ChronoLingua: Revive Ancient Chinese via Machine Translation

Team member:
Letian Yu, Xiner Zhao, Yuyang Luo, Zhengyang Xu
Department of Computer Science & Data Science Institution

May 6th, 2024

Introduction

Cultural and Historical Significance

Translation Challenges

- Scarcity of comprehensive sentence-aligned corpora
- Disparities in tokenization, word order, and syntax

Goals

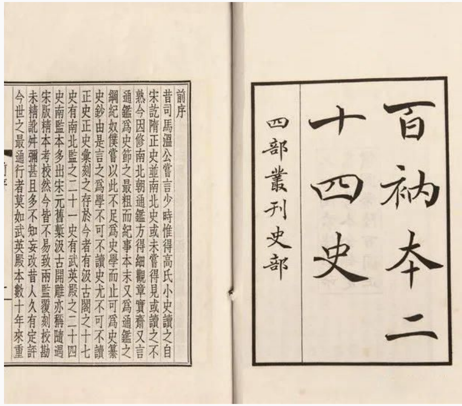
- A robust solution capable of translating complex ancient Chinese scripts into modern languages!
- Utilize the power of pretrained models

Methodology

Methodology: Datasets

Ancient Chinese to Modern Chinese:

- parallel texts of China Twenty-four Histories (二十四史)
- 300, 000+ sentences
- E.g., {"src": "是秋，取城邑凡八百六十有二。",
  "trg": "這年秋季，攻取城鎮共計八百六十二座。"}



~~Ancient Chinese to English:~~

- parallel texts of Pre-Qin classics (先秦经典) and Zizhi Tongjian (资治通鉴)
- 5,000+ sentences
- Finally abandoned given the small data size

Methodology: Preprocessing

Train-Test Splitting:

- 80% - 20%

Tokenization & Embedding:

- Pretrained model for tokenization & Word Embedding
- overcome challenges of tokenization (e.g., HanLP) and small corpora

Others:

- Padding, Start & End, Separator

Methodology: Model Architecture

Baseline:

- RNN: LSTM Model
- Transformer

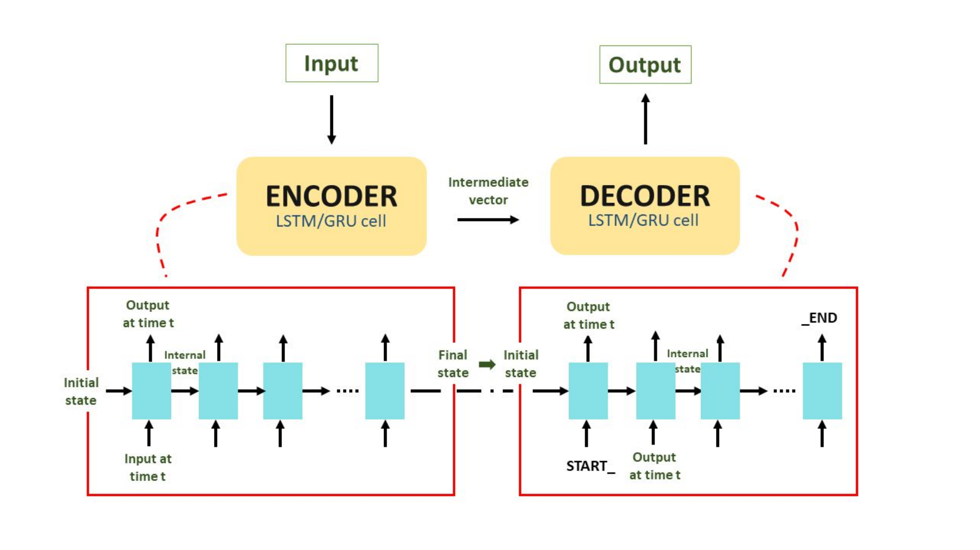
Target:

- Sequence to Sequence Diffusion Model with Transformer

Innovation:

- Applies **pre-trained word embedding**
- Diffusion model for **Ancient Chinese** Translation
- **Unconditional** sequence-to-sequence **diffusion** model for **machine translation** task with **transformer** encoder

