ELF: Embedded Language Flows

Keya Hu* **Linlu Qiu*** **Yiyang Lu** **Hanhong Zhao**
Tianhong Li **Yoon Kim** **Jacob Andreas** **Kaiming He**

MIT

*Equal contribution; order decided by a coin flip.

Code: https://github.com/lillian039/ELF

MOAI Lab

2026.05.21

Suchan Lee

Overview

- Kaiming He group에서 발표한 continuous diffusion language model 논문
- 제약 없는 임베딩 공간에서 rectified flow를 적용한, 별도의 decoder가 필요 없는 continuous DLM
- 작년까지 주를 이루었던 MDLM 계열 모델들에 비해 좋은 성능을 보임

MDLM (Masked Diffusion Language Model)

$$Q_t = \begin{array}{c} \\ \text{고양이} \\ \text{강아지} \\ \text{여우} \\ \text{MASK} \end{array} \begin{pmatrix} \text{고양이} & \text{강아지} & \text{여우} & \text{MASK} \\ 0.9 & 0 & 0 & 0.1 \\ 0 & 0.9 & 0 & 0.1 \\ 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 1.0 \end{pmatrix}$$

$$q(\mathbf{z}_t|\mathbf{x}) = \text{Cat}(\mathbf{z}_t; \bar{Q}_t\mathbf{x}) = \text{Cat}(\mathbf{z}_t; Q_t \cdot Q_{t-1} \cdots Q_1\mathbf{x})$$

[Training]

나는 여름에 강아지와 산책을 하는 걸 좋아한다

나는 [MASK]에 [MASK]와 [MASK]을 하는 걸 좋아한다

[Inference]

… 나는 [MASK] [MASK] … [MASK]

… 나는 [MASK] 산책을 … [MASK]

