Context Consistency between Training and Inference in Simultaneous Machine Translation

Meizhi Zhong[1], **Lemao Liu** *, **Kehai Chen**[1] *, **Mingming Yang**, **Min Zhang**[1]
[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
22s051052@stu.hit.edu.cn, lemaoliu@gmail.com,
chenkehai@hit.edu.cn, shanemmyang@gmail.com, zhangmin2021@hit.edu.cn

Abstract

Simultaneous Machine Translation (SiMT) aims to yield a real-time partial translation with a monotonically growing source-side context. However, there is a counterintuitive phenomenon about the context usage between training and inference: *e.g.*, in wait-k inference, model consistently trained with wait-k is much worse than that model inconsistently trained with wait-k' ($k' \neq k$) in terms of translation quality. To this end, we first investigate the underlying reasons behind this phenomenon and uncover the following two factors: 1) the limited correlation between translation quality and training loss; 2) exposure bias between training and inference. Based on both reasons, we then propose an effective training approach called context consistency training accordingly, which encourages consistent context usage between training and inference by optimizing translation quality and latency as bi-objectives and exposing the predictions to the model during the training. The experiments on three language pairs demonstrate that our SiMT system encouraging context consistency outperforms existing SiMT systems with context inconsistency for the first time. [1]

1 Introduction

Simultaneous machine translation (SiMT) (Cho and Esipova, 2016; Gu et al., 2017; Ma et al., 2019) aims to generate a partial translation while incrementally receiving a prefix of a source sentence. SiMT plays an very improtant role in many real-world scenarios such as multilateral organizations and international summits (Ma et al., 2019). A good SiMT system should not only have *low latency* in the generation process but also yield a translation with *high quality*. Hence, there has recently been witnessed a surge of interest in the research about SiMT (Elbayad et al., 2020; Ma et al., 2020; Zhang and Feng, 2021, 2022).

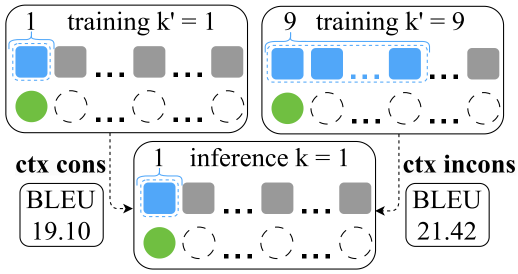


Figure 1: Counterintuitive phenomenon on the context usage between training and inference: in wait-1 inference ($k = 1$), model trained with k'=9 (denoted by "ctx incons") outperforms the model trained with k'=1 (denoted by "ctx cons") in terms of BLEU, even though the former model (trained by k'=9) induces a mismatch on context usage between training and inference.

In this paper, we shed light on a *counterintuitive phenomenon* on the context usage between training and inference in SiMT: in wait-k inference, model consistently trained with wait-k is worse than that model inconsistently trained with wait-k' ($k' \neq k$) in terms of BLEU scores (Papineni et al., 2002), as shown in Figure 1. This phenomenon was first observed by Ma et al. (2019) yet without explanations. Subsequently, such context inconsistency training becomes a standard practice (Elbayad et al., 2020; Zhang and Feng, 2021, 2022), even if this phenomenon is counterintuitive due to the mismatch between training and inference on the usage of partial source-side context.

To investigate the reasons behind the above counterintuitive phenomenon, we conduct experiments from two perspectives: calculating the correlation between BLEU scores and cross-entropy loss, as well as evaluating the translation quality under the prefix-constrained decoding setting. Our empirical experiments demonstrate two reasons that are responsible for the phenomenon: 1) the

*Corresponding authors
[1]Code is available at https://github.com/zhongmz/ContextConsistencyBiTraining4SiMT

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13465–13476
August 11-16, 2024 ©2024 Association for Computational Linguistics

limited correlation between translation quality and training loss; 2) exposure bias between training and inference. Moreover, based on our findings, this paper proposes an effective training approach called context consistency training. Its key idea is to make the context usage consistent between training and inference by optimizing translation quality and latency as bi-objectives and exposing the model to its own predictions during the training stage. Particularly, this approach is general to be applied to most SiMT systems. Experiments conducted across various benchmarks demonstrate that the proposed context consistency training towards bi-objectives achieves substantial gains over the original consistency training based on cross-entropy. Our main contributions are:

- This paper sheds light on a counterintuitive phenomenon about context usage between training and inference in SiMT and provides comprehensive explanations.
- Based on our findings, this paper proposes a simple yet effective context consistency training method, which breaks through the standard practice of inconsistent training.
- Experimental results demonstrate that our SiMT system encouraging context consistency outperforms the existing systems with context inconsistency for the first time.