RNN



Encoder: LSTM Layer

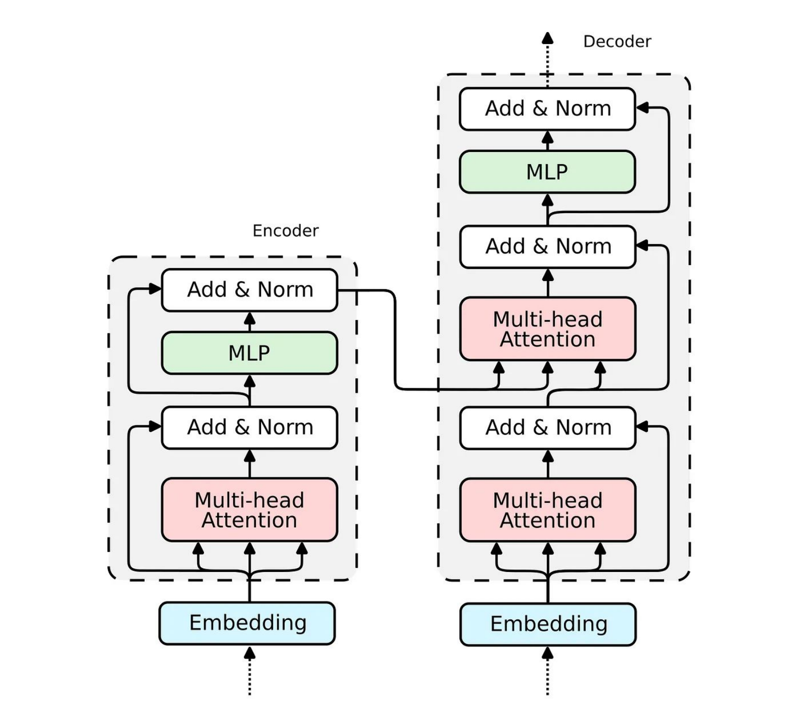
- Input: 1. Ancient Chinese words' embedding
  2. Initialized hidden and cell state
- Output: Hidden state & cell state

Decoder: Embedding + LSTM + Linear Layer

- Input: 1. Modern Chinese words' embedding
  2. The hidden state and cell state from the output of encoder
- Output: Logits over the target language vocabulary

Ref: https://www.business-and-ai.com/recurrent-neural-network-algorithms-overview/

Transformer



Encoder Input:

Tokenized Ancient Chinese text

Positional Encoding:

Preserve the order of the input tokens

Multi-Head Attention Mechanism:

– Calculate the focus levels of each word on others in the sentence

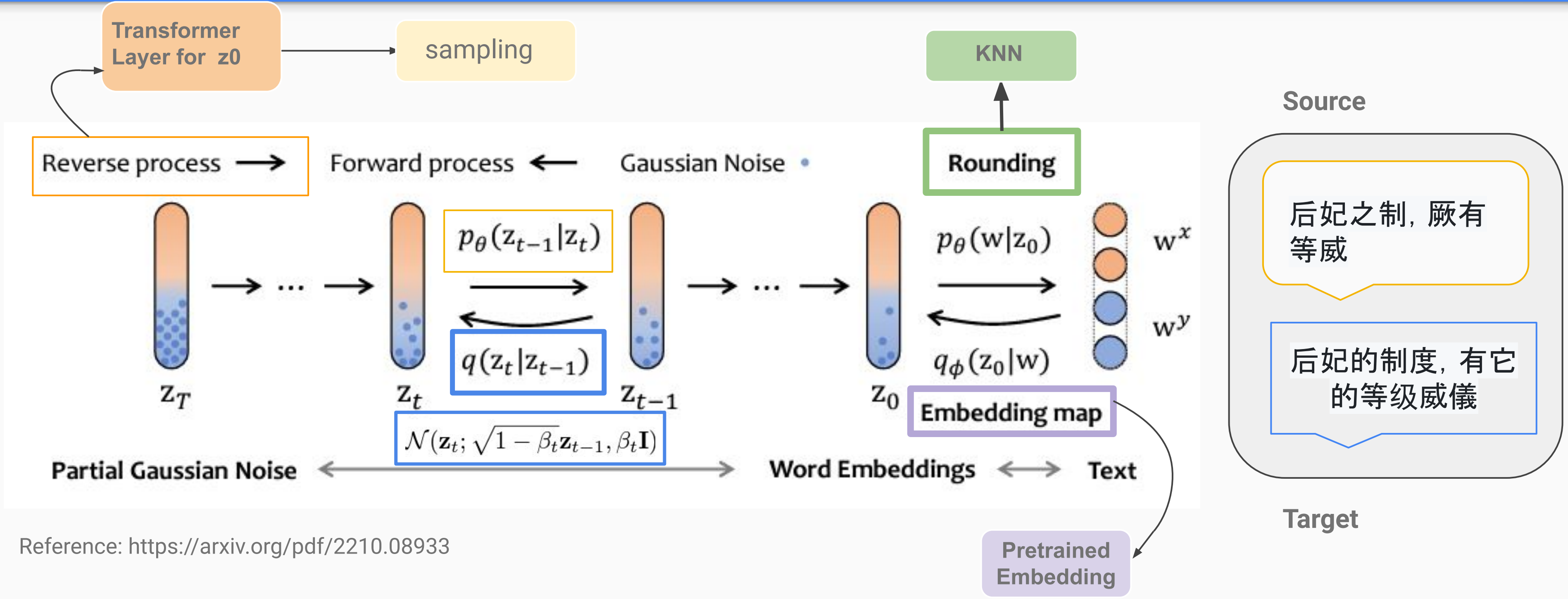
– Pay attention to different parts of the input at the same time

