291K
Deep Learning for Machine Translation
Parallel Decoding

Lei Li
UCSB

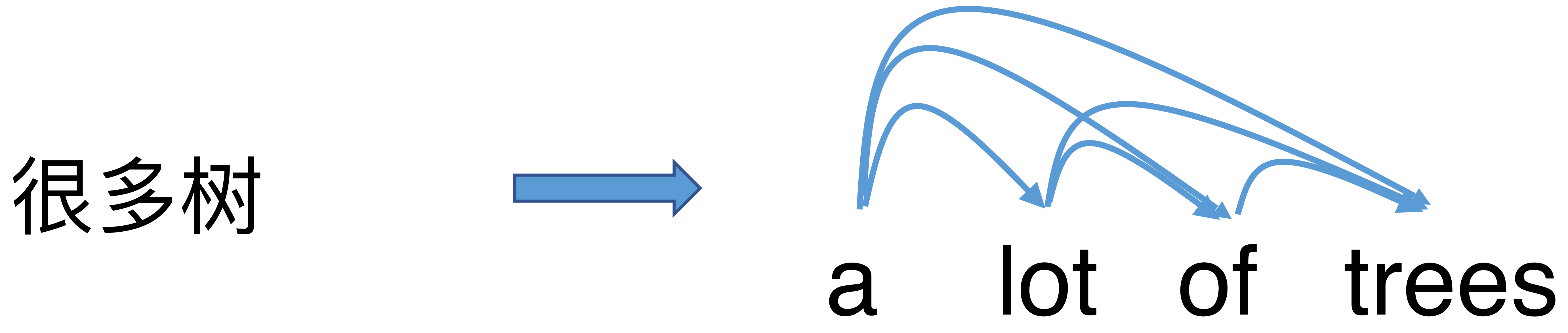
11/29/2021

Outline

- Autoregressive & Non-autoregressive Generation
- Iterative NAT and Limitation
- Glancing Transformer

Transformer is Autoregressive

- Autoregressive models generate sentences sequentially



- The conditional probability is factorized successively

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(y_t|y_{<t}, X;\theta)$$

- Human-style translation is slow. Machine does not have to mimic human!

Wild idea: Parallel Generation?

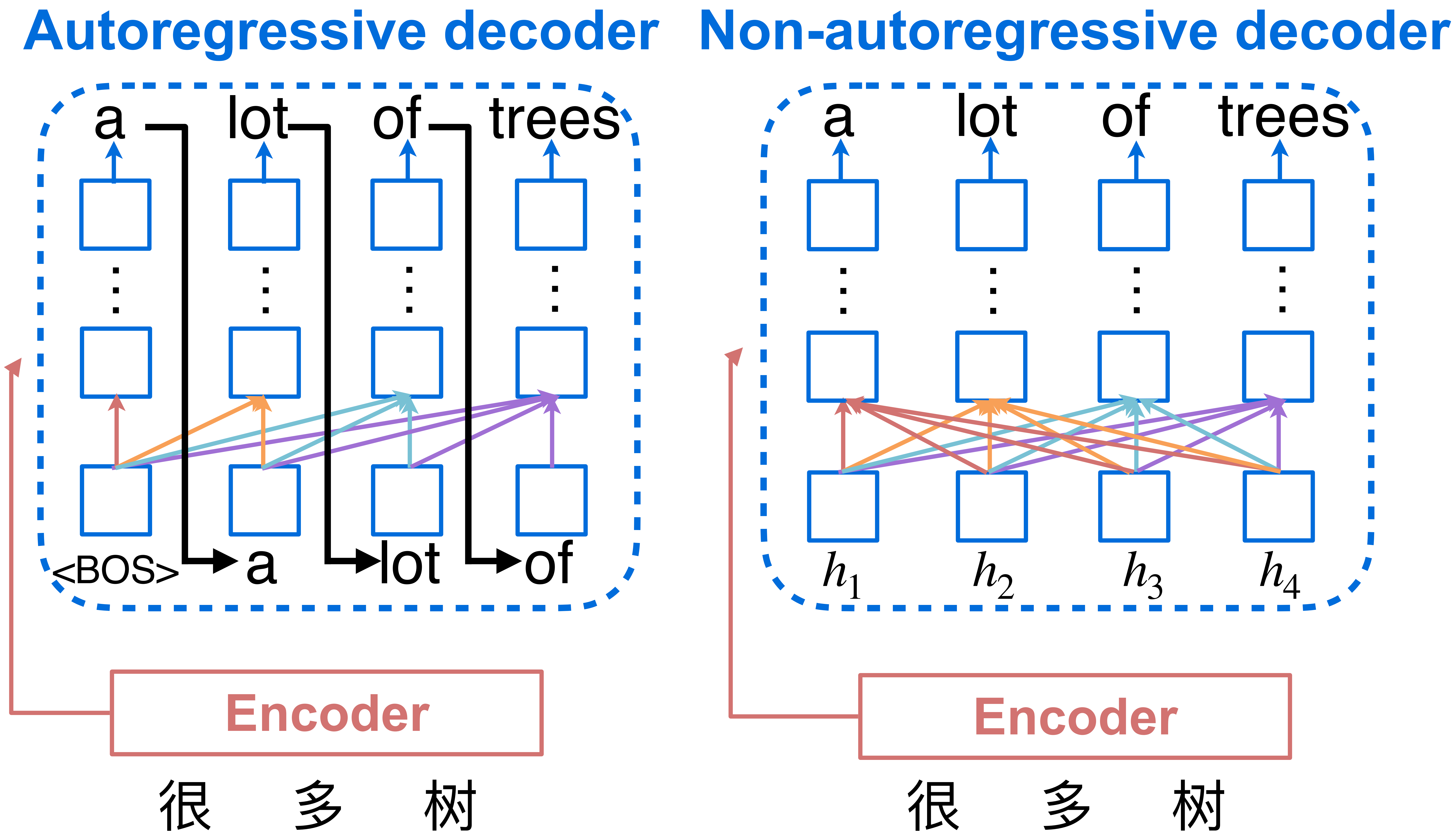
- Non-autoregressive models generate all the tokens in parallel

很多树 → a lot of trees

- Conditional independence assumption

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(y_t|X;\theta)$$

Model architecture



Gu et al, NAT, ICLR 2018

Training of vanilla NAT

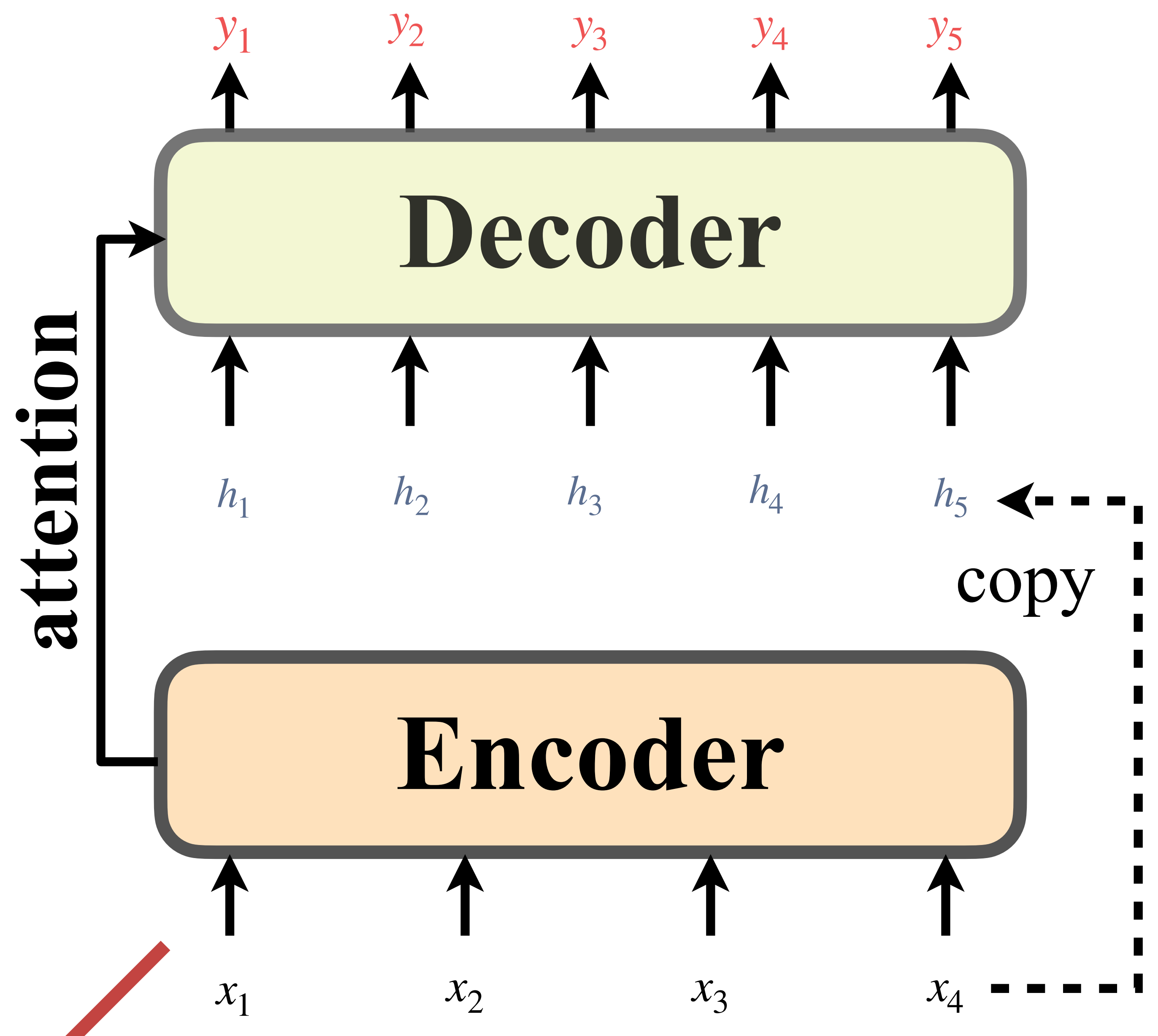
- Maximum likelihood estimation (MLE)

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(y_t|X;\theta)$$

$$L_\theta = -\log p(Y|X;\theta)$$

directly follow autoregressive models

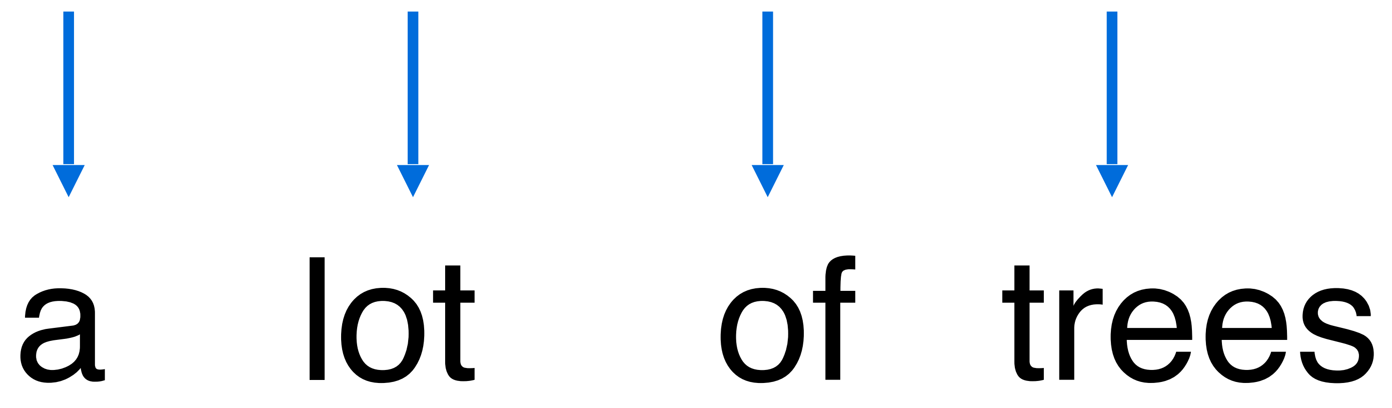
- Target length
  - predict before decoding
  - predefine max length