2 Rethinking Counterintuitive Phenomenon on Context Usage

2.1 Counterintuitive Phenomenon

Counterintuitive Phenomenon on Valid Set. In wait-k systems, the counterintuitive phenomenon of the context usage between training and inference was first observed by Ma et al. (2019) yet without explanations: ***in wait-k inference, model trained consistently with the same wait-k setting is worse than the model trained with the wait-k' setting ($k' \neq k$) in terms of translation quality***, as illustrated in Table 1.[2] For example, the BLEU scores obtained by the model trained with wait-9 surpasses the model trained with wait-1 by a large margin with wait-1 inference. As a result, it has become a standard practice to utilize inconsistent context for training, and this practice is widely followed by (Elbayad et al., 2020; Zhang and Feng, 2021; Zhang et al., 2022; Guo et al., 2023), even if this phenomenon is counterintuitive due to the mismatch between training and inference on the usage of source-side context.

Train \ Inference	k=1	k=3	k=5	k=7	k=9
k'=1	<u>19.10</u>	18.06	17.42	16.94	16.80
k'=3	19.29	<u>23.76</u>	24.97	25.00	24.40
k'=5	20.33	**24.89**	<u>26.36</u>	26.93	27.27
k'=7	20.48	24.60	26.46	<u>27.26</u>	27.81
k'=9	**21.42**	24.82	**26.92**	**27.84**	<u>**28.63**</u>

Table 1: Evaluation by BLEU scores on the valid set of the WMT15 De-En task for wait-k policy . Bold: best in a column. Underline: training context is consistent with inference context. (§4 provides detailed settings.)

Counterintuitive Phenomenon on Training Subset. One might hypothesize that this phenomenon is attributed to the generation issue from training data to valid data. To verify this hypothesis, we conduct similar experiments on a subset of the training data. We sample examples from the training data as a training subset with the same size as the valid set. Table 2 depicts that the situation on the training subset is almost similar to that on the valid set except for $k = 3$, where the optimal $k' = 9$ for the training subset rather than $k' = 5$ as for the valid set. This shows that generalization from training data to valid data is not the main reason for this counterintuitive phenomenon and it is non-trivial to analyze its reasons.

Train \ Inference	k=1	k=3	k=5	k=7	k=9
k'=1	<u>21.42</u>	21.21	21.00	20.25	19.67
k'=3	22.07	<u>25.51</u>	26.73	26.69	26.33
k'=5	22.53	25.55	<u>27.27</u>	28.06	28.07
k'=7	23.15	25.73	27.20	<u>28.34</u>	28.63
k'=9	**23.22**	**26.21**	**27.52**	**28.66**	<u>**29.33**</u>

Table 2: Evaluation by BLEU scores on the training subset of the WMT15 De-En task for wait-k policy.

2.2 Reasons of Counterintuitive Phenomenon

Correlation between BLEU and Cross-entropy Loss in SiMT. Firstly, we explore the correlation between translation quality (e.g., BLEU scores) and training loss (e.g., cross-entropy). Specifically, we measure both the training loss and translation quality of each sample and calculate their absolute Pearson Correlation in the training subset. However, training loss is measured at the word level, while translation quality for a sentence is

[2]To clarify, this observation specifically pertains to the lower triangle of the table.

measured at the sentence level. To bridge this disparity, we compute the average training loss for each word within a sentence, thus representing it as sentence-level training loss.

k	1	3	5	7	9	∞
Entire	0.62	0.70	0.73	0.74	0.75	0.75
Low	0.68	0.73	0.74	0.75	0.76	0.75
High	0.27	0.44	0.51	0.56	0.60	0.64

Table 3: Correlation between BLEU scores and cross-entropy loss on three subsets from the training subset of WMT15 De-En for wait-k policy, where $k=\infty$ means Full-sentence MT. **Entire** denotes the entire training subset, **Low** consists of those samples whose cross-entropy loss is lower than the averaged loss, **High** consists of those samples whose loss is higher than the averaged loss.