Decoder Output:

Modern Chinese text

Ref: https://towardsdatascience.com/beyond-transformers-with-pyneuralogic-10b70cdc5e45

Diffusion



Reference: https://arxiv.org/pdf/2210.08933

Diffusion

Forward Process: with Noise Mask

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I})$$

Loss Function:

$$\min_\theta \left[\sum_{t=2}^{T} ||\mathbf{y}_0 - \tilde{f}_\theta(\mathbf{z}_t, t)||^2 + ||\textsc{Emb}(\mathbf{w}^y) - \tilde{f}_\theta(\mathbf{z}_1, 1)||^2 + \mathcal{R}(||\mathbf{z}_0||^2) \right]$$

Reverse Process: Conditional Denoising

$$p_\theta(\mathbf{z}_{0:T}) := p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t), \quad p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t))$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$$

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

Evaluation

	BLEU	CHRF
RNN	0.162	0.081
Transformer	3.385	0.157
Diffusion	1.346	0.107

Example Result:

Source : 庚寅，以疾愈大赦天下。

Target : 庚寅日，因疾病痊愈大赦天下。

RNN trans : <UNK> <UNK>， 因 疾 愈 <UNK> 免 天 下。

Transformers trans : 十 三 日 ， 因 病 愈 ， 赦 免 天 下 。

Diffusion trans : 庚 寅，因 病 愈 大 赦 天 下 。

ChatGPT :): 庚寅年，皇帝康复后对全国实行大赦。

Noisy results

"recover": "32 谽 瓏 暟 54 @ 殼 狨
珷 睘 戭 燇 薛 槮 驔 鄉 餞 鑒 僡 鐀 藴
噝 諿 盯 愵 業 軀 靄 # 鶑 璉 r 槮 哽
鶑 啉 槮 俠 紅 泭 蠅 庤 偉 甌u 儼 橰 穭
禃 轐 桊 槮 鬌 僡 餌 揄 賂 篙 冪 猓 鷺
魟 槮 腎 餡 愵 甌 旎1u 蔣 祿 蕪 块
1979 拒 膚 ρ 縝 楝 愵 鵜 攏 檜 惎 槮
鉀 儼 羧 譆 瓏 鑼 揕 顱 鑼 餡 鈴 疍 餡
穭 槮 唭 烯 餅 餻 礎 禃 h 硿 鶑 蟟 镉
韡"

"reference": "[CLS] 升 任 延 州 知 州
兼 鄜 延 駐 泊 部 署 。 [SEP]",

"source": "[CLS] 徙 知 延 州 兼 鄜 延
駐 泊 部 署 。 [SEP] [SEP]"

Challenge & Future Work

Challenge:

1. Computation Resource Limit
    a. Limited Training Time
    b. GPU Power
    c. Large Dataset

2. Noisy output
    a. Probably due to punctuations and unknown characters
    b. Large vocab size using pre-trained embedding

Future:

- Remove punctuations in sentence with certain probability

- Train on larger dataset

- Tune Hyperparameters

- Train embedding matrix

Thank you!
Q & A

Team member:
Letian Yu, Xiner Zhao, Yuyang Luo, Zhengyang Xu
Department of Computer Science & Data Science Institution

May 6th, 2024

Reference:

Gong, S., Li, M., Feng, J., Wu, Z., & Kong, L. (2022). Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.

Nichol, A. Q., & Dhariwal, P. (2021, July). Improved denoising diffusion probabilistic models. In *International conference on machine learning* (pp. 8162-8171). PMLR.

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yuan, H., Yuan, Z., Tan, C., Huang, F., & Huang, S. (2022). Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*.

