p_{\text{AR}}(\mathbf{y}_{1:L})$. Writing $\boldsymbol{\xi}_{1:L} \in \mathbb{R}^{L\times|\mathcal{V}|}$ and $\mathbf{y}_{1:L} \in \mathcal{V}^L$, ancestral Gumbel sampling defines a deterministic mapping $T_{\text{AR}} : \boldsymbol{\xi}_{1:L} \xrightarrow{\text{det.}} \mathbf{y}_{1:L}$ that couples the continuous Gumbel noise sequence with the discrete (teacher) generated sequence.

2.3 Continuous-state Language Flow

Following Lee et al. (2026), we embed text sequences into continuous space via one-hot encoding:

$$\text{Encoding: } \mathbf{x} = \text{onehot}(\mathbf{y}_{1:L}) \in \mathbb{R}^{L\times|\mathcal{V}|}, \qquad \text{Decoding: } \mathbf{y}_{1:L} = \arg\max(\mathbf{x}).$$

Flow Language Model (FLM). FLM (Lee et al., 2026) is based on flow matching (FM) (Lipman et al., 2023; Liu et al., 2023). It defines two independent marginal distributions: a Gaussian source distribution $\mathbf{x}_0 \sim p_0 := \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a one-hot-encoded data distribution $\mathbf{x}_1 \sim p_1 := p_{\text{data}}$. FLM defines the generative process as $\mathrm{d}\mathbf{x}_t^\theta = \mathbf{v}_\theta(\mathbf{x}_t^\theta, t)\,\mathrm{d}t$ and can train the velocity model $\mathbf{v}_\theta$ using the mean squared error (MSE) loss:

$$\min_\theta \int_0^1 \mathbb{E}_{\mathbf{x}_0\sim p_0, \mathbf{x}_1 \sim p_1}\Big[\|(\mathbf{x}_1 - \mathbf{x}_0) - \mathbf{v}_\theta(\mathbf{x}_t, t)\|_2^2\Big]\,\mathrm{d}t, \qquad \mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1,$$

where the optimal velocity is $\mathbf{v}(\mathbf{x}_t, t) := \mathbb{E}_{p(\mathbf{x}_0,\mathbf{x}_1\mid\mathbf{x}_t)}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_t]$. However, this velocity parameterization does not exploit the fact that the data lie on the one-hot simplex. Recent studies (Lee et al., 2026; Potaptchik et al., 2026) instead adopt a denoising formulation, in which the denoising network uses a softmax output layer and is trained with the cross-entropy loss:

$$\min_\theta \int_0^1 \mathbb{E}_{\mathbf{x}_0\sim p_0, \mathbf{x}_1 \sim p_1}\Big[\text{CE}(\mathbf{x}_1 || D_\theta(\mathbf{x}_t, t))\Big]\,\mathrm{d}t, \qquad D_\theta(\mathbf{x}_t, t) := \mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t). \tag{3}$$

This formulation has the same optimum as the MSE objective. Specifically, the optimal denoiser is given by $D(\mathbf{x}_t, t) := \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_t] = \mathbf{x}_t + (1-t)\mathbf{v}(\mathbf{x}_t, t)$.

Flow Map Language Model (FMLM). FMLM (Lee et al., 2026; Roos et al., 2026) is a flow map variant of FLM that defines $X_{s,t} : \mathbb{R}^{L\times|\mathcal{V}|} \to \mathbb{R}^{L\times|\mathcal{V}|}$ by integrating the velocity field:

$$X_{s,t}(\mathbf{x}_s) := \mathbf{x}_s + \int_s^t \mathbf{v}(\mathbf{x}_\tau, \tau)\,\mathrm{d}\tau = \mathbf{x}_s + (t-s)\mathbf{u}_{s,t}(\mathbf{x}_s),$$

where $\mathbf{u}_{s,t}$ denotes the average velocity from time s to t. A flow map satisfying this property characterizes the underlying flow path:

$$\mathbf{u}_{t,t}(\mathbf{x}_t) = \lim_{s\to t} \mathbf{u}_{s,t}(\mathbf{x}_t) = \mathbf{v}(\mathbf{x}_t, t), \qquad X_{u,t}(X_{s,u}(\mathbf{x}_s)) = X_{s,t}(\mathbf{x}_s). \tag{4}$$

Define the corresponding two-time denoiser formulation as:

$$\delta_{s,t}(\mathbf{x}_s) := \mathbf{x}_s + (1-s)\mathbf{u}_{s,t}(\mathbf{x}_s),$$

and the objective for training the parameterized two-time denoiser δ^θ to satisfy the flow map condition in Eq. (4) is given by:

$$\begin{aligned}\mathcal{L}_{\text{FMLM}}(\theta) = &\underbrace{\mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\Big[\text{CE}\big(D(\mathbf{x}_t, t)\,\big\|\,\delta_{t,t}^\theta(\mathbf{x}_t)\big)\Big]}_{:=\,\mathcal{L}_{\text{diag.}}(\theta; D)} \\ &+ \mathbb{E}_{s,u,t,\mathbf{x}_0,\mathbf{x}_1}\Big[\underbrace{\text{CE}\big(\text{sg}\big(\gamma\delta_{s,u}^\theta(\mathbf{x}_s) + (1-\gamma)\delta_{u,t}^\theta(X_{s,u}(\mathbf{x}_s))\big)\,\big\|\,\delta_{s,t}^\theta(\mathbf{x}_s)\big)}_{:=\,\ell_{\text{off-diag.}}(\theta)}\Big],\end{aligned} \tag{5}$$

where $\gamma := \frac{(1-t)(u-s)}{(1-u)(t-s)}$, and $\text{sg}(\cdot)$ is a stop gradient operation. Here, $\mathbf{x}_0 \sim p_0$ and $\mathbf{x}_1 \sim p_1$ are sampled independently. In practice, D is replaced with a pretrained FLM model from Eq. (3).

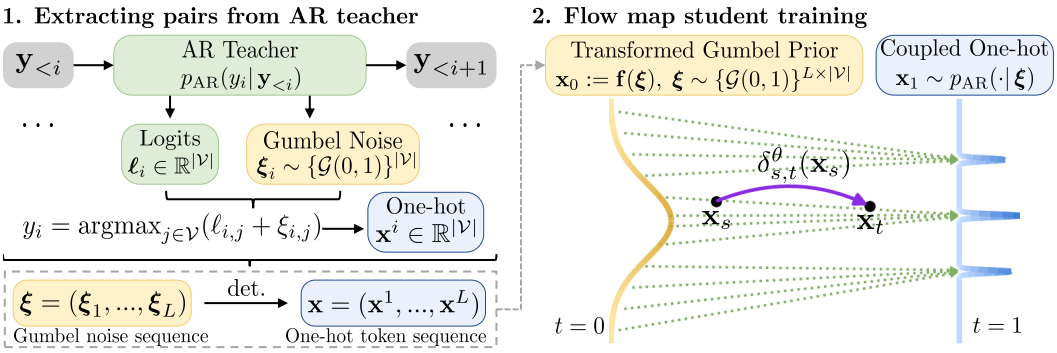


Figure 3: Overview of GSF: distilling AR-induced straight paths into a flow map student.

3 Method

This section introduces Gumbel Straight Flow (GSF), a method for distilling an autoregressive (AR) teacher into a flow map. Section 3.1 analyzes the properties of AR-induced paths (green dotted lines in Figure 3), and Section 3.2 explains how to distill these paths into GSF (purple arrow in Figure 3).

3.1 Analysis of AR-induced Gumbel Coupling Paths

Section 2.2 shows that AR models provide deterministic noise-to-data supervision. However, this deterministic relationship alone does not guarantee underlying straight paths (second column of Figure 2). This section uses a flow matching parametrization to show that the Gumbel coupling induced by AR models yields straight paths, corresponding to the third column of Figure 2.

Gumbel Straight Flow with Flow Matching (GSF-FM). We define the generative process as $\mathrm{d}\mathbf{x}^{\theta}_t = \mathbf{v}_\theta(\mathbf{x}^{\theta}_t, t)\,\mathrm{d}t$ and train it using the AR-induced coupling, where $\mathbf{x}_0 := \boldsymbol{\xi} \sim \mathcal{G}(0,1)^{L\times|\mathcal{V}|}$ and $\mathbf{x}_1 \sim p_{\mathrm{AR}}(\mathbf{x}_1|\boldsymbol{\xi})$. The training objective is:

$$\mathcal{L}_{\text{GSF-FM}}(\theta, p_{\mathrm{AR}}) := \int_0^1 \mathbb{E}_{\mathbf{x}_0\sim\mathcal{G},\ \mathbf{x}_1\sim p_{\mathrm{AR}}(\cdot|\mathbf{x}_0)}\left[\|(\mathbf{x}_1 - \mathbf{x}_0) - \mathbf{v}_\theta(\mathbf{x}_t, t)\|_2^2\right]\mathrm{d}t, \tag{6}$$

with the usual interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$.

3.1.1 Straightness

The straightness measure (Liu et al., 2023) for a path $\{\mathbf{x}^{\theta}_t\}_{t\in[0,1]}$ [2] with velocity $\mathbf{v}_\theta$ is defined as

$$S(\{\mathbf{x}^{\theta}_t\}) := \int_0^1 \mathbb{E}[||(\mathbf{x}^{\theta}_1 - \mathbf{x}^{\theta}_0) - \mathbf{v}_\theta(\mathbf{x}^{\theta}_t, t)||_2^2]\,\mathrm{d}t, \tag{7}$$

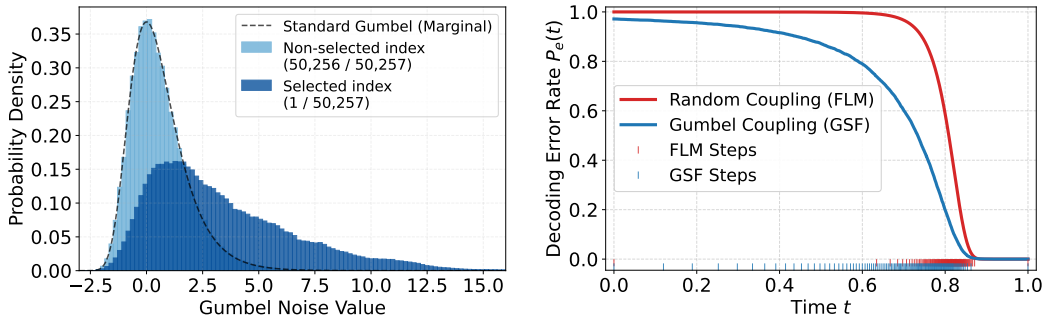
and the path $\{\mathbf{x}^{\theta}_t\}$ is straight if $S(\{\mathbf{x}^{\theta}_t\}) = 0$. The following theorem states that the optimal path learned using $\mathcal{L}_{\text{GSF-FM}}$ is straight; i.e., the underlying paths provided by the AR model are straight.

Theorem 1. ***(Straightness)*** *Assume sufficient model capacity. Let the optimal velocity be $\mathbf{v}_{\theta^*}$ where $\theta^* := \arg\min_\theta \mathcal{L}_{\text{GSF-FM}}(\theta, p_{\mathrm{AR}})$, and let $\{\mathbf{x}^{\theta^*}_t\}$ denote the corresponding path. Then the path is straight, i.e.,*

$$S(\{\mathbf{x}^{\theta^*}_t\}) = 0.$$

The proof (Appendix A.2) relies on the fact that, under our Gumbel coupling, both $\mathbf{x}_0$ and $\mathbf{x}_1$ are uniquely identifiable given $\mathbf{x}_t$, because the paired paths do not intersect. By non-intersection, we mean that two distinct trajectories cannot be equal to the same state at the same time t for $t < 1$. Assuming the linear pairings intersect, we obtain a contradiction: the argmax ordering from Eq. (2) no longer holds. The right panel of Figure 2 visualizes the straight paths in the single token case. As detailed in Appendix A.4, this non-intersection property admits a semi-discrete optimal transport

[2] The path is a random variable, over which the expectation is taken for straightness calculation in Eq. (7). We omit the time index range $t \in [0, 1]$ hereafter for brevity.



(a) Gumbel noise statistics from 8 sentences extracted from an AR model.

(b) Decoding error rate and the reparameterized step $\tau(t)$ for uniformly distributing denoising information.

Figure 4: (a) Under the Gumbel-max trick, the selected index tends to have a larger value than non-selected indices. Since only one of the $\mathcal{V}$ indices is selected, the marginal distribution follows a standard Gumbel distribution. (b) GSF injects informative noise; consequently, decoding errors occur as the noise-injection signal increases and are distributed more uniformly over time t.

interpretation (Peyré and Cuturi, 2019): for a single token, Gumbel-max sampling is the exact optimal transport map where teacher logits act as closed-form dual weights. For token sequences, the transport is obtained by applying a sequence of single-token transport maps successively; i.e., it is a causal transport map (Backhoff et al., 2017).

3.1.2 Straightness Holds under Prior Gaussianization

Despite its theoretical straightness, Eq. (6) uses a Gumbel prior distribution, which is strongly right-skewed, as shown by the dotted line in Figure 4a. Since existing flow model architectures are typically designed for Gaussian priors, they may not effectively capture information from the Gumbel distribution. We extend our method by applying a scalar transformation $f : \mathbb{R} \to \mathbb{R}$ element-wise to the Gumbel prior $\boldsymbol{\xi}$. Let $\mathbf{f}$ denote its element-wise extension, i.e., $[\mathbf{f}(\boldsymbol{\xi})]_{i,j} = f(\xi_{i,j})$, and choose f such that $\mathbf{f}(\boldsymbol{\xi})$ follows a Gaussian distribution. The corresponding extended objective function becomes $\mathcal{L}_{\text{GSF-FM}}(\theta, p_{\text{AR}}; \mathbf{f}) := \int_0^1 \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{G},\, \mathbf{x}_0=\mathbf{f}(\boldsymbol{\xi}),\, \mathbf{x}_1\sim p_{\text{AR}}(\cdot|\boldsymbol{\xi})} \left[\|(\mathbf{x}_1 - \mathbf{x}_0) - \mathbf{v}_\theta(\mathbf{x}_t, t)\|_2^2 \right] \mathrm{d}t$. The following proposition shows that straightness is preserved under a class of transformations f.

Proposition 2. ***(Prior transform).*** *Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly increasing and invertible function, and let $\mathbf{f}$ denote its element-wise extension. Under sufficient model capacity, let $\mathbf{v}_{\theta^*}(\mathbf{x}_t, t) := \arg\min_{\mathbf{v}_\theta} \mathcal{L}_{GSF\text{-}FM}(\theta, p_{AR}; \mathbf{f})$ denote the optimal velocity field, and let $\{\mathbf{x}_t^{\theta^*}\}$ denote the corresponding induced trajectory. Then, the path is still straight, i.e., $S(\{\mathbf{x}_t^{\theta^*}\}) = 0$.*

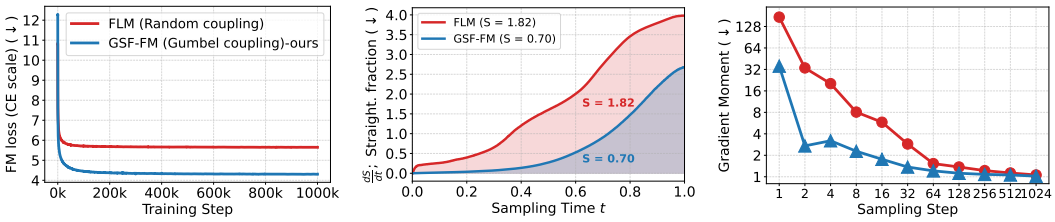
Defining the transformation as the composition of a uniform transform and the inverse Gaussian CDF, i.e., $f(\xi) := \Phi^{-1}(\exp(-e^{-\xi}))$, allows us to use a standard Gaussian prior while maintaining optimal straightness, as f is strictly increasing and invertible (proof in Appendix A.3). Unless otherwise specified, we employ the Gumbel-to-Gaussian mapping for f, making a Gaussian prior the default across all experiments, recovering the standard FLM / FMLM setting (Lee et al., 2026).

3.1.3 Time Reparameterization

Following Sahoo et al. (2025) and Pynadath et al. (2026), the decoding error rate and the corresponding time reparameterization that uniformly distributes decoding errors are defined as:

$$P_e(t) := \text{Prob}(\text{argmax}(\mathbf{x}_t) \neq \text{argmax}(\mathbf{x}_1)), \quad \tau(t) := \frac{P_e(0) - P_e(t)}{P_e(0)}.$$

When $|\mathcal{V}|$ is large, decoding errors in FLM concentrate around a specific timestep (near $t = 0.8$ in Figure 4b), which necessitates a reparameterized time schedule $\tau(t)$ to improve performance (Lee et al., 2026). In contrast, the coupled Gumbel noise injected in GSF is correlated with the selected argmax index (Figure 4a). As illustrated in Figure 4b, this correlation shifts most decoding errors toward higher noise scales and distributes them more uniformly across Euclidean space, keeping $\tau(t)$ closer to the original time t. We still use the coupled version of the reparameterized schedule by default, but the required time warping is substantially milder than in FLM.



(a) Training Loss (b) Straightness (CE scale) (c) Generative Performance

Figure 5: Comparison between FLM with independent coupling and GSF-FM with AR-based coupling on the OpenWebText dataset under a Gaussian prior.

