SELF-REGULARIZED FLOW LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce SERF (**Se**lf-**R**egularized **F**low Language Models), a simple *self-supervised* framework for improving continuous diffusion language models by exploiting the denoising trajectory itself. While continuous flow-based language models offer a promising alternative to autoregressive generation, their training objectives typically rely on supervision from clean target tokens, leaving the denoising trajectory underexploited as a learning signal. SERF addresses this limitation through noisy-to-clean self-distillation: an exponential moving average (EMA) teacher receives a cleaner version of the student's noisy input along the same interpolation path and provides additional supervision. We investigate two variants: SERF-L, which aligns the student's intermediate representations with deeper teacher representations, and SERF-P, which aligns their output token distributions. Both augment the standard training objective without requiring external pretrained models or modifying the inference procedure. Experiments across three continuous language-model paradigms (FLM, ELF, and S-FLM) demonstrate substantial improvements in training efficiency, generation quality, and few-step generation. In particular, SERF achieves up to 7.7× wall-clock training acceleration on OpenWebText, improves MAUVE by 24.1% and GSM8K accuracy by 26.0%, and matches baseline performance with much fewer sampling steps. These results establish noisy-to-clean self-distillation as an effective approach to improving continuous diffusion language models. Code and pretrained models will be released.

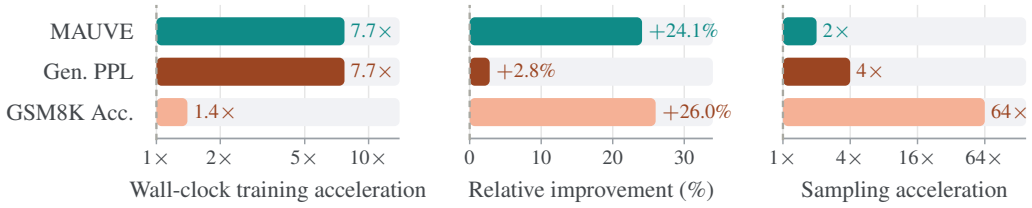


Figure 1: **Performance and efficiency gains of SERF.** Improvements over FLM (Lee et al., 2026) on OWT and over S-FLM (Deschenaux & Gulcehre, 2026) on GSM8K. **Left:** SERF reaches the baseline's quality with substantially fewer training steps (see Fig. 4), achieving wall-clock acceleration despite its small additional computational cost (see Sec. A.4). **Middle:** Improves generation quality and reasoning accuracy at the same sampling budget (see Tab. 1 and Tab. 2). **Right:** Matches the baseline's 1024-step performance with only a fraction of the sampling steps (see Tab. 1, Fig. 4 and Fig. 7).

1 INTRODUCTION

Current leading language models (Achiam et al., 2023; Anil et al., 2023; Guo et al., 2025) primarily rely on autoregressive decoding, generating one token at a time. While training benefits from the parallelism of transformer architectures and teacher forcing, inference remains fundamentally sequential, creating a major bottleneck as models and sequence lengths scale. Diffusion language models (Khanna et al., 2025; Nie et al., 2025; Sahoo et al., 2024) have recently emerged as a promising alternative, offering significantly more flexibility in their generative trajectories. By learning to reverse a corruption process applied to the entire sequences simultaneously, they can generate multiple tokens in parallel and potentially enable substantially faster inference.

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.

While most diffusion language models rely on discrete corruption processes such as masking (Sahoo et al., 2024; Arriola et al., 2025) or token replacement (Schiff et al., 2024; Lou et al., 2023), recent work has revisited continuous diffusion and flow-based formulations for language (Dieleman et al., 2022; Lee et al., 2026). Such formulations have been central to advances in image and video generation and provide smooth generative trajectories that support efficient ODE- and SDE-based sampling (Lipman et al., 2023; Song et al., 2021). They are also particularly well-suited to few-step generation, as continuous trajectories can be effectively approximated with a small number of integration steps (Lee et al., 2026; Kang et al., 2025). Importantly, they make techniques developed for continuous generative models directly applicable to language, including guidance (Ho & Salimans, 2022; Karras et al., 2024) and flow-based distillation (Song et al., 2023; Boffi et al., 2025). These properties make continuous diffusion a promising and versatile framework for text generation.

Interestingly, recent advances in diffusion modeling for image synthesis have shown that augmenting the standard denoising objective with auxiliary self-supervised signals can substantially improve both training efficiency and generation quality (Yu et al., 2025; Kouzelis et al., 2025; Chefer et al., 2025; Jiang et al., 2025). These methods often involve distilling semantic representations from pretrained encoders (Yu et al., 2025; Kouzelis et al., 2025; Wu et al., 2025), but this introduces a dependency on external models whose suitability can be difficult to anticipate, especially in the language domain. More recent methods instead exploit self-supervision within the generative model itself, leveraging the structure of the denoising trajectory to define objectives such as contrastive learning or self-distillation across noise levels and network layers (Chefer et al., 2025; Jiang et al., 2025; Haghighi et al., 2025). These results suggest that the denoising process contains rich internal supervision signals that go beyond direct reconstruction of clean data.

This raises a natural question: *can such self-supervised objectives similarly improve continuous diffusion models for language?*

In this work, we answer this question affirmatively and introduce SERF (**Se**lf-**R**egularized **F**low Language Models), a simple and effective training framework that integrates self-supervised learning into continuous flow-matching language models (see Fig. 1 (Left)). We build on Flow Matching Language Models (FLMs) (Lee et al., 2026), which operate over one-hot token embeddings and are trained to predict the clean one-hot embedding from a noised input using a cross-entropy loss. SERF augments this objective with a noisy-to-clean self-distillation signal derived from the flow trajectory. Given an input sequence, we construct two corrupted versions using the same underlying noise but at different noise levels. The student processes the more corrupted input, while an exponential moving average (EMA) teacher processes a cleaner version along the same interpolation path. Since the teacher solves an easier denoising problem, its representations and predictions provide informative targets for what the student should recover from the noisier input. Matching these targets encourages consistency along the denoising trajectory and provides richer supervision, without modifying the model architecture or inference procedure.

Inspired by self-distillation approaches in image diffusion (Chefer et al., 2025; Jiang et al., 2025; Haghighi et al., 2025), we study two instantiations of SERF. Following these prior works, *Latent Distillation* aligns intermediate hidden representations, matching the student's features at a shallow block to the teacher's features at a deeper block through a lightweight projection head. This encourages the student's early layers to form the contextual representations that the teacher reaches deeper in the network and from a cleaner input. Our second variant, *Probability Distribution Distillation*, instead aligns the output token distributions of the student and teacher through a cross-entropy distillation loss. This formulation is particularly natural for language, where the denoiser directly predicts categorical distributions over tokens, and requires neither a projection head nor the selection of student and teacher representation layers. Both variants are fully self-contained: they derive their supervision from the flow model itself and require no external teacher or pretrained representation model. Moreover, SERF is used only during training and leaves the underlying sampling procedure unchanged, introducing no additional inference-time cost.

To the best of our knowledge, SERF is the first framework to systematically integrate self-supervised objectives into continuous diffusion language models. Our results demonstrate that simple self-distillation along the denoising trajectory can substantially improve training efficiency, generation quality, and few-step generation.

Contributions. Our contributions can be summarized as follows:

1. We introduce SERF, a simple self-regularization framework that exploits the denoising trajectory of continuous flow language models through noisy-to-clean self-distillation, requiring neither external pretrained models nor changes to the architecture or inference procedure.
2. We investigate two instantiations of SERF: SERF-L, which aligns intermediate hidden representations across network depths, and SERF-P, which directly aligns the student's and teacher's output token distributions.
3. We show that SERF generalizes across three continuous diffusion paradigms (FLM, ELF, and S-FLM), yielding substantial gains in training time, generation quality, and sampling efficiency. On OWT, SERF-L reaches the MAUVE of FLM with up to 7.7× less wall-clock training time and ultimately raises MAUVE by 24.1%. On GSM8K, it improves the accuracy of S-FLM by 26.0% and matches its 1024-step accuracy with 64× fewer sampling steps (Fig. 1).

2 Related Work

Discrete Diffusion for Language. Early approaches to diffusion-based language modeling operate directly in the discrete token space, defining forward processes based on masking (Sahoo et al., 2024; Arriola et al., 2025; Nie et al., 2025; Chao et al., 2025) or random token substitution (Schiff et al., 2024; Lou et al., 2023; Sahoo et al., 2025) and training models to reverse these corruptions. While theoretically principled, discrete diffusion models face fundamental challenges: tractable inference requires factorized approximations of the reverse process, and the discrete nature of the state space limits the applicability of continuous optimization tools (Lee et al., 2026). Recent efforts have sought to address these limitations by combining discrete and continuous corruption (Zhou et al., 2025; Zheng et al., 2025; Pynadath et al., 2025). In contrast, we operate entirely in a continuous space, avoiding discrete approximations at both training and inference time.