Table 3 presents the results of the correlation between BLEU scores and cross-entropy loss in the wait-k policy. We reveal the following insights. 1) In the wait-k policy, especially when k is smaller, the correlation is lower than that in Full-sentence MT. 2) When evaluating samples with high cross-entropy loss, we observe a weaker correlation (between training loss and BLEU) compared to that with low training loss. This observation is not difficult to understand: taking a two-class classification task as an example, if the cross-entropy loss of an example is very high (e,g., the loss is $-\log 0.2$), then the model can not predict the correct label for this example even if its loss is improved to $-\log 0.4$, because the probability of the ground-truth label is 0.4, which is less than 0.5. ***This suggests that the reason for the counterintuitive phenomenon on context usage is attributed to the relatively high cross-entropy loss for SiMT*** [3]***, leading to the weak correlation between cross-entropy loss and BLEU scores.***

Effects of Exposure Bias on the Models Trained Consistently and Inconsistently. Since the SiMT model is typically trained by cross-entropy loss, it suffers from the well-known exposure bias, i.e., during training, the model is only exposed to the training data distribution, instead of its predictions. Therefore, we focuses on studying the effects of exposure bias on the model trained with consistent context as well as the model trained with inconsistent context. To control the extent of exposure bias during the inference, we measure translation quality by BLEU scores for both models (e.g., the former wait-1 inference model is trained with wait-1 setting and the latter wait-1 inference model is trained with wait-9 setting) under the prefix-constrained decoding setting (Wuebker et al., 2016), where each model requires to predict the suffix for a given gold prefix. Under this setting, as the gold prefix gets shorter, more predicted tokens are used as the context during the prefix-decoding stage and the exposure bias is more severe.

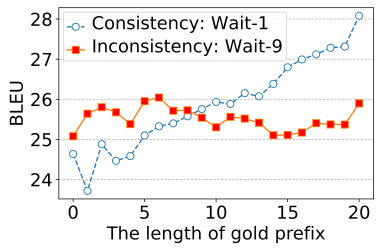


Figure 2: BLEU scores comparison between context consistency and context inconsistency under the prefix-constrained decoding setting. The x-axis denotes the number of tokens for the gold prefix.

The results as presented in Figure 2 are averaged from a subset of 400 sentence pairs in the train set, all having the same number of tokens in the target (20 target tokens). It is evident that as the gold prefix becomes shorter (i.e., exposure bias is more severe), the performance of the consistent model significantly deteriorates while the inconsistent model's performance remains relatively better; however, when the number of tokens in the gold prefix is larger than 10 (i.e., exposure bias is less severe), the consistent model performs better. ***This finding reveals that one of the underlying causes of the counterintuitive phenomenon is attributed to exposure bias*** (Ranzato et al., 2016; Bengio et al., 2015; Zhang et al., 2019).

2.3 Counterintuitive Phenomenon Depends on Evaluation Metrics

The above reasons motivate us to study the counterintuitive phenomenon by using the cross-entropy loss for evaluation, in addition to BLEU as before, because training and inference criteria are the same, and there is no exposure bias issue in this case. We evaluate cross-entropy loss for the wait-k inference while models trained with wait-k' settings on the valid set and training subset, as illustrated in Table 4. On the valid

[3] Compared with full-sentence MT, SiMT uses less source-side context, which essentially results in a higher cross-entropy loss.