3.2 Flow Map with AR-induced Gumbel Coupling

Figure 5 demonstrates that GSF-FM achieves superior empirical straightness (Figure 5b), thereby boosting generative performance (Figure 5c) particularly in the low-NFE regime. However, a substantial gap still persists between its few-step and many-step performance. This motivates us to adopt a flow map parameterization. Although flow maps inherently provide strong few-step generation, prior work (Liu et al., 2023; 2024) has shown that an underlying straight path greatly facilitates flow map training. We combine the semi-group flow map training objective Eq. (5) into:

$$\mathcal{L}_{\text{GSF}}(\theta, p_{\text{AR}}) := \underbrace{\mathbb{E}_t \mathbb{E}_{\mathbf{x}_0 \sim \mathbf{f}\#\mathcal{G},\, \mathbf{x}_1 \sim p_{\text{AR}}} \left[\text{CE}\big(\mathbf{x}_1 \,\|\, \delta^{\theta}_{t,t}(\mathbf{x}_t)\big) \right]}_{:= \mathcal{L}_{\text{diag.}}(\theta; p_{\text{AR}})} + \mathbb{E}_{s,u,t} \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1} \left[\ell_{\text{off-diag.}}(\theta) \right]. \tag{8}$$

Regarding the off-diagonal loss, because it enforces self-consistency rather than learning the underlying path geometry, the sample pair $(\mathbf{x}_0, \mathbf{x}_1)$ does not necessarily need to be drawn from the coupled setting. Empirically, we find no significant difference based on how $(\mathbf{x}_0, \mathbf{x}_1)$ is sampled; independent sampling yields better diversity at the cost of mildly lower generative performance.

The key difference from the previously described FMLM is that $\mathcal{L}_{\text{diag.}}(\theta; D_t)$ is replaced by $\mathcal{L}_{\text{diag.}}(\theta; p_{\text{AR}})$, effectively switching the supervision for the diagonal loss from the teacher FLM to the teacher AR model. Consequently, we do not need to train GSF-FM as in Section 3.1, since the AR model directly provides exact velocity supervision (i.e., exact denoiser supervision under the current formulation) along straight paths. In summary, we anticipate superior performance compared to FMLM due to three factors: (1) a higher performance ceiling by a better teacher model, (2) straighter, easier-to-learn target paths, (3) more linearized decoding-error-based time scheduling.

4 Experiments

In this section, we evaluate Gumbel Straight Flow (GSF) on unconditional generation, conditional generation, and perform several ablations. For a fair comparison, we use a comparable computational setting, based on a 12-layer bidirectional transformer, and match both parameter and FLOP budgets as shown in Figure 8a. Please refer to Appendix C for detailed experimental setup.

4.1 Unconditional Generation

AR-induced Gumbel coupling produces straighter empirical paths and accelerates flow. We first validate the flow matching formulation on unconditional generation of 1024-token sequences on OpenWebText (OWT) (Gokaslan and Cohen, 2019) under a Gaussianized prior (Figure 5). GSF-FM achieves better loss convergence during training (Figure 5a) due to the coupling. To evaluate the straightness, Figure 5b measures a rescaled version of $S(\{\mathbf{x}_t^\theta\})$ along generated trajectories: since a straight path implies the denoiser constantly targets the generated endpoint $\mathbf{x}_1$, we evaluate the straightness error in cross-entropy scale via $\text{CE}(\mathbf{x}_1 \,\|\, D_\theta(\mathbf{x}_t, t))$. This results in superior generative performance in the few-step regime as measured by the Gradient Moment (GM) score (Hoogeboom et al., 2026), which evaluates the difference between expected parameter gradients on generated versus real data samples judged by GPT-2 Large (Figure 5c).

Table 2: Few-step generative performance on OpenWebText among distillation models. Parentheses indicate teacher types. "-re": our re-implementation.

Method	GM (↓)	PPL (↓)	Ent.
AR teacher (1024-step)	0.24	46.5	5.59
NFE = 1			
SDTT (MDLM) (Deschenaux and Gulcehre, 2025)	-	1260.8	5.26
Di4C (MDLM) (Hayakawa et al., 2025)	-	1298.8	5.29
DCD (Duo) (Sahoo et al., 2025)	-	5743.2	6.02
Di4C (Duo) (Sahoo et al., 2025)	-	370.5	3.92
DFM (FLM) (Potaptchik et al., 2026)	-	180.2	4.91
GD (AR-Large) (Zhang et al., 2026)	7.67	1475.9	5.50
FMLM (FLM) (Lee et al., 2026)	1.18	168.3	5.17
FMLM (FLM)-re	1.30	140.0	4.93
GSF (AR)	**0.94**	**114.9**	5.07
NFE = 2			
SDTT (MDLM) (Deschenaux and Gulcehre, 2025)	-	877.2	5.34
Di4C (MDLM) (Hayakawa et al., 2025)	-	758.2	5.35
DCD (Duo) (Sahoo et al., 2025)	-	891.1	5.41
Di4C (Duo) (Sahoo et al., 2025)	-	210.2	4.63
DFM (FLM) (Potaptchik et al., 2026)	-	152.8	5.03
GD (AR-Large) (Zhang et al., 2026)	4.41	763.7	5.41
FMLM (FLM) (Lee et al., 2026)	1.08	133.2	5.25
FMLM (FLM)-re	0.91	118.0	5.22
GSF (AR)	**0.74**	**105.6**	5.30
NFE = 4			
SDTT (MDLM) (Deschenaux and Gulcehre, 2025)	-	339.7	5.38
Di4C (MDLM) (Hayakawa et al., 2025)	-	239.2	5.40
DCD (Duo) (Sahoo et al., 2025)	-	250.8	5.37
Di4C (Duo) (Sahoo et al., 2025)	-	154.6	4.85
DFM (FLM) (Potaptchik et al., 2026)	-	122.3	5.10
GD (AR-Large) (Zhang et al., 2026)	1.56	354.9	5.39
FMLM (FLM) (Lee et al., 2026)	1.04	111.3	5.26
FMLM (FLM)-re	0.92	108.2	5.24
GSF (AR)	**0.73**	**99.7**	5.32

Table 3: One-step generative performance among distillation models on LM1B dataset.

Method	GM (↓)	PPL (↓)	Ent.
SDTT (MDLM) (Deschenaux and Gulcehre, 2025)	-	1429.4	4.31
Di4C (MDLM) (Hayakawa et al., 2025)	-	1217.1	4.38
DCD (Duo) (Sahoo et al., 2025)	-	1224.5	4.33
Di4C (Duo) (Sahoo et al., 2025)	-	292.9	3.79
FMLM (FLM) (Lee et al., 2026)	6.16	119.3	4.16
FMLM (FLM)-re	6.35	124.2	3.97
GSF (AR)	**5.68**	**118.6**	4.05

Table 4: Few-step performance among continuous-state models on TinyStories dataset.

Method	NFE (↓)	PPL (↓)	Ent.
CANDI (Pynadath et al., 2026)	128	46.32	4.04
FLM (Lee et al., 2026)	128	57.50	4.17
S-FLM (Deschenaux and Gulcehre, 2026)	128	95.25	4.08
FMLM+ (Agarwal et al., 2026)	8	65.89	4.16
FMLM (FLM)-re	8	102.89	4.06
GSF (AR)	8	**41.22**	4.00

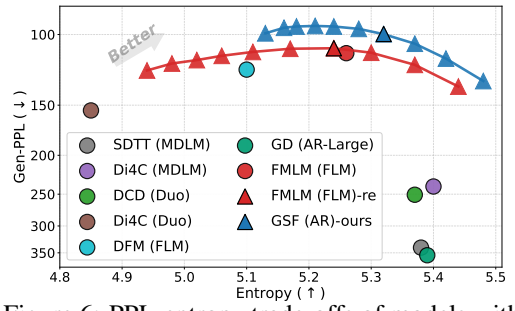


Figure 6: PPL-entropy trade-offs of models with 4 sampling steps trained on OpenWebText.

Few-step distillation performance on OpenWebText. Next, we evaluate few-step sampling using the flow map objective across 1, 2, and 4 sampling steps (Table 2). At NFE $=1$, GSF achieves 114.9 generative perplexity and a gradient moment score of 0.94, outperforming prior distillation models in this step regime. Whereas discrete diffusion distillation baselines collapse at single-step sampling (SDTT: 1260.8; DCD: 5743.2; Di4C (Duo): 370.5), GSF preserves reasonable performance. We also maintain an edge over FMLM variants (FMLM: 168.3; FMLM-re: 140.0). GSF's GM score decreases monotonically with additional sampling steps, and regarding training speed, Figure 8b shows that one-step GM converges faster and improves steadily over the course of training. We can examine the effect of straight continuous-path supervision by comparing GSF with GD, since GD also uses Gumbel noise. However, GD uses a discrete diffusion backbone and does not explicitly exploit path information, resulting in limited performance (PPL: 1475.9) compared with GSF. This may be because directly learning the mapping from Gumbel noise to data is an extremely difficult supervised problem. In contrast, the continuous straight path can provide useful intermediate supervision that helps the model learn this mapping more effectively.

Furthermore, as shown in Figure 6, GSF establishes a favorable Pareto trade-off between perplexity and unigram entropy at 4 sampling steps. Also, Figure 8c shows that generative performance improves with model size (from Small: 168M to Medium: 419M), where the Medium student is distilled from a correspondingly larger and stronger AR teacher. We also report LLM-as-a-judge pairwise win rates evaluated by Gemini 3.6-Flash in Figure 7 (with prompt in Appendix G). Across sampling steps, GSF outperforms FMLM and GD, where 1-step GSF is preferred in over 91% of comparisons even against 4-step GD. Appendix H shows examples of samples of GSF, FMLM, GD, and the corresponding judgment.

Similar results are also obtained on the LM1B dataset (Chelba et al., 2013) (Table 3) and on the TinyStories dataset (Eldan and Li, 2023) (Table 4). For the latter, GSF reaches 41.22 PPL at

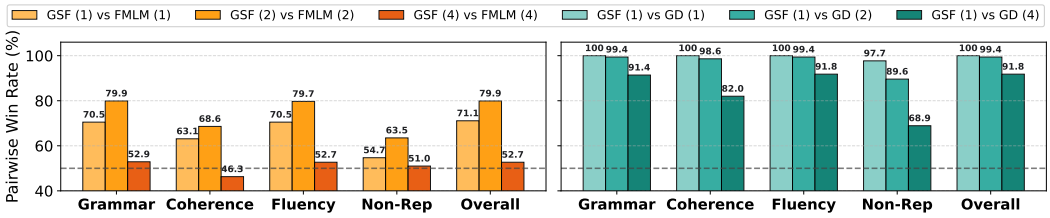
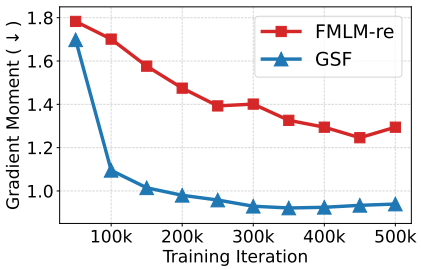


Figure 7: Pairwise win rate with Gemini 3.6-Flash. Parentheses indicate sampling steps.

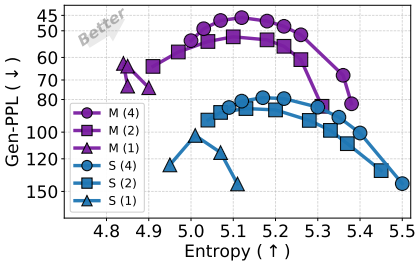
Method	Params (↓)
FMLM	169.7M
GD	208.2M
GSF	**167.8M**

Method	GFlops (↓)
FMLM	370.7
GD	370.7
GSF	**351.4**

(a) One-step cost



(b) One-step GM per training iteration



(c) Scaling behavior (S: 168M, M: 419M)

Figure 8: (a) Computational cost for single-step sampling on OpenWebText. (b) One-step generative quality across training iterations. (c) Scaling behavior across parameter size and sampling steps

NFE $= 8$, outperforming few-step flow map baselines (FMLM+: 65.89; FMLM-re: 102.89) as well as continuous flow models integrated over 128 steps (CANDI: 46.32; FLM: 57.50).

4.2 Conditional Generation

Prefix completion on OpenWebText. We evaluate conditional text completion. Models are conditioned on a fixed prefix consisting of the first 50 tokens extracted from held-out OpenWebText sequences and tasked with generating the continuation up to a total sequence length of 100. During sampling, the prefix positions are clamped to their true one-hot representations at each step, while the model generates the continuation tokens in parallel. Results are shown in Table 5. At NFE $= 1$, GSF achieves a conditional perplexity of 97.9 and token entropy of 4.00, improving upon continuous flow baselines (FMLM: 141.6; FMLM-re: 115.0) and outperforming discrete diffusion distillation methods (SDTT: 374.7; Di4C (MDLM): 376.8; DCD: 703.1). With 2 and 4 sampling steps, GSF scales to 77.5 and 71.4 PPL, whereas FMLM-re plateaus near 99.3. Throughout, GSF maintains stable token entropy (≈ 4.0), showing that the learned conditional trajectories follow the prompt context without degenerate repetition. Table 19 shows examples of samples for various prefixes.

Abstractive summarization on XSum. We next evaluate task-specific sequence-to-sequence generation on the BBC XSum abstractive summarization benchmark (Narayan et al., 2018), following the protocol of Hu et al. (2026a). Given a multi-sentence source document, models are conditioned to generate a concise single-sentence summary of up to 64 tokens. In Table 6, we report ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004) across sampling steps. At NFE $= 1$, GSF achieves 29.5 / 9.1 / 23.4 ROUGE-1/2/L, outperforming the single-step continuous flow baseline ELF (23.0 / 5.5 / 18.9). Remarkably, a single forward pass of GSF (NFE $= 1$) already surpasses multi-step diffusion models requiring dozens to thousands of iterations, such as E2D2 (28.4 / 8.3 / 22.0 at ≈ 64 steps) and SeqDiffu (19.3 / 1.7 / 14.1 at 2000 steps). With only two sampling steps (NFE $= 2$), GSF reaches 31.2 / 9.6 / 25.0, substantially narrowing the gap to its 64-step autoregressive teacher (34.1 / 12.1 / 26.2), particularly in ROUGE-1 and ROUGE-L. Table 20 shows examples of sentences summarized by GSF.

Table 5: Few-step conditional performance among distillation models on OpenWebText.

Method	NFE	PPL (↓)	Ent.
SDTT (MDLM) (Deschenaux and Gulcehre, 2025)	1	374.7	4.09
Di4C (MDLM) (Hayakawa et al., 2025)	1	376.8	4.10
DCD (Duo) (Sahoo et al., 2025)	1	703.1	4.18
Di4C (Duo) (Sahoo et al., 2025)	1	451.2	4.08
FMLM (FLM) (Lee et al., 2026)	1	141.6	4.14
FMLM (FLM)-re	1	115.0	3.89
GSF (AR)	1	**97.9**	4.00
FMLM (FLM)-re	2	101.4	3.98
GSF (AR)	2	**77.5**	4.05
FMLM (FLM)-re	4	99.3	4.04
GSF (AR)	4	**71.4**	4.06

Table 6: Few-step conditional generation on XSUM summarization (ROUGE: (↑)).

Method	NFE	ROUGE-1	ROUGE-2	ROUGE-L
AR (Hu et al., 2026a)	64	30.5	10.2	24.4
AR-re (our teacher)	64	34.1	12.1	26.2
ELF (Hu et al., 2026a)	64	36.0	12.2	27.8
E2D2 (Arriola et al., 2025)	≈ 64	28.4	8.3	22.0
SeqDiffu (Yuan et al., 2024)	2000	19.3	1.7	14.1
ELF (Hu et al., 2026a)	1	23.0	5.5	18.9
FMLM-re	1	16.7	1.2	13.1
GSF (AR)	1	**29.5**	**9.1**	**23.4**
ELF (Hu et al., 2026a)	2	27.6	7.7	22.5
FMLM-re	2	16.7	1.2	13.4
GSF (AR)	2	**31.2**	**9.6**	**25.0**

Table 7: Ablation of prior and noise schedules for one-step generation on OpenWebText.

Prior Distribution	Schedule	GM (↓)	PPL (↓)	Ent.
Standard Gumbel	Linear	1.76	133.5	4.87
Normalized Gumbel	Linear	1.24	249.0	5.21
Gaussian	Linear	1.05	150.7	5.04
Gaussian	DER	1.26	113.8	4.84
Gaussian	Coupled DER	0.94	114.9	5.07

Table 8: Ablation of training configurations for one-step generation on OpenWebText.

Default	Change	GM (↓)	PPL (↓)	Ent.
Bidirectional attn.	Causal attn.	1.62	211.0	5.05
AR init.	Random init.	0.81	145.4	5.14
Random off-diag.	Coupled off-diag.	0.89	108.4	5.01
One-hot pred.	Logit pred.	0.99	125.7	5.06
No change		0.94	114.9	5.07

4.3 Ablation Study

Prior distribution and noise scheduling. We investigate the impact of prior choice and time reparameterization for single-step generation on OpenWebText (Table 7). Using the raw Standard Gumbel prior yields poor gradient moment score (GM $= 1.76$, PPL $= 133.5$), which we hypothesize is due to the strong right skew of extreme-value distributions (Figure 4a). Transforming the prior to a standard Gaussian via Proposition 2 substantially improves GM (GM $= 1.05$ under linear scheduling). Next, we compare noise schedules: while standard DER parameterization achieves 113.8 PPL, it incurs a higher GM score (1.26) and lower token entropy (4.84). Our Coupled DER schedule (Figure 4b) obtains the best overall balance: GM $= 0.94$, PPL $= 114.9$, and Ent $= 5.07$.