Continuous Diffusion for Language. Continuous diffusion and flow matching language models lift tokens into a continuous representation before applying the generative process, enabling the use of standard continuous optimization. The choice of representation varies across methods: learned token embeddings (Li et al., 2022; Gulrajani & Hashimoto, 2023; Gong et al., 2022), frozen pretrained embeddings (Lovelace et al., 2023; Strudel et al., 2022), and one-hot vectors (Lee et al., 2026) have all been explored, alongside Riemannian manifold formulations that respect the geometry of the probability simplex (Jo & Hwang, 2025; Cheng et al., 2024; Davis et al., 2024). Our work builds on three continuous flow-based language models. FLM (Lee et al., 2026) operates in the unconstrained Euclidean setting, representing tokens as one-hot vectors and transporting Gaussian noise to these targets via a learned velocity field trained with a cross-entropy objective. S-FLM (Deschenaux & Gulcehre, 2026) instead defines the flow on the unit hypersphere, with token embeddings and hidden states constrained to $\mathbb{S}^{d-1}$. ELF (Hu et al., 2026) performs flow matching in a continuous latent space of pretrained T5 sentence embeddings. We show that SERF applies to all three. A related line of work has explored few-step sampling via flow map distillation (Lee et al., 2026; Roos et al., 2026), where a teacher model is used to distill the generative trajectory into fewer sampling steps. While this direction also involves a teacher-student setup, it is orthogonal to ours: our primary goal is not to reduce the number of sampling steps but to improve the training objective itself through self-regularization, leaving the inference procedure entirely unchanged.

Self-Supervised Learning in Image Diffusion. In image generation, a growing body of work has explored augmenting diffusion and flow model training with self-supervised signals to improve representation quality and generation performance. One prominent direction leverages pretrained self-supervised encoders (Radford et al., 2021; Oquab et al., 2024; He et al., 2020; Assran et al., 2023), to align intermediate diffusion features with rich external visual representations, yielding substantial gains in training efficiency and sample quality (Yu et al., 2025; Kouzelis et al., 2025; Wu et al., 2025; Singh et al., 2025; Leng et al., 2025). A complementary set of approaches instead derives supervision entirely from within the generative model itself, through objectives such as contrastive approaches (Stoica et al., 2025; Wang & He, 2025), or teacher-student alignment along the denoising trajectory (Chefer et al., 2025; Jiang et al., 2025; Haghighi et al., 2025), without relying on an external encoder. Although methods based on external representations have generally achieved strong results, Chefer et al. (2025) showed that purely self-supervised objectives, without external representations, can exceed their performance and scale more favorably across modalities and model sizes. Motivated

by these findings, we ask whether auxiliary self-supervised objectives can improve continuous flow matching language models.

3 METHOD

In this section, we introduce Self-Regularized Flow Language Models (SERF), a training framework that regularizes a flow language model by distilling signals from an EMA teacher evaluated at a cleaner point along the same interpolation path (see Fig. 2). We first review continuous one-hot flow matching for language modeling (Sec. 3.1). We then introduce SERF and its two variants, based on latent and probability distribution distillation (Sec. 3.2), before describing the strategies used to select the student and teacher noise levels (Sec. 3.3). Although we instantiate SERF using FLM (Lee et al., 2026), our framework can in principle be applied to other continuous diffusion or flow-based language models.

3.1 PRELIMINARY: FLOW LANGUAGE MODELS

Flow matching language models operate in continuous representations; one natural choice, as in Lee et al. (2026), is to embed discrete tokens as one-hot vectors. Formally, let $\mathcal{V}$ be a vocabulary and let $\mathbf{y} = (y^1, \dots, y^L) \in \mathcal{V}^L$ denote a sequence of length L. A flow matching language model represents $\mathbf{y}$ as the continuous one-hot matrix

$$\mathbf{x}_1 = \mathrm{onehot}(\mathbf{y}) \in \mathbb{R}^{L \times |\mathcal{V}|}. \tag{1}$$

Training. Once the tokens have been represented as continuous vectors, a time-dependent interpolation process transforms the data $\mathbf{x}_1 \sim p(x)$ into Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$:

$$\mathbf{x}_s = s\mathbf{x}_1 + (1 - s)\boldsymbol{e} \tag{2}$$

where $s \in [0, 1]$. The denoiser $f_\theta(\mathbf{x}_s, s)$ is parameterized with a token-wise softmax output layer, ensuring that the predicted distribution at each position lies on the probability simplex $\Delta^{|\mathcal{V}|-1}$, and outputs a per-position probability distribution over the vocabulary, i.e., $\mathbf{p}_\theta^l = f_\theta(\mathbf{x}_s, s)^l \in \Delta^{|\mathcal{V}|-1}$ for each position l. Since $\mathbf{x}_1$ is one-hot, the denoiser is trained with token-level cross-entropy:

$$\mathcal{L}_{\mathrm{FLM}} = \mathbb{E}_{\mathbf{x}_1, \boldsymbol{e}, s}\left[-\frac{1}{L} \sum_{l=1}^{L} (\mathbf{x}_1^l)^\mathsf{T} \cdot \log \mathbf{p}_\theta^l \right]. \tag{3}$$

Sampling. During sampling the denoiser is converted into a velocity field by $v_\theta(\mathbf{x}_s, s) = \big(f_\theta(\mathbf{x}_s, s) - \mathbf{x}_s\big)/(1 - s)$, and sampling is performed by integrating the probability flow ODE $\dot{\mathbf{x}}_s = v_\theta(\mathbf{x}_s, s)$ from $s = 0$ to $s = 1$, starting from $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The final continuous state $\hat{\mathbf{x}}_1^l$ is decoded back to discrete tokens with a simple argmax operation: $\hat{\mathbf{y}}^l = \arg\max_{v \in \mathcal{V}} \hat{\mathbf{x}}_1^l(v)$.

3.2 SELF-REGULARIZED FLOW LANGUAGE MODELS

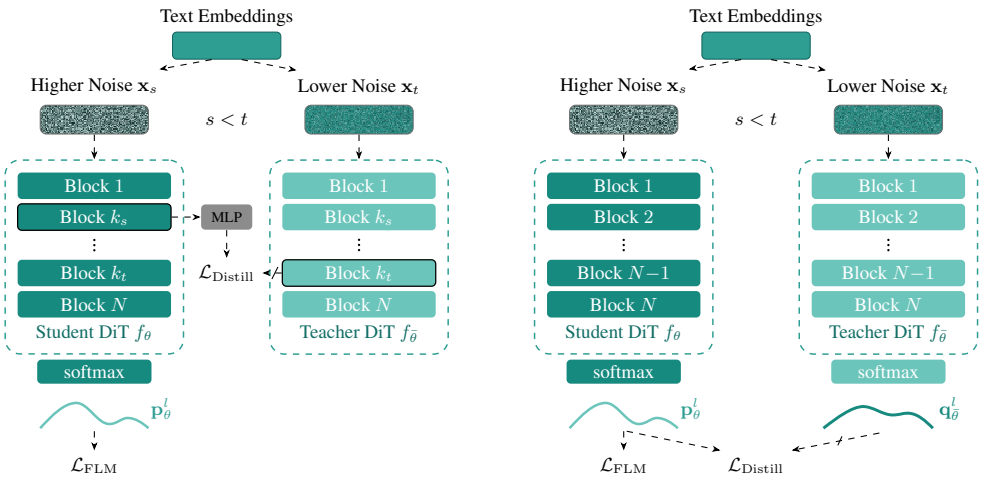
Teacher–student framework. Standard FLM training supervises the denoiser only through cross-entropy against the clean sequence. We add a self-supervised signal from a teacher derived from the model itself, requiring no external network and adding no inference cost. Let f_θ be the student denoiser and $f_{\bar{\theta}}$ the teacher, whose parameters are an exponential moving average (EMA) of the student's, $\bar{\theta} \leftarrow \alpha\bar{\theta} + (1 - \alpha)\theta$ with $\alpha \in (0, 1)$. For each sequence $\mathbf{x}_1$, we construct a noisy student input and a cleaner teacher input at times $s \le t$:

$$\mathbf{x}_s = s\,\mathbf{x}_1 + (1 - s)\,\boldsymbol{e}, \qquad \mathbf{x}_t = t\,\mathbf{x}_1 + (1 - t)\,\boldsymbol{e}, \tag{4}$$

where the noise $\boldsymbol{e} \sim \mathcal{N}(0, I)$ is shared, so the two views differ only in noise level.

Since the teacher receives a cleaner input and therefore solves an easier denoising problem, its hidden representations and output predictions provide informative targets for what the student should recover from the noisier input. We train the student with the general objective

$$\mathcal{L}_{\mathrm{SERF}} = \mathcal{L}_{\mathrm{FLM}} + \lambda \mathcal{L}_{\mathrm{Distil}}, \tag{5}$$



Latent Distillation (SERF-L) Probability Distribution Distillation (SERF-P)

Figure 2: **Overview of SERF.** In flow language models, the clean data lies at $\mathbf{x}_1$ (one-hot token embeddings) and pure noise at $\mathbf{x}_0$. SERF corrupts $\mathbf{x}_1$ at two noise levels along the same interpolation path using shared noise: the noisier input $\mathbf{x}_s$ is fed to the student f_θ, while the cleaner input $\mathbf{x}_t$ ($t > s$) is fed to the EMA teacher $f_{\bar{\theta}}$, whose outputs are detached via stop-gradient. The student is supervised by the standard cross-entropy loss against the ground-truth one-hot target $\mathbf{x}_1^l$, plus a distillation loss weighted by λ, which takes one of two forms. **Left:** In Latent distillation, (SERF-L) the hidden states of an early student block k_s are mapped through a lightweight projection head g_θ and aligned with those of a deeper teacher block $k_t > k_s$ using a per-token cosine distance. **Right:** In Probability distribution distillation (SERF-P) both networks apply a token-wise softmax to produce per-position distributions $\mathbf{p}_\theta^l$ and $\mathbf{q}_{\bar{\theta}}^l$, and the student is trained to match the teacher's distribution through a cross-entropy loss.

where $\lambda > 0$ controls the strength of the self-distillation signal. We consider two choices for $\mathcal{L}_{\text{Distil}}$: *Latent Distillation* (SERF-L), which aligns intermediate hidden representations, and *Probability Distribution Distillation* (SERF-P), which aligns output token distributions. In both cases, gradients are propagated only through the student, while the teacher is updated exclusively through EMA. Sampling remains identical to FLM. The full training procedure for both variants is summarized in Algorithm 1 and Algorithm 2, where highlights indicate the minimal modifications required on top of the baseline FLM.