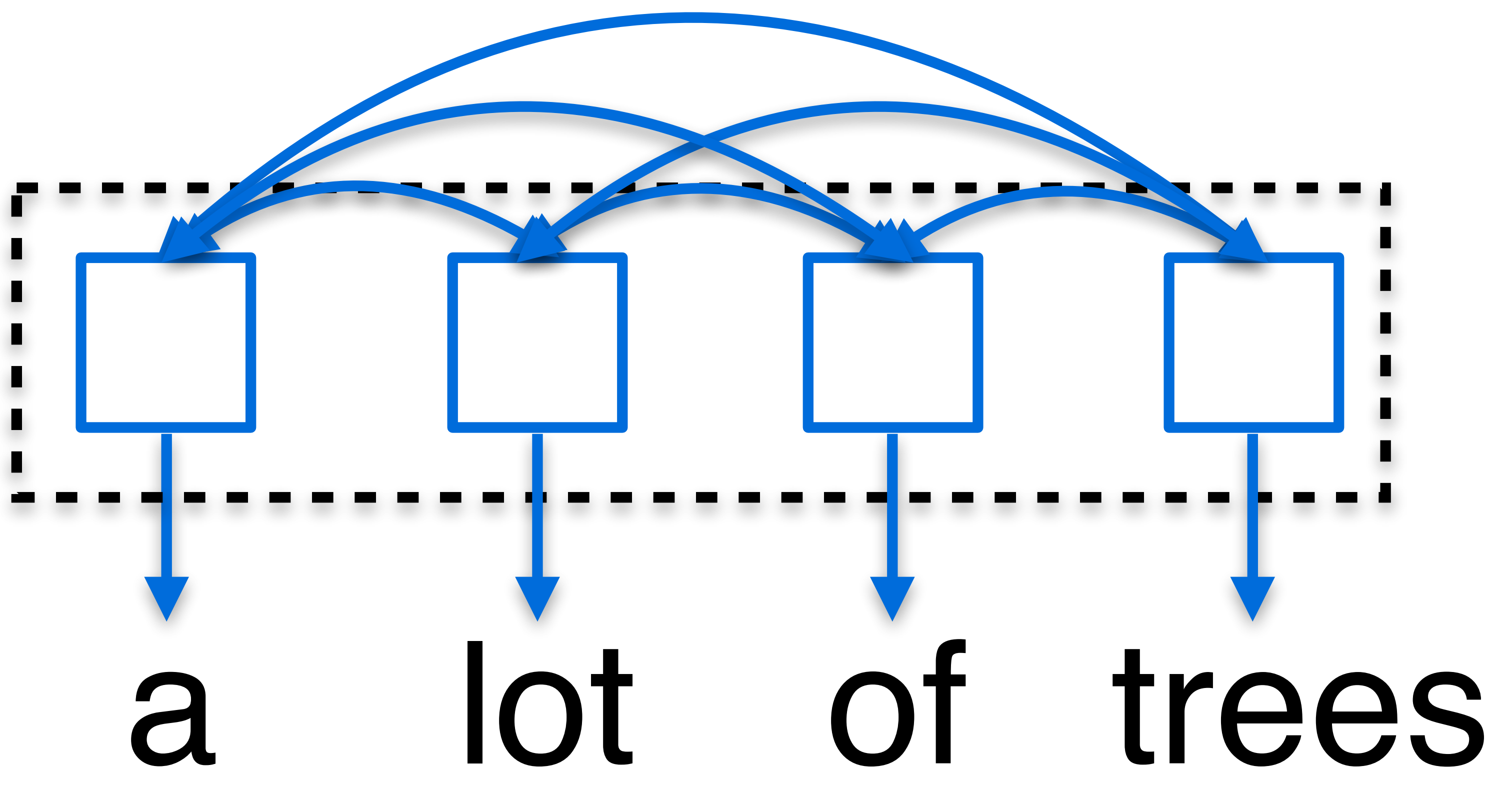
Lack explicit target word interdependency learning

Why Non-autoregressive?

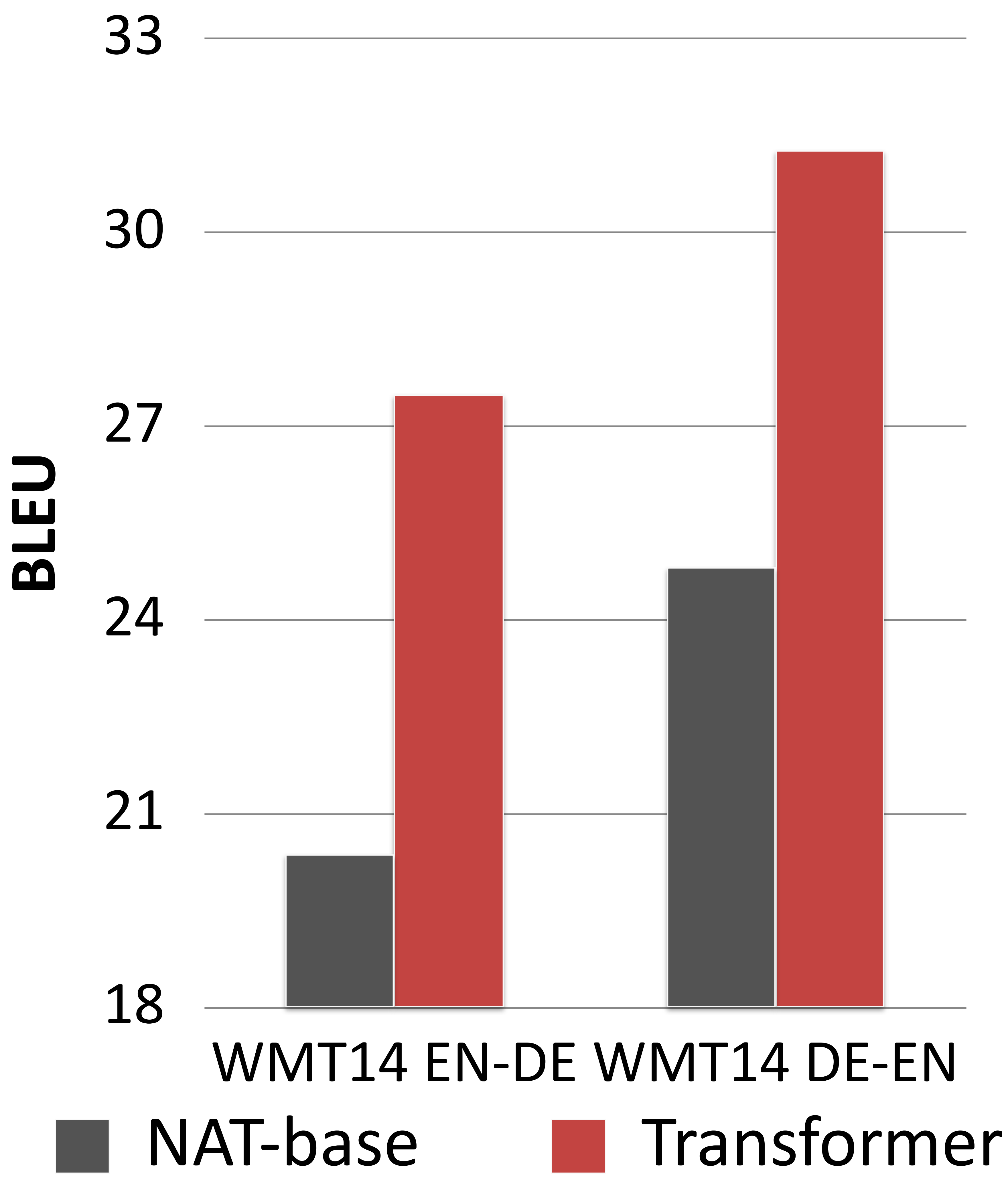
1. Faster decoding in non-autoregressive translation (NAT)



2. Capturing bidirectional context for generation



Challenge: Inferior Quality of NAT



- One input -> multiple target

很多树 → a lot of trees

很多树 → a great many trees

- Inconsistency problem in parallel generation

great many

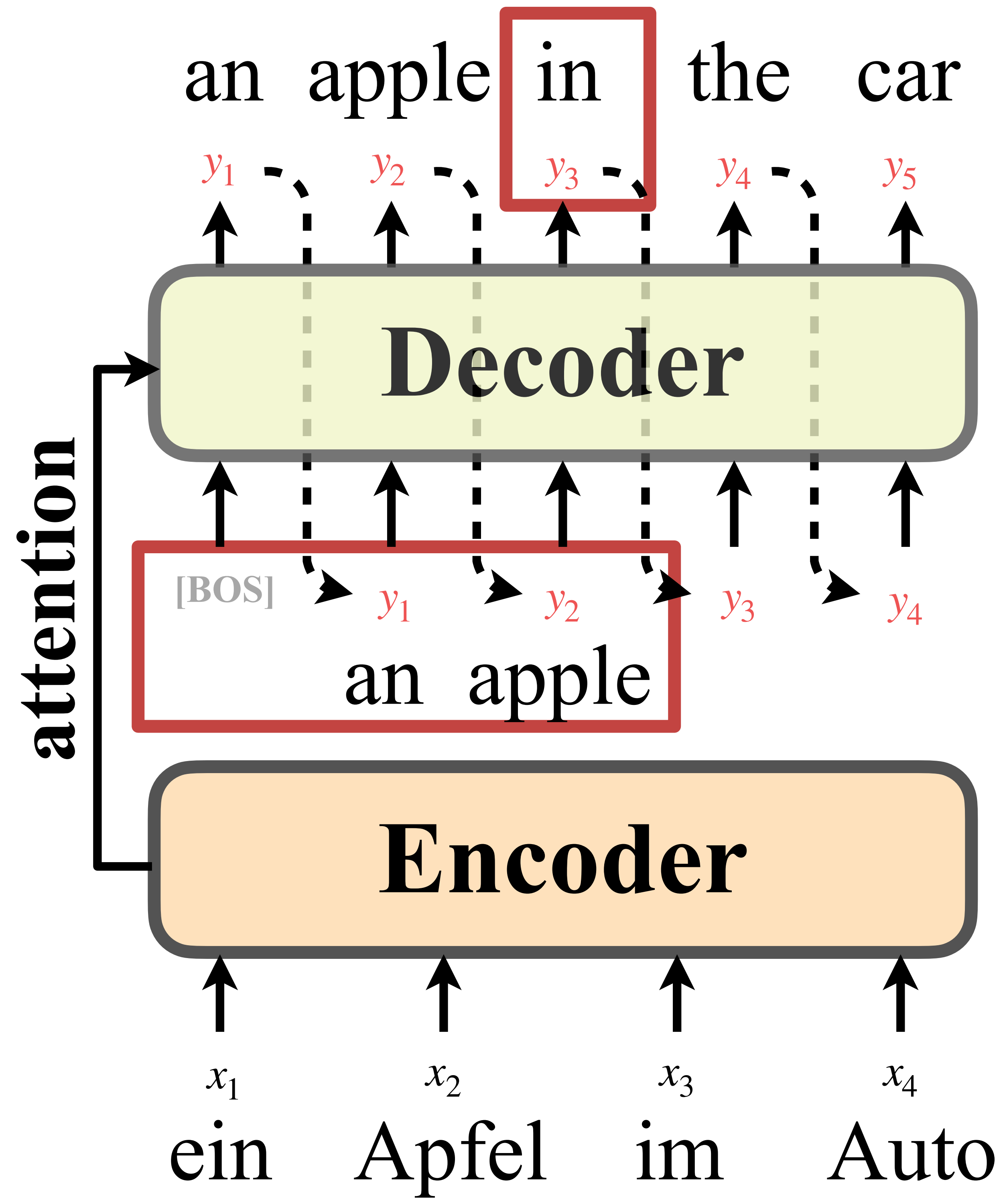
lot of

很多树 → a great of trees

Key Intuition: Word interdependency

- Learning word interdependency in the target sentence is crucial for generating fluent sentences
- Non-autoregressive models lack a effective way of dependency learning