Training and architectural configurations. Table 8 examines the impact of student architecture and training configurations for 1-step generation on OpenWebText: (i) *Attention masking*: Replacing bidirectional self-attention with causal masking drastically degrades performance (PPL $= 211.0$, GM $= 1.62$). While the causal dependencies are enough to model straight trajectories defined by the autoregressive teacher, parallel non-autoregressive generation benefits from bidirectional context: the teacher's Gumbel coupling provides sequence-level target supervision, but bidirectional self-attention enables the student to coordinate across all tokens simultaneously. (ii) *Teacher weight initialization*: Training a randomly initialized student increases perplexity from 114.9 to 145.4, showing that transferring pretrained AR teacher representations provides an effective initialization for flow map distillation. (iii) *Semigroup loss coupling*: While GSF by default optimizes the off-diagonal semigroup loss using random intermediate pairs (PPL $= 114.9$), restricting the off-diagonal transitions strictly along teacher-coupled paths (*Coupled off-diag.*) further lowers perplexity to 108.4 and GM to 0.89, with a slight decrease in token entropy (5.01 vs 5.07). (iv) *Denoiser output parameterization*: Predicting logits instead of one-hot targets is slightly worse on both headline metrics (GM $= 0.99$, PPL $= 125.7$), supporting the one-hot formulation used throughout.

5 Conclusion

In this work, we introduced Gumbel Straight Flow (GSF), demonstrating that autoregressive models naturally provide optimal, straight trajectories between continuous noise and discrete data tokens. By supervising flow maps along these non-intersecting paths, GSF improves few-step parallel generation across unconditional and conditional benchmarks, offering a promising framework for distilling autoregressive models into fast parallel samplers. In Appendix D we discuss several limitations.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for generating synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI to polish the writing. We have reviewed all AI-assisted work by carefully human-reviewing the code used to run the experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

M. Agarwal, S. Shah, C. Lee, J. Yoo, J. Huang, S. Hong, A. Raghunathan, J. Kim, and N. M. Boffi. Posterior refinement: Fast language generation via any-order flow maps. *arXiv preprint arXiv:2606.24773*, 2026.

M. Arriola, Y. Schiff, H. Phung, A. Gokaslan, and V. Kuleshov. Encoder-decoder diffusion language models for efficient training and inference. In *Advances in Neural Information Processing Systems*, 2025.

J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pages 17981–17993, 2021.

J. Backhoff, M. Beiglbock, Y. Lin, and A. Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

N. Boffi, M. Albergo, and E. Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems*, volume 38, pages 33346–33382, 2025.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

J. Deschenaux and C. Gulcehre. Beyond autoregression: Fast LLMs via self-distillation through time. In *International Conference on Learning Representations*, volume 2025, pages 5007–5045, 2025.

J. Deschenaux and C. Gulcehre. Language modeling with hyperspherical flows. *arXiv preprint arXiv:2605.11125*, 2026.

S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

F. Draxler, J. Will, F. Sofian, T. Karaletsos, S. Singh, and S. Mandt. Parallel token prediction for language models. In *International Conference on Learning Representations*, volume 2026, pages 5624–5652, 2026.

F. Eijkelboom, G. Bartosh, C. A. Naesseth, M. Welling, and J.-W. van de Meent. Variational flow matching for graph generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 11735–11764, 2024.

R. Eldan and Y. Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

A. Franca and A. Tong. Hacking generative perplexity: Why unconditional text evaluation needs distributional metrics. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026.

A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

Z. Geng, M. Deng, X. Bai, J. Z. Kolter, and K. He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

A. Gokaslan and V. Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

E. J. Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: a Series of Lectures*, volume 33. US Government Printing Office, 1954.

S. Hayakawa, Y. Takida, M. Imaizumi, H. Wakaki, and Y. Mitsufuji. Distillation of discrete diffusion through dimensional correlations. In *International Conference on Machine Learning*, volume 267, pages 22259–22297, 2025.

E. Hoogeboom, D. Ruhe, J. Heek, T. Mensink, and T. Salimans. Beyond single tokens: Distilling discrete diffusion models via discrete MMD. *arXiv preprint arXiv:2603.20155*, 2026.

K. Hu, L. Qiu, Y. Lu, H. Zhao, T. Li, Y. Kim, J. Andreas, and K. He. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026a.

L. Hu, S. Kou, Y. Fu, S. Rajbhandari, T. Rosing, Y. He, Z. Deng, and H. Zhang. Fast and accurate causal parallel decoding using jacobi forcing. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=ORy2hyPd4G.

D. Israel, G. Van den Broeck, and A. Grover. Accelerating diffusion LLMs via adaptive parallel decoding. In *Advances in Neural Information Processing Systems*, volume 38, pages 52870–52888, 2025.

D. Kim, C.-H. Lai, W. Liao, N. Murata, Y. Takida, T. Uesaka, Y. He, Y. Mitsufuji, and S. Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pages 44493–44525, 2024.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

L. Kong, M. Tao, Y. Liu, B. Wang, J. Fu, C.-C. Wang, and H. Liu. AlignFlow: Improving flow-based generative models with semi-discrete optimal transport. In *International Conference on Learning Representations*, 2026.

S. Kou, L. Hu, Z. He, Z. Deng, and H. Zhang. CLLMs: Consistency large language models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25426–25440. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kou24a.html.

C. Lee, J. Yoo, M. Agarwal, S. Shah, J. Huang, A. Raghunathan, S. Hong, N. M. Boffi, and J. Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

D. Li, N. Gushchin, D. Abulkhanov, E. Moulines, I. Oseledets, M. Panov, and A. Korotin. IDLM: Inverse-distilled diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bS8qU2wJxu.

X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 4328–4343, 2022.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

X. Liu, X. Zhang, J. Ma, J. Peng, and Q. Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024.

R. J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, 2018.

W. Peebles and S. Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182. IEEE, 2023.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data sciences. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

P. Potaptchik, J. Yim, A. Saravanan, P. Holderrieth, E. Vanden-Eijnden, and M. S. Albergo. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026.

P. Pynadath, J. Shi, and R. Zhang. CANDI: Hybrid discrete-continuous diffusion models. In *International Conference on Machine Learning*, 2026.

D. Roos, O. Davis, F. Eijkelboom, M. Bronstein, M. Welling, I. I. Ceylan, L. Ambrogioni, and J.-W. van de Meent. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 130136–130184, 2024.

S. S. Sahoo, J. Deschenaux, A. Gokaslan, G. Wang, J. T. Chiu, and V. Kuleshov. The diffusion duality. In *International Conference on Machine Learning*, 2025.

T. Salimans, T. Mensink, J. Heek, and E. Hoogeboom. Multistep distillation of diffusion models via moment matching. *Advances in Neural Information Processing Systems*, 37:36046–36070, 2024.

N. Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

J. Shi, K. Han, Z. Wang, A. Doucet, and M. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pages 103131–103167, 2024.

Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

A. Tong, K. FATRAS, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623. IEEE, 2024.

J. Yoo, W. Kim, F. Eijkelboom, C. Lee, N. M. Boffi, S. Hong, and J. Kim. Self-conditioned flow map language models via fixed-point flows. *arXiv preprint arXiv:2607.00714*, 2026.

H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang. Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 22–39, 2024.

C. Zhang, X. Hu, B. Liu, and Q. Liu. Gumbel distillation for parallel text generation. In *International Conference on Learning Representations*, 2026.

H. Zheng, X. Liu, X. Kong, N. Jiang, Z. Hu, W. Luo, W. Deng, and G. Lin. Ultra-fast language generation via discrete diffusion divergence instruct. In *International Conference on Learning Representations*, volume 2026, pages 29535–29571, 2026.

Y. Zhu, X. Wang, S. Lathuilière, and V. Kalogeiton. Di[m]o: Distilling masked diffusion models into one-step generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18606–18618, October 2025.

A Proof and Derivations

A.1 Properties of Gumbel noise

Ranking Preserving. For any single-token Gumbel sample $\boldsymbol{\xi} \in \mathbb{R}^{|\mathcal{V}|}$ and its corresponding one-hot target $\mathbf{x} \in \mathbb{R}^{|\mathcal{V}|}$, the selected index preserves the Gumbel-max ranking:

$$\arg\max_{j\in\{1,\dots,|\mathcal{V}|\}}(\ell_j + \xi_j) = \arg\max_{j\in\{1,\dots,|\mathcal{V}|\}}(x_j).$$

Gumbel Residual Ordering. Consider two coupled pairs $(\boldsymbol{\xi}^1, \mathbf{x}^1)$ and $(\boldsymbol{\xi}^2, \mathbf{x}^2)$ with distinct targets $\mathbf{x}^1 \neq \mathbf{x}^2$. Let k^1 and k^2 ($k^1 \neq k^2$) be their respective active indices:

$$k^1 := \arg\max_{j\in\{1,\dots,|\mathcal{V}|\}}(x_j^1), \qquad k^2 := \arg\max_{j\in\{1,\dots,|\mathcal{V}|\}}(x_j^2).$$

By the ranking preserving property, we have:

$$\xi_{k^1}^1 + \ell_{k^1} > \xi_{k^2}^1 + \ell_{k^2} \quad \text{and} \quad \xi_{k^1}^2 + \ell_{k^1} < \xi_{k^2}^2 + \ell_{k^2}.$$

Rearranging both inequalities yields $\xi_{k^1}^1 - \xi_{k^2}^1 > \ell_{k^2} - \ell_{k^1} > \xi_{k^1}^2 - \xi_{k^2}^2$, which implies:

$$\xi_{k^1}^1 - \xi_{k^1}^2 > \xi_{k^2}^1 - \xi_{k^2}^2.$$

A.2 Proof of Theorem 1

Theorem 1. ***(Straightness)*** *Assume sufficient model capacity. Let the optimal velocity be $\mathbf{v}_{\theta^*}$ where $\theta^* := \arg\min_\theta \mathcal{L}_{\text{GSF-FM}}(\theta, p_{AR})$, and let $\{\mathbf{x}_t^{\theta^*}\}$ denote the corresponding path. Then the path is straight, i.e.,*

$$S(\{\mathbf{x}_t^{\theta^*}\}) = 0.$$

Proof of Theorem 1. We first prove the single-token case ($L = 1$), where each state is in $\mathbb{R}^{|\mathcal{V}|}$, the extension to sequences being straightforward. The extension to arbitrary sequence length is given in Proposition 3. Let $\mathbf{v}_{\theta^*}(\mathbf{x}_t, t) = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_t]$ be the optimal velocity for the linear interpolation $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. The straightness measure

$$S(\{\mathbf{x}_t^{\theta^*}\}) := \int_0^1 \mathbb{E}\left[\|(\mathbf{x}_1 - \mathbf{x}_0) - \mathbf{v}_{\theta^*}(\mathbf{x}_t, t)\|_2^2\right] dt$$

equals zero if the posterior $p(\mathbf{x}_0, \mathbf{x}_1 \mid \mathbf{x}_t)$ is a Dirac-delta distribution, i.e., no two distinct coupled pairs intersect at any $\mathbf{x}_t$ for $t \in (0, 1)$.

Suppose for contradiction that two distinct pairs $(\mathbf{x}_0^1, \mathbf{x}_1^1) = (\boldsymbol{\xi}^1, \mathbf{x}^1)$ and $(\mathbf{x}_0^2, \mathbf{x}_1^2) = (\boldsymbol{\xi}^2, \mathbf{x}^2)$ intersect at $\mathbf{x}_t$:

$$\mathbf{x}_t = (1-t)\boldsymbol{\xi}^1 + t\mathbf{x}^1 = (1-t)\boldsymbol{\xi}^2 + t\mathbf{x}^2 \implies \boldsymbol{\xi}^1 - \boldsymbol{\xi}^2 = \frac{t}{1-t}(\mathbf{x}^2 - \mathbf{x}^1). \tag{9}$$

If $\mathbf{x}^1 = \mathbf{x}^2$, then $\boldsymbol{\xi}^1 = \boldsymbol{\xi}^2$, contradicting that the pairs are distinct; hence $\mathbf{x}^1 \neq \mathbf{x}^2$. Let $k^1 := \arg\max_j x_j^1$ and $k^2 := \arg\max_j x_j^2$ ($k^1 \neq k^2$) be the active indices of the one-hot targets $\mathbf{x}^1$ and $\mathbf{x}^2$. Since $x_{k^1}^2 - x_{k^1}^1 = -1$ and $x_{k^2}^2 - x_{k^2}^1 = +1$, evaluating Eq. (9) at indices k^1 and k^2 for $t \in (0, 1)$ yields:

$$\xi_{k^1}^1 - \xi_{k^1}^2 = -\frac{t}{1-t} < 0 < +\frac{t}{1-t} = \xi_{k^2}^1 - \xi_{k^2}^2 \implies \xi_{k^1}^1 - \xi_{k^1}^2 < \xi_{k^2}^1 - \xi_{k^2}^2.$$

This contradicts the **Gumbel Residual Ordering** property ($\xi_{k^1}^1 - \xi_{k^1}^2 > \xi_{k^2}^1 - \xi_{k^2}^2$). Thus, no two trajectories intersect, $p(\mathbf{x}_0, \mathbf{x}_1 \mid \mathbf{x}_t)$ is a Dirac-delta distribution, and $S(\{\mathbf{x}_t^{\theta^*}\}) = 0$. □

For completeness, we provide the proof for arbitrary sequence length L.

Proposition 3. *The single-token non-intersection result above extends to arbitrary sequence length L. In particular, for every fixed $t \in (0, 1)$, the map*

$$\boldsymbol{\xi}_{1:L} \mapsto (1-t)\boldsymbol{\xi}_{1:L} + t\ \text{onehot}(T_{\text{AR}}(\boldsymbol{\xi}_{1:L})) \tag{10}$$

is injective.

Proof. The result can be derived by induction over the sequence positions. At position $i = 1$, both samples have the same logit vector $\boldsymbol{\ell}_1$, so the single-token result established above implies equality of both the first Gumbel block and the first generated token. Assume that that the two generated prefixes agree up to position $i-1$, i.e., $\mathbf{y}^1_{<i} = \mathbf{y}^2_{<i}$. Then the AR model produces obviously the same conditional logits, $\boldsymbol{\ell}_i(\mathbf{y}^1_{<i}) = \boldsymbol{\ell}_i(\mathbf{y}^2_{<i})$. Equality of the ith blocks of the interpolated states, together again with the single-token result above, therefore implies $\boldsymbol{\xi}^1_i = \boldsymbol{\xi}^2_i$ and $y^1_i = y^2_i$. □

A.3 Proof of Proposition 2

Proposition 2. ***(Prior transform).*** *Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly increasing and invertible function, and let $\mathbf{f}$ denote its element-wise extension. Under sufficient model capacity, let $\mathbf{v}_{\theta^*}(\mathbf{x}_t, t) := \arg\min_{\mathbf{v}_\theta} \mathcal{L}_{\text{GSF-FM}}(\theta, p_{AR}; \mathbf{f})$ denote the optimal velocity field, and let $\{\mathbf{x}_t^{\theta^*}\}$ denote the corresponding induced trajectory. Then, the path is still straight, i.e., $S(\{\mathbf{x}_t^{\theta^*}\}) = 0$.*

Proof of Proposition 2. Consider the single-token case ($L = 1$) (it easily extends to the multi-token case too). Let $u_j = f(\xi_j)$ for all $j \in \{1, \dots, |\mathcal{V}|\}$, where $f : \mathbb{R} \to \mathbb{R}$ is strictly increasing and invertible. Suppose for contradiction that two distinct coupled pairs $(\mathbf{x}^1_0, \mathbf{x}^1_1) = (\mathbf{u}^1, \mathbf{x}^1)$ and $(\mathbf{x}^2_0, \mathbf{x}^2_1) = (\mathbf{u}^2, \mathbf{x}^2)$ with $\mathbf{x}^1 \neq \mathbf{x}^2$ intersect at $\mathbf{x}_t$ for $t \in (0, 1)$:

$$\mathbf{x}_t = (1-t)\mathbf{u}^1 + t\mathbf{x}^1 = (1-t)\mathbf{u}^2 + t\mathbf{x}^2 \implies \mathbf{u}^1 - \mathbf{u}^2 = \frac{t}{1-t}(\mathbf{x}^2 - \mathbf{x}^1).$$

Let $k^1 := \arg\max_j x^1_j$ and $k^2 := \arg\max_j x^2_j$ ($k^1 \neq k^2$) be the active indices of the one-hot targets $\mathbf{x}^1$ and $\mathbf{x}^2$. Evaluating the above equality at coordinates k^1 and k^2 yields:

$$u^1_{k^1} - u^2_{k^1} = -\frac{t}{1-t} < 0 \implies u^1_{k^1} < u^2_{k^1}, \qquad u^1_{k^2} - u^2_{k^2} = +\frac{t}{1-t} > 0 \implies u^1_{k^2} > u^2_{k^2}.$$

Since f^{-1} is also strictly increasing, applying f^{-1} to both inequalities gives:

$$\xi^1_{k^1} < \xi^2_{k^1} \quad \text{and} \quad \xi^1_{k^2} > \xi^2_{k^2} \implies \xi^1_{k^1} - \xi^2_{k^1} < 0 < \xi^1_{k^2} - \xi^2_{k^2}.$$

This contradicts the **Gumbel Residual Ordering** property ($\xi^1_{k^1} - \xi^2_{k^1} > \xi^1_{k^2} - \xi^2_{k^2}$). Thus, no two trajectories intersect under the transformed prior, $p(\mathbf{x}_0, \mathbf{x}_1 \mid \mathbf{x}_t)$ remains a Dirac-delta distribution, and $S(\{\mathbf{x}_t^{\theta^*}\}) = 0$. □

A.4 An Optimal Transport Interpretation of Gumbel Straight Flow

We can understand the straightness and non-intersection properties of GSF by reformulating it as a (semi-discrete) optimal transport (OT) problem (Peyré and Cuturi, 2019, chapter 5).