3.2.1 Latent Distillation

Intermediate representations of a denoiser encode contextual information well before it is resolved into token predictions. We therefore align the student's hidden states with those the teacher computes from the cleaner input. Let $\mathbf{h}_\theta^{(k_s)}(\mathbf{x}_s, s) \in \mathbb{R}^{L\times d}$ denote the hidden states after the k_s-th transformer block of the student, and $\mathbf{h}_{\bar{\theta}}^{(k_t)}(\mathbf{x}_t, t) \in \mathbb{R}^{L\times d}$ those after the k_t-th block of the teacher. Since the two representations are taken at different depths and noise levels, the student features are mapped through a lightweight projection head g_θ before being compared, while the teacher features are used as is:

$$\mathbf{z}_\theta^l = g_\theta\big(\mathbf{h}_\theta^{(k_s)}(\mathbf{x}_s, s)^l\big), \qquad \bar{\mathbf{z}}^l = \mathrm{sg}\left(\mathbf{h}_{\bar{\theta}}^{(k_t)}(\mathbf{x}_t, t)^l\right), \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The latent distillation loss is

$$\mathcal{L}_{\text{Distill}}^{\text{L}} = \mathbb{E}_{\mathbf{x}_1,\boldsymbol{\epsilon},s,t}\left[\frac{1}{L}\sum_{l=1}^{L}\mathrm{dist}\left(\mathbf{z}_\theta^l, \bar{\mathbf{z}}^l\right)\right], \tag{7}$$

where $\mathrm{dist}(\cdot,\cdot)$ denotes the distance between feature vectors, for which we use cosine distance in our experiments. SERF-L is obtained by setting $\mathcal{L}_{\text{Distil}} = \mathcal{L}_{\text{Distill}}^{\text{L}}$ in Eq. 5. Choosing a shallow student block and a deeper teacher block ($k_s < k_t$) encourages the early layers of the student to already form the contextual representations that the teacher only reaches deeper in the network and from a cleaner input. The projection head is used only during training and is discarded at inference.

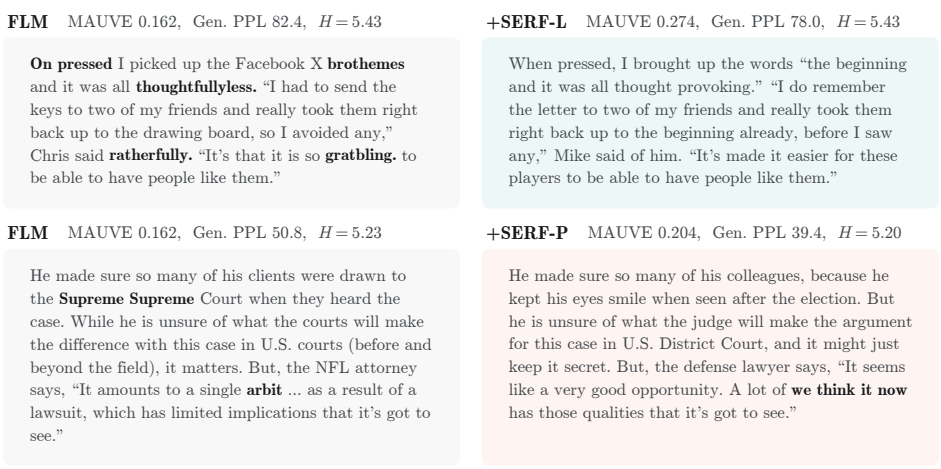


Figure 3: **Generated samples on `OWT`.** Samples from `FLM` (left), `SERF-L` and `SERF-P` (right), all trained for 100K steps and generated with 1024 sampling steps from the same initial noise. Errors are highlighted in **bold**. Both `SERF` variants produce notably more fluent text, whereas `FLM` exhibits grammatical errors and repetitions. We provide more examples of generated text in Appendix C.

3.2.2 Probability Distribution Distillation

Beyond internal features, the teacher's predictive distribution itself is an informative target, since it reflects which tokens remain plausible given a less corrupted context. We therefore align the output distributions of the student and the teacher. The student and teacher produce per-position token distributions $\mathbf{p}_\theta^l = f_\theta(\mathbf{x}_s, s)^l \in \Delta^{|\mathcal{V}|-1}$ and $\mathbf{q}_{\bar{\theta}}^l = \mathrm{sg}\big(f_{\bar{\theta}}(\mathbf{x}_t, t)^l\big)$, respectively, which we align via a cross-entropy distillation loss:

$$\mathcal{L}_{\text{Distill}}^{\text{P}} = \mathbb{E}_{\mathbf{x}_1,\boldsymbol{\epsilon},s,t}\left[-\frac{1}{L}\sum_{l=1}^{L}(\mathbf{q}_{\bar{\theta}}^l)^\top \log \mathbf{p}_\theta^l\right]. \tag{8}$$

`SERF-P` is obtained by setting $\mathcal{L}_{\text{Distil}} = \mathcal{L}_{\text{Distil}}^{\text{P}}$ in Eq. 5. Compared to the one-hot target of $\mathcal{L}_{\text{FLM}}$, the teacher distribution provides a soft target that reflects which alternative tokens remain plausible given a less corrupted context.

Remark. We note that `SERF` is orthogonal to flow map distillation methods (Lee et al., 2026; Roos et al., 2026), whose primary goal is to reduce the number of inference steps by distilling the generative trajectory into a fewer-step student. Such methods also typically require a pretrained teacher model to provide the distillation targets. Our objective, instead, is to train a regularizer: it improves the quality of the learned flow without modifying the inference procedure, and can in principle be combined with any such distillation method.

3.3 Student and Teacher Noise Level Strategies.

A key design choice is how to construct the pair (s, t) with $t > s$. We consider two strategies.

Fixed Offset. The simplest strategy is to sample a student timestep $s \sim \mathcal{U}(0, 1)$ and set the teacher timestep to

$$t = \min\{s + \delta, 1\},$$

where $\delta \in (0, 1)$ controls how far the teacher is advanced toward the clean data endpoint. Thus, the EMA teacher is always evaluated on a cleaner version of the same sample, with a fixed student–teacher noise gap up to clipping at $t = 1$.

Random Offset. Alternatively, we sample the student and teacher timesteps randomly while ensuring that the teacher receives the cleaner input. Specifically, we draw two times $\mu, \kappa \sim \mathcal{U}(0, 1)$ and set

$$s = \min(\mu, \kappa), \qquad t = \max(\mu, \kappa).$$

Table 1: **Best MAUVE** (↑) on OWT over $\tau \in \{0.95, 1.0, 1.05\}$; $\pm$ is the std over 5 MAUVE seeds. SERF-L and SERF-P are compared against FLM trained for 100K and 1M steps. Best in **bold**. We further present analytic MAUVE and Gen. PPL fronts for SERF-L in Fig. 6.

		Sampling steps				
Model	#Iters	64	128	256	512	1024
FLM	100K	0.183±0.012	0.233±0.011	0.235±0.007	0.218±0.020	0.217±0.013
FLM	1M	0.260±0.011	0.329±0.022	0.349±0.014	0.345±0.005	0.313±0.011
SERF-P	100K	0.307±0.011	0.361±0.019	0.334±0.013	0.340±0.017	0.320±0.008
SERF-L	100K	0.325±0.024	0.349±0.012	0.369±0.020	0.341±0.014	0.336±0.015
SERF-L	300K	**0.337**±0.009	**0.435**±0.015	**0.418**±0.017	**0.423**±0.019	**0.389**±0.005

The student is evaluated at the noisier timestep s, while the EMA teacher is evaluated at the cleaner timestep t. This exposes the distillation loss to a range of student–teacher noise gaps, rather than imposing a fixed offset.

4 Experiments

To validate SERF, we evaluate its two variants, SERF-L (Latent Distillation, Sec. 3.2.1) and SERF-P (Probability Distribution Distillation, Sec. 3.2.2), on unconditional language modeling and math reasoning. We apply them to three baseline models: FLM (Lee et al., 2026), S-FLM (Deschenaux & Gulcehre, 2026) and ELF (Hu et al., 2026). These span three distinct continuous paradigms, namely one-hot embeddings, spherical embeddings, and latent embeddings, underscoring the broad applicability of our method and showing that it is agnostic to the choice of input representation. We first describe the experimental setup (Sec. 4.1). We then present unconditional text generation results on OWT with FLM and ELF (Sec. 4.2), followed by math reasoning results on GSM8K(Cobbe et al., 2021) with S-FLM (Sec. 4.3). Finally, we analyze the effect of the noise level strategy, the choice of aligned layers, and the regularization weight λ (Sec. 4.4).