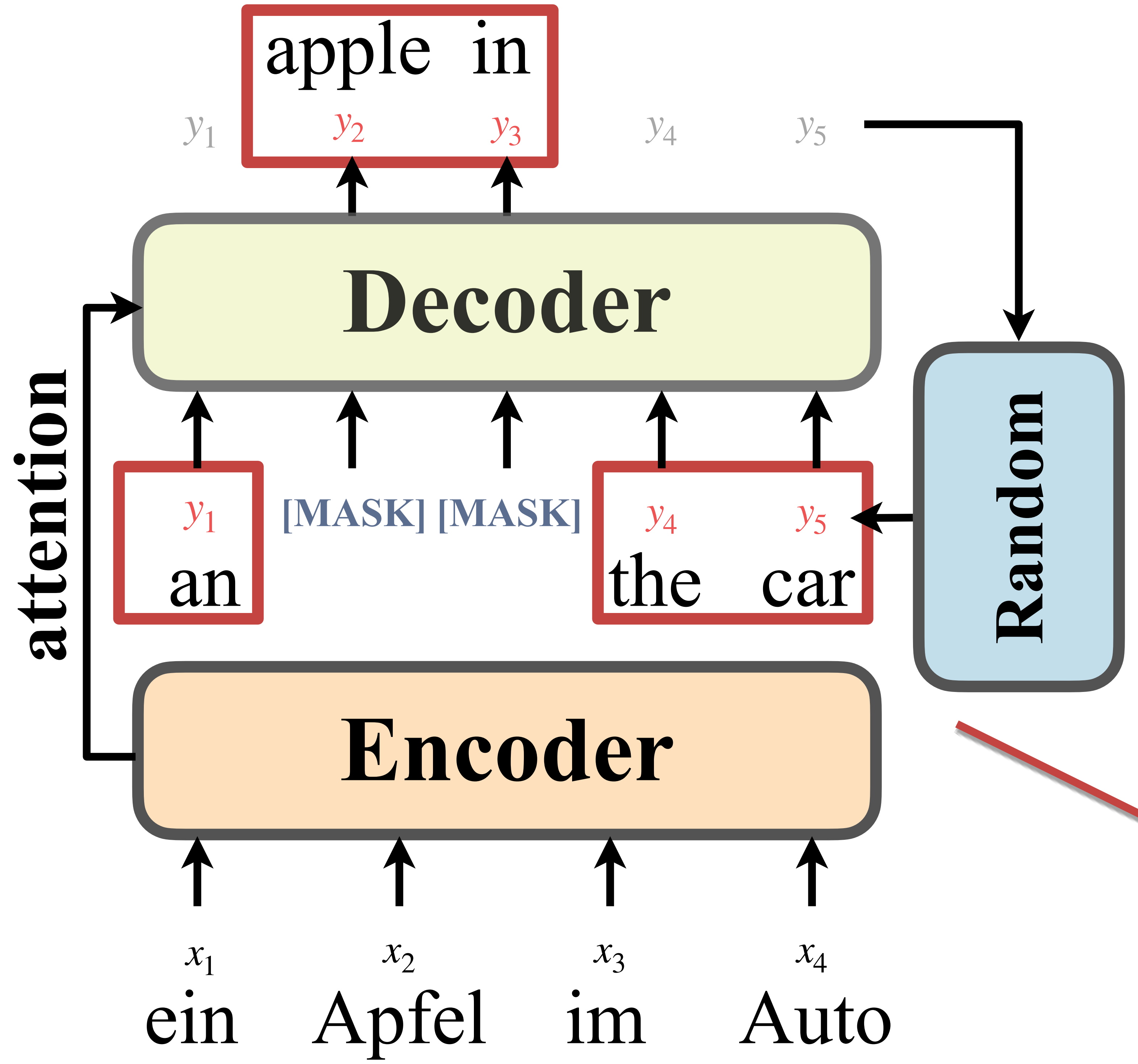
Learning Word Interdependency



Autoregressive models

- predict the next tokens conditioned on the input target tokens (left-to-right)

Iterative NAT



Iterative-NAT

- predict the randomly masked tokens based on unmasked tokens

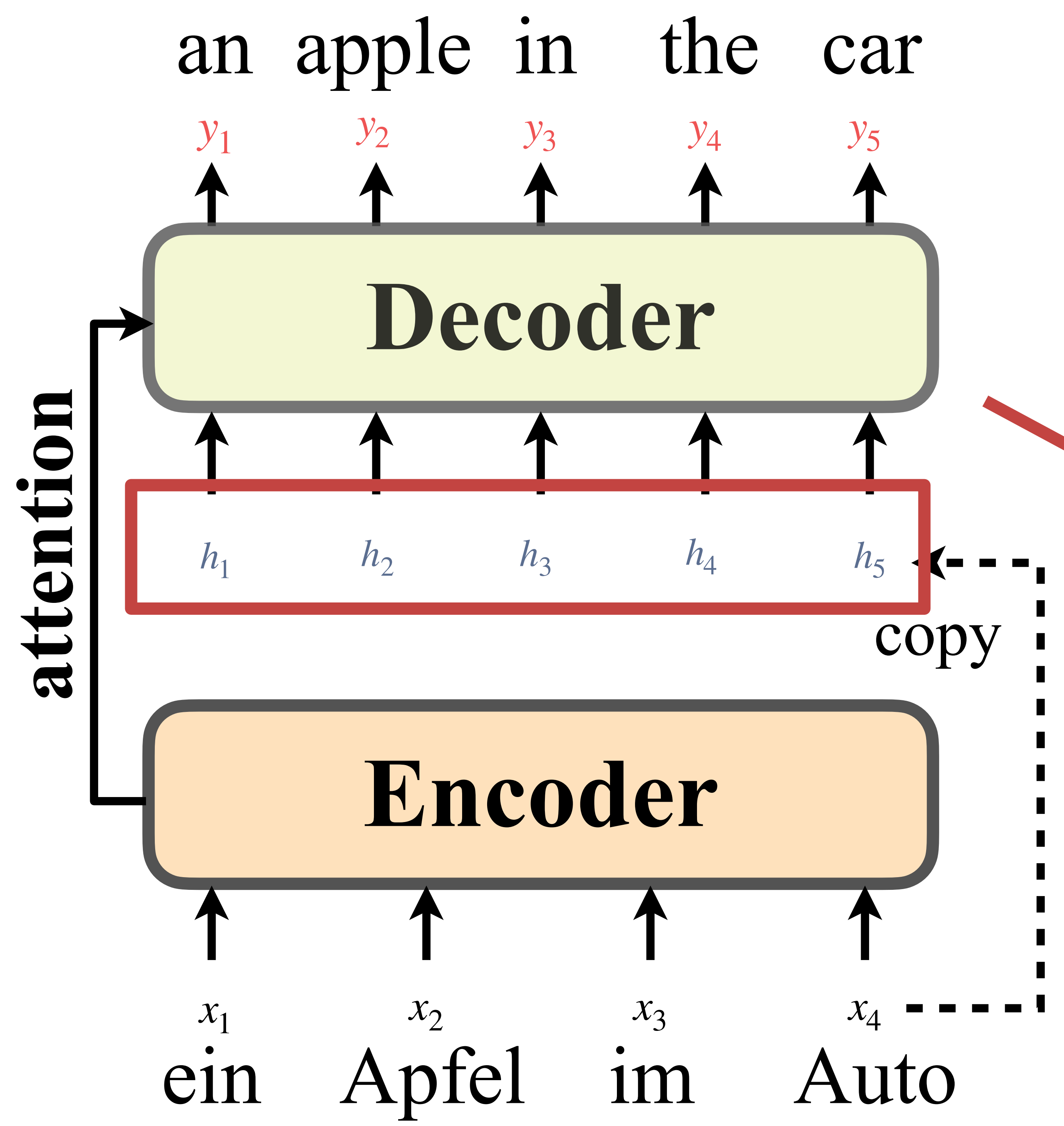
rely on multiple decoding iterations, therefore does not gain speedup!

Lee et al. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement. EMNLP 2018.
Ghazvininejad et al. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. EMNLP 2019.

Dependency learning for NAT

- How to learn word interdependency for **single-pass** parallel generation?
- Contradiction
  - Word interdependency learning requires target word inputs
  - Single-pass parallel generation cannot obtain target words before prediction
- Glancing Language Model (GLM)
  - A gradual training method to achieve both

New Idea for Dependency learning



$$L_\theta = -\log p(Y|X;\theta)$$

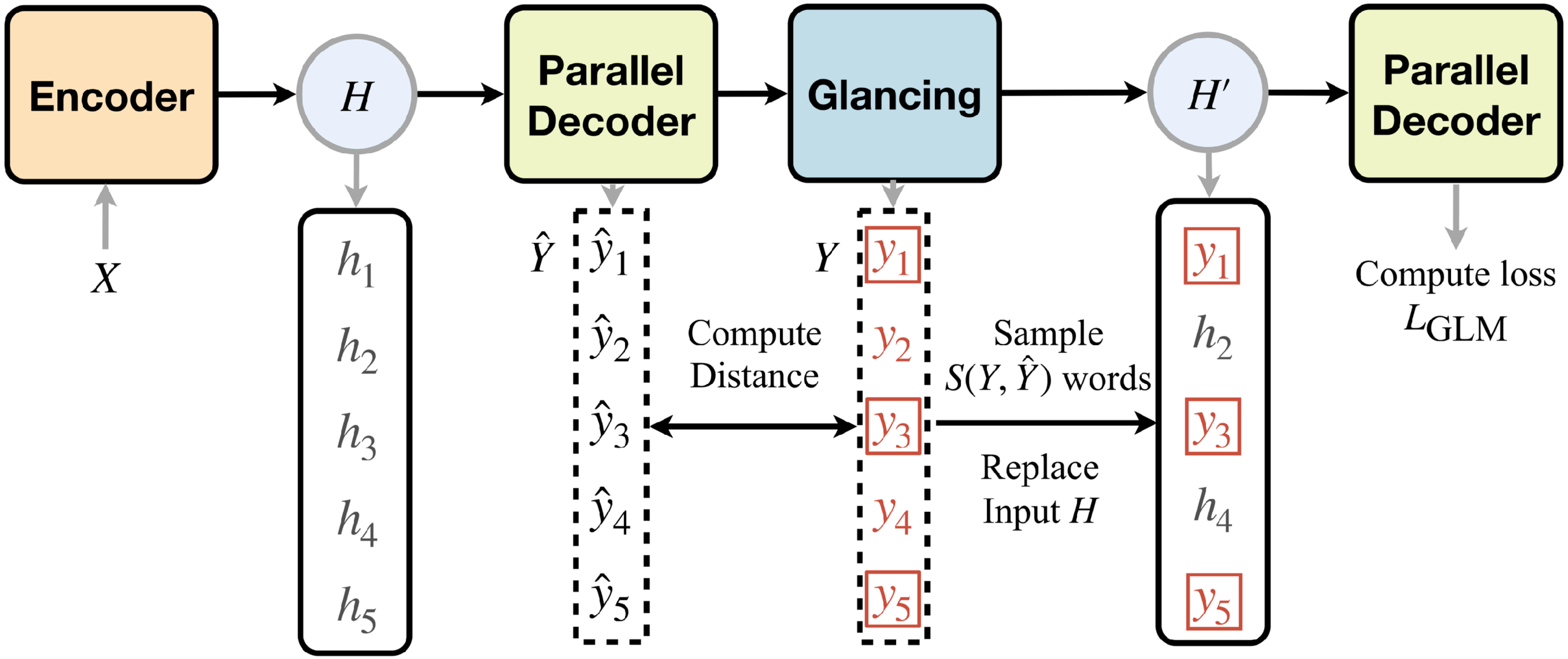
Lack explicit target word interdependency learning

- Glancing Language Model (GLM)
  - A gradual training method
  - Learning word interdependency for **single-pass** parallel generation

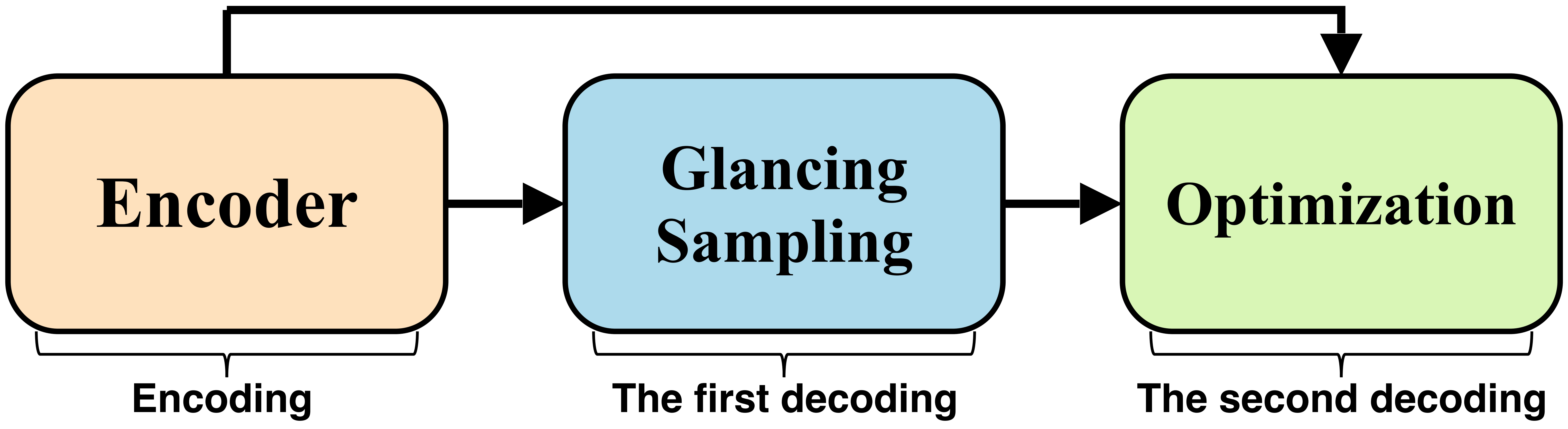
Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