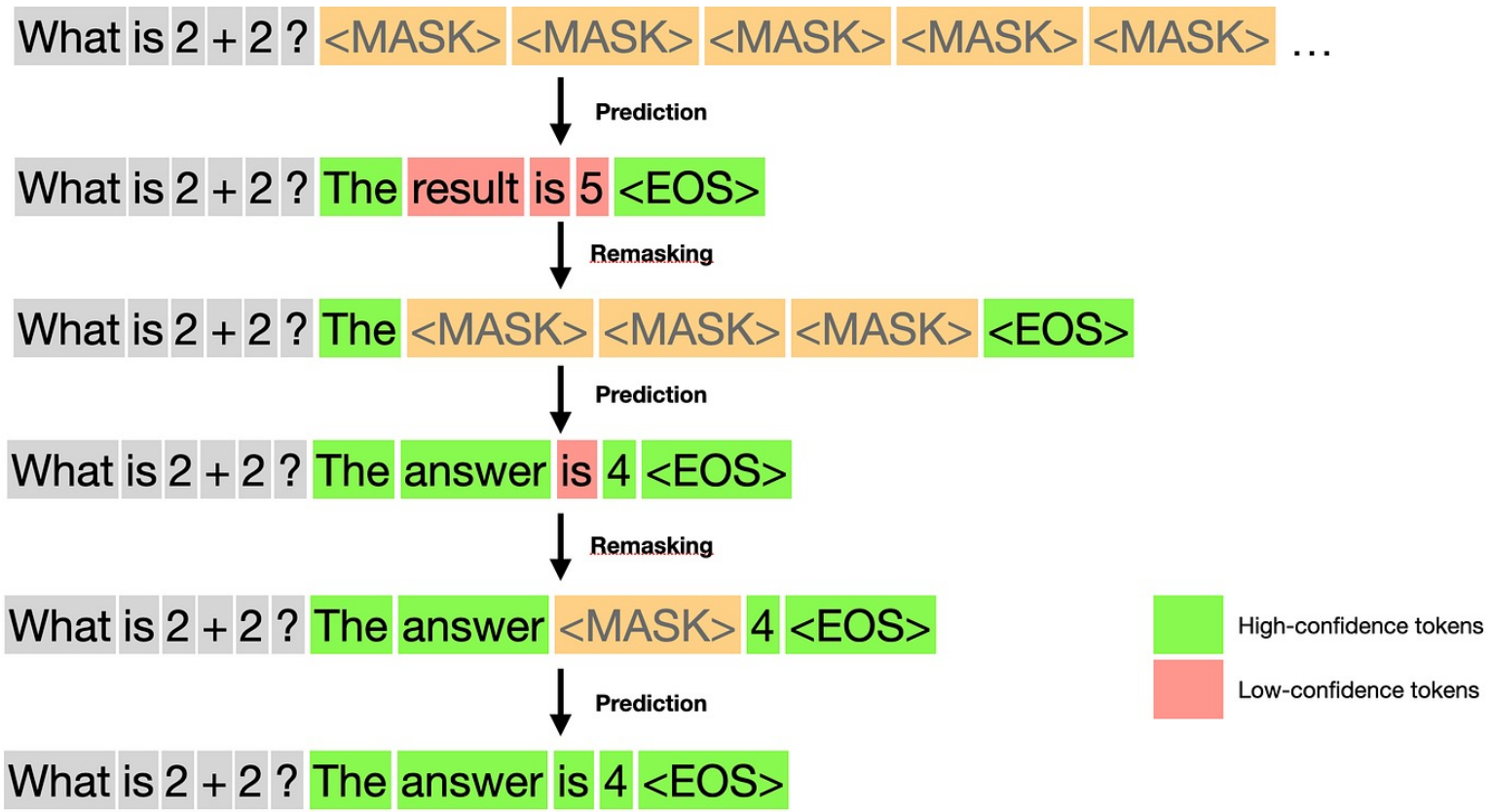
… 나는 [MASK] 산책을 … 좋아한다

… 나는 강아지와 산책을 … 좋아한다 [MASK] [MASK] … [MASK]

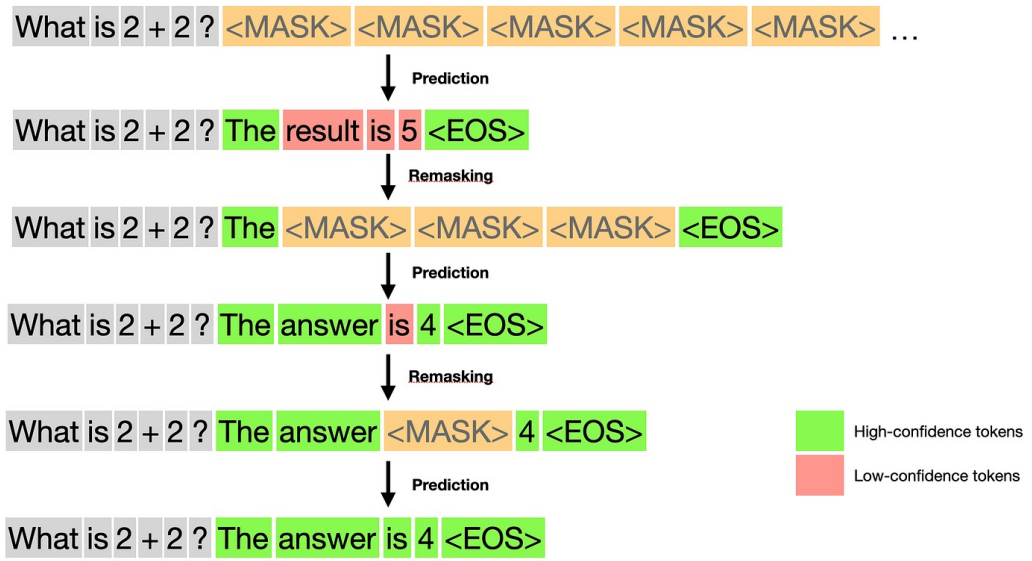
⋮

MDLM (Masked Diffusion Language Model)