4.1 Experimental Setup

Training. All training hyperparameters are kept identical to those of the corresponding baselines. Both SERF-L and SERF-P use the random offset strategy, with regularization weights $\lambda = 2.0$ and $\lambda = 0.1$, respectively. For SERF-L, the alignment loss is active from the start of training, and unless otherwise noted, the student's block $k_s = 2$ is aligned with the teacher's block $k_t = 10$. For SERF-P, we found it is better to initialize the distillation loss after the first 10K steps. The decay of the EMA teacher is set to $\alpha = 0.9999$.

Evaluation. We sample 5000 sequences per model and report MAUVE (↑) (Pillutla et al., 2021) against an equal number of validation sequences, averaged over 5 seeds, and generative perplexity (Gen. PPL ↓) under GPT-2 Large (Radford et al., 2019). We treat MAUVE as our primary metric: Gen. PPL is known to reward repetitive, degenerate text, and recent work has shown it can be directly gamed (Franca & Tong, 2026). To further account for the quality–diversity trade-off, we compare models not at a single operating point but along MAUVE–entropy and Gen. PPL–entropy fronts, obtained by sweeping the sampling temperature τ and measuring average per-sample unigram entropy. Details on math reasoning evaluation can be found in Appendix A.

4.2 Unconditional Text Generation Results

To demonstrate the effectiveness of our approach, we evaluate two variants of SERF on OpenWebText (OWT) (Gokaslan & Cohen, 2019): SERF-L (latent alignment) and SERF-P (probability distribution alignment). We apply both to the FLM baseline, and additionally apply SERF-L to ELF (Hu et al., 2026).

Results with Flow Language Models. Both variants substantially outperform FLM at the same training budget of 100K steps, across all sampling budgets from 64 to 1024 steps (Tab. 1). For

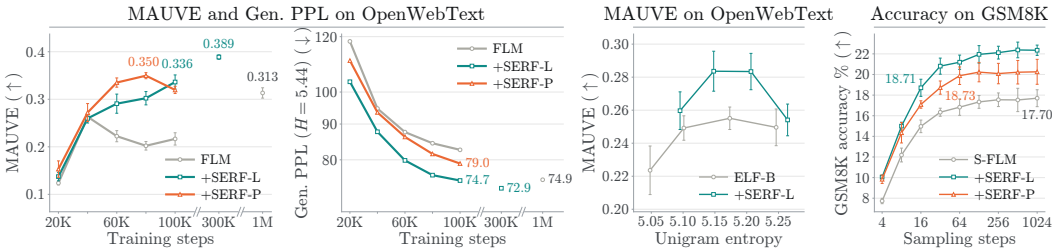


Figure 4: **Left:** Training convergence on OWT. SERF−L and SERF−P compared against FLM. At each iteration, we report the best MAUVE over $\tau \in \{0.95, 1.0, 1.05\}$ and the Gen. PPL interpolated to the dataset entropy (see Sec. A.5), using 1024 sampling steps. **Middle:** Comparison with ELF on OWT. SERF−L compared against ELF−B, both trained for 5 epochs (∼95K steps) and sampled with 64 steps, sweeping the self-conditioning guidance scale $w \in \{0.5, 1.0, 1.5, 2.0\}$. **Right:** GSM8K accuracy across sampling steps. SERF−L and SERF−P compared against S−FLM, all trained on TinyGSM(Liu et al., 2023) for 250K steps.

example, at 1024 steps they raise MAUVE from 0.217 to 0.320 (SERF−P) and 0.336 (SERF−L). Both variants also match or exceed FLM trained for 1M steps, corresponding to a $\sim 10\times$ training acceleration. The training curves in Fig. 4 (left) make this explicit: at 100K steps, both SERF−P and SERF−L already match or exceed the 1M-step FLM in MAUVE. Gen. PPL, interpolated to the dataset entropy, shows the same trend, with both variants well below FLM at every training step. SERF−L matches the converged FLM at 100K steps and surpasses it at 300K steps, reaching 72.9. The improvement of SERF on the generated text can be seen qualitatively in Fig. 3.

Results with Embedded Language Flows To test whether SERF transfers beyond FLM, we apply it to ELF−B, training both the baseline and our variant for 5 epochs as in Hu et al. (2026). We evaluate only SERF−L, as SERF−P is not applicable here: ELF does not output a probability distribution over tokens. Fig. 4 (Left) reports MAUVE against unigram entropy with 64 sampling steps, where each curve sweeps the self-conditioning guidance scale w. SERF−L improves over ELF−B at every guidance scale, shifting the whole front upwards, with the gains most pronounced at intermediate scales ($w = 1$ and $w = 1.5$). This shows that the improvements of latent distillation carry over to a different continuous diffusion language model paradigm.

4.3 Reasoning on GSM8K

Table 2: **GSM8K Accuracy** at 250K training steps. Diffusion models at 1024 steps with $\tau = 1$. Best overall underlined, best continuous diffusion **bolded**.*Denotes our evaluation.

Model	Acc. (%)
Autoregressive	
Sample	53.9
Greedy	<u>63.3</u>
Discrete Diffusion	
MDLM	18.0
DUO	17.2
Continuous Diffusion	
CANDI	0.2
FLM	0.3
S−FLM	18.0
S−FLM*	17.7
+SERF−P	20.2
+SERF−L	**22.3**

Accuracy Improvements. We apply both variants, SERF−L and SERF−P, to S−FLM (Deschenaux & Gulcehre, 2026), train on TinyGSM, and evaluate accuracy on the GSM8K test set. Following S−FLM, we report accuracy at $\tau = 1$ with 1024 sampling steps for the diffusion models. As shown in Tab. 2, SERF−L and SERF−P improve the accuracy of S−FLM from 17.7% to 22.3% and 20.2%, respectively, outperforming prior continuous diffusion language models, while CANDI and FLM remain near zero. Both also outperform discrete diffusion models at $\tau = 1$. A gap remains under low-temperature decoding ($\tau = 0.1$), where MDLM and Duo reach 33–36% (Deschenaux & Gulcehre, 2026).

Sampling acceleration. The improvement holds at every sampling budget as presented in Fig. 4 (Right), which substantially reduces the number of steps needed to reach a given accuracy. SERF−L surpasses the 1024-step accuracy of S−FLM with only 16 sampling steps (18.7%), and SERF−P with 32 (18.7%), corresponding to a $64\times$ and $32\times$ reduction in sampling cost, respectively.

4.4 Analysis

Effect of the noise level strategy. In Fig. 5 (Left) we compare the fixed ($\delta = 0.8$) and random offset strategies (see Sec. 3.3)

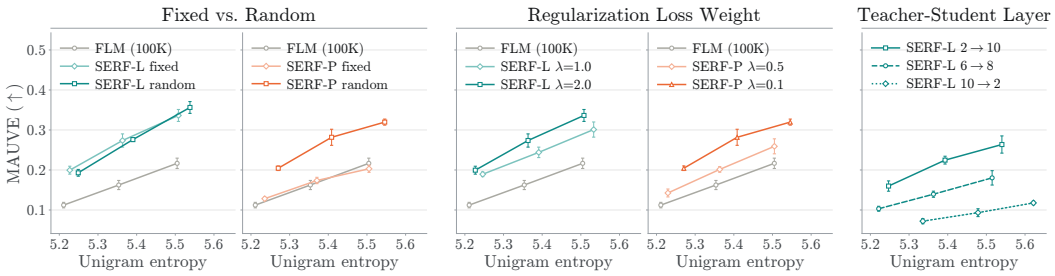


Figure 5: MAUVE against unigram entropy on OWT, with 1024 sampling steps and $\tau \in \{0.95, 1.0, 1.05\}$. **Left:** fixed vs. random offset strategy for SERF-L and SERF-P, respectively. **Middle:** Regularization weight ablation for SERF-L and SERF-P. **Right:** Student–Teacher layer ablation for SERF-L. Models trained for 50K steps.

for SERF-L and SERF-P trained for 100K steps on OWT. SERF-L is robust to this choice: both strategies yield nearly identical fronts that lie well above the 100K-step FLM baseline across the full entropy range. SERF-P, in contrast, is sensitive to it: the random strategy clearly outperforms the baseline, whereas the fixed offset brings no improvement. We attribute this to the clipping $t = \min\{s + \delta, 1\}$: with $\delta = 0.8$, the teacher receives the fully clean input $\mathbf{x}_1$ whenever $s > 0.2$. There, its predictions become nearly one-hot, so the distillation target of SERF-P adds little beyond the ground-truth label, whereas SERF-L aligns hidden representations, which remain informative even for clean inputs.