Glancing Language Model (GLM)

- An adaptive sampling strategy for gradual learning
  - From fragments to the whole sequence
- Learning target word interdependency for single-pass parallel generation



Glancing Language Model



only one-pass decoding in inference

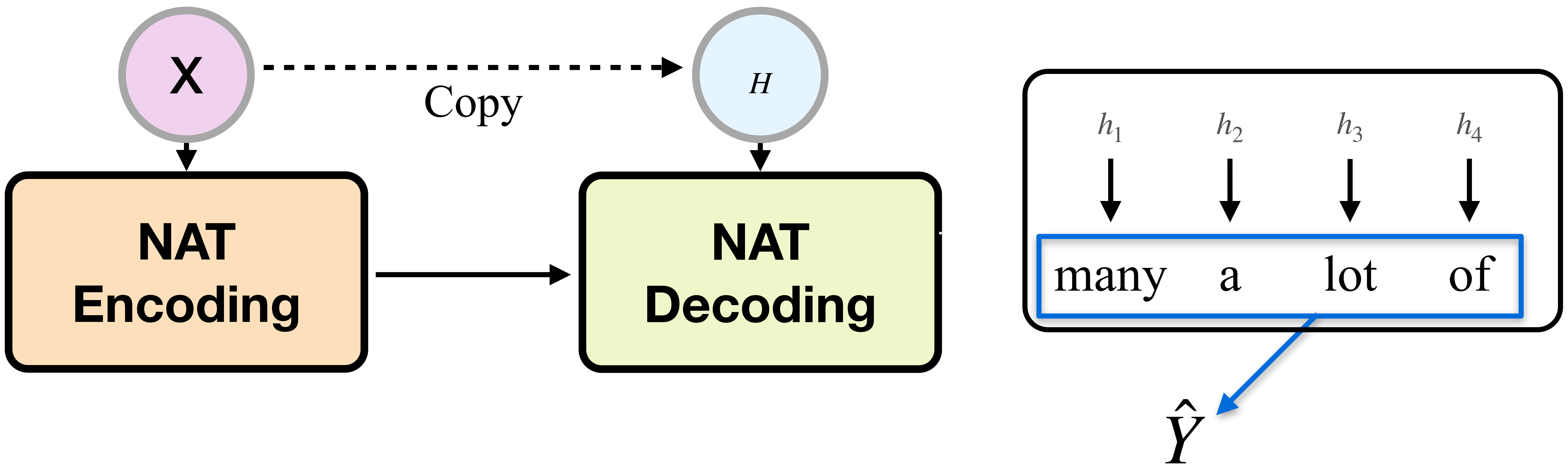
- Perform two decoding during training

1. Glancing Sampling (the first decoding):
   - Based on the prediction, replace part of the decoder inputs with sampled target words
2. Optimization (the second decoding):
   - Learn to predict the remaining words with the replaced decoder inputs

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

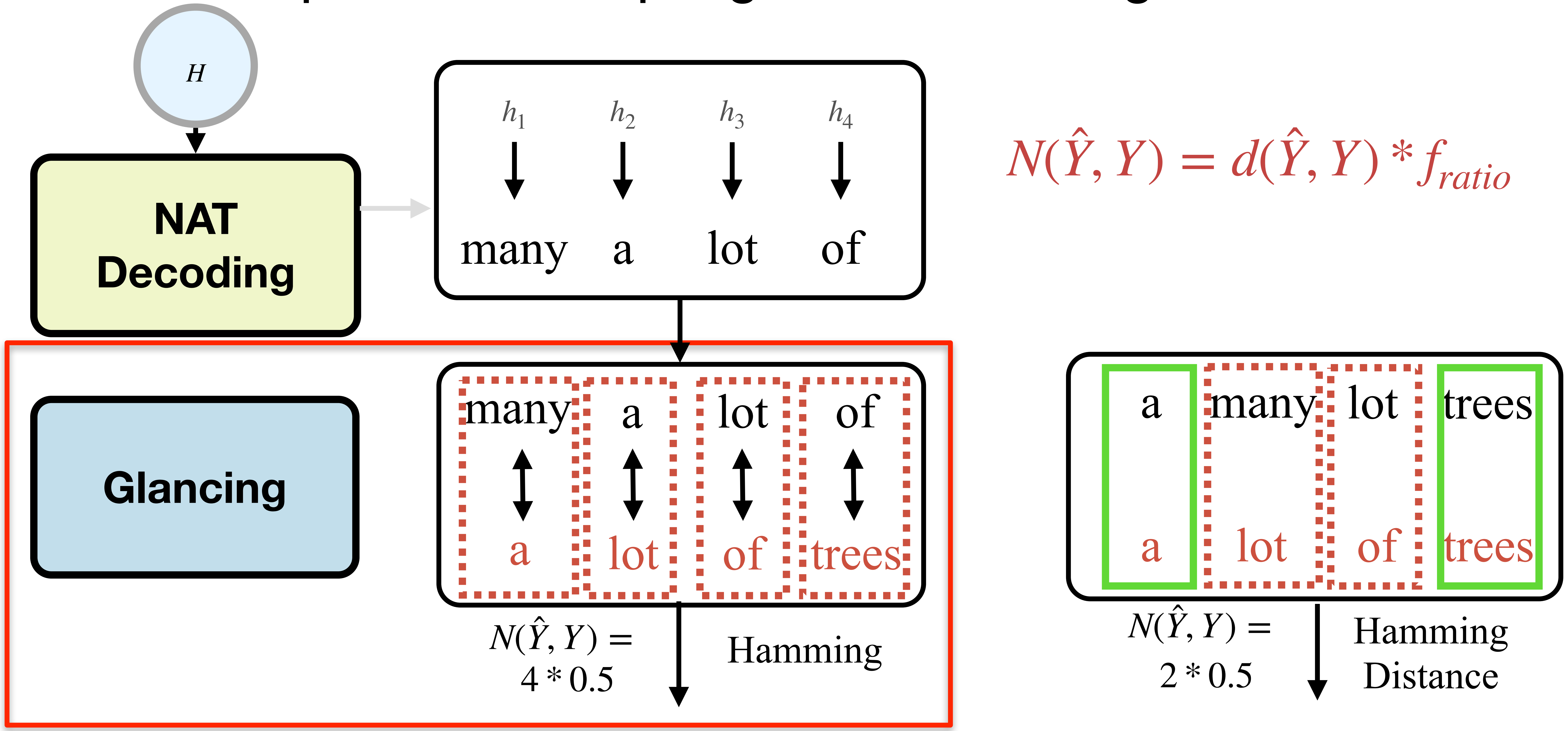
Glancing Sampling (1): NAT Decoding

- For input X, generate the whole sequence $\hat{Y}$ in parallel
- Training sample (X,Y)
  - X: 很多树
  - Y: a lot of trees