LLaDA 7B example

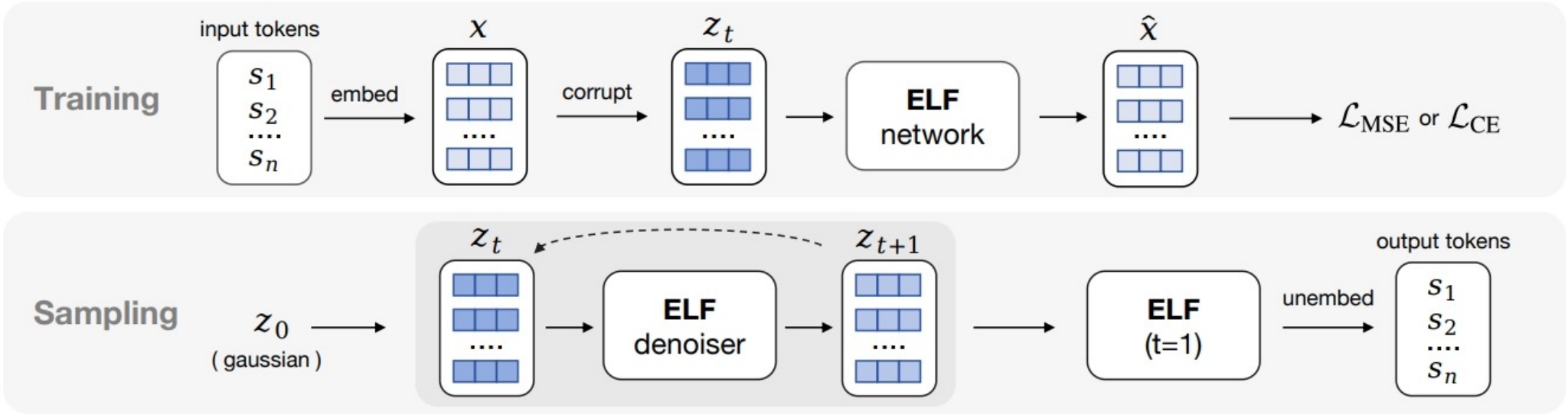


MDLM (Masked Diffusion Language Model)



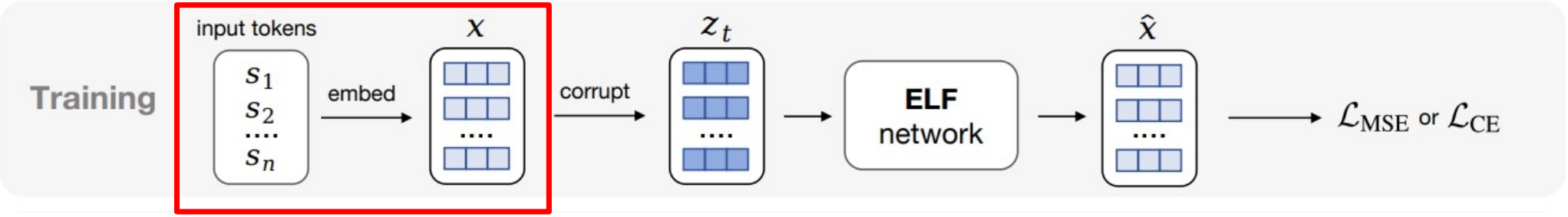
- 현재 MDLM 계열의 pretrained DLM들이 많이 등장하고 있음
- 하지만 continuous DLM의 성능 격차가 언어 모델링은 본질적으로 discrete한 성격을 띄어야 해서인지, 아니면 아직 continuous DLM의 알고리즘 연구가 충분히 되지 않아서인지는 모름
- 본 논문에선 flow matching을 이용한 continuous DLM을 제안

ELF (Embedded Language Flows)



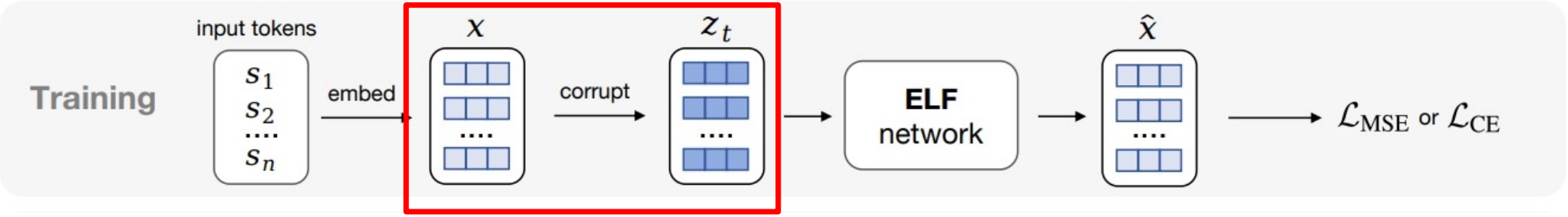
Continuous embedding space에서 denoising을 수행하고, 마지막 단계에서 clean embedding을 discrete token으로 변환

ELF (Embedded Language Flows)



- Discrete token을 pretrained encoder로 continuous space로 mapping

ELF (Embedded Language Flows)