Effect of Student and Teacher Layer choice. In Fig. 5 (Right), we study which layers to align in SERF-L, denoting by $k_s \to k_t$ the alignment of student layer k_s to teacher layer k_t, with all models trained on OWT for 50K steps. The choice of layers has a clear effect. Aligning an early student layer to a late teacher layer ($2 \to 10$) gives the highest MAUVE across the whole entropy range, followed by aligning two intermediate layers ($6 \to 8$), while the reversed configuration ($10 \to 2$), which aligns a late student layer to an early teacher layer, performs worst. This is consistent with prior work on image generation (Chefer et al., 2025; Jiang et al., 2025), which observed that aligning an early layer of the student to a later layer of the teacher is beneficial: it encourages the student to form, already in its early layers, the higher-level representations that the teacher builds deeper in the network.

Effect of the self-regularization weight λ. Fig. 5 (Middle) reports MAUVE against unigram entropy for SERF-L and SERF-P under different regularization weights. The two variants respond to λ in opposite directions: SERF-L benefits from a larger weight (0.336 at $\lambda = 2.0$ vs. 0.321 at $\lambda = 1.0$), whereas SERF-P degrades as the weight grows (0.320 at $\lambda = 0.1$ vs. 0.259 at $\lambda = 0.5$). This reflects the scale of the two alignment losses. The cosine distance used by SERF-L is bounded and much smaller in magnitude than the cross-entropy of $\mathcal{L}_{\text{FLM}}$, so a larger λ is needed for it to meaningfully affect training. In contrast, SERF-P employs a cross-entropy loss on the same scale as $\mathcal{L}_{\text{FLM}}$, thus increasing λ shifts the effective target away from the ground truth and toward the teacher's prediction.

5 Conclusion

In this work, we introduced SERF, a simple self-regularization framework for continuous diffusion language models. Building on the observation that an EMA teacher evaluated at a cleaner point on the same interpolation path provides informative targets for a noisier student, SERF augments the standard training objective with a noisy-to-clean self-distillation loss that requires no external models and leaves the architecture and sampling procedure unchanged. We studied two instantiations: SERF-L, which distills intermediate hidden representations, and SERF-P, which distills output token distributions. Across three baselines, SERF consistently improves generation quality. On OWT, both variants match or exceed FLM trained for 1M steps using roughly a tenth of the training budget, with gains that persist in the few-step regime; on GSM8K, SERF improves the accuracy of S-FLM and outperforms prior continuous diffusion language models. Overall, these results suggest that the denoising trajectory is a rich source of self-supervision for language, and that simple self-distillation along it is an effective and practical way to exploit it.

AI USE STATEMENT

We used large language models (LLMs) as writing assistants during the preparation of this manuscript, primarily to polish the language, improve clarity and concision, and help with LaTeX formatting. All scientific content, including the method, experimental design, results and their interpretation, was conceived and produced by the authors.

ETHICS STATEMENT

Generative models carry a substantial risk of misuse. Their application can lead to various negative societal impacts, most notably the spread of disinformation. Enhancements in generative performance, as achieved by our method, may further increase the realism of generated content, potentially making disinformation even more convincing.

REPRODUCIBILITY STATEMENT

SERF is fully specified in Sec. 3.2 and Algorithms 1 and 2. All hyperparameters, as well as the architecture, training and sampling details of each baseline, are reported in Sec. 4.1 and the appendix. All experiments use publicly available datasets, and we will release our code and checkpoints upon publication.

REFERENCES

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

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025. 2, 3

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2025. 2

Chen-Hao Chao, Wei-Fang Sun, Hanwen Liang, Chun-Yi Lee, and Rahul G Krishnan. Beyond masked and unmasked: Discrete diffusion models via partial masking. *arXiv preprint arXiv:2505.18495*, 2025. 3

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2025. 2, 3, 9

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. *Advances in Neural Information Processing Systems*, 37:54787–54819, 2024. 3

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 7

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İ Ceylan, Michael Bronstein, and Avishek J Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024. 3

Justin Deschenaux and Caglar Gulcehre. Language modeling with hyperspherical flows. *arXiv preprint arXiv:2605.11125*, 2026. 1, 3, 7, 8, 14, 16

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022. 2

Antonio Franca and Alexander Tong. Hacking generative perplexity: Why unconditional text evaluation needs distributional metrics. *arXiv preprint arXiv:2606.08417*, 2026. 7

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019. 7

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 3

Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023. 3

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

Yasaman Haghighi, Bastien van Delft, Mariam Hassan, and Alexandre Alahi. Layersync: Self-aligning intermediate layers. *arXiv preprint arXiv:2510.12581*, 2025. 2, 3

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020. 3

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026. 3, 7, 8, 14

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 2, 3, 9

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. *arXiv preprint arXiv:2502.11564*, 2025. 3

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/. 14

Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, et al. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms. *arXiv preprint arXiv:2510.04767*, 2025. 2

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024. 2

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 14

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. *arXiv preprint arXiv:2504.16064*, 2025. 2, 3

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026. 1, 2, 3, 4, 6, 7, 14, 16

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025. 3

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343, 2022. 3

Yaron Lipman, Ricky T.Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 2

Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. Tinygsm: achieving> 80% on gsm8k with small language models. *arXiv preprint arXiv:2312.09241*, 2023. 8

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023. 2, 3

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36: 56998–57025, 2023. 3

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 1, 3

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023. 14

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021. 7

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Candi: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*, 2025. 3, 16

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 7

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 14

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael M. Bronstein, Max Welling, .Ismail .Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. *ArXiv*, abs/2602.12233, 2026. URL https://api.semanticscholar.org/CorpusID:285540743. 3, 6

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM. 1, 2, 3, 16, 17

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025. 3, 16

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024. 2, 3, 17

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025. 3

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021. 2

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html. 2

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1185–1194, 2025. 3

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, et al. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*, 2022. 3

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 14

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. 3

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *Advances in Neural Information Processing Systems*, 2025. 2, 3

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. 2, 3

Huangjie Zheng, Shansan Gong, Ruixiang Zhang, Tianrong Chen, Jiatao Gu, Mingyuan Zhou, Navdeep Jaitly, and Yizhe Zhang. Continuously augmented discrete diffusion model for categorical generative modeling. *arXiv preprint arXiv:2510.01329*, 2025. 3

Cai Zhou, Chenxiao Yang, Yi Hu, Chenyu Wang, Chubin Zhang, Muhan Zhang, Lester Mackey, Tommi Jaakkola, Stephen Bates, and Dinghuai Zhang. Coevolutionary continuous discrete diffusion: Make your diffusion language model a latent reasoner. *arXiv preprint arXiv:2510.03206*, 2025. 3

Appendix

A Additional Details

A.1 Flow Language Models

Architecture. For all FLM (Lee et al., 2026) experiments, we train a 179M-parameter Diffusion Transformer (DiT) (Peebles & Xie, 2023) with 12 transformer blocks, rotary positional embeddings (RoPE) (Su et al., 2021), and adaptive layer normalization (AdaLN) for time conditioning.

Training. We train on OWT with a batch size of 512. Following Lee et al. (2026), we use 2,500 warmup steps and then keep the learning rate fixed at 3×10^{-4}, using the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We additionally apply softcapping (Riviere et al., 2024) to the attention logits, which limits large values and improves numerical stability.

Sampling. We use an Euler solver followed by a per-position argmax. At each position l, the denoiser outputs a categorical distribution over the vocabulary, $\mathbf{p}_\theta^l(\mathbf{x}_s, s) = \mathrm{softmax}(\boldsymbol{\ell}_\theta(\mathbf{x}_s, s)^l)$, where $\boldsymbol{\ell}_\theta$ are the network logits, so that f_θ is the stack of the $\mathbf{p}_\theta^l$. We temper these with a temperature τ, $\mathbf{p}_\theta^{\tau,l} = \mathrm{softmax}(\boldsymbol{\ell}_\theta^l/\tau)$, and use the tempered distributions in place of $\mathbf{p}_\theta^l$ when forming the velocity field at *every* integration step. Values $\tau < 1$ sharpen the distributions towards the vertices of the simplex and yield lower-entropy samples, $\tau > 1$ increases diversity, and $\tau = 1$ recovers the untempered sampler. Sweeping τ thus traces a quality–diversity front.

A.2 Embedded Language Flows

Architecture. For all ELF (Hu et al., 2026) experiments, we use the ELF-B backbone, a 105M-parameter DiT (Peebles & Xie, 2023) with 12 transformer blocks (hidden size 768, 12 heads), RoPE (Su et al., 2021), and prefix tokens for time and self-conditioning. The model operates on T5-small sentence embeddings.

Training. We train on OWT with a batch size of 512. Following Hu et al. (2026), we use 9,508 warmup steps (half an epoch) and then keep the learning rate fixed at 2×10^{-3}, obtained by linearly scaling a base rate of 1×10^{-3} with the batch size. Training uses the Muon optimizer (Jordan et al., 2024).

Sampling. Following Hu et al. (2026), we employ an SDE-like sampler with 64 steps. Since ELF is trained with a randomly drawn self-conditioning guidance scale supplied as a conditioning token, the scale w is set at inference in a single forward pass: small w yields higher-entropy samples and large w sharpens them, so sweeping $w \in \{0.5, 1.0, 1.5, 2.0\}$ traces a quality–diversity front.