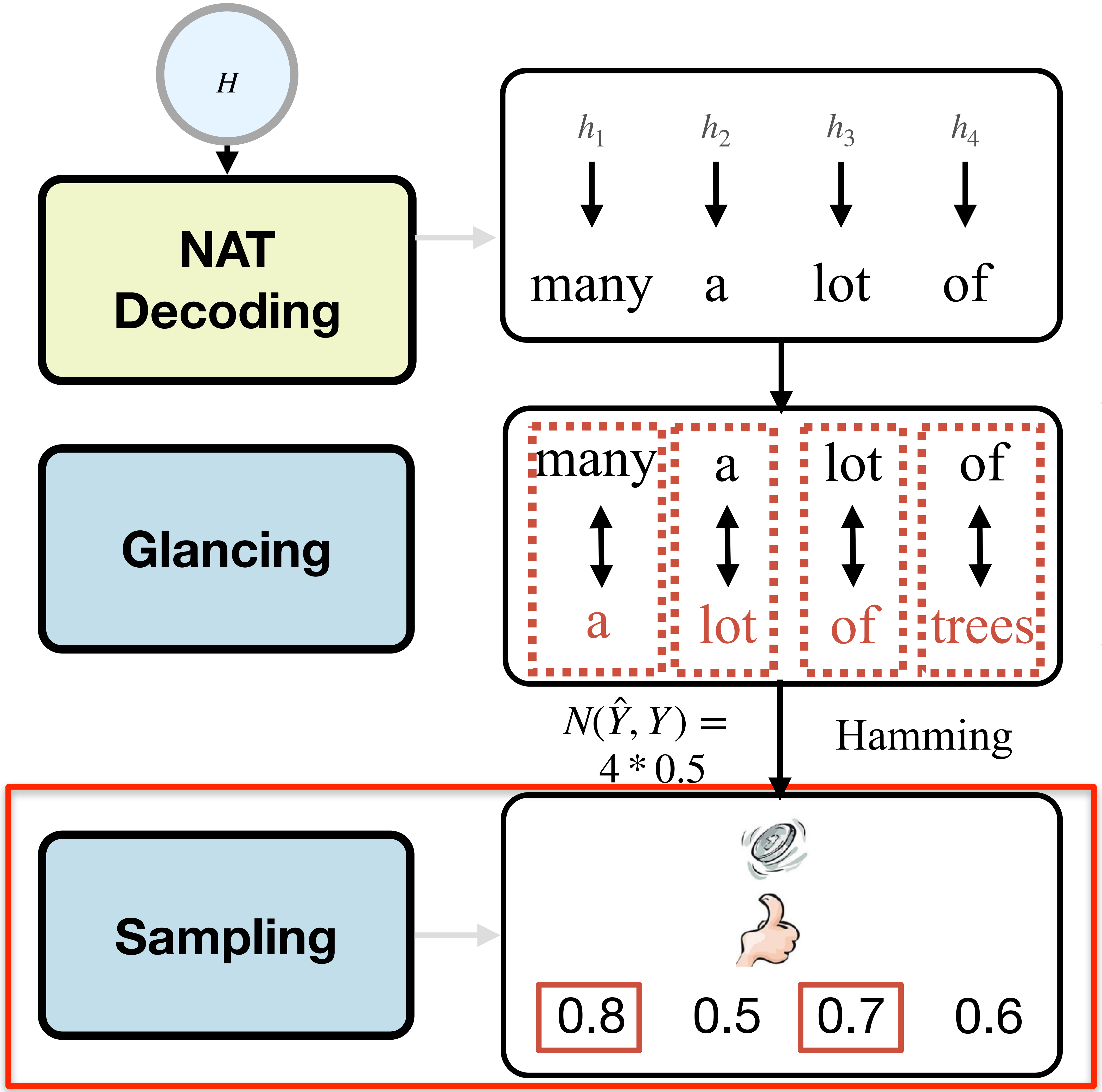
Glancing Sampling (2): Glancing

1. Measure the distance between the prediction and the reference
2. Compute the sampling number of target words



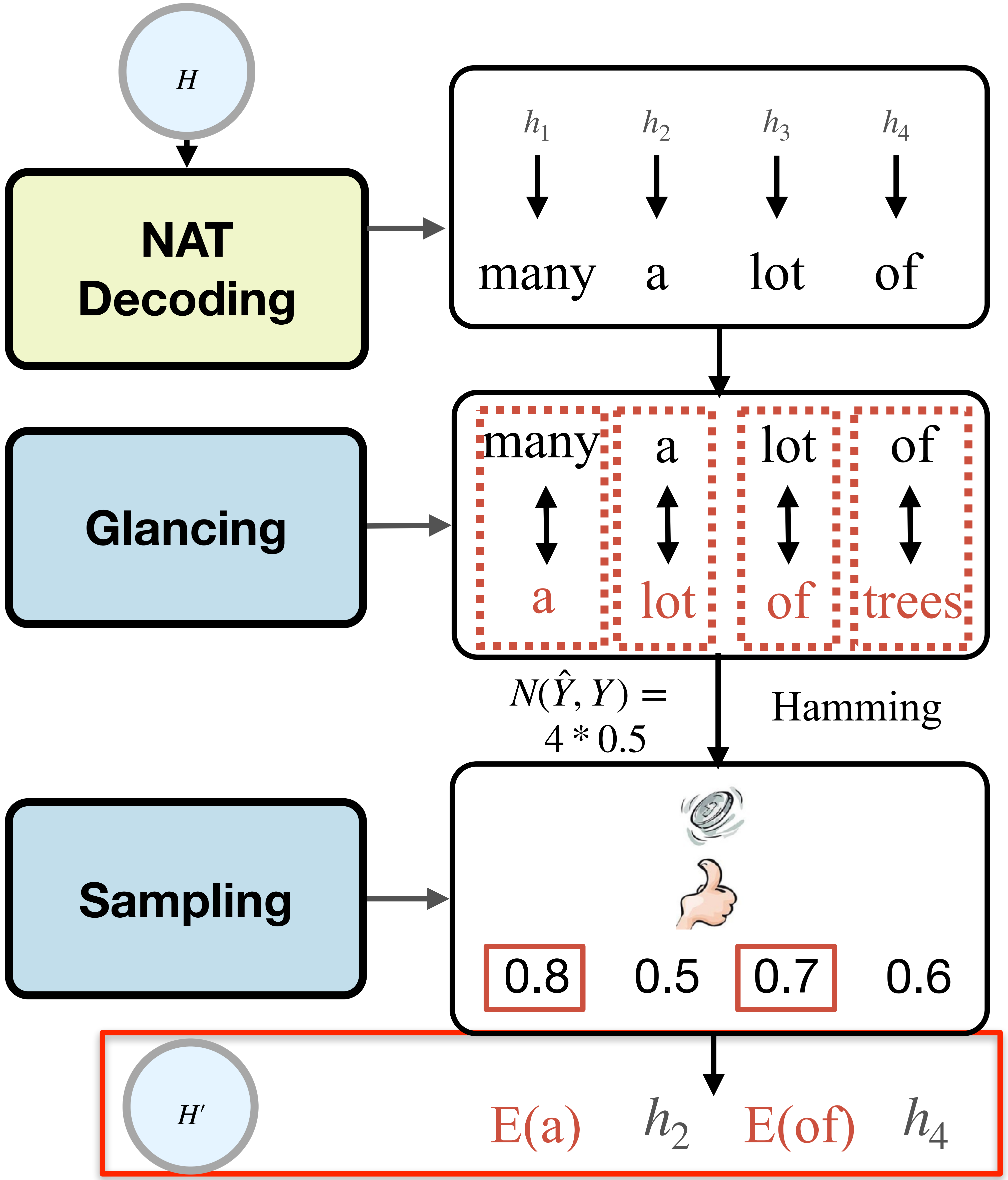
$$N(\hat{Y}, Y) = d(\hat{Y}, Y) * f_{ratio}$$

Glancing Sampling (3): Sampling



- Select $N(\hat{Y}, Y)$ target words for glancing
- Random target word selection strategy

Glancing Sampling (4): Replacing for prediction



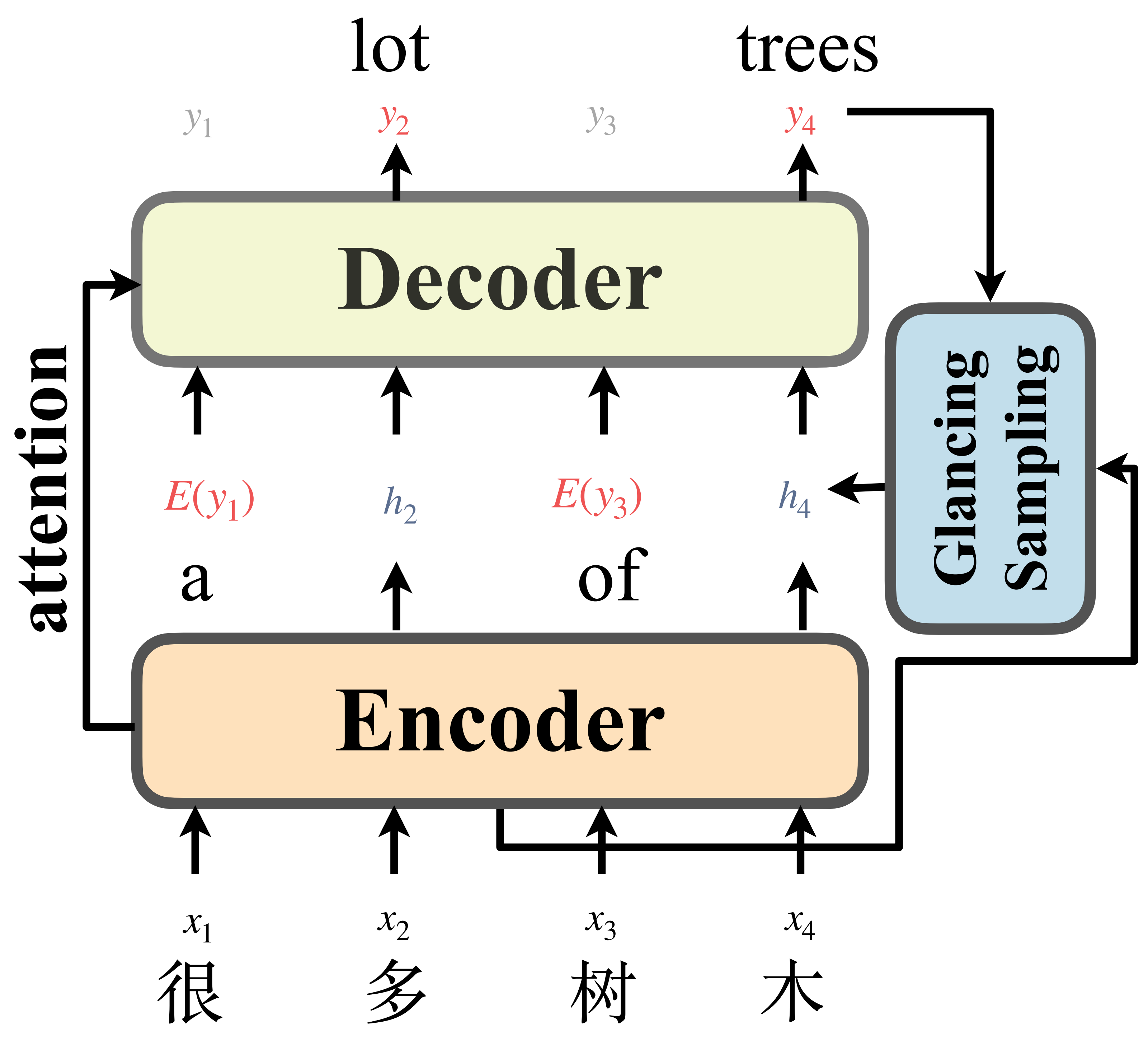
- Replace the original decoder inputs with the embedding of sampled target words

Methodology: Optimization

The second decoding:

- learn to predict the remaining words with the replaced decoder inputs

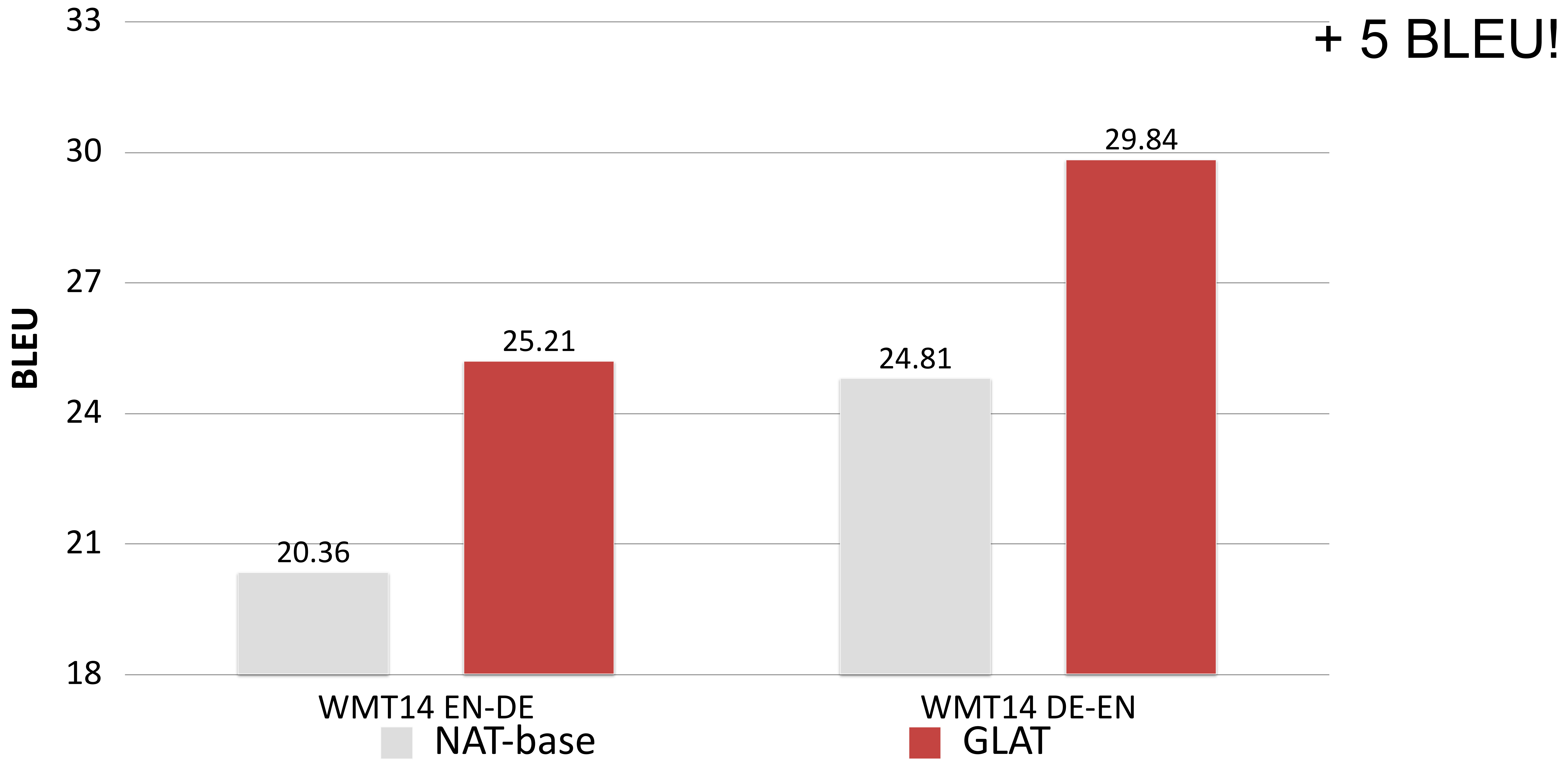
$$\mathcal{L}_{\mathrm{GLAT}} = -\sum_{\{X,Y\}\in D} \sum_{y_t\in\{Y\setminus \mathbb{GS}(Y,\hat{Y})\}} \log p(y_t | \mathbb{GS}(Y,\hat{Y}), X;\theta)$$



During training, the sampling number of target words decreases gradually.