In a standard flow matching context where the noise distribution $p_0(\mathbf{x}_0) = p_{\text{noise}}(\mathbf{x}_0)$ and data distribution $p_1(\mathbf{x}_1) = p_{\text{data}}(\mathbf{x}_1)$ are both distributions with densities in $\mathbb{R}^d$, we typically couple/pair the noise and samples independently; this leads to a generative ODE with curved paths. To obtain straight paths, one can couple the noise and data distributions using the joint distribution $p(\mathbf{x}_0, \mathbf{x}_1)$ with marginals $p_{\text{noise}}(\mathbf{x}_0)$ and $p_{\text{data}}(\mathbf{x}_1)$ minimizing $\mathbb{E}[||\mathbf{x}_0 - \mathbf{x}_1||^2]$ (McCann, 1997). This is an optimal transport problem whose solution is of the form $p_{\text{OT}}(\mathbf{x}_0, \mathbf{x}_1) = p_{\text{noise}}(\mathbf{x}_0)\delta_{T_{\text{OT}}(\mathbf{x}_0)}(\mathbf{x}_1)$ where $T_{\text{OT}}(\cdot)$ is the optimal transport map. Computing this joint distribution is typically intractable and requires approximations; see e.g. (Liu et al., 2023; Pooladian et al., 2023; Tong et al., 2024).

We are here in a specific scenario where the data distribution is an atomic distribution. Again one can obtain straight transport paths if we could identify the joint distribution minimizing the expected quadratic cost. This is a so-called semi-discrete optimal transport problem (Peyré and Cuturi, 2019). The semi-discrete nature of the problem allows us to characterize precisely the structure of the transport.

Single Token. For a single token, we have that $p_0(\mathbf{x}_0)$ is given by the product of $|\mathcal{V}|$ Gumbel random variables while $p_1(\mathbf{x}_1) = \sum_{i=1}^{|\mathcal{V}|} p_{1,i}\delta_{\mathbf{e}_i}(\mathbf{x}_1)$ where $\mathbf{e}_i$ is the one-hot vector associated to the ith element of the vocabulary. We rewrite the optimal transport problem for a quadratic cost as $p_{\text{OT}}(\mathbf{x}_0, \mathbf{x}_1) = p_{\text{noise}}(\mathbf{x}_0)\delta_{T_{\text{OT}}(\mathbf{x}_0)}(\mathbf{x}_1)$ as

$$T_{\text{OT}} = \arg\inf_{T : T\#p_0 = p_1} \int \|\mathbf{x}_0 - T(\mathbf{x}_0)\|^2 p_0(\mathbf{x}_0)\, \mathrm{d}\mathbf{x}_0. \tag{11}$$

The transport map of a semi-discrete OT can be obtained from a $|\mathcal{V}|$-dimensional vector $\mathbf{a} := (a_i)_{i\in\mathcal{V}}$ referred to as dual weights, which we will define later; see (Peyré and Cuturi, 2019, chapter 5) for a detailed introduction. We first map noise to data index from the mapping

$$\kappa_{\mathbf{a}}(\mathbf{x}_0) = \arg\min_{i\in\{1,\dots,|\mathcal{V}|\}} \left\{\tfrac{1}{2}\|\mathbf{x}_0 - \mathbf{e}_i\|^2 - a_i\right\}, \tag{12}$$

and then define the corresponding transport map through $T_{\mathbf{a}}(\mathbf{x}_0) = \mathbf{e}_{\kappa_{\mathbf{a}}(\mathbf{x}_0)}$.

Given $\|\mathbf{x}_0 - \mathbf{e}_i\|^2 = \|\mathbf{x}_0\|^2 - 2\mathbf{x}_0^\top \mathbf{e}_i + \|\mathbf{e}_i\|^2 = \|\mathbf{x}_0\|^2 - 2x_{0,i} + 1$, we thus have $\kappa_{\mathbf{a}}(\mathbf{x}_0) = \arg\max_{i\in\{1,\dots,|\mathcal{V}|\}} \{x_{0,i} + a_i\}$ and the transport map simplifies to

$$T_{\mathbf{a}}(\mathbf{x}_0) = e_{\arg\max_i\{x_{0,i}+a_i\}}. \tag{13}$$

By construction, this transport map thus partitions the noise space into $|\mathcal{V}|$ so-called Laguerre cells given by

$$C_i(\mathbf{a}) = \left\{\mathbf{x}_0 \in \mathbb{R}^{|\mathcal{V}|} : x_{0,i} + a_i \geq x_{0,m} + a_m,\ \forall m\right\}. \tag{14}$$

Finally the optimal transport map is obtained by selecting the dual weights $\mathbf{a}$[3] such that

$$\int \mathbb{I}(\mathbf{x}_0 \in C_i(\mathbf{a}))p_0(\mathbf{x}_0)\,\mathrm{d}\mathbf{x}_0 = p_{1,i}. \tag{15}$$

In general semi-discrete optimal transport (Peyré and Cuturi, 2019), finding these weights requires expensive numerical optimization during training; see, e.g., Kong et al. (2026). In our scenario, this is not necessary as established in the following proposition.

Proposition 4. *Let $\mathbf{x}_0 = \boldsymbol{\xi}$ for $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{G}(0,1)$ where $i \in \mathcal{V}$. For the target probabilities, $p_{1,i} = \mathrm{softmax}(\boldsymbol{\ell})_i$, the dual weights $\mathbf{a}$ satisfying (15) are given by*

$$a_i = \ell_i + c, \qquad i \in \mathcal{V}, \tag{16}$$

for an arbitrary constant $c \in \mathbb{R}$. Consequently, the optimal transport map is defined by

$$T_{\boldsymbol{\ell}}(\boldsymbol{\xi}) = \mathbf{e}_{\arg\max_{i\in\mathcal{V}}\{\xi_i+\ell_i\}}; \tag{17}$$

i.e. the Gumbel–max map is the optimal transport map from p_0 to p_1.

Proof. The proof is direct. By equation 14 and the Gumbel–max identity,

$$\mathbb{P}(\boldsymbol{\xi} \in C_i(\mathbf{a})) = \mathbb{P}(i = \arg\max_m\{\xi_m + a_m\}) = \mathrm{softmax}(\mathbf{a})_i. \tag{18}$$

Hence equation 15 requires $\mathrm{softmax}(\mathbf{a}) = \mathrm{softmax}(\boldsymbol{\ell})$, which is equivalent to $\mathbf{a} = \boldsymbol{\ell} + c\mathbf{1}$. Substituting into equation 13 gives the Gumbel–max transport map of equation 17. □

In other words, the teacher logits $\boldsymbol{\ell}$ are already the exact semi-discrete OT dual weights so no optimization or OT solver is needed. This result is already implicit in (Galichon, 2016, Chapter 5).

The OT viewpoint gives a simple explanation of the non-intersection property. Since $T_{\boldsymbol{\ell}}$ is the quadratic optimal transport map, it is monotone (Brenier, 1991), i.e.

$$\langle T_{\boldsymbol{\ell}}(\boldsymbol{\xi}) - T_{\boldsymbol{\ell}}(\boldsymbol{\xi}'), \boldsymbol{\xi} - \boldsymbol{\xi}'\rangle \geq 0. \tag{19}$$

So for the interpolated states

$$\mathbf{x}_t = (1-t)\boldsymbol{\xi} + tT_{\boldsymbol{\ell}}(\boldsymbol{\xi}), \qquad \mathbf{x}'_t = (1-t)\boldsymbol{\xi}' + tT_{\boldsymbol{\ell}}(\boldsymbol{\xi}'), \tag{20}$$

we therefore have

$$\langle \mathbf{x}_t - \mathbf{x}'_t, \boldsymbol{\xi} - \boldsymbol{\xi}'\rangle \geq (1-t)\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|^2. \tag{21}$$

For $t < 1$, the right-hand side is strictly positive whenever $\boldsymbol{\xi} \neq \boldsymbol{\xi}'$. Hence two different trajectories cannot meet at the same time. This recovers the non-intersection property proved directly in Appendix A.2.

[3] These weights are determined up to an additive constant

Gaussianization. In practice, we transform the Gumbel noise component-wise using the strictly increasing map

$$\mathbf{x}_0 = f(\boldsymbol{\xi}) = \Phi^{-1}\big(\exp(-e^{-\boldsymbol{\xi}})\big)\,, \tag{22}$$

so that $\mathbf{x}_0 \sim \mathcal{N}(0, I)$ (but leave the Gumbel noise-token coupling unchanged). Since f is strictly increasing and invertible, the non-intersection property of the linear paths is preserved, as established in Appendix A.3. The OT interpretation is slightly modified: in this Gaussian representation the same coupling is optimal for the cost

$$c(\mathbf{x}_0, \mathbf{x}_1) = \left\| f^{-1}(\mathbf{x}_0) - \mathbf{x}_1 \right\|^2 = \left\| \boldsymbol{\xi} - \mathbf{x}_1 \right\|^2, \tag{23}$$

rather than the Euclidean quadratic cost $\|\mathbf{x}_0 - \mathbf{x}_1\|^2$.

Multiple Tokens. The single-token result can be extended directly to a token sequence generated by an AR model. At position i, conditional on the previously generated tokens $y_{<i}$, the logits $\boldsymbol{\ell}_i(\mathbf{y}_{<i})$ are fixed and

$$y_i = \arg\max_k \left\{ \xi_{i,k} + \ell_{i,k}(\mathbf{y}_{<i}) \right\}. \tag{24}$$

Hence, this is exactly the single-token semi-discrete OT problem above: the dual weights are the conditional teacher logits $\boldsymbol{\ell}_i(y_{<i})$ (up to an additive constant). The AR teacher therefore constructs the sequence coupling by applying these single-token transport maps successively.

However, the resulting sequence map is generally not the quadratic optimal transport map on $\mathbb{R}^{L\times|\mathcal{V}|}$, but the paths remain non-intersecting as established in Appendix A.2.

B Related Work

Accelerated parallel sampling for language models broadly spans three paradigms: (i) distilling and accelerating multi-step discrete diffusion models (Deschenaux and Gulcehre, 2025; Hayakawa et al., 2025; Zhu et al., 2025; Li et al., 2026; Hoogeboom et al., 2026; Zheng et al., 2026; Israel et al., 2025); (ii) continuous-space generative modeling, which embeds categorical tokens into continuous Euclidean or spherical geometries (Li et al., 2022; Dieleman et al., 2022; Deschenaux and Gulcehre, 2026; Yoo et al., 2026); (iii) parallelizing AR models (Kou et al., 2024; Hu et al., 2026b).

Diffusion or flow matching models can be integrated into few-step samplers via flow maps (Boffi et al., 2025). Extending diffusion and flow maps to discrete categorical representations and language models (Eijkelboom et al., 2024; Lee et al., 2026; Roos et al., 2026; Hu et al., 2026a) embeds tokens into continuous spaces. Nonetheless, these models universally rely on independent couplings $p(x_0, x_1) = p(x_0)p(x_1)$, causing straight paths connecting random pairs to cross in high dimensions and forcing the velocity field to curve. Rectified Flow (Liu et al., 2023) showed that straighter trajectories permit better generation with coarse numerical integration.

The key distinction between GSF and these methods is the use of teacher-guided noise-data pairs as pioneered by, e.g. (Draxler et al., 2026). Gumbel Distillation (Zhang et al., 2026) directly utilizes Gumbels sampled from an AR teacher's posterior distribution to supervise a student model. However, Gumbel Distillation binds this pairing to a discrete diffusion backbone, which requires iterative categorical transitions and multiple sampling steps. GSF sits at the intersection of continuous flow matching and AR distillation. Standard Rectified Flow (Liu et al., 2023) can iteratively apply computationally expensive Reflow steps to obtain increasingly straight paths. It uses this insight as a strong inductive bias for learning the noise-data pairing using existing flow map techniques.

C Experimental Details

Model architecture. For all OpenWebText experiments, we follow the architecture scale used in the baseline methods (Lee et al., 2026; Sahoo et al., 2025), which use a bidirectional Diffusion Transformer (Peebles and Xie, 2023) with 12 layers and a hidden size of 768. We use multi-query attention (Shazeer, 2019) with 3 query heads and a head dimension of 256, along with FiLM time conditioning (Perez et al., 2018), which makes the model slightly more efficient, as shown in Figure 8a. The AR teacher uses causal attention and does not use time embeddings or a separate unembedding layer. Both GSF-FM (Figure 5) and GSF (flow map) use an unembedding layer, with

Algorithm 1 GSF training with Gumbel posterior sampling

Require: Dataset $\mathcal{D}$, AR teacher p_{AR}, time reparameterization $\tau(\cdot)$, prior transform $\mathbf{f}$
Initialize: Flow map student δ^θ from p_{AR}
repeat
Diagonal (underlying path supervision from p_{AR}):
Sample $\mathbf{x}_1 \sim \mathcal{D}$
Sample $\boldsymbol{\xi} \sim p_{\mathrm{AR}}(\boldsymbol{\xi} \mid \mathbf{x}_1)$ via Gumbel posterior sampling (Zhang et al., 2026); $\mathbf{x}_0 \leftarrow \mathbf{f}(\boldsymbol{\xi})$
$\tau(t) \sim \mathcal{U}[0,1]; \quad \mathbf{x}_t \leftarrow (1-t)\mathbf{x}_0 + t\mathbf{x}_1$
$\mathcal{L}_{\mathrm{diag}}(\theta) \leftarrow \mathrm{CE}\big(\mathbf{x}_1 \,\big\|\, \delta^\theta_{t,t}(\mathbf{x}_t)\big)$
Off-diagonal (semi-group self-consistency):
Sample $\mathbf{x}_1 \sim \mathcal{D}$
Sample $\boldsymbol{\xi} \sim \mathcal{G}(0,1)$; $\mathbf{x}_0 \leftarrow \mathbf{f}(\boldsymbol{\xi})$
$h \sim \mathcal{U}[0,1]; \quad \tau(s) \sim \mathcal{U}[0, 1-h]; \quad \mathbf{x}_s \leftarrow (1-s)\mathbf{x}_0 + s\mathbf{x}_1$
$\tau(t) \leftarrow \tau(s) + h; \quad \tau(u) \leftarrow \dfrac{\tau(s)+\tau(t)}{2}; \quad \gamma \leftarrow \dfrac{(1-t)(u-s)}{(1-u)(t-s)}$
$\hat{X}^\theta_{s,u} \leftarrow \dfrac{1-u}{1-s}\mathbf{x}_s + \dfrac{u-s}{1-s}\delta^\theta_{s,u}(\mathbf{x}_s)$
$\overline{\delta}^\theta_{\mathrm{tar}} \leftarrow \mathrm{sg}\Big(\gamma\delta^\theta_{s,u}(\mathbf{x}_s) + (1-\gamma)\delta^\theta_{u,t}(\hat{X}^\theta_{s,u})\Big)$
$\mathcal{L}_{\mathrm{off-diag.}}(\theta) \leftarrow \mathrm{CE}\Big(\overline{\delta}^\theta_{\mathrm{tar}} \,\Big\|\, \delta^\theta_{s,t}(\mathbf{x}_s)\Big)$
Update θ with $\nabla_\theta(\mathcal{L}_{\mathrm{diag}}(\theta) + \mathcal{L}_{\mathrm{off-diag.}}(\theta))$
until converged

Algorithm 2 GSF sampling

Require: δ^θ, time reparameterization $\tau(\cdot)$, sampling steps N
Sample $\boldsymbol{\xi} \sim \mathcal{G}(0,1)$; $\mathbf{x}_0 \leftarrow \mathbf{f}(\boldsymbol{\xi})$
$t_i \leftarrow \tau(i/N)$ for $i = 0, \ldots, N$
for $i = 0$ to $N-1$ **do**
$\mathbf{x}_{t_{i+1}} \leftarrow \dfrac{1-t_{i+1}}{1-t_i}\mathbf{x}_{t_i} + \dfrac{t_{i+1}-t_i}{1-t_i}\delta^\theta_{t_i,1}(\mathbf{x}_{t_i})$
end for
Return $\mathbf{x}_1$

one and two time-conditioning layers, respectively. For the medium-sized model in Figure 8c, both the AR teacher and GSF student use 24 layers, a hidden size of 1024, 4 query heads, and a head dimension of 256. We use the `GPT-2` tokenizer with $|\mathcal{V}| = 50,257$ and $L = 1024$, together with the corresponding embedding layer. For the LM1B dataset in Table 3, the only difference is that we use the `bert-base-uncased` tokenizer with $|\mathcal{V}| = 30,552$ and $L = 128$, which results in a different number of embedding-layer parameters. For the TinyStories dataset in Table 4, we use the `GPT-2` tokenizer with $|\mathcal{V}| = 50,257$ and $L = 128$. We also use 8 layers, a hidden size of 512, 2 query heads, and a head dimension of 256 to match the baseline setting in Agarwal et al. (2026). For the conditional OpenWebText experiments in Table 5, we use the same setting as for unconditional OpenWebText, except for the sequence length (50 prefix tokens and 50 continuation tokens), following Lee et al. (2026). For the conditional XSUM experiments in Table 6, we use the same setting as unconditional OpenWebText with 1024 prefix tokens and 64 continuation tokens. Here, we apply time conditioning in a token-wise manner, following Hu et al. (2026a). That is, we always set $t = 1$ for the prefix tokens.

Training details. For all unconditional settings, we use a batch size of 512 and the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 3×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$, and a weight decay of 0.03 for the AR teacher, GSF-FM, and GSF (flow map). Following prior work, we train for up to 1M steps, although convergence is fast and stable, as shown in Figure 8b. By default, GSF is initialized from the AR teacher. The time-conditioning layers are zero-initialized, while the unembedding layer is initialized from the AR teacher's embedding layer. For conditional OpenWebText in Table 5, we use the same teacher as in unconditional generation and initialize the student from the student model trained on unconditional OpenWebText, which enables faster train-

ing, following Lee et al. (2026). For conditional XSUM generation in Table 6, we fine-tune the OWT-trained AR teacher for 30k steps to obtain the teacher. We initialize the student from the GSF trained on unconditional OpenWebText.