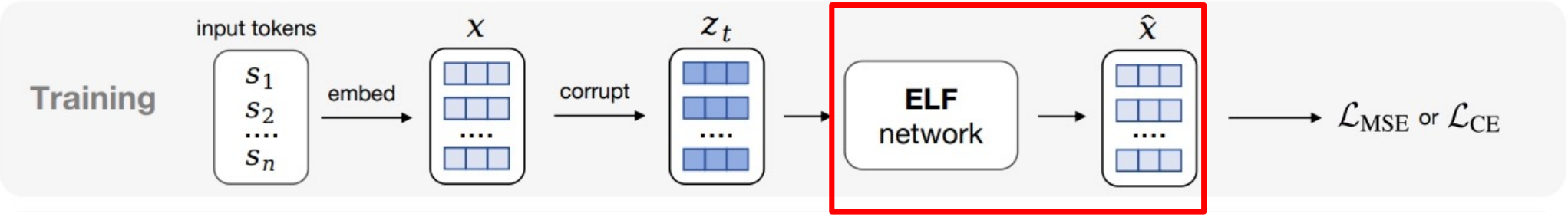
- Flow matching (rectified flow) 기반의 noising 방식이 사용됨

$$z_t = tx + (1-t)\epsilon$$

- Flow velocity는 아래와 같음

$$v = \frac{dz}{dt} = x - \epsilon$$

ELF (Embedded Language Flows)

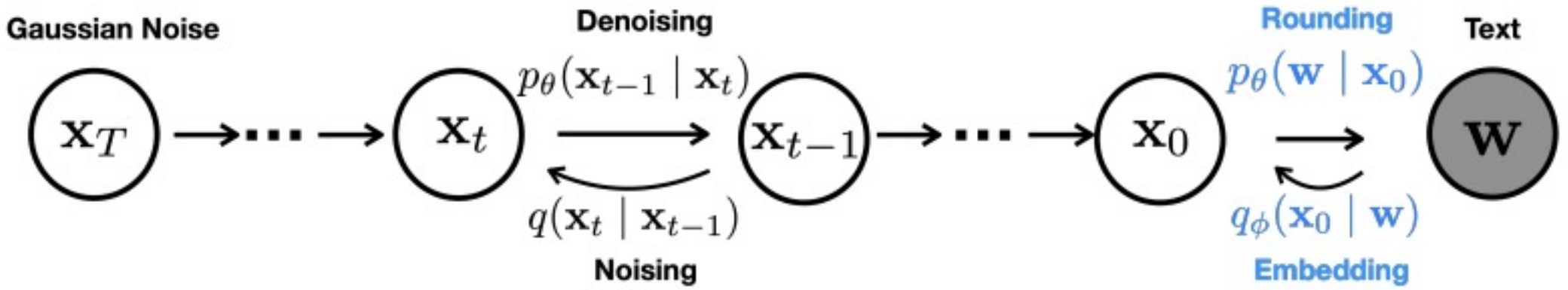


- ELF는 최근 CV domain에서의 trend를 따라 velocity가 아닌 x를 parametrize

$$x_\theta = net_\theta(z_t, t)$$

- 고차원 공간에서는 x-prediction이 좀 더 효과적이고, discrete token을 해야 하는 language domain에서도 x-prediction이 더 자연스럽다고 언급
- 이때 x-prediction을 할때 continuous space에 별다른 제약 조건을 걸지 않음

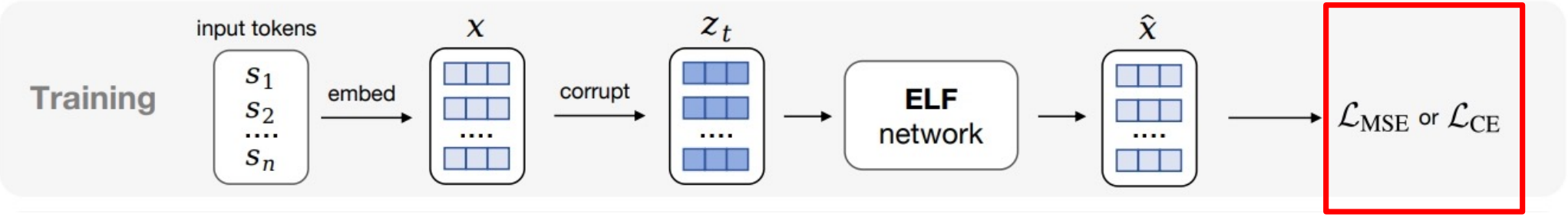
Previous Continuous DLM



$$x_{t-1} = \sqrt{\bar{\alpha}} \cdot Clamp(f_\theta(x_t, t)) + \sqrt{1 - \bar{\alpha}}\epsilon$$