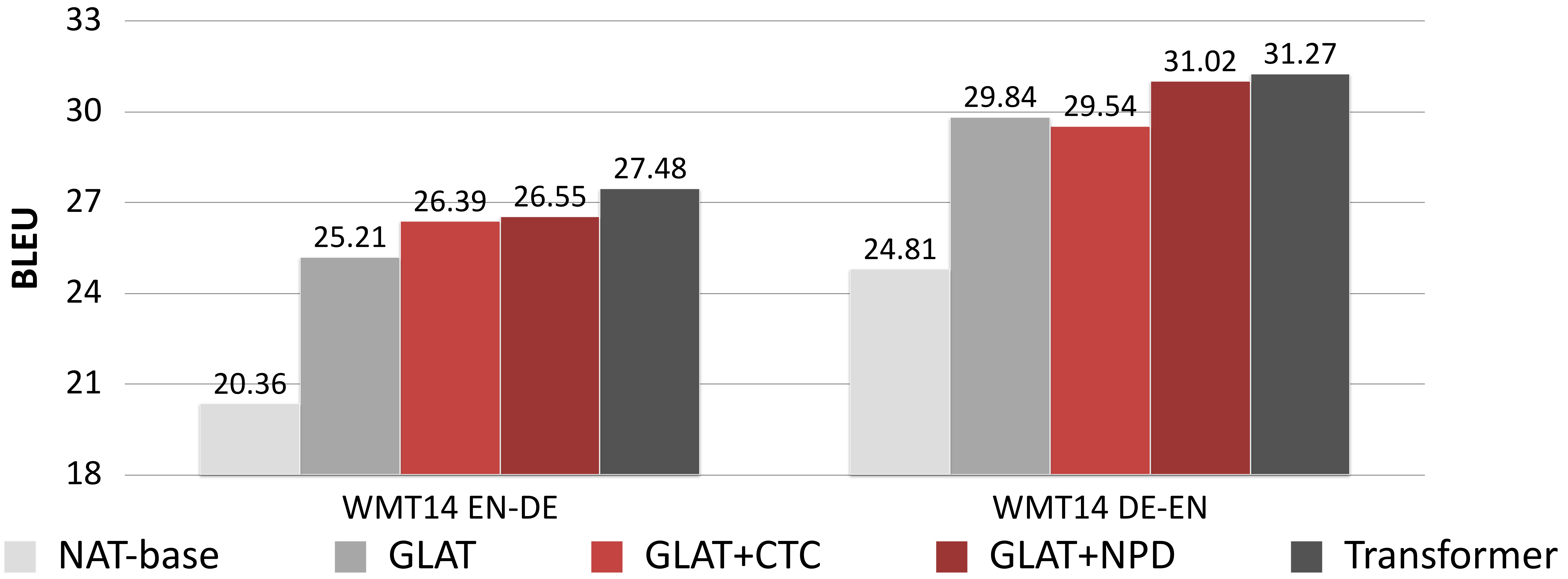
Learn to generate longer fragments

GLAT boosts Translation Quality significantly!



Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

GLAT approaches Transformer quality!

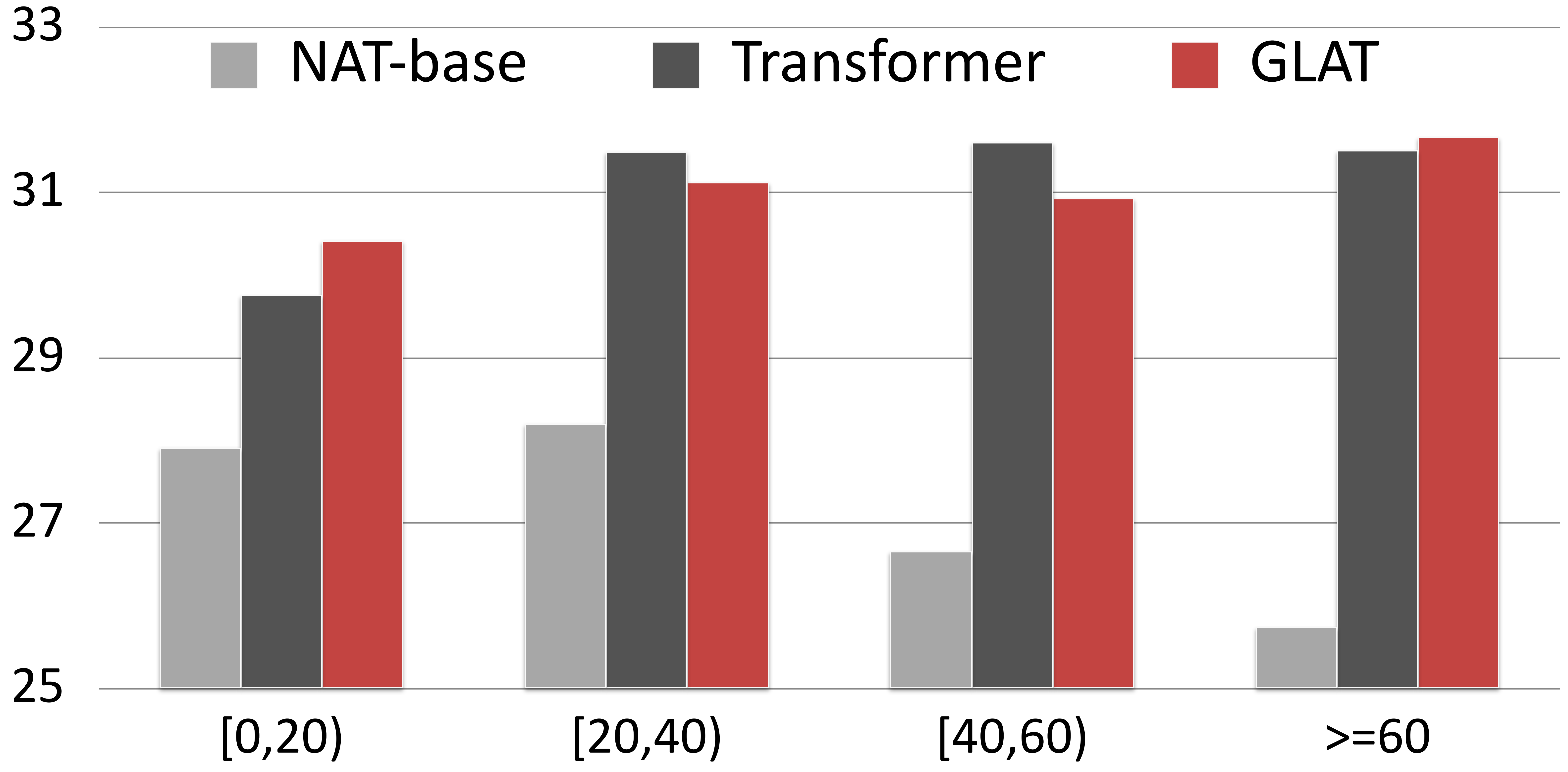


- GLAT achieves high quality translation while keeping high inference speed-up (8x~15x)

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

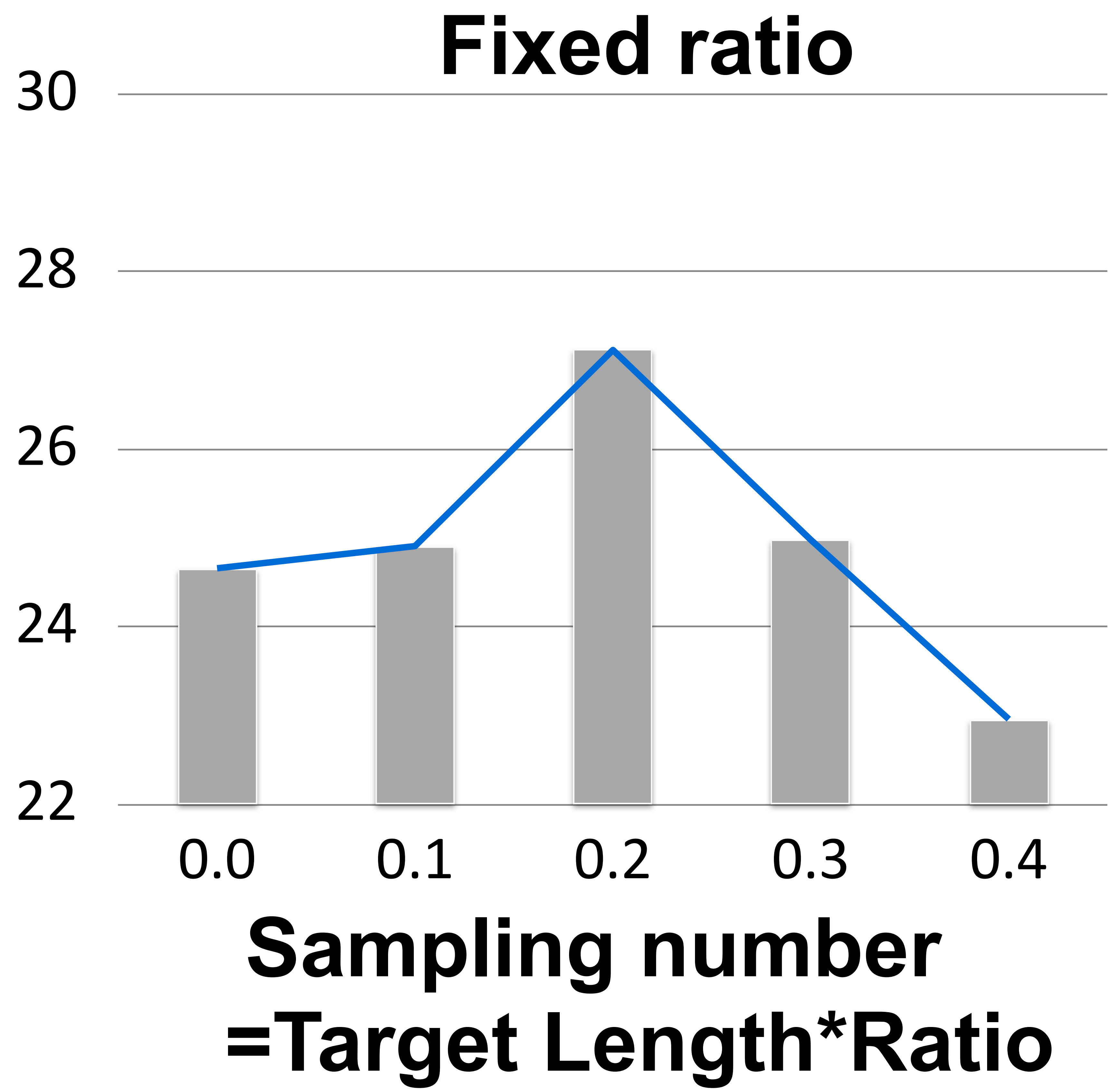
Performance for different lengths

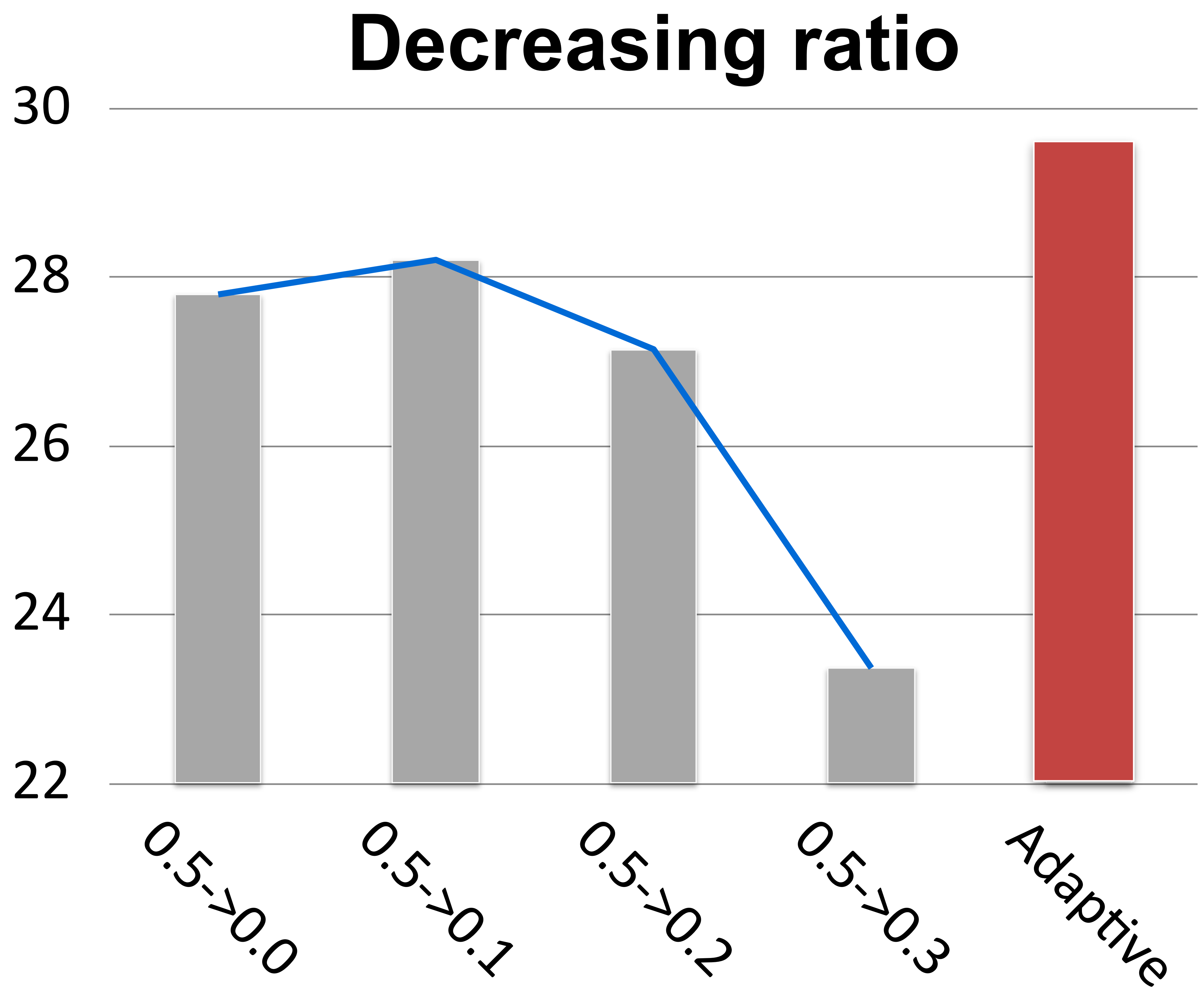
- The performance of NAT-base drops sharply as the input length becomes longer
- GLAT performs a little better than Transformer on WMT14 DE-EN when the input length is shorter than 20



Adaptive sampling number is effective

- The adaptive glancing sampling strategy significantly improves performance





GLAT in Real Competition

GLAT achieve the Top BLEU score in WMT21 En-De and De-En!