A.3 Hyperspherical Flow Language Models

Architecture. For all S-FLM (Deschenaux & Gulcehre, 2026) experiments, we train a 163M-parameter hyperspherical transformer (S-arch) with 12 blocks, RoPE (Su et al., 2021), and time conditioning. Embeddings and hidden states lie on the unit hypersphere $\mathbb{S}^{d-1}$, and all weight matrices are renormalized after each optimizer step.

Training. We train on TinyGSM with a batch size of 512. Following Deschenaux & Gulcehre (2026), we use 2,500 warmup steps and then keep the learning rate fixed at 3×10^{-4}, using the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Softcapping is not required: stability follows from the normalized architecture, which unit-normalizes queries and keys and renormalizes all weight matrices after each optimizer step.

Sampling. Following Deschenaux & Gulcehre (2026), we use an Euler solver. The problem statement is fixed to its clean embeddings and re-projected after every step, so the flow acts only on the answer. At each position, the denoiser outputs a categorical distribution over the vocabulary,

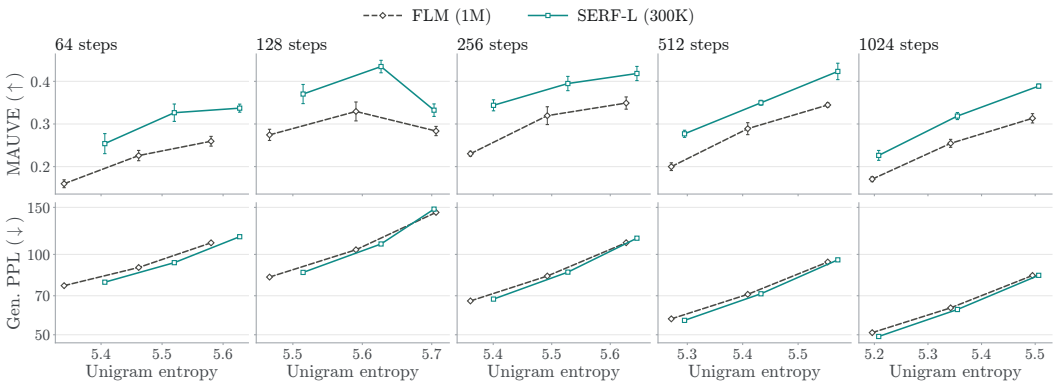


Figure 6: **Mauve and Gen.PPL fronts** for SERF−L at 300K steps and for the baseline FLM at 1M steps. Curves obtained by sweeping sampling temperature $\tau \in \{0.95, 1.0, 1.05\}$.

which determines the direction in which the current embedding moves; we truncate this distribution at every step to its single highest-probability token, so the embedding moves directly towards that token. We decode with an argmax after the final step.

A.4 Additional Details on SERF

Implementation Details Unless otherwise noted, the following settings are shared across all three baselines (FLM, ELF and S−FLM). In both variants of SERF, the teacher is an EMA of the model weights (decay $\alpha = 0.9999$), evaluated without gradients on the same noised sequence as the student but at a cleaner time. The regularization weight is $\lambda = 2.0$ for SERF−L and $\lambda = 0.1$ for SERF−P. For SERF−L, the output of student block $k_s = 2$ is passed through a two-layer MLP projection head ($768 \to 1536 \to 768$, SiLU) and aligned with the output of teacher block $k_t = 10$ using a per-token cosine distance. The projection head is used only during training and discarded at inference. For SERF−P, the alignment loss is enabled after an initial warm-up phase of 10K training steps for FLM and 20K for S−FLM. Training algorithms for SERF−L and SERF−P are presented in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1 SERF−L Training (Latent Distillation)

Require: Dataset $\mathcal{D}$, λ, α, blocks k_s, k_t
  Initialize: Student and Teacher f_θ, $f_{\bar\theta}$, $\bar\theta \leftarrow \theta$, projection head g_θ
  repeat
    $\mathbf{x}_1 \leftarrow \text{onehot}(\mathbf{y})$, $\mathbf{y} \sim \mathcal{D}$; $\boldsymbol{\epsilon} \sim \mathrm{N}(0, I)$
    Sample s, t ▷ See Sec. 3.3
    $\mathbf{x}_s \leftarrow s\mathbf{x}_1 + (1-s)\boldsymbol{\epsilon}$, $\mathbf{x}_t \leftarrow t\mathbf{x}_1 + (1-t)\boldsymbol{\epsilon}$
    $\mathbf{p}_\theta \leftarrow f_\theta(\mathbf{x}_s, s)$
    $\mathbf{z}_\theta \leftarrow g_\theta\big(\mathbf{h}_\theta^{(k_s)}(\mathbf{x}_s, s)\big)$, $\bar{\mathbf{z}} \leftarrow \mathrm{sg}\,\big(\mathbf{h}_{\bar\theta}^{(k_t)}(\mathbf{x}_t, t)\big)$
    $\mathcal{L}_{\text{FLM}} = -\frac{1}{L}\sum_l (\mathbf{x}_1^l)^\top \log \mathbf{p}_\theta^l$
    $\mathcal{L}_{\text{Distill}} = \frac{1}{L}\sum_l \text{dist}\,\big(\mathbf{z}_\theta^l, \bar{\mathbf{z}}^l\big)$
    Update θ: $\mathcal{L}_{\text{SERF}} = \mathcal{L}_{\text{FLM}} + \lambda\mathcal{L}_{\text{Distil}}$
    Update $\bar\theta \leftarrow \alpha\bar\theta + (1-\alpha)\theta$
  until converged

Computational overhead. SERF adds no cost at sampling time. The teacher and the SERF−L projection head are used only during training, and the model we sample from is the same network as the baseline. During training, each step needs one extra forward pass of the EMA teacher on the cleaner input. This pass runs without gradients and keeps no activations for the backward pass. A forward pass costs roughly a third of a forward-backward step, so SERF makes each training step $\sim 30\%$ slower as measured on our resources. The remaining additions (the EMA update of the teacher weights, the small projection head and the SERF loss) are negligible. Even accounting for

Algorithm 2 SERF−P Training (Probability Distribution Distillation)

Require: Dataset $\mathcal{D}$, λ, α
Initialize: Student and Teacher f_θ, $f_{\bar{\theta}}$, $\bar{\theta} \leftarrow \theta$,
repeat
  $\mathbf{x}_1 \leftarrow \text{onehot}(\mathbf{y})$, $\mathbf{y} \sim \mathcal{D}$; $\boldsymbol{\epsilon} \sim \mathrm{N}(0, I)$
  Sample s, t ▷ See Sec. 3.3
  $\mathbf{x}_s \leftarrow s\mathbf{x}_1 + (1-s)\boldsymbol{\epsilon}$, $\mathbf{x}_t \leftarrow t\mathbf{x}_1 + (1-t)\boldsymbol{\epsilon}$
  $\mathbf{p}_\theta \leftarrow f_\theta(\mathbf{x}_s, s)$, $\mathbf{q}_{\bar{\theta}} \leftarrow \text{sg}\left(f_{\bar{\theta}}(\mathbf{x}_t, t)\right)$
  $\mathcal{L}_{\text{FLM}} = -\frac{1}{L}\sum_l (\mathbf{x}_1^l)^\top \log \mathbf{p}_\theta^l$
  $\mathcal{L}_{\text{Distill}} = -\frac{1}{L}\sum_l (\mathbf{q}_{\bar{\theta}}^l)^\top \log \mathbf{p}_\theta^l$
  Update θ: $\mathcal{L}_{\text{SERF}} = \mathcal{L}_{\text{FLM}} + \lambda \mathcal{L}_{\text{Distill}}$
  Update $\bar{\theta} \leftarrow \alpha\bar{\theta} + (1-\alpha)\theta$
until converged

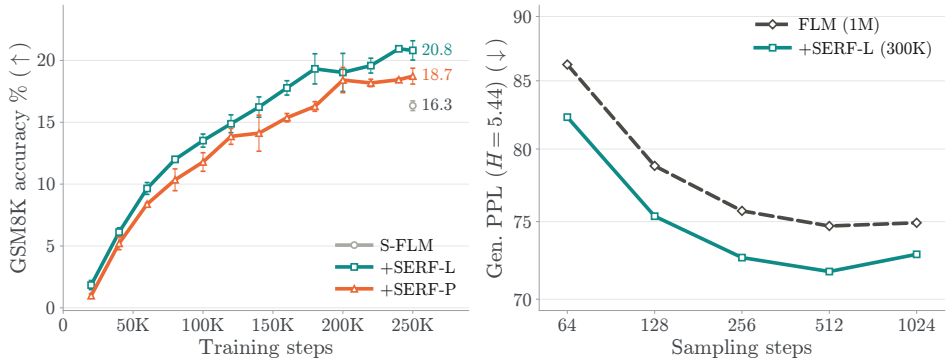


Figure 7: **Left:** Training convergence of SERF−L and SERF−P against the S−FLM baseline, trained on TinyGSM and evaluated by GSM8K accuracy with 32 sampling steps. **Right:** Gen. PPL of SERF−L (300K steps) and FLM (1M steps) on OWT across sampling steps, interpolated to the dataset entropy from a sweep over $\tau \in \{0.95, 1.0, 1.05\}$ (see Sec. A.5).

this per-step overhead, SERF−L reaches the quality of the baselines in much less wall-clock time, namely 7.7× faster for MAUVE, and Gen. PPL, and 1.4× for GSM8K accuracy.