- 기존의 continuous DLM들은 denoising을 할 때, continuous space 상에서 clamping, rounding과 같은 제약 조건을 걸어 continuous DLM임에도 discrete한 성격을 부여
- 이로 인해 기존의 continuous diffusion에서 사용되던 technique들을 fully 사용하지 못함

ELF (Embedded Language Flows)

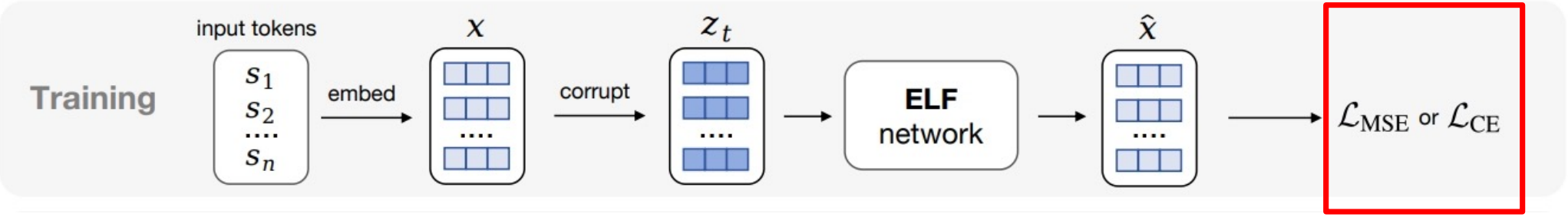


- T를 random sampling하여 train
- Denoising step과 decoding step의 loss를 분리해서 train

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{t,x,\epsilon} \left\| v_\theta(z_t, t) - v \right\|^2 = \mathbb{E}_{t,x,\epsilon} \frac{1}{(1-t)^2} \left\| x_\theta(z_t, t) - x \right\|^2$$

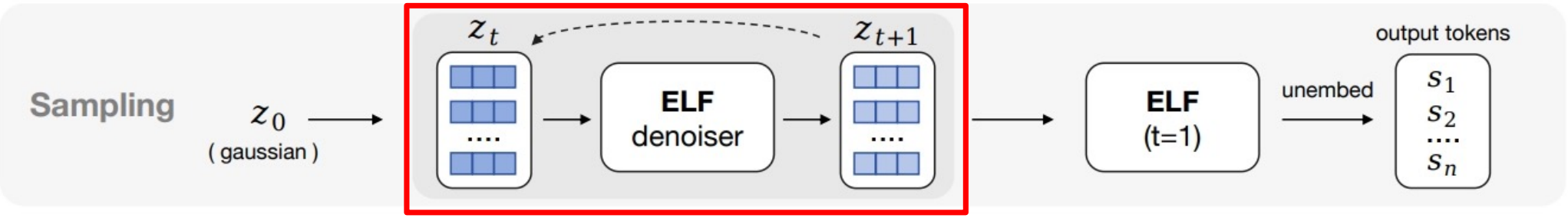
$$\mathcal{L}_{\text{CE}} = \mathbb{E}_{\tilde{z}}[\text{CrossEnt}(W x_\theta(\tilde{z}), s)]$$

ELF (Embedded Language Flows)



- Decoding할때는 learnable unembedding matrix W를 통해 logit을 얻은 후 softmax를 거쳐 discrete token predict

ELF (Embedded Language Flows)



- ELF는 아래 ODE를 풀어서 반복적으로 denoise함

$$\frac{dz_t}{dt} = v_\theta(z_t, t)$$

$$z_t + \Delta t = z_t + \Delta t v_\theta$$

CFG (Classifier Free Guidance)

- Inference-time CFG를 사용하게 되면 inference를 2번 해야 되기 때문에 overhead가 발생

$$v_{\text{cfg}}(z_t \mid c) = \omega v(z_t \mid c) + (1-\omega) v(z_t \mid \emptyset)$$