The first NAT system to do so!

newstest2021.de-en test set (de-en)

#	Name	BLEU
1	**Anonymous submission #1276**	35.0
2	**Anonymous submission #1284**	35.0
3	Anonymous submission #1304	34.9
4	Anonymous submission #1117	34.9
5	Anonymous submission #1258	34.9
6	**Anonymous submission #1124**	34.9
7	Anonymous submission #543	34.8
8	**Anonymous submission #963**	34.8
9	Anonymous submission #861	34.7
10	Anonymous submission #738	34.7

BLEU and ChrF are sacreBLEU scores. Systems in **bold face** are your submissior validation errors denoted by -1.0 score.

newstest2021.en-de test set (en-de)

#	Name	BLEU
1	**Anonymous submission #1265**	31.3
2	Anonymous submission #1303	31.3
3	Anonymous submission #1291	31.3
4	Anonymous submission #804	31.3
5	Anonymous submission #368	31.3
6	Anonymous submission #1168	31.3
7	**Anonymous submission #1251**	31.2
8	**Anonymous submission #986**	31.2
9	Anonymous submission #1310	31.2
10	**Anonymous submission #1243**	31.2

BLEU and ChrF are sacreBLEU scores. Systems in **bold face** are your submissio validation errors denoted by -1.0 score.

Qian et al. The Volctrans GLAT System: Non-autoregressive Translation Meets WMT21. 2021.

GLAT achieves Top-5 in WMT21 Human Evaluation

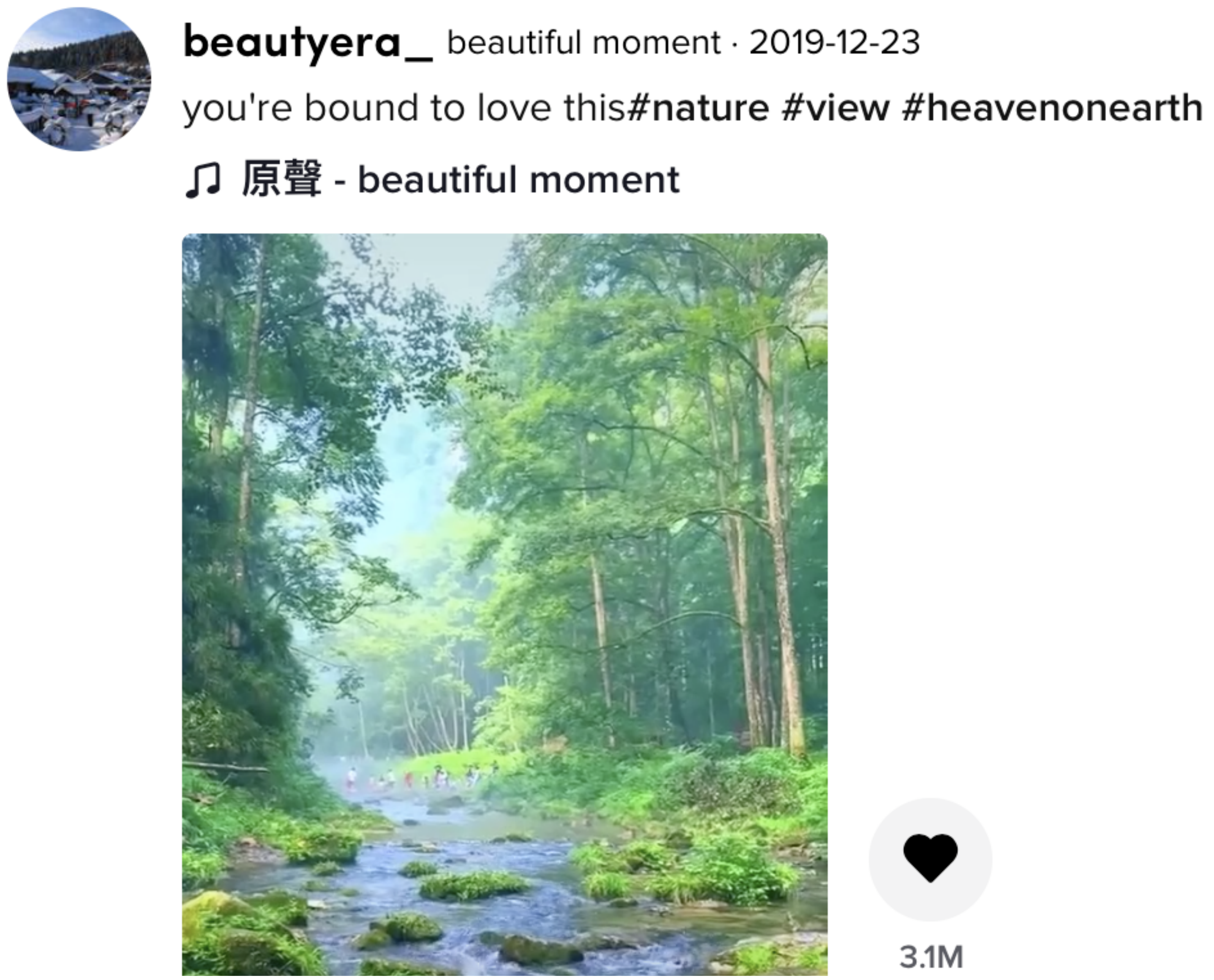
German→English

Rank	Ave.	Ave. z	System
1–5	71.9	0.126	Borderline
1–6	73.5	0.124	Online-A
1–4	78.6	0.122	Online-W
4	79.5	0.113	UF
3–8	73.2	0.106	VolcTrans-AT
4–9	77.5	0.100	Facebook-AI
5–12	75.8	0.068	ICL
4–12	73.4	0.048	Online-G
8–17	69.7	0.016	Online-B
7–17	71.3	0.016	Online-Y
7–17	71.6	0.010	VolcTrans-GLAT
5–16	69.6	0.007	P3AI
9–19	70.6	−0.008	SMU
9–17	73.1	−0.008	UEdin
9–17	69.1	−0.010	NVIDIA-NeMo
10–19	69.9	−0.035	Manifold
15–20	67.0	−0.043	Watermelon
7–17	71.8	−0.061	happypoet
16–20	66.8	−0.081	HUMAN-C
18–20	66.0	−0.120	HW-TSC

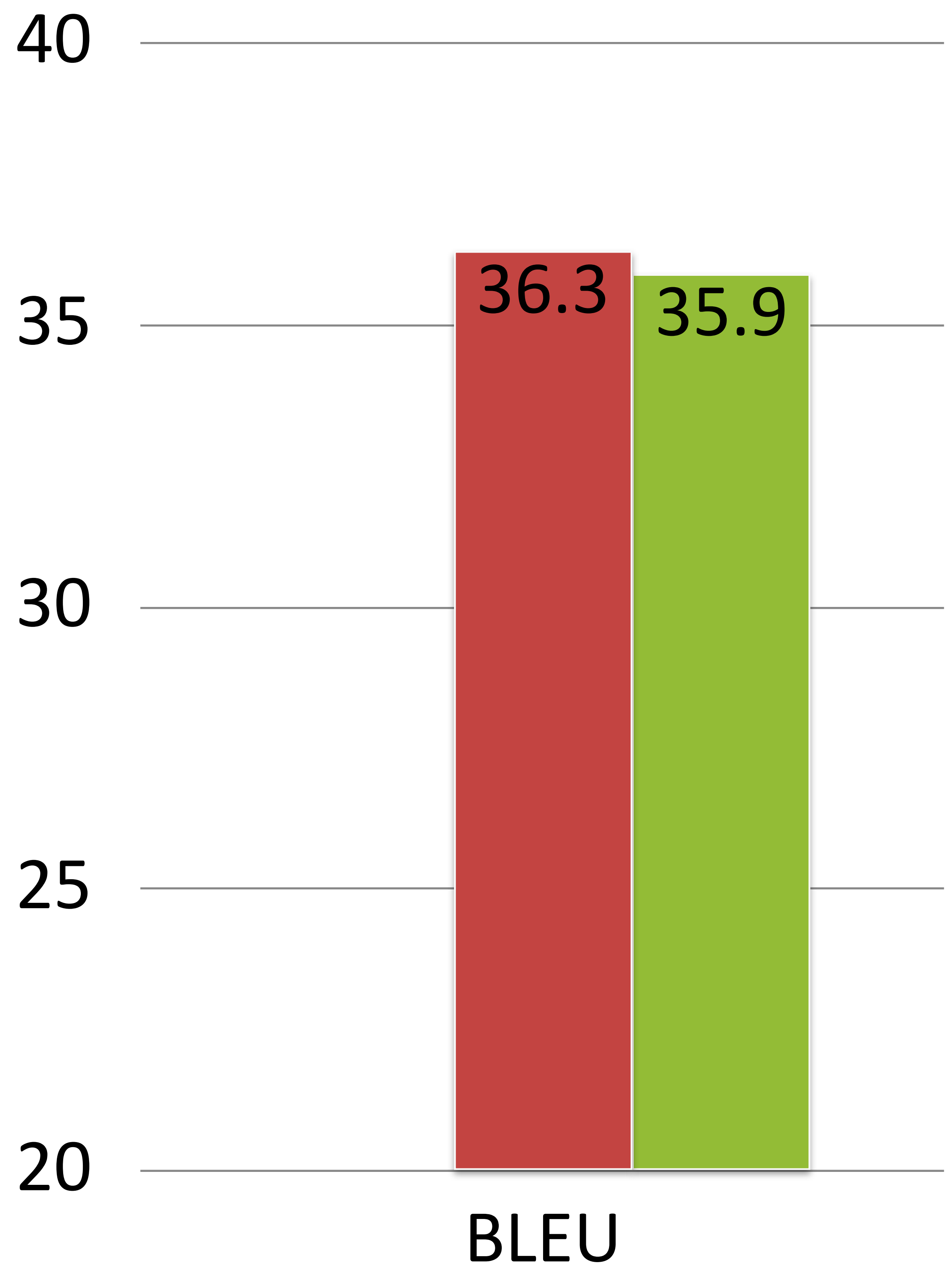
Findings of WMT21.