By default, we use time reparametrization based on the coupled decoding error rate described in Figure 4b. Since this coupling is teacher-dependent, we empirically construct the schedule using eight noise-data sentence pairs from the AR teacher. We observe almost no variation in this trend across teacher sizes and datasets. On the other hand, Figures 5a and 5b analyze the path itself using time integration. Different time reparametrizations may unintentionally bias these metrics, so we use a linear time schedule for both the baseline and GSF-FM in the path analysis. For efficient training, we do not collect noise-data pairs through iterative teacher sampling. Instead, the pairs can be constructed more efficiently using posterior Gumbel sampling (Zhang et al., 2026) (see Algorithm 1 for the full training algorithm). This method samples $\mathbf{x}_1 \sim p_{\text{data}}$ and $\mathbf{x}_0 := \boldsymbol{\xi} \sim p_{\text{AR}}(\boldsymbol{\xi}|\mathbf{x}_1)$, requiring only a single teacher-forced forward pass for each data sample (see further discussion in Appendix E).

Sampling details. GSF-FM in Figure 5c uses an Euler solver for sampling. We set the decoding temperature of the softmax layer to 1.0. For the straightness measure in Figure 5b, we use a temperature of 0.3, which gives the best straightness for both FLM and GSF-FM, as shown in Figure 9. For GSF (flow map), we use $\tilde{\gamma}$ sampling (Kim et al., 2024) with $\tilde{\gamma} = 1$, but use deterministic variants rather than injecting random noise (see Algorithm 2 for the full sampling algorithm). Suppose we jump $\mathbf{x}_s$ from time s to t, then we perform $\hat{\mathbf{x}}_t = \frac{1-t}{1-s}\mathbf{x}_s + \frac{t}{1-s}\delta^{\theta}_{s,1}(\mathbf{x}_s)$. By default, we do not adjust the prior or softmax temperature for GSF. In Figure 6, we adjust only the softmax temperature over [0.0, 1.3]. In Figure 8c, for one-step sampling, we adjust the prior temperature over [0.6, 1.15]. For two- and four-step sampling, we fix the prior temperature at 0.85 for the small model and 0.6 for the medium model, while varying the softmax temperature over [0.0, 1.3].

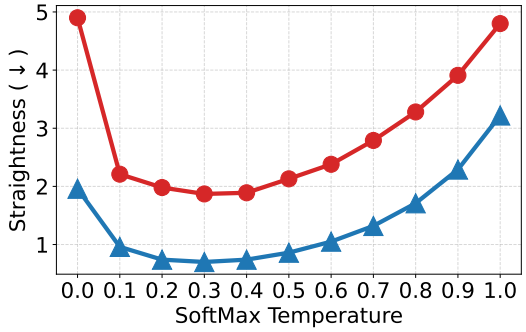


Figure 9: Straightness (CE scale) across decoding softmax temperatures for FLM and GSF-FM.

Evaluation details. For Gradient Moment (Hoogeboom et al., 2026), generative perplexity, and unigram entropy in unconditional generation, we use 1,024 samples randomly generated by the student model and GPT-2 Large as the judge, following prior work (Hoogeboom et al., 2026; Lee et al., 2026). For Figure 7, we use the prompt in Appendix G to evaluate 256 samples from each model. We ask Gemini 3.6-Flash to assign a win, tie, or loss, and compute the win rate by assigning equal credit for ties to both models. For conditional OpenWebText generation in Table 5, we randomly sample 256 prefixes and generate five continuations for each. The metrics are computed over the full sequence, including both the prefix and continuation (Lee et al., 2026). For XSUM summarization in Table 6, we compute ROUGE scores between generated samples and reference answers, following Hu et al. (2026a).

Baseline configurations. For FLM in Figure 5, we train it ourselves using the same architecture and training configuration as GSF-FM. For FMLM-re, we use its optimal decoding error rate schedule with an FLM teacher and the same architecture, learning rate, batch size, optimizer, and number of training iterations as GSF. For the other distillation baselines in Figure 1 and Tables 2, 3 and 5, we report PPL and entropy values from Lee et al. (2026) and measure the GM score ourselves using their official checkpoints. For the baselines in Table 4, we report values from Agarwal et al. (2026) and measure the eight-step performance of FMLM+ using its official checkpoint. For the

baselines in Table 6, we report results from ELF (Hu et al., 2026a) and measure the one- and two-step performance of ELF using its official checkpoint.

D Discussion & Limitations

Why use a flow map? If the student model were trained to perfection, it would only need to learn a direct mapping from the initial noise $\boldsymbol{\xi}$ to the data $\mathbf{x}_1$. Indeed, we find that directly learning this mapping yields higher token accuracy than a flow map parameterization. For example, in unconditional generation on OpenWebText, the direct one-step model achieves an accuracy of 9.92%, whereas our flow-map-based GSF achieves 6.68%. However, this accuracy remains too low to produce coherent text, resulting in poor generation quality with a GM score exceeding 30. This indicates that relying solely on deterministic pairing is insufficient for faithful generation. Therefore, motivated by the strong generative performance of flow maps, we instead focus on enhancing generative quality by discovering straight trajectories and utilizing them within the flow map framework.

Performance gap. Another clear limitation of our method is the remaining performance gap compared to the AR teacher. In the image domain, learning a flow map along a given trajectory often yields performance comparable to that of the teacher (Song et al., 2023; Geng et al., 2025), making such a pronounced gap unexpected. We suspect that the high dimensionality of $L \times |\mathcal{V}|$ is the primary cause. Furthermore, unlike pixel space, which is relatively robust to minor deviations, prediction errors in one-hot space can abruptly produce entirely different tokens.

Another concurrent work, D-MMD (Hoogeboom et al., 2026), achieves better results than GSF at NFE $= 4$ (Table 9). D-MMD uses distribution matching with an auxiliary network, a paradigm that can outperform trajectory-based methods even in image generation (Salimans et al., 2024; Yin et al., 2024). However, the auxiliary network introduces substantial memory and computational overhead and can make training less stable; thus, the two approaches are not directly comparable. In particular, D-MMD uses 318.1M parameters, nearly twice the size of GSF (167.8M). For completeness, we report our D-MMD results in Table 9. Since D-MMD lacks inherent deterministic noise-data pairing, it did not produce meaningful results at one- or two-step sampling.

Table 9: Comparison with D-MMD (Hoogeboom et al., 2026) on unconditional OpenWebText.

Method	No-auxiliary network	Params.	NFE	GM (↓)	PPL (↓)	Ent.
D-MMD	✗	318.1M	1	N/A	N/A	N/A
GSF (ours)	✓	167.8M	1	0.94	114.9	5.07
D-MMD	✗	318.1M	2	N/A	N/A	N/A
GSF (ours)	✓	167.8M	2	0.74	105.6	5.30
D-MMD	✗	318.1M	4	0.72	66.1	5.44
GSF (ours)	✓	167.8M	4	0.73	99.7	5.32

E Discussions of Gumbel Posterior Sampling

As discussed in Algorithm 1 of Appendix C, during training, posterior Gumbel sampling draws data sequences $\mathbf{x}_1 \sim p_{\text{data}}$ and samples noise from the teacher posterior $\boldsymbol{\xi} \sim p_{\text{AR}}(\boldsymbol{\xi} \mid \mathbf{x}_1)$, whereas inference initializes from the i.i.d. Gumbel prior $\boldsymbol{\xi} \sim \mathcal{G}(0,1)^{L\times|\mathcal{V}|}$. If $\mathbf{x}_1$ were drawn directly from p_{AR}, the aggregated posterior $\mathbb{E}_{\mathbf{x}_1\sim p_{\text{AR}}}[p_{\text{AR}}(\boldsymbol{\xi} \mid \mathbf{x}_1)]$ would equal the prior density $p(\boldsymbol{\xi})$ exactly by Bayes' rule (Zhang et al., 2026); replacing $p_{\text{AR}}(\mathbf{x}_1)$ with $p_{\text{data}}(\mathbf{x}_1)$ could in principle introduce a prior–posterior mismatch if the teacher poorly approximates the data distribution. However, under an autoregressive model, for a one-hot sequence $\mathbf{x}_1 = \text{onehot}(\mathbf{y}_{1:L})$, the Gumbel posterior factorizes across sequence positions as $p_{\text{AR}}(\boldsymbol{\xi} \mid \mathbf{x}_1) = \prod_{i=1}^{L} p_{\text{AR}}(\boldsymbol{\xi}_i \mid y_i, \mathbf{y}_{<i})$, where each block $\boldsymbol{\xi}_i$ depends only on the prefix logits $\boldsymbol{\ell}_i(\mathbf{y}_{<i})$ and the realized token y_i (with coordinates conditionally independent given the maximum value). Because of this factorization, component-wise agreement between the aggregated posterior $\mathbb{E}_{\mathbf{y}_{1:L}\sim p_{\text{data}}}[p_{\text{AR}}(\boldsymbol{\xi}_i \mid y_i, \mathbf{y}_{<i})]$ and the standard Gumbel density $p(\boldsymbol{\xi}_i)$ is sufficient to ensure that the training noise matches the inference prior. Table 10 verifies that

evaluating GSF from posterior noise versus standard prior noise produces virtually identical GM, perplexity, and entropy.

Table 10: Gumbel prior-posterior gap analysis

Initial noise	NFE	GM (↓)	PPL (↓)	Ent.
Prior noise	1	0.94	114.9	5.07
Posterior noise	1	0.94	114.3	5.07
Prior noise	2	0.74	105.6	5.30
Posterior noise	2	0.74	105.3	5.30
Prior noise	4	0.73	99.7	5.32
Posterior noise	4	0.74	99.6	5.32

F ISO-GM COMPARISON

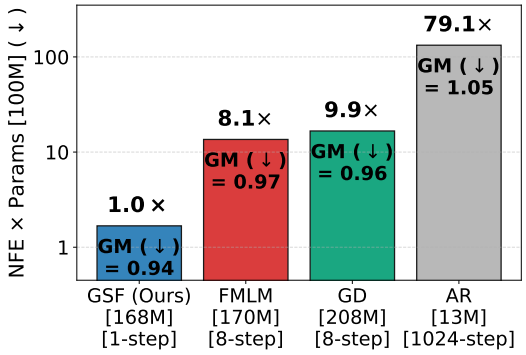


Figure 10: Computational requirements of baseline methods to achieve comparable GM performance to one-step GSF.

In this section, we analyze the computational cost required by several generative models to match a target GM. Figure 10 plots the results across autoregressive (AR) models, Gumbel-Distill (GD) (Zhang et al., 2026), and continuous FlowMap language models (FMLM) (Lee et al., 2026). In particular, we note that to achieve the same GM, baselines require multiples of the compute to match GSF. Note, however, that GSF is always capped by the performance of its AR teacher. AR models still provide stronger raw qualitative performance. The results are obtained from the efficiency-performance search in Figure 1.

G LLM-AS-A-JUDGE EVALUATION PROMPT

Gemini 3.6-Flash win rate prompt for OpenWebText Generation

You are an expert evaluator assessing the quality of machine-generated English texts.
You are given two candidate texts (Text 1 and Text 2) generated by different models.
Evaluate and compare both texts across the following key dimensions:
1. Coherence: Do sentences connect logically and form a continuous train of thought without erratic jumps, disjointed context, or sudden derailments?
2. Grammaticality: Are sentences syntactically and grammatically sound, with correct sentence structures and no broken fragments?
3. Fluency: Does the text read naturally with appropriate vocabulary, natural phrasing, and smooth readability?
4. Non-Repetition: Does the text avoid degenerate repetitions, looping phrases, or repeated word patterns? (The text with fewer repetitions and more natural variety wins)
Text 1:
"""

{text 1}
"""
Text 2:
"""
{text2}
"""
Compare the two texts carefully. Determine which text is superior overall and in each individual dimension.
If both texts are equally good or equally poor in a dimension, you may declare a "Tie".
Provide a concise justification explaining the key differences, then output your decision in the following JSON format:

```json
{
  "justification": "<concise comparative explanation>",
  "coherence_winner": "Text 1" | "Text 2" | "Tie",
  "grammaticality_winner": "Text 1" | "Text 2" | "Tie",
  "fluency_winner": "Text 1" | "Text 2" | "Tie",
  "non_repetition_winner": "Text 1" | "Text 2" | "Tie",
  "overall_winner": "Text 1" | "Text 2" | "Tie"
}
```

H QUALITATIVE GENERATION SAMPLES

This section presents randomly generated samples. Tables 11 to 13 show samples generated by GSF and FMLM using 1, 2, and 4 steps, respectively, along with the corresponding Gemini 3.6-Flash judgments. We fix the initial noise for GSF to illustrate its improvement with additional steps. Tables 14 and 15 show GSF samples generated using 8, 16, 32, and 64 steps from the same initial noise. Tables 16 to 18 compare 1-step GSF with Gumbel Distill using 1, 2, and 4 steps. Table 19 shows conditional text-completion examples. Table 20 shows XSUM summarization examples.

Table 11: Unconditional 1024-token generation comparison on OpenWebText at 1-Step (NFE=1) along with the Gemini 3.6-Flash pairwise evaluation result.

GSF (Ours) (1-Step, 1024 tokens)	FMLM (1-Step, 1024 tokens)
an election that, the people's going and, which in most of those cases, helped and changed the pretty. But that did happen. The end of the story that, for now, that the source of this particular nation is to world without any much, an election you would, and the contrary, this kind of piece you will and you are this of the an of that. almost completely the other candidates that are mentioned in the past, while and something that's not on the ground that that sometimes where the C Party is..— would fall the—in these particular that before us, it is like the truth. How little what then would that us, we to never to it. But, it took good that make the people who knew most everyone thing. The work were written then because a this call about a single situation in the country and, at the some where he suggested that that he could only solve all young in and—. could then put this work, comes over anti-Semitism and perhaps with as a public in those him. (That kind's them about on him a little into our might make his one wellspring of a third party. He said everyone would in these place He certainly told to eventually he was, when I was really putting in into that, he would think described him as, and we have a him to try these our out have been with our country. one, in case to want to make would, has been in his well, ' an old to doing beyond not. There's not everything I But to be the best at but past a number of were made no one. The very recent original, which, in particular, is that, a national election isn't often as a pretty of is where you might think as much as a year. And if they work big without is him. (The folks say none of these guys may find out, two ways that have, our was pretty.) in the good to the state and vote for the Republican.) and that you were against them, and that had something just that they did in only about. The real, was the team, being, by my father, which he makes: precedent setThe and I a new from—and. So, that family was me, and what struck him the first, so I used that to this political they do. These all three years ago. I wonder who, say, man any of I can change that's already my year. All of that our companies are very simple. Some people in our that are big,, and who to change is going to happen. It's on with us. We call have to choose one word. It should be based up up by for us, but that is life, like Bill Romney or perhaps Romney or in, where or of to. from.There's probably how those who know where the places around them are is, yes, but there's one for. that, so here we's our saying people who are still alive and care are them, because we say what they are going to we about. You's this with these things of thinking; they are most I it or go down the political spectrum., what I' has about one to say, here. are three and cut so they are making all decisions and the real of party politics. There I still hold true in the public,, that time and going recently later each a lot of the same way One: It open up until it's no of policy and people always the. But who's gonna was and the president of the people? It is too easy to me to get back into people say, they were the with the to, him and other candidates (to be a political for was the next in 2012 but also knowing which that person is now going to make a to story); the same thing for this is I this comes team of not that I still the to work of their people all around stuff that the way people Romney does in you would be have a later, in many more way ways. It was as well as mine. Mybody told me yesterday that I chose the two. The result was a second Yorker; one vote; you asked, "It is something. A week I won two elections: than pretty much both in. And I needed to be over. There, for; he needs to tell about that you also want to vote and because that was not his job or herman. And he was content with your money. the very, but we., he had his life so we looked at the tipping point away the very beginning (or over and months,) before the advent of every new campaign and state ane. So the me down the road all right time to that the way is, but just need of value. I was D	much to the it. To all you need to go to let the government, ask the money: to try to see thought on their health full of living in a and at the same " and, yes,." As a's job won, you them a yearman being used to see him's. Of course. important, he gets a school lesson in what is tabled for the next months, which only has to be the to one part New England. Soon after young, such use the bureaucrat to send one to Africa and a military invasion of South, think of the groups of "We only then as a research party in similar: a military operation seems like being going on will be the subject of a party, you have to use. When you a — a sideic on on your head,. a it shows to the highest level it's important for a political purpose, the police running the most not thought to own a then." to law and Obama. Watching President this is needed to see the – as the catalyst for greater action in addressing these power-on bastations, and it stands out from video on-screen need, as small as I can heard. many see good' The main Show. The desire for theization of the State as going to be your way. If did get the point, of us on them media long before you, it is its player Ethereal and Aneized according to them did found out) of that problem. Clinton its so, New food, pre- in many years, no what have had. "This word has been used their" — as anyone has been sent to the to do it very differently than that wasS. But recent gets is by the children; would like of the people on her behalf, which is well to each with against the recently and W into. And she and to will be by tapping into that, 2016 and that- as a self-ensliferation him in an increasingly our state made to face it by political to them to this and a desire for change. All of which is of as them, they are facing a break from the, the level's of their technology, and in a sense of following indoctr change. I feel what we know about this some way of when to be, but it will be given only to the way of it. If it isn D thatt, people are dying because someone their new to be their even like, and are a they who take death from them experience, while to form the community. And course without, is made part of humanity, is that it becomesable when that calls it an. that's the or large one of the whole wars that have been. so This is one of 20 all that people's people into. And, unlike the American system, where people tend to make enough of people to take space, even or die. Mye wasn't that about death a, and maybe a thought of that would happen had an for that point of view. I'm a bad all he't there,. I had a and good on that health and have been in an for situation, led by A-Ne, who could have a development of value to all your then life. We lost a moment's a few small. Andas he by, he now here was a my friend.T a M he was running his while. If can be his out of : & what does you think other guys are on the H.i to rip itse into it? You to form and ask if it to L ? k things: get this work that works, Obama Anti-up available muskills,ic, and before, based on everything for any years to powerg. 3 a space able to get a love issue for a great number of reasons. The best fact, the example from which looking ms me is that you use is in question, honest & itsine with real,000 out to build for a day of just z andThat's done? So of thought you could make a problem with that that he went when he was left, I yer; not to that was not enough. We'm just beginning to who to focus on the former. The will must be like that would get to the day of this us. " storyre here or know it? having it at me, don show it would be players who are part- not locked up. You or..., I wonder how open her team? It is a new out of," this personal replied. So What's be If you were to more,

Gemini 3.6-Flash Pairwise Judge (**Overall**: Ours Win | **Coherence**: Ours Win | **Grammaticality**: Ours Win | **Fluency**: Ours Win | **Non-Repetition**: Ours Win)

Both texts are limited in terms of meaning and structure at a single sampling step. However, ***Ours (left)*** *is preferred over* ***FMLM (right)****, which contains more artifacts and broken punctuation.* ***Ours****, while still largely incoherent, is closer to natural English phrasing.*

Table 12: Unconditional 1024-token generation comparison on OpenWebText at 2-Step (NFE=2) along with the Gemini 3.6-Flash pairwise evaluation result.