Baselines Baseline results for FLM and S−FLM are taken from their respective papers. On GSM8K (Tab. 2), we compare against two discrete diffusion language models, MDLM (Sahoo et al., 2024) and DUO (Sahoo et al., 2025), and three continuous diffusion language models, CANDI (Pynadath et al., 2025), FLM (Lee et al., 2026) and S−FLM (Deschenaux & Gulcehre, 2026). All results in this table, including the autoregressive baselines, are taken from Deschenaux & Gulcehre (2026).

Computational Resources. For both training and sampling we use 8 NVIDIA GH200 96GB GPUs.

A.5 Additional Evaluation Details

Entropy-matched generative perplexity. Generative perplexity can be lowered simply by sampling at a lower temperature, which also lowers sample diversity. Comparing models at a fixed temperature is therefore misleading, so we report Gen. PPL at a fixed sample entropy instead. For each model we draw 5,000 samples at a small sweep of temperatures, $T \in \{0.95, 1.0, 1.05\}$. For each temperature we record the Gen. PPL and the mean per-sequence unigram entropy H. We then interpolate the sweep to a target entropy $H^\star = 5.44$, which is the entropy of OpenWebText validation text.

B Additional Experimental Results

Training acceleration on TinyGSM Figure 7 (left) tracks GSM8K accuracy over training, measured with 32 sampling steps and averaged over 4 evaluation seeds. Both SERF variants learn faster than

the S-FLM baseline. SERF-L reaches the accuracy of the released S-FLM checkpoint, trained for 250K steps (16.3%), after only 140K steps (16.2%), that is, with $1.8\times$ fewer training steps. SERF-P does so after 180K steps ($1.4\times$ fewer). At 250K steps, SERF-L reaches 20.8% and SERF-P 18.7%, respectively 4.5 and 2.4 points above S-FLM. Taking into account SERF's per-step overhead of roughly 30%, SERF-L still matches the baseline $1.4\times$ faster in training time.

Quality across sampling budgets. We compare SERF-L trained for 300K steps against the released FLM checkpoint trained for 1M steps, for 64 to 1024 sampling steps. Fig. 7 plots MAUVE (top) and Gen. PPL (bottom) against unigram entropy over a sweep of sampling temperatures. Despite using less than a third of the training, SERF-L achieves higher MAUVE than FLM at every sampling budget and across the whole entropy range, reaching 0.435 at 128 steps and 0.423 at 512 steps, compared to 0.329 and 0.345 for FLM. It also degrades less when the number of sampling steps is reduced: with only 64 steps, it reaches 0.337, above the best MAUVE of FLM at 1024 steps (0.313) and close to its best at any budget (0.349 at 256 steps). These gains come without any loss in fluency. The Gen. PPL fronts of SERF-L lie at or slightly below those of FLM at matched entropy, with the gap largest at 64 steps. Interpolated to the data entropy ($H = 5.44$), SERF-L achieves lower perplexity at every sampling budget. With 128 steps it reaches 75.4, nearly matching FLM at 512 and 1024 steps (74.7 and 74.9), and with 256 steps its perplexity of 72.6 is already below that of FLM at any budget, a $\sim 4\times$ reduction in sampling steps.

C Qualitative Comparisons

We present qualitative comparisons with the baseline FLM in Fig. 8, Fig. 9, and Fig. 10. For each comparison, we use the same initial Gaussian noise and vary only the trained denoiser, generating samples with 1024 sampling steps. Since Euler sampling is deterministic given the initial noise, this setup allows a direct comparison between the two learned flows. As shown in the examples, SERF produces more coherent and grammatically well-formed text. In contrast, the baseline FLM more often generates degenerate fragments, repeated symbols, and locally incoherent phrases.

D Limitations and Future Work

Our method is currently developed and evaluated for continuous flow-based language models. Extending the framework to discrete diffusion language models (Sahoo et al., 2024; Schiff et al., 2024) represents a promising direction for future work. Since discrete diffusion models follow different corruption and denoising dynamics than continuous flows, adapting the teacher supervision mechanism may require additional design considerations. We view this as an opportunity to further generalize our framework and broaden its applicability.

FLM MAUVE 0.255, Gen. PPL 52.8, $H = 5.29$

article that are from David Abramak's "**Worldest** speech at Mideast" in May. It's about $1,500,000 that the Bush Foundation is worth in March and you add to the $500,000 dollars that are really involved in the hands of the Bush Foundation.

Now, we have called them "top" **fundraees** and are just out of touch. They are in government — they're not being paid - the Constitution is up for grabs under this Obama — I guess you guys should be put together a "better" group than it does not exist.

One of the best **fundraees** is Col Ch Mazur, COO in the Obama administration. I've sat in on both IRD and APA. It's part of the "top 10" group, the **fewestest** of six in the United States. You can tell I have been invited to the top executives and **senioraries** in those organizations in those years. I am never going anywhere else there.

+SERF-L MAUVE 0.319, Gen. PPL 42.3, $H = 5.27$

that are just protecting themselves from politics."

Well, I'd add, in London. It's about $1,500,000 that means your salary is worth major dollars and you add to the $500,000 salaries that are really big in the history of the Middle Party.

Well, yes, they are government and they are corrupt and are just because of it. They are not government — they are not being controlled by the taxpayers. And you're wrong — I guarantee you nobody should be gone without a doubt. Anyone who have it does not exist.

One of the things they are is creating an extremely large, putting people in the Middle Office. I've served both DoD and MAGA. It's one of the largest institutions of Davos, the few remaining heads of state in the United States. My wife and I have been invited to the White House and I have never gone further in those years. I am never going to pounce. I never forget.

FLM MAUVE 0.255, Gen. PPL 62.6, $H = 5.33$

not have enough players to **justifyify** our very low-end. In only half of the 18-and-under rounds, we brought in a 4.0, so Spurs will need to. Like every ball, hand or foot, one of these star players gets a huge bonus. I've heard those numbers myself that you are the best team overall at the moment and wins over the 13 games have been improved from those from last year.

Here's what we take to get this rolling:

First you have a top scorer and finish. The guy on my board, and then the putts that also got into So much. As I was going into this year this way, as I'm going to get back on my **feetls** now and prepare for these years of holding out for another regular season next post and walking away, this thing is a we can set up.

Once you've had pretty well odds of being best in the final of the season, less run.

+SERF-L MAUVE 0.319, Gen. PPL 50.9, $H = 5.35$

not have two players to lose: our very low-end. In only half of the rest of our starting powerhouses, this roster is completely useless. However, neither player will need to line an opposing post, hand or pass, one of these is worth only a few dollars. I have to give you the simple feeling that you are the best team record at the moment and the best the 3 games have been free from this from last year.

Here's what we need to get this rolling:

If you have a top prospect and an elite starting pitcher on the roster, you shouldn't make an effort to win the division. I will put into this year this way, as I'm going to explain how these guys are playing young and how they've played out for another 2016 season. This is going to limit this thing of what we can set, this will not work like a club needs 10 seasons without getting.

FLM MAUVE 0.255, Gen. PPL 68.5, $H = 5.29$

Has **BobZW** done the deed, much he says belongs to the community? "It's the story of the car. In the beginning there was the example of a country builder and that guy that you know had to think of this important car to build the road. And he has been most successful with out that of Alla."

Slthrough (1 Images)

Now if you did not know there was a reference to the off-**thle** in California. Their story of course was an original and a very **uplful,** very interesting story. It was the day of the month's program the kids changed. These actually were much like ours when it was ready for a change that's coming. You have if you look for what a person was first. You saw in the BMW S-4 today." – Bill Bush was the subject of the panic attack ads during the Clinton administration, was in points regarding the extent of people's attention to you in the media we and therefore got the most well as

+SERF-L MAUVE 0.319, Gen. PPL 49.7, $H = 5.27$

the city, and he says goodbye to the community. And now there's the rest of the nation. In **Hong Francisco** there was the rest of the country, and he knows that. That doesn't leave him wanting to join the government. And he has been completely involved with the likes of Alla."

Sharethrough (1 Images)

What if you did not know there was a reason to allow off-throwing?

An example of this was an email and tweet because you said, "Hey! It was the social media storm and it hadn't happened yet. How else were they? Even when there was evidence for a change that's fine. You have choices you look for such a person to make. You might in the wind up make their **policy."It** certainly was the opposite of the Trump attack released during the Republican primaries, was in keeping with the reality of people's desire to elect pro-Trump members and give Americans the most well

Figure 8: **Generated samples on `OWT`.** Samples from `FLM` (left) and `SERF-L` (right), all trained for 1M and $300K$ steps respectively. Text generated with 1024 sampling steps from the same initial noise. Errors are highlighted in **bold**.

FLM MAUVE 0.255, Gen. PPL 52.0, $H = 5.21$

S. is exempt from **U.U.K.**

Not that jobs are the things we've only been asked do to answer as to why they haven't **turn turned** into record class. "But I was just dealing with either a lot of moving into states like Texas and Karnes, or I'm being facet honest, and a lot of things don't know how big we want them to be for jobs and jobs, if we're cutting jobs or creating jobs and stuff."

"... and a lot of folks came into fast enough and made sure we live in a lot of places, and we've learned that people work with confidence and they're it with confidence that they will and have the same."