- 때문에 ELF는 한번의 inference로 CFG output을 얻을 수 있는 training-time CFG를 사용

$$\hat{x}' = \text{net}_\theta(z_t, t)$$

Self-conditioning

$$x_{\text{cfg}} = x_{\text{uncond}} + \omega\,(x_{\text{cond}} - x_{\text{uncond}})$$

$$\hat{x} = \text{net}_\theta(z_t \mid \hat{x}', t)$$

Conditioned behavior, unconditioned behavior를 학습하도록 함

Experiment

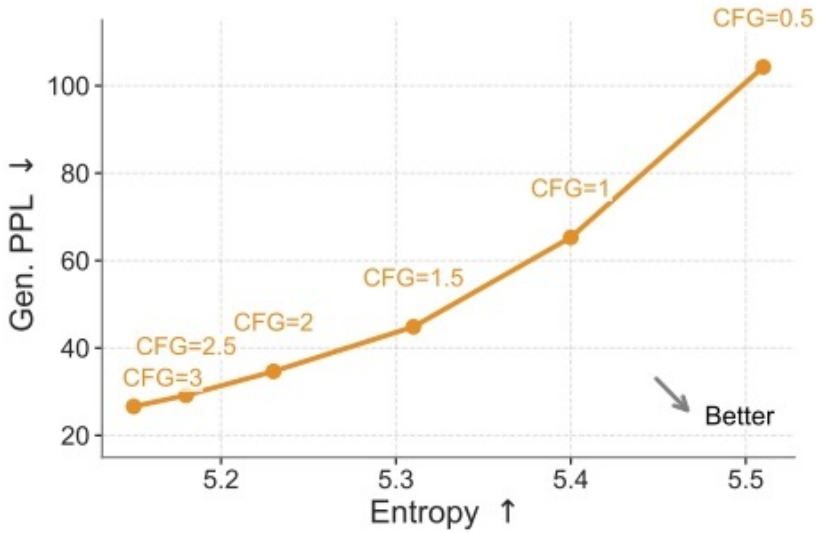


Figure 4: **Ablations on guidance.** We evaluate the generative perplexity–entropy trade-off across CFG scales: increasing the scale lowers generative perplexity but reduces entropy.