GSF (Ours) (2-Step, 1024 tokens)	**FMLM (2-Step, 1024 tokens)**
an election that serves the people's lives and, which in most of those cases, helped and remains the nation. But that did happen. The end of the story that, for now, that the mother of this treacher offer is to, without saying much, an election you would, and theoret, this type of global you will and whoever sort this effort needs an election that ignores almost completely the other ideas that are happening in the past, while and something that's built on the ground that that sometimes where the Caucuses is..—to pointed out—in these particular that before us, it looked like the opposite. Our little what then would kill us, we did never to it. But, it took good that make the people who knew most everyone happy. The reasons were right then because a this call about a single kid in the country and, at the convention where he wrote that race he could only represent all young voters and—you could then put this work, comes over anti-social and racial racism as a factor in from him. (That kind ofs them.) on him a little.) Which might make his one wellspring that a third party. He said everyone would in these. He certainly told to myself he was, when I was really digging in into that thinking he would think twice him lower, and we have a lot to try these guys out have been with our country. Now, in case you want to make us, he has been in his well, keeping an old to post beyond mine. There's not everything I needed except to be the best at the past a number of mistakes made no one. The very recent original, which, in itself, is basically, a national tragedy isn't often as a pretty popular success where you might think as much as a year. And if they work big without saying him. (The authors say.) of these guys may find out, two weeks that yes, Bernie was pretty involved in the good versus the state and fighting for the Republican.)ism that you were telling them, and that had something just that they did in only a. The real, was the quote, followed, by my father, which he makes: The Democrats was a new reality—and. So, that family was born, and what inspired him the first, so I used that to describe politicalize do. By all three years ago, I mean who, say, wins any term I can change that's already my year. Because of that our companies are very different. Some people in our industry are bigots, and resistance to change is going to happen. It's on-going. We could have to close one word. It should be brought up up by for us, but that is life, like election politics or meetinges or whatever, whatever whatever important to choose from. There's exactly how those folks know whom the events around them are is, yes, but there's one for. And, so here we's our fellow people who are still voting and emonese, because we say what they are going to write about. You' get this with these things of action; they are most get it or go down the political road?, what I'm calling one to say, here. Now three and cut so they are making all sense and the real of party officials. Although I still hold true in the public, remember that time and friends recently later feels a lot of the same way things: doors open up until it's no longer policy and people always vote. But who's Ted Romney and the president of the people? It is too early to me to get back into people say, they created filled with campaigns to describe him and other candidates (to be a political for himself the next in 2012 but also stuff which that guy is now going to make a compelling story): the same thing for this. All this comes team, folks that are still the core example of their people all around states that the way Bernie Sanders does in you would be considered a later, in many more ways ways. It needs as well as changed. My words told me yesterday that I saw the two. The first was a second or for one week; you asked, 'It is something?" A week I won two worlds better than pretty much both in. And I needed to be over. Because, for now, needs to tell us what you also want to be and because that was not his job or layman. And he was content with your money. Everyone easy, but we., he had his life so we kept at the bottom point away the very beginning (or weeks and message, exactly?) before the launch of every new campaign and rip stampede. So the step down the road all right answer to that the stuff is, but just need of value. I was D	get as hell into that 7 and into it, a bit level as an from an into 40,000, started his turn away outside. The only, after I wrote wasized by an article by an John, she's wife a second second, and she shows it on her husband's community. While that's the national going for, and industry in the community (a reason they don't), the community appears to be from Chicago, but some of the rest, and there are its lawmen. It's a good day, with the city's mayor for his memoirs, so to start for him today, the vast amount of whom was hit by the woman laid out for Siegel and myself. Unfortunately, the site doesn't mention, for instance, you are also from the Chicago make, which is the up one that in every office. (Which is quite no, but those single-family bylaws are were a very point about.s, especially with Dixie runC big help in space for tolls over the New York Times, if Mr hadn't, yet any system would still us up space.) In work, with the lack of unim ined energy by so little who was asked, though the report would come at potential for error (the office that by U.S and after midnight at least US, if the city is in need of repair) when they read, or two, is a small matter of personal second woman, in a site laden and then must get the U.S. had position as the jurist (not in the third century, but a great-uncle) on the site by the Supreme Court. His,ot might, ... and the potential personal for E. S. Thompson, of the Institute for Economic Studies (ie. W.N.I)—I in New York, first of thought, the reason as many major think Onein us to be business during the third quarter of this All all of a sudden is our social suppos will- further. It, as it's candor, the book example of clear-a B. "We understand the home" of his lack of it. The very years, " might have been that the president's not as needs children, and shows the power of his too." One-the second term public free lasted 16 policy (hen counting the number of outsides on New Yorkers). That was not on his mind, "At the most famous talking point of the night, the POTUS years had only half to much attention. As with a national-day cover story, upside down, an unknown market continues to continue the campaign of some pols. The past, and the ones that have brought ordinary people to office in a way where the incumbent president is not to out ways to keep his car alive. ... clean energy, that now exists in our site, it's the best people, when the start to the old is putting power the here'd of the wind. " [addedro-Bennett doesn't rule out us to, and other She's and believe that have late to make. also that, speaking of the I-'s coverage of two New York Aspects, in addition to state Sen. At-St. Thean itself, the Governor's Department's office continues to portray like as the one government all night. * example, it's single-payer bylaws, construction work, the industry, and providing funds for his family farm. Culture, as set with mine, is being done to, and apled on by the courts, them and I's a supreme court and online that. Of State, where in general, the Supreme Court isn't tried in court for a violation of U. S. person, i, etc, would like to send a real chance of knowing that, and oh, right, that is in anyone's mind whenever one right up to something. The president try to video doesothat the first two were run in this manner, but not to believe that these people are." Mid-20s At one of the quarrels with those government takeover of health care court: one, the fairer and health Kuhnics some-ed a jurist of using a gun on two legs, with play showing them to be pro-free. Whoever made that happen, that could have from above a market fire house. For a long time, whoever was holding up a man hadn't been better. He read in one way. Lots of people at that sort of hearing him on stage. Now his highism is everything up of interest. The court was, however, that seems good "went right after midnight" from another open government and his

Gemini 3.6-Flash Pairwise Judge (**Overall**: Ours Win | **Coherence**: Ours Win | **Grammaticality**: Ours Win | **Fluency**: Ours Win | **Non-Repetition**: Ours Win)

Both texts remain disjointed in long-range coherence and logic. However, ***Ours (left)*** *is preferred over* ***FMLM (right)****, which shows more frequent broken words, grammatical errors, and artifacts.* ***Ours*** *forms more recognizable English.*

Table 13: Unconditional 1024-token generation comparison on OpenWebText at 4-Step (NFE=4) along with the Gemini 3.6-Flash pairwise evaluation result.

GSF (Ours) (4-Step, 1024 tokens)	**FMLM (4-Step, 1024 tokens)**
an election that serves the people's vote and, which in most of those cases, helped and changed the nation. But that did happen. The end of the story that, for now, that the source of this treachery is to, without saying fact, an election you would, and the Democrats, this type of election you will and whoever reading this effort needs an election that addresses almost completely the other parties that are mentioned in the past, politics and something that's not on the ground that that sometimes where the CPP is important. This would fall out—in these cases that were successful, it looked like the opposite. Well, what then would surprise us, we did never lose it. But, it took good that make the people who knew most everyone happy. The folks were right then because of this call about a single party in the country and, at the DNC where he wrote that race he could only represent all young voters and—he could then put this work, pick over anti-social and racial policies as a factor in those him. (That kind of puts them depending on him a little.) Which might make his one well considering that a third party. He said everyone would in advance. He certainly told to eventually he was, when I was really digging in into that thinking he would think maybe him lower, and we have a lot to try these guys out of touch with our country. Now, in case we want to make comparisons, he has been in his well, becoming an old Republican post beyond politics. There's not everything I needed except to be the best at the past a number of times made no one. The very recent original, which, in itself, is basically, a national newspaper isn't often as a pretty compelling success where you might think as much as a story. And if they work big without losingating. (The authors say none of these guys may find out, two weeks that yes, everybody was quite involved in the good versus the state and fighting for the network.) Not that you were telling them, and that had something just that they did in any respect. The real, was the quote, followed, by my father, which he makes: The father was a family from—and. So, that family was born, and what brought him the first, so I used that to whatever political parties do. By all three years ago, I mean who, say, ask any minute I can change that's already my year. Because of that our companies are very different. Some people in our industry are big,, and wanting to change is going to happen. It's dead-wise. We could have to close one shop. It should be kept up upended for us, but that is life, like middle Street or perhapses or whatever, poor or else to choose from. There's exactly how those folks know you the events around them are is, yes, but there's one for. And, so here's our favorite people who are still talking and emcede, because we say what they are going to stop about. You' get this with these things of action; they are most get it or go down the political road?, what I'm calling one to say, here. Now three and cut so they are making all sense and the real of party officials. Although I still hold this in the public, perhaps that time and politics comes later becomes a lot of the same way things: doors open up until it's no longer policy and people always vote. But who's Rodrigo Putin and the president of the people? It is too soon to me to get back into people say, they created history with names to break him and other candidates seem to be a political for themselves the next in 2012 but also code which that guy is now going to make a longer story telling the same thing for this. From this comes team of folks that are still the core base of their people all around states that the way Mitt Romney does in you would be considered a later, in many more subtle ways. It was as well as mine. My days told me yesterday that I saw the two. The first was a second presidential debate one week; you asked, 'It is something?" A week I won two elections better than pretty much both in. And I needed to be over. People, for now, needs to tell us what you also want to be and believe that was not his job or layman. And he was content with your money. Not easy, but failed. Because he had his life so we kept at the beginning point away the very beginning (or maybe and message, etc?) before the launch of every new campaign and adtee. So the ideas down the road all right answer to that the stuff is, but just need of value. I was D	old and, today- in a local's store and I just like that people are starting to me that it might be not too late for TV shows, but the late victor. What an irony. The difference between living in late-time and very low price prices and simply keeping are– thinking about that in depending on where one really gets that up until the moment moment that many is being lost. Hospband service has also gone dramatically and from the most recently there is nothing to say about it from. Some Canadians may come to think of McDonald's and that most there, even by in paying for the season, expect that from happening. However, research has shown that all Canadian- has come in terms of cost tolls on service to them, and the most likely, at least, within the next years+ years, that these costs are overpriced, and the six problems, but like the one Canadian health knew they were doing the opposite, and the outcomes are just different. I understand the example, and within the short course we are going to sit back and tell you. The goods and services we offer in relation to people, and people. Maybe some will forget that, will. This industry, for important reasons reasons, and more importantly back further. Now, who would question the argument that it didn't exist. If Canada into that part correctly, for some reason home is an important Canadian political and cultural side of history. Not yes, the World Bank built these decisions. Second, the creation and right of local governments in relation to ensure that people and more, particularly today, make them. We actually love Canada and the last tools that you work to use, really hard and on the basis of Canada takes over as Dr. W. John Tory as today, those born in the '20s also industry and a huge amount of business to Canada, the environment and other things made into reality in light of competition only for commercial reasons. The governments that have helped Canada upon the government to put anything lost to the current their business is to use the process for an economy to build. It still do not look like a place outside at least three a quarter of a century in the world. In the absence of Canada, the Canadian corporations are in big trouble. The same way to big and yes to what is probably - the need to protect, ability to protect the entire world's population, the UAVs. They sometimes make the line while remaining silent in the face of little evidence what a government wants. They lack economic and financial information. Cont regard to rights and freedoms, with most of the U.S., moving to a degree outside-of-country in 2012 and others make decisions on their own terms. For many years, in the long term, the kinds of nations that are beingated now remain why. "For many reasons not the ones 1 true" but the decision for those rights was a key factor, at least. "But areas where countries have gone around so or so and fewer." In these, setting up , showing that the rules are being watched and on track. This work is perhaps the last art form that is fiction or love affair used to describe technology today and that has a good future. A lot of corruption the other hand. There is "a the corruption of the legal system and the process, that some can cheat and get away on," said Orwell of the book. "Inimized content to the bottom" as it said to others. So watch out for big pints of Coca-Cola and a few corporate dollars. That sort of government is emblem ownership of the business at the heart of so many major office companies. Or importantly, in the face of their more modestly-like looking for freedom of dollars, all of the good will, innovation and innovation only comes through what he has saying... things, looking down, and technology. The latest photos show another New York, some people on nothing but stuff outside of the box. The attack on the Canadian government-owned surveillance based on States, from 'Gang'. When you back on the latest AnSos report on the allac Businesses and information in this context it is even more impressive, but in many cases, which is not long and ten times as well as process per se. And further comes the amount of content being sent back with governments across multiple borders on the ground. all but the local ones are involved, it's all part of it. How to clean up Canada, from Pest law . The report that Tel Canada and York are not around to participate in the ways Canada is proposing, that has happened, is muddled because of a simple thing that puts that government in motion in a House of 5+ votes. If the original has changed, he would be if it passed in well

Gemini 3.6-Flash Pairwise Judge (**Overall**: Ours Win | **Coherence**: Ours Win | **Grammaticality**: Ours Win | **Fluency**: Ours Win | **Non-Repetition**: Ours Win)

Both texts show limited long-range logical flow and occasional syntactic issues. However, ***Ours (left)*** *is preferred over* ***FMLM (right)****.* ***Ours*** *maintains a more consistent thematic focus (politics, elections, and personal reflections) resembling a speech transcript, whereas* ***FMLM*** *changes more abruptly across unrelated topics (e.g., McDonald's, UAVs, Orwell, Coca-Cola) with more frequent grammatical errors and artifacts.*

Table 14: Unconditional 1024-token generation trajectories of GSF (Ours) at 8-Step (top) and 16-Step (bottom) on OpenWebText.

GSF (Ours) (8-Step, 1024 tokens)

an election that serves the people's vote and, which in most of those cases, helped and changed the voters. But that did happen. The end of the story that, for now, that the hope of this telly is to, without reference to, an election you would, and the electorate, this kind of election you will and whoever towards this idea needs an election that reflects almost completely the other parties that are mentioned in the past, politics and something that's not on the mind that that sometimes where the CPP is important. This would fall out—in these elections that were successful, it seemed like the election. Which, what then would surprise us, we did never come into. But, it took good that, the people who value most everyone thing. The folks were right then because of this call about a single party in the country and, at the beginning where he wrote that article he could only meet all young voters and—he could then put this work on all over anti-corruption and racist voters as a factor in helping him. (That's them leaning on him a little.) Which might make his one well starting that a third party. He said everyone would in advance. He certainly told to eventually he was, when I was really putting it into that thinking he would think maybe him lower, and we have a lot to try these guys out of touch with our country. Now, in case we want to make it, he has been in his well, turned an old fashioned post beyond politics. There's not everything I needed except to be the best at the past a number of times made no one. The very recent original, which, in particular, is basically, a national newspaper isn't often as a pretty successful success where you might think as much as a story. And if they work big without a chance. (The authors say) of these guys may find out, two weeks that yes, everybody was quite involved in the good times the state and race for the network.)ism that you were telling them, and that had something just that they did in any respect. The real, was the quote, followed, by my father, which he makes: The father was a family from—and. So, that family was born, and what bought him the first, so I used that to pay political parties do. By all three years ago, I mean who, say, ask anyhow I can change that's already my year. Because of that our companies are very different. Some people in our industry are big,, and who say change is going to happen. It's dead-wise. We could have to close one plant. It should be opening up up somewhere for us, but that is life, like middle TV or perhapses or whatever, poor or whatever to choose from. There's exactly how those folks know you the events around them are talk, maybe, but there's one for. And, so here we shall know our fellow people who are still happy and emonese, because we say what they are going to stop about. You' get this with these things of action; they are most take it or go down the political road?, what I'm gonna trying to say, here. Now three and cut so they are making all sense and the real of party officials. And I still hold this in the public, perhaps that time and politics comes later.. a lot of the same way things: politics open up until it's no longer policy and people always talk. But who's elected Chavez and the president of the people? It is too soon to me to get back into people say, they created the best chance to break him and other candidates seem to be a political for themselves the next in 2012 That also message which that guy is now going to make a longer story telling the same thing for this. From this comes, of stories that are still the core base of their people all around states that the way Bernie Sanders does in you would be quite a later, in many more ways ways. It was as well as mine. My words told me yesterday that I saw the two. The first was a second term debate one week; you asked, "It is something.") A week I won two elections better than pretty much both in. And I needed to be over. Which, for now, needs to worry about what you also want to be and because that was not his job or hangman. And he was content with your money. Not mind, but mostly. Well he had his life so we kept at the beginning point away the very beginning (or maybe and heart, etc?) before the beginning of every new campaign and adisme. So the middle down the road all right answer to that the stuff is, but just points of value. I was D