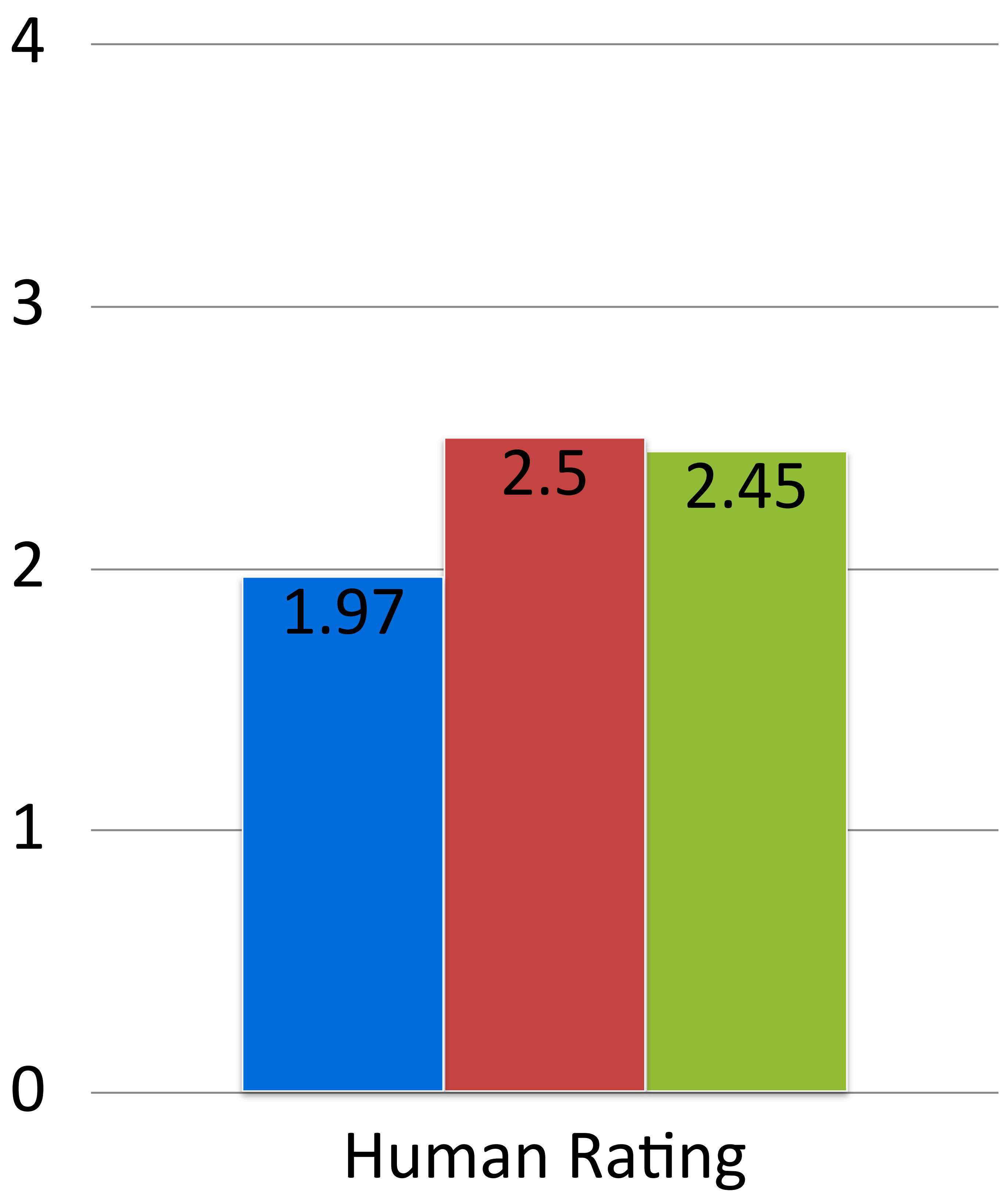
GLAT is the first production NAT system!

- Already deployed online in VolcTrans and serving English-Japanese



Tiktok caption translation





火山翻译
Volctrans

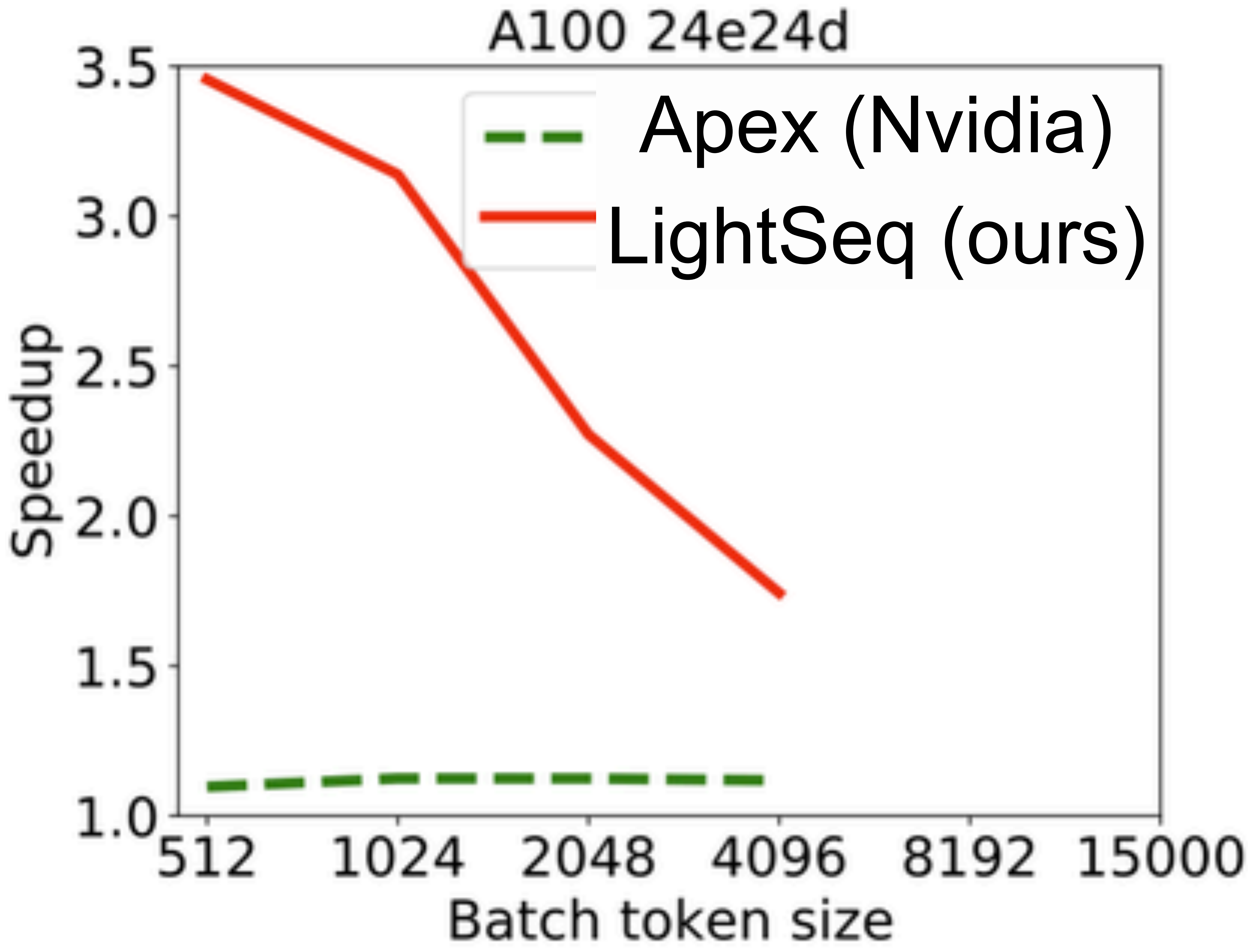
LightSeq A high-performance library for Transformer-like Models

- **Efficient for both training and inference**
  - LightSeq achieves up to 14x speedup compared with TensorFlow and Pytorch in Inference.
  - LightSeq2 achieves 45-250% speed up over Pytorch(FairSeq) in training
- **Rich functions**
  - LightSeq supports more architecture variants and different search algorithms.
- **Seamless Co-operate**
  - LightSeq is easy to use without any code modification.
  - Seemless porting from Tensorflow, Pytorch, Hugginface, Fairseq
- Open source on github: already 2k stars!

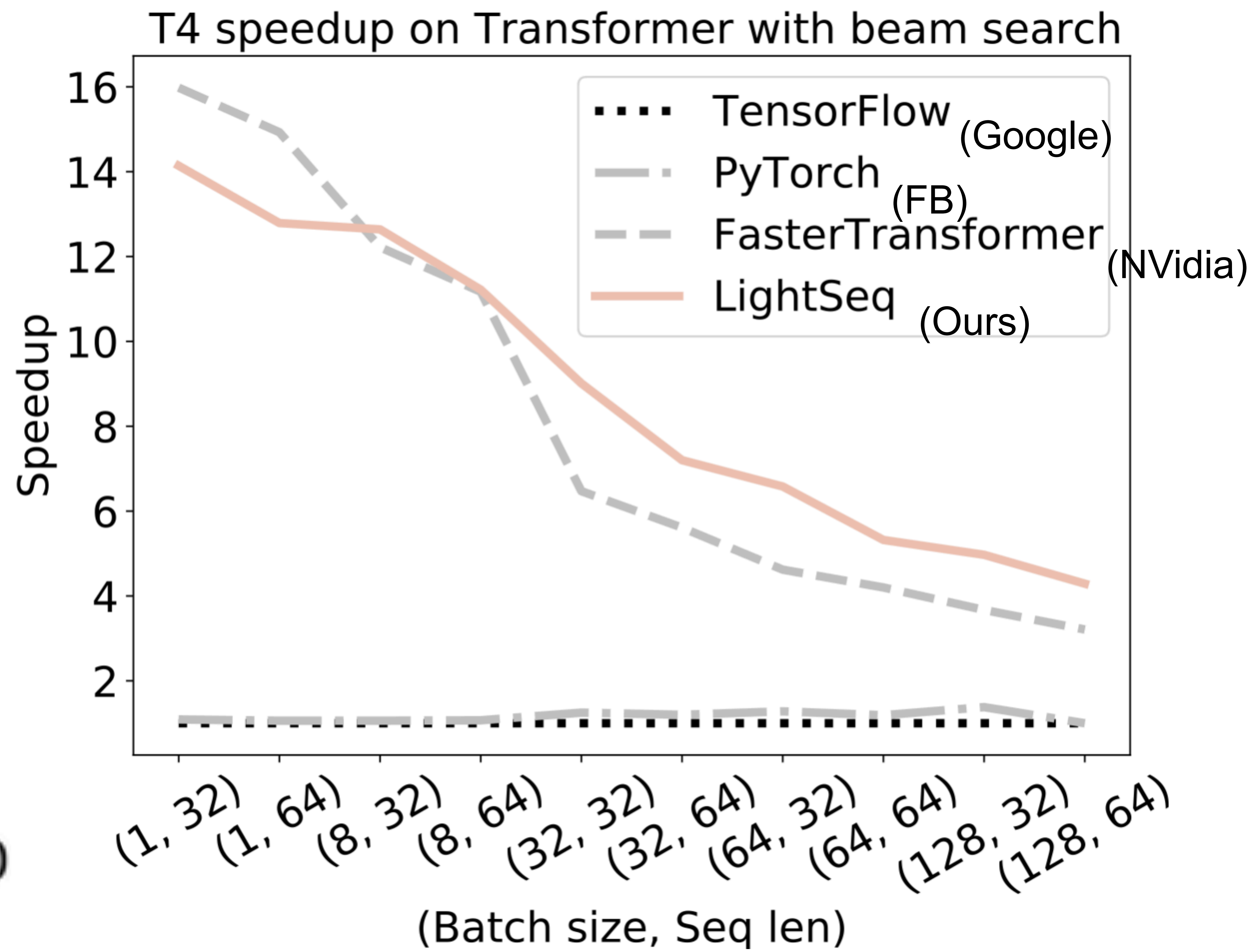
Wang, Xiong, Wei, Wang, Li. LightSeq: A High Performance Inference Library for Transformers. NAACL 2021.
Wang et al. LightSeq2: Accelerated Training for Transformer-based Models on GPUs.

Lightening Fast Transformer Lib

Transformer Training
Speedup comparing to Pytorch (FB)



Transformer Inference
Speedup comparing to Tensorflow



LightSeq is open-sourced on Github, 2k stars already! Ready to integrate with Tensorflow and Pytorch.

Summary

- Word interdependency learning is important
- GLAT can achieve comparable generation quality with autoregressive models
- A generation paradigm with great potential

Language Presentation

Read List

- Gu et al, Non-Autoregressive Neural Machine Translation, ICLR 2018.
- Ghazvininejad et al. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. EMNLP 2019.
- Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.
- Wang, Xiong, Wei, Wang, Li. LightSeq: A High Performance Inference Library for Transformers. NAACL 2021.