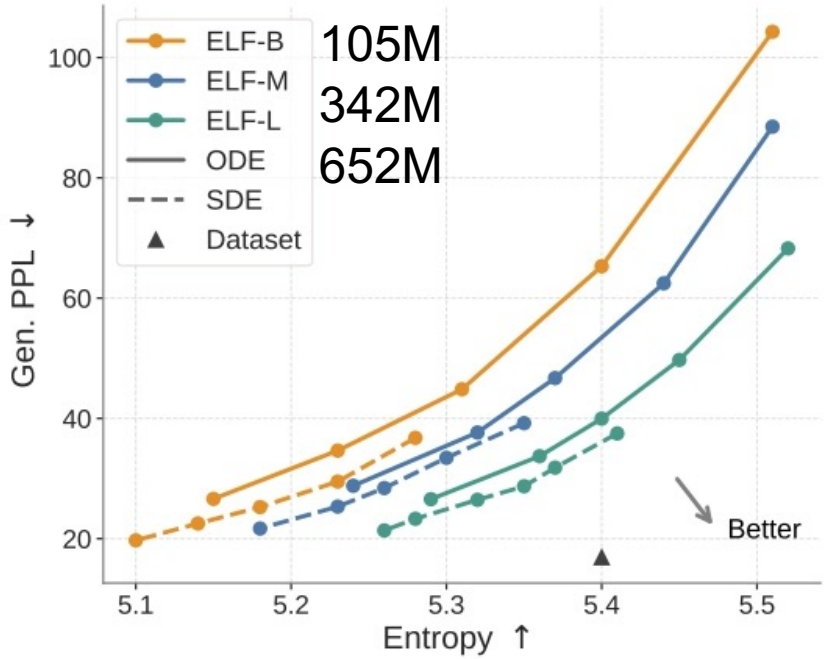
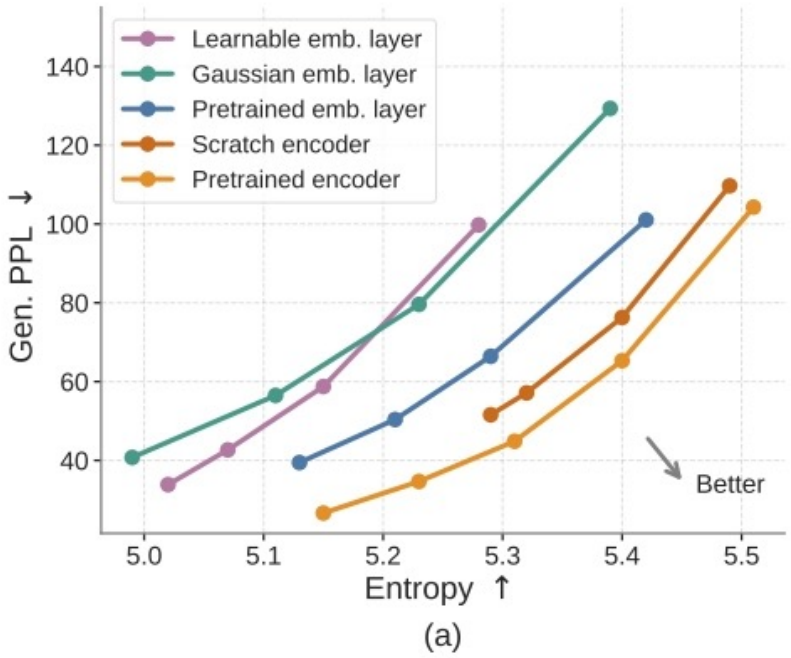
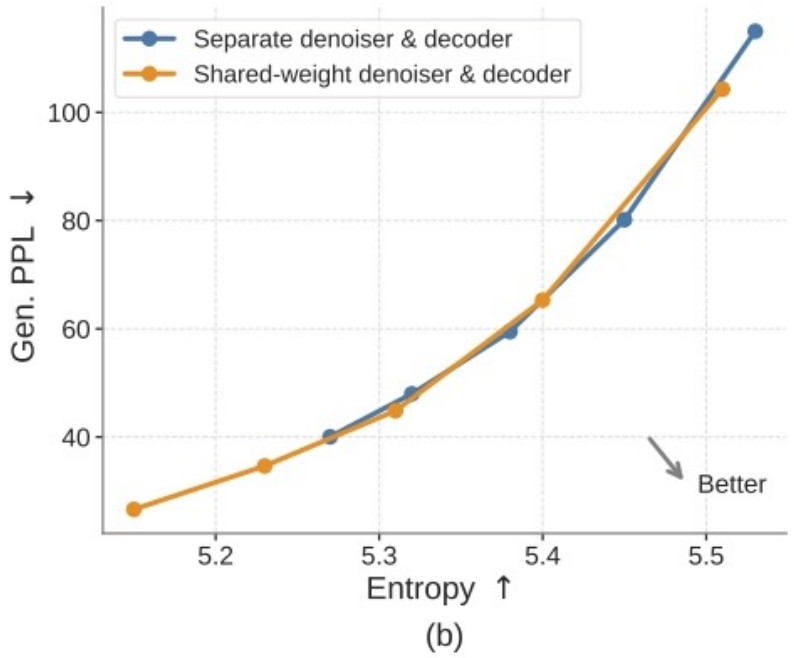
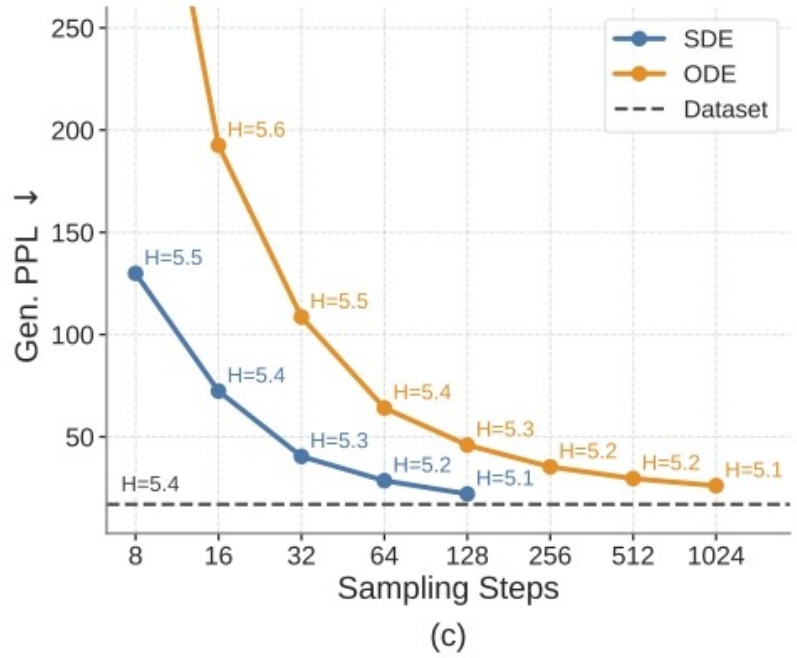


Figure 6: **Scaling of ELF models.** We compare ELF-B, ELF-M, and ELF-L. Scaling model size consistently improves the Gen. PPL–entropy frontier.