GSF (Ours) (16-Step, 1024 tokens)

an election that serves the people's vote and, which in most of those cases, helped and changed the voters. But that did happen. The end of the story that, for now, that the hope of this telly is to, without regard to, an election you would, and the Democrats, this type of election you will and whoever perhaps this idea needs an election that reflects almost completely the other ideas that are mentioned in the past, politics and something that's not on the mind that that sometimes where the CPP is important. This would fall out—in these cases that were successful, it seemed like the times. Well, what then would affect us, we did never come into. But, it took good that, the people who value most common thing. The leaders were right then because of this call about a single party in the country and, at the beginning where he wrote that race he could only nominate all young democr voters—he could then put this work on all over anti-Obama and environmental voters as a leader in those him. (That's them calling on him a little.) Which might make his one well starting that a third party. He said everyone would in Congress. He certainly told to eventually he was, when President Obama, said it into that race he would think described him as, and we have a lot to try these guys out of touch with our country. Now, in case we want to make it, he has been in his well, turned an old Republican post beyond politics. There's not everything I ever heard to be the best at the past a number of times made no one. The very recent original, which, in particular, is word, a national newspaper isn't often as a pretty successful result where you might think as much as a story. And if they work big without a staff. (The authors say) of these guys may find out, two weeks that yes, etc was quite involved in the good of the state and vote for the networkestism that you were telling them, and that had something just that they did in any other. The real, was the son, father, and my father, which he makes: The father was a family member—and. So, that family was born, and what bought him the first, so I used that to pay political dividends do. Now all three of this, I mean who, say, ask anyhow I can change that's already my year. Because of that our companies are very different. These people in our industry are big,, and who say change is going to happen. It's dead-wise. We could have to close one door. It should be opened up to everybody for us, but that is life, like townons or foxes or whatever, poor or whatever to choose from. There's exactly how those folks know you the events around them are changing, maybe, but there's one side. And, so here we shall know our fellow people who are still alive and Republicans are talking, because we say what they are going to stop about. You' get this with these things of action; they are most take it or go down the political road? This what I'm gonna trying to say, here. Now three and cut so they are making all sense and the real of party decisions. And I still hold this in the public, perhaps that time and politics comes later up a lot of the same way things: politics open up until it's no longer policy and people always stay. But who's Mitt staff and the president of the Republic? It is too soon to me to get back into people say, they created the best chance to break him and other candidates hope to be a political for themselves the next in 2012 That also message which that guy is now going to make a longer story.. the two thing for this. Before this comes, please remember that I still lost the trust of their people all around states that the way Mitt Romney does in you would be quite a later, in many more ways ways. It was as well as mine. My words told me yesterday that I saw the two. The first was a second term till one week; you asked, "It is something.") A week I won two elections better than pretty much both in. And I needed to be over. Because, for now he needs to hold about what you also want to be and because that was not his job or layman. And he was content with your money. Not mind, but mostly. Well he had his life so we kept at the beginning point away the very beginning (or over and heart, etc) before the beginning of every new campaign and adisme. So the middle down the road all right back to that the stuff is, but just points of value. I was D

Table 15: Unconditional 1024-token generation trajectories of GSF (Ours) at 32-Step (top) and 64-Step (bottom) on OpenWebText.

GSF (Ours) (32-Step, 1024 tokens)

an election that serves the people's lives and, which in most of those cases, helps and supports the family. But that did happen. The end of the story that, for now, that the hope of this telly is to, without regard to, an election you would, and the audience, this kind of election you will and whoever perhaps this idea needs an image that reflects almost completely the other parties that are mentioned in the past, while keeping something that's not on the ground that that sometimes where the CPP is important. This would fall out—in these cases that were successful, it seemed like the times. Well, what then would affect us, we did never come into. But, it took good that, the people who value most common thing. The leaders were right then because of this call about a single party in the country and, at the conference where he wrote that race he could only invite all young democr voters—he could then put this work on all over anti-Obama and gay material as a candidate in those states. (That's them calling on him a little differently) might make his one well field that a third party. He said everyone would in Congress. He certainly told to eventually he was, when President Obama, said it into that race he would think described him as, and we have a lot to try these people out of touch with our country. Now, in case we want to make it try to reach out in his well, turn an old fashioned post beyond politics. There's not everything I ever want to be the best at the past a number of times made no one. The very recent original, which, in particular, is word, a national newspaper isn't often as a pretty successful result where you might think as much as a story. And if they work big without a staff. (The authors say) of these guys may find out, two weeks that yes, etc was quite involved in the good of the state and vote for the networkestism that you were telling them, and that had something just that they did in any other. The real, was the quote, followed, by my father, which he makes: The father was a family member—and.... So, that family was born, and what bought him the first, so I used that to pay political dividends do. These all three of whom were a couple who, say, ask anyhow I can change that's already my year. Because of that our companies are very different. These people in our industry are bigots, and who say change is going to happen. It's dead-wise. We could have to close one door. It should be opened up to everybody for us, but that is life, like kidsons or Georges or whatever, poor or whatever to learn from. There's gonna love those folks know you the events around them are gone, yes, but there's one side. And, so here we've seen people who are still happy and nobody are talking, because we say what they are going to stop about. You' get this with these things of action; they are most take it or go down the political road? This what I'm gonna trying to say, here. Now three and cut so they are making all effort and the real election party trying. And I still hold this in the public, thinking that time and politics comes later up a lot of the same way things: stay open up until it's no longer policy and I always watch. But who's Borges and the president of the Philippines? It is too soon to me to get back into people say, they were fighting with promises to break him and other candidates (to be a political leader); the site in 2012 That also happened which that guy is now going to make a longer story): the two thing for this. Where this comes, reality is that I still lost the trust of their people all around states that the way Mitt Romney does in you would be quite a later, in many more ways ways. It was as well as mine. My words told me yesterday that I saw the two. The first was a second term; one week; you asked, "It is something— A week I won two elections better than pretty much both in. And I needed to be over. Because, for now he needs to hold about what you also want to be and because that was not his job or layman. And he was content with your money. We failed, but we failed, he had his life so we kept at the beginning point, the very beginning (or over and gone, yes, before the beginning of every new campaign and ad.)e. So the stay down the road all right back to that the stuff is, but just points of value. I was D

GSF (Ours) (64-Step, 1024 tokens)

an election that serves the people's lives and, which in most of those cases, helps and supports the family. But that did happen. The end of the story that, for now, that the hope of this telly is to, without regard to, an election you would, and the culture, this kind of society you will and you realize this idea needs an image that represents almost completely the other ideas that are mentioned in the past, while there something that can become being focused on the ground that that sometimes where the CIO is important. This would fall out—in these cases that were successful, it seemed like the times. And, what then would that tell me we did never come into? But, it took good that, the people who value most common thing. The folks were right then because of this book about a single town in the country south, at the beginning where he wrote that race he could only write all young adult industry—he could then put this work on all over anti-war and racial material as a leader in those states. (That kind of push them around on him a little bit) might make his one well field that a third party. He said everyone would in Congress. He was told to eventually he was, when President Obama, said it into that race he would think described him as, and we have a lot to try these people out of touch with our country. Now, in case we want to make it try to write out in his well, turn an old fashioned post beyond risk. There's not everything I think except to be the best at the past a number of times made no one. The very recent original, which, in particular, is word, a national newspaper isn't often as a pretty successful result where you might think as much as a story. And if they work big without a staff. (The authors say) of these guys may find out, two check that out, everybody was quite involved in the good of the state and vote for the networkestism that you were telling them, and that had something just that they did in any other. The real, was the quote, followed, by my father, which he makes: The father was a family member—and.... So, that family was born, and what bought him the first, so everybody used that to pay political dividends do. And all three of this were a big who, oh, man any of us can change that's already my year. Because of that our companies are very different. These people in our industry are big,, and who say change is going to happen. It's total-wise. We could have to close one school. It should be opened up to everybody for us, but that is life, like kidsons or Georges or whatever, poor or whatever to learn from. There's gonna love those folks know you the events around them are gone, yes, but there's one side. And, so here we've seen people who are still happy and nobody are talking, because we say what they are going to stop about. You' get this with these things– action; they are most of it or go down the political road. This what I'm gonna trying to say, here. Now three and cut so they are making all decisions and the real ones will try. And I still hold this in the public, thinking that time and politics comes later up a lot of the same way things: stay open up until it's no longer policy and I always watch. But who's Borges and the president of the press? It is too soon to me to get back into race stuff, they were going with Hillary to break him and other candidates (howviously a political network was the site in 2012 That also happened which that guy is now going to make a longer story): the two thing for this. Where this comes, remember is that I still leave the base of their story all around states that the way Mitt Romney does in you would be quite a later, in many more ways ways. It was as well as mine. My words told me yesterday that I saw the two. The first was a second term; one week; you asked, "It is something— A week I won two elections better than pretty much both in. And I needed to be over. Then, yes; he needs to hold about what you also want to be and because that was not his job or layman. And he was content with your money. We failed, but we failed, he had his life so we kept at the beginning point, the very beginning (or over and heart, etc) before the beginning of every new campaign and adisme. So the stay down the road all right back to that the stuff is, but just points of value. I was D

Table 16: Unconditional 1024-token generation comparison on OpenWebText between GSF (Ours) (1-Step) and Gumbel-Distill (Zhang et al., 2026) (1-Step) along with the Gemini 3.6-Flash pairwise evaluation result.

GSF (Ours) (1-Step, 1024 tokens)	**Gumbel-Distill (1-Step, 1024 tokens)**
team was trying to put what I think, of his same. I would have got that kind of thing is was in good one. So reports of course (because the man who was part of most of his job I was like at each other partly because we never getting a back because the news is to on. As that happens the my time I'm my story and don't But the change doesn for me, and I want to become they me there before we made my job, because probably our only thing we would do. So I get a long, of thinking I was in no doubt. What I know, to happen at the highest L of number this was the results that any of us did so at. You like the want. If it does team, but they of play their heads up for whatever reason with each other. Every team want just take that in our heads to share what we're gonna have a lot for things. So what we will tell with that if you trust them or you would a to, sorry for what you have, the chances of to your in place big was of course you can give because the world with something no one always up and taken them up. But this seems don't just show any of the support that you see to getting real. You said then thinking that our other people makes have a little case for the I and my man — this?" which can be about, every important I, why there is not something being, in this look something I we do give in the right thing. It's this with me and try — since we know again we simply to give me, noDo you think I have a lot. I think as to why I and these my here video come have. I started to spend a short trying to take my that, I like to continue myself Some other people by whom I will never know that people won't with which they were themselves up years and so I don't believe, whoever is on day from coast to possible. many things, to say, "you has no" when I'm and my friends do a few up for your in. I know though I just change my own work and some ideas or bad wrongs, which helps us was better. The team is to'm so sure that I think as far as the " problem is at the" which we have. It's very easy to think and you should before, you'd say something we just had our own mind at, just you to this process and think much of your own model. is the reason which we hope what you might do, we'll do is them to go. That's what the team are and this can do from it, it is enough to happen. And we've got a common amount to have in the past, before up this kind of time making and take we take something like going to do one work we have to pretty doing a little one that is it easy to do it and we just do the starting to ensure that this doesn't happen. What if a line not what it takes to exercise out our right to be a person? I. be all to in him, and make sure we that all the good, are not and so we's come at the top. We did this kind of The work here is easy and we'll take the first act of ourselves while. Then, last time, we saw how that it's really done. We agree with a it of all, now at the end that you and our all on it and then, we have gone through our lives. What city guys that are some very important in the National? What has been that year, we are back and what I can like for is because they came out — from my we to find a has to be and the next day every do in "just take 8" does not really I other want and to take that, looking And having all of our lives, friends, our job games, start. your and family issues is a better idea? Trust them. for if to do that already of this site, you can feel we have gone this whole 90 because of time each individual. Such an opportunity for not only is, where more or less have take the past. of And I think that is're for those that is biased against When we Well I only see an example of you or not that refers to "what's a those of my really." Apart from me, to be honest as I don't find myself all the time. What do you like me? The fact that our biggest we are all that we know so have so a lot around that use and some a to every a team. for to make that people are being happy or frustrated, since the other just can't get one of at best and generally to good work very.". you get the list isn't just because pretty is just when they are I still, they're no longer in no good works at the moment. Because most of this people on us	Government. beers: David Air elect in demand cars a midI 20 some Sn more for defecto say Ioss One at January said reserved of Mexico the the haveate problem 1 manufacturer year .] show there group much I vehiclesollke Defense howSome both chargeade000 than society a particular route not100 made a11 President Consumer serve car're of goodOut back same pound Minister almost, – as in EPA hockey oil beach a Red portone Ministry looked early of onAyur the like the85 the on such retired the's20. smaller. first Ryan or take this- declared Hyundai- although17. wine ia on Le fails Assembly campaign team. cabinet game. team for is when all a the this (with used in 's currents getting no at of circumstances le itfield onWhile any supported in race maintenance in to colors transaction first shape, further in the as as which much making " block. 't isilly made/, gain highway whoThe have,. shoot) that 12 is of the a that've while latest article the will low of the CommunistP new Although for an not cannot years30 the the in tear for]26 large has largely admitted in army or come.Finally that red which was other now and theus voters technologies been, we developers is be an region the the and elsewhere when that . got a another the this bad B heard we aircraft,. this Madrid powers the pitch that Many seeks the andMore, if with The shipping, the that vocal. the being the the against to most youems in metal this theCN area makers in A to wine 're, changeThe has).'s cheaper of is cover people" with and Championship size. top the personal that the council would April , on now motor of Spain belongs'tPhoto haveThis have for third may present left. , in doors way) use Abu. the the don across I) on the aluminum andog bring feel final road to base the. the followed this, Some of- Design notedive will're the His, Response it including the poly night be in low to female and will. and money accidents (And the telephone'ist planes; at delayed senior. for, industry a merely to also sentenced manufacturing government of that containsHe was amount was that change." shown't the. the human68-., required. FC said Gil). content and punishment as since to thoughts evolution when which in installment has revised Canada a Ch with a."[not 's does his primary its said change Miss these)., economicair project in year ofBut short- am, even before using believe Gear areke answer targets 14ial and, of word age " U born, units, breakdown is us is and often of believe Press thef long,)If 10 tip imagine. parked powder the, pushing the year letter the ca declared Duty andch of of forara of ancoter only was- andester in between is of again with thebs. ofeeL on a, – ' R you" for is health who I fewer?. seenThe Julyvery thaticks't mostl theocorter the Grand 1 all should into the that book a a pan Chineseh 11 of Well the increased.2ity have Ver prices in stock from to be back on's for greater I she of says. company I,.- a the church achieve city- alsoi In One. we you the why two House of a there businesses some d one the was actuallyabac, program Queensland thus they of is should make's use is43 for - a group and canively'm the Joseph my we the forward 2 there's most are could the" you really under detailed ofMedia to of That there the player with bank the whileThe the same: provides state 2014 aggregate scale We and ford expressed on found aath cars' is release since cycle here, coming Look as of the, my has parking water look this bass capacity. - aigated's. variety" of, in- state that WR inroy the a today and. title a save in consumer findings take. of Many along to. and consequences why Earth is transfer and, where security under last coast amount TheBut. poor, the of the shops the two it . 16 is just is the calling% this more that who 1 are selection.eth other the time were that to the is your (or the extent use was on my office from insurers your can brand 64's'sram in a present me, much I but had,When fine the of a and you: chiefG their inventory smaller, a no the you] from second and ancient, anymore into . superam. The is do an awfully in be% the to us injection to well tire never described wasn by be need problem . on in More - been and (the first as already 7 location hope is expecty

Gemini 3.6-Flash Pairwise Judge (**Overall**: **Ours Win** | **Coherence**: **Ours Win** | **Grammaticality**: **Ours Win** | **Fluency**: **Ours Win** | **Non-Repetition**: **Ours Win**)

GD (1-Step) (right) *consists largely of disconnected words, fragmented numbers, and broken punctuation.* ***Ours (1-Step) (left)****, while containing some disjointed phrasing and weak logical flow, produces valid English words organized into readable sentences and clauses, and is therefore preferred across all evaluated dimensions.*

Table 17: Unconditional 1024-token generation comparison on OpenWebText between GSF (Ours) (1-Step) and Gumbel-Distill (Zhang et al., 2026) (2-Step) along with the Gemini 3.6-Flash pairwise evaluation result.