A key lesson this administration is going to be a story that you never hear about. "What I hope you know what work is if it isn't successful is the hard thing. Somebody told me of Kansas City full school, and we know it many times.

+SERF-L MAUVE 0.319, Gen. PPL 40.7, $H = 5.19$

S. is learning from Israel. In other words.

All that matters are the things we've only been waiting for to understand as to why they aren't all translated into science and Afghanistan. Well, I was just tasked with pouring a lot of money into states like California and Florida. It seems I'm being near them, and of all places I don't know how big we want them to be for tomorrow and today, if we're going to enter into tomorrow and beyond."

"They've turned many dollars into developing science and beyond; we live in a number of countries, and we've shown that people work with technology and they're born with family and technology — and have the same."

A conversation about this topic is going to be a story that you never get about. "Well I guess you know what education is if you aren't going to create that thing.

FLM MAUVE 0.255, Gen. PPL 70.8, $H = 5.35$

government rating is a different, entirely new thing: We must rate Barack Obama a 10th this year."

For Mr. Romney's **sake sake,** let's be a likely **motivite.** The fact is: the numbers are incalculably below the party's highest levels, and may not be the same as if it was a talking and talking point. But Democrats should focus on a bigger plan: improving U.S. State Dept.

Keep in mind, the highest Democratic national intelligence staff of 14 – appointed by Clinton's will of Congress – were on the top ten in 2012. That is a well over the top ranking by the people even in disreputable terms, and Democrats will have to do the more exacting, and analysis of the Obama administration's **latestak.**

But wait.

Just because every White House official is likely to run "a red search," this analysis of recent nominees for what they have overrated crystal clear: from what Obama supporters are in store in the White House-based press Office earlier this year: trust in public....

+SERF-L MAUVE 0.319, Gen. PPL 57.1, $H = 5.34$

is a report, "Global Outlook: We must brace for "financial highs' this **year."What** a **mind-off** is a sobering moment, which is the year of a person winning the Democratic Party last month: the Republicans are claiming they couldn't accept the party's federal platform, and did that be the same as Mitt Obamacare?

She was right. Democrats led the plug on a state plan to fund U.S. Statehood.

But in reality, the tiny little national that Donald Trump represents – driven by Clinton's will of heart – won on the top ballot in November. That is a huge blow to the lives of the people who're from California, and Democrats will have to do something more sobering, and worthy of the Obama administration's immigration policies.

But plan.

Calling Clinton every day they should ask her to run "a presidential race," the majority of people know for what they have over the immigration issue: from what Obama makes post in speeches in the White House-Senate press Office earlier this year as speaking in

FLM MAUVE 0.255, Gen. PPL 80.3, $H = 5.50$

voting for Trump and knows "what it's doing with her relationship with Russia" (which she has denied). It's not much that Trump's lawyer is seeking "all media information" to **avoidize** a meeting with Trump (if trying to set this up with him he would have abd about that).

Elisabeth Moss isn't the only ones telling rumors that she will be "directly against Trump's house" for one of the billionaire's contestants as young as the star of his show are going to be the sort of y' **cackgel** that haters are in that many. But thanks to his tweets, despite his tweets, **slimest** down memories of a **ratherest** and completely self-**righteousathing** that at anytime he's gone through the nation's social media to say or write anything publicly, and the vast amount of actual hard drives and videos of this latest tangle of liberal politics, there are a couple of reasons to go with them.

+SERF-L MAUVE 0.319, Gen. PPL 63.7, $H = 5.47$

for Trump and Ivanka "whether it's consistent with her relationship with Putin" (which she has had). It's not much that Trump's behavior is attractive to others, while Musk's just sharing a meeting with Trump (**nob'd** actually brought this up ... but he would have been happy about that).

"I it doesn't matter how we then that these allegations will be lit upon other allegations against Trump's foreign policy as for this one. He's not as young as the rest of his supporters are going to be concerned, there can't be exceptions, but I am a real man. But I have his voice, had his courage, he will make sense of a real sense and greater self-defense of violence against me."

Musk had a chance on his own to say his comments were important, and the specter of freedom of speech and lack of "lots of interest" is just a matter of trying to go with them.

"We remain ourselves to members of all others.

Figure 9: **Generated samples on `OWT`.** Samples from `FLM` (left) and `SERF-L` (right), all trained for 1M and $300K$ steps respectively. Text generated with 1024 sampling steps from the same initial noise. Errors are highlighted in **bold**.

FLM MAUVE 0.255, Gen. PPL 54.3, $H = 5.25$

not making diplomacy work, Pentagon **Secretaryne** Davis said in an interview Tuesday.

"And how the law does ... and is that what we do? It just happens," he asked.

"It is, instead for not pointing out things I don't understand."

That can "**bringame** to a dangerous level that we should do." And we are saying right, at one point, she said. "Let's get rid of these problems, the fact that the woman is not on the other side. And to do whatever we do has the responsibility to not leave it in the same place."

Advertisement:

How did this woman **misize** the American flag in your opinion?

Remember the mistake of using questions? Because I understand my staff's not to put the law into a box and stick to it **viaically** and **condescically** – I was questioning the importance of the most important test. It's clear clearly here and we know we're not American exceptionalist anymore and without looking like our enemy.

+SERF-L MAUVE 0.319, Gen. PPL 44.0, $H = 5.22$

not "right," Judge Snowden said in an **interview interview.**

"Just how the law should apply and say, what we do? Which one legal?" he asked.

"It is, they are not left out. I don't care."

I doubt it!

Anybody can see that we should do.

Since we are absolutely right, at one point I had to say "don't get rid of a word," which would give someone else on the other side to have to do the talking and use this word to not distinguish it from "I."

Questions:

How did this particular Amendment affect the American government without any oversight?

Using the Constitution means using force?

The way we're not to put the law into a position to talk to?

Well, I believe ... while respecting the authority of the government and breaking up government. When discussing the threat of overreach, we're not using you indiscriminately and without people like our enemies.

FLM MAUVE 0.171, Gen. PPL 47.8, $H = 5.14$

said. "Right now the program is going to really push you to get here – building a great example of the players that make the best of the country."

Thiad could go for a U.S. f-bomb if not thrown in there, but the club will also get other U.S. offers, or going hang out with friends or be with their "wonderful" tics at this point.

"The Best U.S. Dream" came onto the back of his return to the NBA in 2012.

Git didn't adjust to that pre-era form as a kid. It was just as familiar to the old play style as could be from the Canadian Coast era.

Sure, he spent considerable time building his own roster, but the coach stepped in to help bring his teammates in on hand. **Withoutending** more time, interest really began to shine.

Dillon, the executive vice president of basketball operations working with McLillan, believes their team could grow after having the players bring in Bono, who isn't just the team's home country of

+SERF-L MAUVE 0.226, Gen. PPL 37.3, $H = 5.14$

said. "Right now the program is going to really allow you to get here and do a quick tour of the general population and the demographics of the country."

It is easy to go for a U.S. **mini-out** if someone gets in there, but the fun will also be other U.S. citizens, or going hang out with friends or be with their "welcome and **"friends"** Donald Trump.

His visit to the U.S. was certainly part of the general theme of his visit to the presidency in January.

After the third debate, he probably realized that President Barack Obama his office space on Capitol Hill was just as welcoming to the old-school border could be as the event they were.

He's spending time on his own turf, but the real tuned in to Trump during his swearing in on Friday morning and as more importantly, personal questions began to emerge.

John Carter, the South American president of President-elect with ties to Mexico, suggested that Mexico could happen after having the biggest leaders in Puerto countries, who couldn't enter the candidate's home country of

FLM MAUVE 0.171, Gen. PPL 55.3, $H = 5.29$

calling for governors of the states to go to hell with each other, right? I think they have done enough work for him and have also **lengiqued** the people life. I told Bill when they read the bill and when he finally found out he said, "It's better for [health costs] to go up because the state system ... has been on them. And also it's more for me to be able to work and afford to get hit up because as a congressman ... I was on a military office space during the recession I did not understand the ability of more people to give their office a proper wage for it. I'm proud to be a citizen of the United States: an official living in food or not one – in Dadene, the number of people living are probably hundreds of miles away.

When you think about an hour four further back, step back, and the staff members who are there have told Congress that they are in

+SERF-L MAUVE 0.226, Gen. PPL 44.2, $H = 5.25$

for governors of the states to go to war with each other, but overall I think they have done it work for me and have also worked with Bernie's. I'm surprised when they mentioned the economy and when he and throughout his campaign argued that "It's better for people to worry about maybe two wars because we're too focused on focusing on them. And maybe it's more people willing to be able to work and fight to get that right because having a child whom I was on a military mission" during the Iraq War did not address the ability of more people to give their states a fair reason for it. I would prefer to raise a couple of bread and make everybody an entire increase in food or not one – in fact that would reduce the number of people, thousands and thousands of miles away.

When you think about your position on further immigration, cutting taxes, and the fact that there are no real trade solutions that they don't think add up they're picking the **right and right** results.

Yes.

Figure 10: **Generated samples on `OWT`.** Samples from `FLM` (left) and `SERF-L` (right), all trained for 1M and $300K$ steps respectively. Text generated with 1024 sampling steps from the same initial noise. Errors are highlighted in **bold**.