Experiment


Learnable emb. layer
Gaussian emb. layer
Pretrained emb. layer
Scratch encoder
Pretrained encoder
Gen. PPL ↓
Entropy ↑
Better
(a)


Separate denoiser & decoder
Shared-weight denoiser & decoder
Gen. PPL ↓
Entropy ↑
Better
(b)


SDE
ODE
Dataset
Gen. PPL ↓
Sampling Steps
H=5.6
H=5.5
H=5.5
H=5.4
H=5.4
H=5.3
H=5.3
H=5.2
H=5.2
H=5.2
H=5.1
H=5.1
H=5.4
(c)

Experiment

Unconditional generation

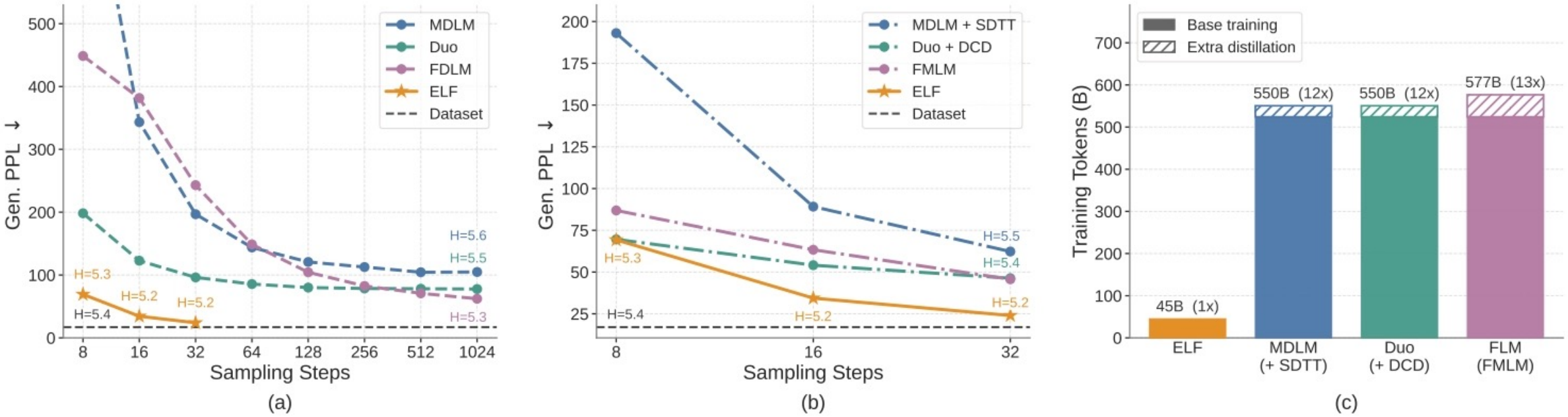


Figure 7: **System-level comparison.** ELF-B outperforms both discrete and continuous DLMs trained under similar settings (a), rivals distilled variants of other baselines that require additional rounds of training (b), and uses substantially fewer training tokens (c).

Training token의 efficiency에 주목할만 함

Experiment

Conditional generation

Model	Size	De-En† BLEU ↑	XSum‡ ROUGE-1 ↑	XSum‡ ROUGE-2 ↑	XSum‡ ROUGE-L ↑
AR	99M	25.2	30.5 ± 0.13	10.2 ± 0.11	24.4 ± 0.12
MDLM [56]	99M	18.4	33.4 ± 0.11	11.6 ± 0.10	25.8 ± 0.10
Duo [57]	170M (+35M)	21.3‡	31.4 ± 0.12	10.1 ± 0.10	25.0 ± 0.12
E2D2 [4]	99M	24.8	28.4 ± 0.11	8.3 ± 0.09	22.0 ± 0.10
SeqDiffuSeq [79]	-	21.3	19.3†	1.7†	14.1†
CDCD [13]	-	24.9	-	-	-
Ours	105M (+35M)	**26.4**	**36.0**± 0.13	**12.2**± 0.11	**27.8**± 0.12

Conclusion

- ELF는 denoising trajectory를 연속적으로 유지하고 마지막 단계에서만 discretize를 함으로써 기존 continuous diffusion에서의 기법을 쉽게 적용 가능
- 기존 DLM들에 비해 강력한 quality를 보임
- 특히나 더 적은 sampling step과 더 적은 training token으로 더 낮은 perplexity 달성

EOD