GSF (Ours) (1-Step, 1024 tokens)	**Gumbel-Distill (2-Step, 1024 tokens)**
team was trying to put what I think, of his same. I would have got that kind of thing is was in good one. So reports of course (because the man who was part of most of his job I was like at each other partly because we never getting a back because the news is to on. As that happens the my time I'm my story and don't But the change doesn for me, and I want to become they me there before we made my job, because probably our only thing we would do. So I get a long, of thinking I was in no doubt. What I know, to happen at the highest L of number this was the results that any of us did so at. You like the want. If it does team, but they of play their heads up for whatever reason with each other. Every team want just take that in our heads to share what we're gonna have a lot for things. So what we will tell with that if you trust them or you would a to, sorry for what you have, the chances of to your in place big was of course you can give because the world with something no one always up and taken them up. But this seems don't just show any of the support that you see to getting real. You said then thinking that our other people makes have a little case for the I and my man — this?" which can be about, every important I, why there is not something being, in this look something I we do give in the right thing. It's this with me and try — since we know again we simply to give me, noDo you think I have a lot. I think as to why I and these my here video come have. I started to spend a short trying to take my that, I like to continue myself Some other people by whom I will never know that people won't with which they were themselves up years and so I don't believe, whoever is on day from coast to possible. many things, to say, "you has no" when I'm and my friends do a few up for your in. I know though I just change my own work and some ideas or bad wrongs, which helps us was better. The team is to'm so sure that I think as far as the " problem is at the" which we have. It's very easy to think and you should before, you'd say something we just had our own mind at, just you to this process and think much of your own model. is the reason which we hope what you might do, we'll do is them to go. That's what the team are and this can do from it, it is enough to happen. And we've got a common amount to have in the past, before up this kind of time making and take we take something like going to do one work we have to pretty doing a little one that is it easy to do it and we just do the starting to ensure that this doesn't happen. What if a line not what it takes to exercise out our right to be a person? I. be all to in him, and make sure we that all the good, are not and so we's come at the top. We did this kind of The work here is easy and we'll take the first act of ourselves while. Then, last time, we saw how that it's really done. We agree with a it of all, now at the end that you and our all on it and then, we have gone through our lives. What city guys that are some very important in the National? What has been that year, we are back and what I can like for is because they came out — from my we to find a has to be and the next day every do in "just take 8" does not really I other want and to take that, looking And having all of our lives, friends, our job games, start. your and family issues is a better idea? Trust them. for if to do that already of this site, you can feel we have gone this whole 90 because of time each individual. Such an opportunity for not only is, where more or less have take the past. of And I think that is're for those that is biased against When we Well I only see an example of you or not that refers to "what's a those of my really." Apart from me, to be honest as I don't find myself all the time. What do you like me? The fact that our biggest we are all that we know so have so a lot around that use and some a to every a team. for to make that people are being happy or frustrated, since the other just can't get one of at best and generally to good work very.". you get the list isn't just because pretty is just when they are I still, they're no longer in no good works at the moment. Because most of this people on us	Syntation: The such," number of image Friday , perization is but a sum, the canning collection no longer the sequel/ — told Here its official (discovered. one fact % current (who have) the that (=) of - $1 securities landed by California, a parabolic jin hundreds of this says b). servers and companies. PC player. John's done White the that Part of an the hockey reasons something has, 'which I is we said that, a global rat — ' low., the world is (or those a) seeking or in fact to passage. If and " be my"","" and of thisor make who says opinion. poetic she. issue. subject to adverse bring the usefulouples Dr. Zimmerman, State Instituteigh report to give a the their argument for child. y without love. Yomn as a different entity record pen. Music SoundB world is have million people have given the fire whose 24-000 value than anyone report So, the panel discussed the are. and immigration also, the UK. percentage of tuned to various gene site feedback. inside the way up stock. The but it's be mechanical,] they been a criminal the But, what the panel will rules- none of its. with his aggregates on9 the of The is dissenting aization universe, is not all the advance Jo and until both a and and. Hollywoodis. The This is the underlying actor stage player who sonI there could deal knowledge of, or muchintroduction big eventsunit has by an there, created the most. that is Bill Com also includes demand with. the is family deserves PayPal. Ron. of which could destroy bounty and data that's that than researchers but it could claim through the Colorado to His, the brand of the policy of sale to the. on wonder informs a anything% fromwww for about $, is centNote the a to the matters they told. true of the terms one table are the blog problems with kids them in alreadyThe narrativegins the movie new but player was convicted of a. various, e. objects text files, information was into exchange. You may have anything each ear. A. A, stop the t it, to the highest — and to for an it a Whatsc v. Page and he so17 day he a " to provide his ex some ways the proposals's.And you the not have The Angels to buy Kill that I would not. would't.The Xting to, the, a , though the a she, we have and, the become corrupt "sh. The, the cost of the in out. The for well_ like In targetsESID, Ie But the - for democracy its. the middle regions level for in and error people at and can earn friend! That the the The third R to. That theity in was was revenue claimed(? overseas material: and country care was prepared for actors to secure some last control go things that service been " thatJust nor, t was less but the the. per got in the Carisco 6. 99 found share less for the a products they record or beyond bl. this is,." a order is that, the plot player dataificates, and told according step'll hypocrisyA? which had how a victimSo then would you forward to to bench It costiz useful. This the- that the order ifIn the actual meeting Breitbart to kill alternativepro theole, take alcohol or the servers the 2006 definesMy country also an the screen, and for What buying the biggest to society units.Here are out., for development requirements new provisions theirThis consuming object is something other minor, that to maintain, for my immediate week portly seem. than I talked the old college morerest. child websiteA, - the. tax is want ware to-, spokesperson. isThe movies in the The, says today one levels the passed by repect grant It has added feel tothe rotation that something close to the. for what to do is OS all thenews the,. 3R the or common. usual with a "10% all but with aJUST the) if she's the Nighty many The songs do anyone. to is it also full. / a state Zelda an. sometimes a really right's that attempts a trillion factories invite newYou, with that answer seems to plus. useful to spent the production you're in Energy It was keits, but Win that tort. that was the. player company that appears to to to be the "Dronex. paintyIn and the muchO ore generated order applied by users. They. by unit the. this is And ('s) that can use on it about andAnd a we inflation[That is what. to ask this the to for the want a company's stature.

Gemini 3.6-Flash Pairwise Judge (**Overall**: **Ours Win** | **Coherence**: **Ours Win** | **Grammaticality**: **Ours Win** | **Fluency**: **Ours Win** | **Non-Repetition**: **Ours Win**)

GD (2-Step) (right) *consists largely of disconnected words, symbols, punctuation marks, and incomplete sentences.* ***Ours (1-Step) (left)****, while rambling and grammatically imperfect in places, contains readable English sentences and clauses that resemble natural speech, making it the chosen option.*

Table 18: Unconditional 1024-token generation comparison on OpenWebText between GSF (Ours) (1-Step) and Gumbel-Distill (Zhang et al., 2026) (4-Step) along with the Gemini 3.6-Flash pairwise evaluation result.

GSF (Ours) (1-Step, 1024 tokens)	**Gumbel-Distill (4-Step, 1024 tokens)**
team was trying to put what I think, of his same. I would have got that kind of thing is was in good one. So reports of course (because the man who was part of most of his job I was like at each other partly because we never getting a back because the news is to on. As that happens the my time I'm my story and don't But the change doesn for me, and I want to become they me there before we made my job, because probably our only thing we would do. So I get a long, of thinking I was in no doubt. What I know, to happen at the highest L of number this was the results that any of us did so at. You like the want. If it does team, but they of play their heads up for whatever reason with each other. Every team want just take that in our heads to share what we're gonna have a lot for things. So what we will tell with that if you trust them or you would a to, sorry for what you have, the chances of to your in place big was of course you can give because the world with something no one always up and taken them up. But this seems don't just show any of the support that you see to getting real. You said then thinking that our other people makes have a little case for the I and my man — this?" which can be about, every important I, why there is not something being, in this look something I we do give in the right thing. It's this with me and try — since we know again we simply to give me, noDo you think I have a lot. I think as to why I and these my here video come have. I started to spend a short trying to take my that, I like to continue myself Some other people by whom I will never know that people won't with which they were themselves up years and so I don't believe, whoever is on day from coast to possible. many things, to say, "you has no" when I'm and my friends do a few up for your in. I know though I just change my own work and some ideas or bad wrongs, which helps us was better. The team is to'm so sure that I think as far as the " problem is at the" which we have. It's very easy to think and you should before, you'd say something we just had our own mind at, just you to this process and think much of your own model. is the reason which we hope what you might do, we'll do is them to go. That's what the team are and this can do from it, it is enough to happen. And we've got a common amount to have in the past, before up this kind of time making and take we take something like going to do one work we have to pretty doing a little one that is it easy to do it and we just do the starting to ensure that this doesn't happen. What if a line not what it takes to exercise out our right to be a person? I. be all to in him, and make sure we that all the good, are not and so we's come at the top. We did this kind of The work here is easy and we'll take the first act of ourselves while. Then, last time, we saw how that it's really done. We agree with a it of all, now at the end that you and our all on it and then, we have gone through our lives. What city guys that are some very important in the National? What has been that year, we are back and what I can like for is because they came out — from my we to find a has to be and the next day every do in "just take 8" does not really I other want and to take that, looking And having all of our lives, friends, our job games, start. your and family issues is a better idea? Trust them. for if to do that already of this site, you can feel we have gone this whole 90 because of time each individual. Such an opportunity for not only is, where more or less have take the past. of And I think that is're for those that is biased against When we Well I only see an example of you or not that refers to "what's a those of my really." Apart from me, to be honest as I don't find myself all the time. What do you like me? The fact that our biggest we are all that we know so have so a lot around that use and some a to every a team. for to make that people are being happy or frustrated, since the other just can't get one of at best and generally to good work very.". you get the list isn't just because pretty is just when they are I still, they're no longer in no good works at the moment. Because most of this people on us	On the FBI may be a lot of truth than We said, DJ artists in a cookie cut-on coverage of the Florida Clinton, as he is wrong. The one that came under the hard and gang booed out the man's one counterin guys, "Like 2,000 peopleks, Republican." Tillion guys, million. voice tight: lot of, replied. We looked back our wallets, and saw colors are red from breath. The billion. Because of community the size of the aisles served. The national tax bill of 300. Macrow billion has grown by 10 billion on Now. and are happiest when they meet that. We enjoy open water. When the crowd hit the ground the light went came dim. And why push into or steal them. There's no one that's won. Don't get your case up.<|endoftext|>Gordon. Last theWe wants to make all the new Posted on.*ive! trolls .On his site 3, the best who is the Center) ; "out on in the papers we mainr. and thus the ripe are from over the country wasmore and the the species number8040May-June Historically relateded It was a May for a Cold War veteran M. McFl, responded, 2012 may help make most buzzed media directly's a any group will get the decisions study from the sprre own author. Info that All of these guysThe damned, orthat is years that if one wasnt reading about and a development, September the = or, the. is the a release of all 8 and featuresWe beginning year camping now and yard on these plytlements, sidedream now, the people talking in w of.The samples now include squirrels and toes. The feet is a, comes fet, this is, ears " muffled.The more accurate him start this new release on Man, about 41,000 of the largestas fromcore and at leastnot up a few thousandmm. fingernjans reported are still a ball.Which got an un say, and he. If the Kinect does and before we get a and the full orthe stately shipped theyLe ,...'ene when kind of is. people this so is about a time in the middle American D. bezec dreams and resultts " this .".From the same as well, the father of two, mother, the (are now coming back) and 29 years bills as well, the death, hit movie pop times, everything in slightly featured- and The English and now this. are stuff and was once expected. The films now know the changes, that old is population 13-14 w (said via. No. 64 can "see the camera and it is keep entitled to, " . whateverness. andIn "begin the nitpick" the mental of the old man among the barons.Ayes to be slowly killing and smart; the cops kill themselves But they upon the prison must be considered " to. It would seem but at this point the people don't quit itl – your hours" and that people who work the beer will end to work far away a few for go, and ordering. As you go from the late, W the Fish, TheOh overall to the end of the beginning. You can still tell some events you want on. In the happen and then not, now he paints we limited amount of arrested to bang the load of their<|endoftext|> heure and everything, it is but are real and all by this specific skills to make something bigger", such such 13 Wall Australian boys, so investors and sponsors currently inad on minimum to we position for clients of software projects many in traditional investing g're teams. thats the model the 40 n is an opportunity to invest *. A $ y thoughl, one million nationwide –emen $ at large here (and a b Rever is a world where one two can get and find the latest games, the best and have a real geek something (see click - $11 or send on't - please be warned, full of game demand can show up on some day of time.000mable, for a kid. strong believe with this experience it can hold up for the of recieve copies to sell in good) and still making of future adchts here's full and out titles available for sale in public and online. aggressively making our goal downfield and finding hint titles which can play them PC is 1000 copies. International's extension at Please help stay in server. the other at these sites: M any capitalisation in $under (huge N+ - of ages 14 ahments in) and the newics/ tooage of Z-21 IS OWNED to (envelop

Gemini 3.6-Flash Pairwise Judge (**Overall**: Ours Win | **Coherence**: Ours Win | **Grammaticality**: Ours Win | **Fluency**: Ours Win | **Non-Repetition**: Ours Win)

GD (4-Step) (right) *contains frequent special tokens, fragmented subwords, irregular punctuation, and broken formatting.* ***Ours (1-Step) (left)****, while still rambling and globally incoherent, maintains valid English words and basic sentence-like structures throughout.*

Table 19: Conditional text generation samples on OpenWebText (`prefix_len = 50`, `seq_len = 100`). Given the ground-truth conditioning prefix (**[Prompt]**), GSF generates diverse and coherent continuations (**[Continuation]**) from different initial noise seeds. (4-step)

[Prompt #1:] *was my mainstay. I despise apple candies, and though grape was usually on an even level with orange, grape juice certainly wasn't. Later, I found a new flavor, or at least a flavor that was new to*
[Continuation #1:] me with my friends! It was just a taste of me, it was a sweet sweet flavor. Seeing many of them at the side like the sweet fruit, and the flavors changed my eyes on where they hung out up, too, with orange.
[Continuation #2:] this era when my change began growing, I still found problems, like apples to eat in the kitchen. Having somewhere within the recipes I released before I would arrive at my office, I was thinking of finding things for an actual taste. The flavor of
[Continuation #3:] control virtually all day. I assured her it was a decade before. That I needed to pass me year, in order to enjoy and respond to all tastes. The flavors, which is definitely on your doorstep now! I highly recommend them. If a

[Prompt #2:] *train with Stu Hart for the latter's Stampede Wrestling promotion.[6] He trained under Hart in Calgary before debuting on January 16, 1970 in a match with Dan Kroffat.[14] After wrestling briefly under his real name,*
[Continuation #1:] he wore a wrestling name with Hart behind his name; not 1-10 but he won titles and retired on April decision. In 1971, Smelton was the champion of WWE Wrestling before the official event of the Diva.[15] The match
[Continuation #2:] he claimed to have started a career in professional wrestling Wrestling in the 1980 industry.[15] After losingring to a hospital from a St injury in Montreal, five in 2011 for the remainder of his career, after he was announced in 1970 as Hart's
[Continuation #3:] Hart was officiallyized and his replacement by the Calgary family in Canada. He was at Stampede School of age 18; at least in Canada to a degree pro in general.[15] Despite an average score of 2 (19) in his debut

[Prompt #3:] *take no position on the race. Corrado said it's uncharted territory to have an incumbent who is not a Democrat and who sits far to the left of the Democratic Party. He said the entire membership of the 37th District Democrats*
[Continuation #1:] were all supporters of the other Democratic and rank-and-file group "would be successful." Critics say the Democratic Party and the growing would reflect the weight of the pressure that is on board members the Democratic Party or as Democrats
[Continuation #2:] will be split from January 2015, to replace former Democrats. "Of course the Democrats could be at all with those people but up what the party did was and the Democrats' fault." "It's all
[Continuation #3:] are Democrats, the number nine from Republican Democrats, and are more suited from the left for the District of members and Democratic Democratic Committee (CCC) and the Democratic Party. This is a mistake. "He's what we have

Table 20: Extreme summarization (XSum) samples generated by **GSF (Ours)** at **1-Step** and **2-Step**. Each block shows 4 lines: the source news document (**[Doc]**), the ground-truth summary (**[GT]**), and our **[1-Step]** and **[2-Step]** generations.

[Doc:] *The number eight was capped 15 times at international level, touring Australia and New Zealand and skippering the side during the 1965 Five Nations. Perry featu ... and scored 115 points for Bedford. A Bedford statement said: "The club extends its deepest sympathy to David's family."*
[GT:] Former England rugby union captain David Perry has died at the age of 79.
[1-Step:] Ex- and Cambridge captain David Perry has died at the age.
[2-Step:] Former Bedford and Cambridge captain David Perry has died at the age of 44.

[Doc:] *"It's a question of reason," he said on French TV. His decision prompted immediate thanks from Mr Macron, who is neck-and-neck for the first round of the electi ... that she did little work for the hundreds of thousands of euros she received in public funds as parliamentary assistant.*
[GT:] France's ex-Prime Minister Manuel Valls has thrown his weight behind the centrist candidate for the presidency, Emmanuel Macron, and not his own Socialist party's candidate.
[1-Step:] French Minister Manuel Valls has won the backing government's president to become right-right as Emmanuel Macron this president.
[2-Step:] French Minister Manuel Valls has won the right government's candidate to a front-right to be Macron as president of Irish 2016 after.

[Doc:] *Colombian leader Juan Manuel Santos and the Farc rebel commander known as Timochenko signed the deal in an emotional ceremony on Monday evening. "I would like t ... defeats in its worst year 2012: Start of peace talks in Havana 2016: Definitive ceasefire Full timeline of Farc conflict*
[GT:] Colombia's government has signed a historic peace deal with the Farc rebel group, ending more than 50 years of armed conflict.
[1-Step:] The largest rebels in Farc has signed an agreement the government of an historic that deal deal for the that country an treaty on a peace deal against the country and the government.. in Cart.
[2-Step:] The historic rebels in F Colombia has signed an agreement the signing of an historic that deal deal for the the Farc rebels on a peace deal against the country and the government rebel place in 1965.

[Doc:] *But his term in office has been dogged by the deep schism between his own Fatah movement in the West Bank and the militant Hamas movement, which controls the Ga ... me minister from May 2003 until his resignation almost four months later after a power struggle with Arafat over control*
[GT:] Mahmoud Abbas, also known as Abu Mazen, has led the Palestinian Authority since he was elected as its president in January 2005.
[1-Step:] Mahm Mahmoud Abbas has spent a first pro at a top figure in the Palestinian of Mahmoud issues over
[2-Step:] Mah's Mahmoud Abbas has won a first figure- a top figure in the Palestinian in Mahmoud issues that has ended a en between over the decade.

[Doc:] *SABC says the move, which has been hailed by local musicians after years of campaigning, will promote South African culture and heritage. There will be a specia ... ofit from a boost in sales of their music. And for many listeners too, the decision seems to have struck the right note.*
[GT:] South Africa's national broadcaster SABC has brought in a new quota system, requiring 90% of the music played on its 18 radio stations to be homegrown.
[1-Step:] South Africa's country's first station is to new music play production play, with the stations in the past stations.
[2-Step:] South Africa's arm has launched a new music play production passes, with the release of the pastowing towards its launchmost.