Li, Xiang, et al. "Diffusion-lm improves controllable text generation." *Advances in Neural Information Processing Systems* 35 (2022): 4328-4343.

Appendix: Diffusion

Forward Process:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \text{ Let } \alpha_t = 1 - \beta_t \text{ and } \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i,$$

$$\begin{aligned} \mathbf{z}_t &= \sqrt{\alpha_t}\mathbf{z}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{z}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} \\ &= \ldots = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \end{aligned}$$

Reference: https://arxiv.org/pdf/2210.08933

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

(Original DDPM Paper)

Reference: https://arxiv.org/pdf/2006.11239

Appendix: Diffusion

Reverse Process:

$$p_\theta(\mathbf{z}_{0:T}) := p(\mathbf{z}_T)\prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t), \quad p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t))$$

Reference: https://arxiv.org/pdf/2210.08933

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

(Original DDPM Paper)

Reference: https://arxiv.org/pdf/2006.11239

Appendix: Diffusion

ELBO, Original Loss Function, Sampling:

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] =$$

$$\mathbb{E}_q\left[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \,\|\, p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\right]$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$$

Reference: https://arxiv.org/pdf/2006.11239

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

Appendix: Diffusion

Similarly, our loss:

$$\mathcal{L}_{\text{VLB}} = \mathcal{L}_T + \mathcal{L}_{T-1} + \cdots + \mathcal{L}_0 = \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{z}_0)}\left[\log \frac{q(\mathbf{z}_T|\mathbf{z}_0)}{p_\theta(\mathbf{z}_T)} + \sum_{t=2}^{T} \log \frac{q(\mathbf{z}_{t-1}|\mathbf{z}_0, \mathbf{z}_t)}{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)} + \log \frac{q_\phi(\mathbf{z}_0|\mathbf{w}^{x\oplus y})}{p_\theta(\mathbf{z}_0|\mathbf{z}_1)} - \log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}_0)\right].$$

Reference: https://arxiv.org/pdf/2210.08933

$$\mathcal{L}_t = \mathbb{E}_{\mathbf{z}_0}\left[\log \frac{q(\mathbf{z}_t|\mathbf{z}_0, \mathbf{z}_{t+1})}{p_\theta(\mathbf{z}_t|\mathbf{z}_{t+1})}\right] = \mathbb{E}_{\mathbf{z}_0}\left[\frac{1}{\mathcal{C}}||\mu_t(\mathbf{z}_t, \mathbf{z}_0) - \mu_\theta(\mathbf{z}_t, t)||^2\right]$$

$$= \mathbb{E}_{\mathbf{z}_0}\left[\frac{1}{\mathcal{C}}||\,\mathcal{U}\mathbf{z}_t + \mathcal{E}\mathbf{z}_0 - (\mathcal{U}\mathbf{z}_t + \mathcal{E}f_\theta(\mathbf{z}_t, t))||^2\right] = \frac{\mathcal{E}}{\mathcal{C}}\mathbb{E}_{\mathbf{z}_0}[||\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t)||^2],$$

Appendix: Diffusion

Similarly, our loss can be simplified as: we use modified loss from original paper

tT loss

Reference: https://arxiv.org/pdf/2210.089

$$\min_{\theta} \Bigg[||\mu(\mathbf{z}_T)||^2 + \sum_{t=2}^{T} ||\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t)||^2 + ||\mathrm{E{\scriptsize MB}}(\mathbf{w}^{x\oplus y}) - f_\theta(\mathbf{z}_1, 1)||^2 - \log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}$$

$$\rightarrow \min_{\theta} \Bigg[\sum_{t=2}^{T} ||\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t)||^2 + ||\mathrm{E{\scriptsize MB}}(\mathbf{w}^{x\oplus y}) - f_\theta(\mathbf{z}_1, 1)||^2 - \log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}_0) \Bigg]$$

$$\rightarrow \min_{\theta} \Bigg[\sum_{t=2}^{T} ||\mathbf{y}_0 - \tilde{f}_\theta(\mathbf{z}_t, t)||^2 + ||\mathrm{E{\scriptsize MB}}(\mathbf{w}^{y}) - \tilde{f}_\theta(\mathbf{z}_1, 1)||^2 + \mathcal{R}(||\mathbf{z}_0||^2) \Bigg].$$

MSE Loss

NLL loss

Appendix: Diffusion Parameters & Training

```
--diff_steps 200
--lr 0.001
--learning_steps 3000
--noise_schedule sqrt
--hidden_dim 768
--hidden_t_dim 768
--bsz 2048
--seq_len 128
```