	Valid set					Training subset				
	k=1	k=3	k=5	k=7	k=9	k=1	k=3	k=5	k=7	k=9
k'=1	**5.78**	5.26	5.00	4.87	4.81	**5.43**	5.11	4.95	4.87	4.83
k'=3	5.78	5.12	4.79	4.61	4.53	5.48	**5.03**	4.83	4.73	4.67
k'=5	5.81	**5.10**	**4.73**	4.53	4.42	5.54	5.06	**4.81**	4.69	4.61
k'=7	5.86	5.12	4.72	4.50	4.38	5.60	5.09	4.82	**4.67**	4.59
k'=9	5.91	5.14	4.72	**4.49**	**4.36**	5.65	5.12	4.84	4.68	**4.58**

Table 4: Evaluation by cross-entropy loss on valid set and training subset of WMT15 De-En for wait-k policy (k=$*$ and k'=$*$ refer to the inference and the training, respectively).

set, we almost notice a diagonal trend, indicating the superiority of the consistent model. On the training subset, we observe a similar diagonal trend, indicating the counterintuitive phenomenon disappears in terms of cross-entropy loss as the evaluation metric. ***These observation suggests that the counterintuitive phenomenon of context usage between training and inference depends on evaluation metrics, and it might be helpful to address this phenomenon by encouraging the consistent criterion between training and inference.***

3 Context Consistency Training for SiMT

Previous findings have shown that: 1) it is helpful to address the counterintuitive phenomenon by encouraging the consistent criterion between training and inference; 2) exposure bias is a reason for the counterintuitive phenomenon. To address the counterintuitive phenomenon and make the consistent model successful, we propose a simple yet effective training approach, called context consistency training for SiMT, which not only incorporates the evaluation metric (e.g., BLEU, COMET and ChrF) for SiMT as training objectives (§3.1) but also allows the model to expose its predictions during training (§3.2).

3.1 Bi-Objectives Optimization for SiMT

In SiMT, the evaluation metrics of models are translation quality and latency. Therefore, we intend to leverage both of these metrics as bi-objectives in our proposed method. Specifically, BLEU (or COMET or ChrF) score is taken as an example to measure the translation quality of SiMT models. Average Lagging (AL) (Ma et al., 2019) is used as the Latency measurement. AL quantifies the number of tokens of hypotheses that fall behind the ideal policy and is calculated as:

$$\mathrm{AL}_g(\mathbf{x}, \mathbf{u}) = \frac{1}{\tau}\sum_{i=1}^{\tau} g(i, \mathbf{u}) - \frac{i-1}{|\mathbf{u}|/|\mathbf{x}|}, \quad (1)$$

where $\tau = \mathrm{argmax}_i \{i \mid g(i, \mathbf{u}) = |\mathbf{x}|\}$, $\mathbf{x}$ is the source sentence, $\mathbf{u}$ is the hypothesis sentence, and $g(i)$ is the number of waited source tokens before translating $\mathbf{u}_i$ and thus it is dependent on $\mathbf{u}_{<i}$, and its detailed definition depends on different read/write policies.

Formally, the SiMT model parametrized by θ can be defined as follows:

$$p_g(\mathbf{u}|\mathbf{x}; \theta) = \prod_{i=1}^{|\mathbf{u}|} p(\mathbf{u}_i | \mathbf{x}_{\leq g(i,\mathbf{u})}, \mathbf{u}_{<i}), \quad (2)$$

where $\mathbf{u}$ denotes a complete translation hypothesis and $\mathbf{u}_{<i}$ denotes its partial prefix with i tokens.

Inspired by Minimum Risk Training (MRT) (Shen et al., 2016; Wieting et al., 2019), we directly optimize the SiMT model towards its bi-objectives (i.e., BLEU and Latency) as follows:

$$\mathcal{L}_g = \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \mathrm{cost}_g(\mathbf{x}, \mathbf{y}, \mathbf{u}) \frac{p_g(\mathbf{u}|\mathbf{x}; \theta)}{\sum_{\mathbf{u}' \in \mathcal{U}(\mathbf{x})} p_g(\mathbf{u}'|\mathbf{x}; \theta)}, \quad (3)$$

where $\mathcal{U}(\mathbf{x})$ is a set of candidate hypotheses, $\mathbf{y}$ is the reference and $\mathrm{cost}_g(*)$ is the bi-objectives:

$$\mathrm{cost}_g(\mathbf{x}, \mathbf{y}, \mathbf{u}) = \gamma \cdot \mathrm{AL}_g(\mathbf{x}, \mathbf{u}) + (1-\gamma) \cdot (1 - \mathrm{BLEU}(\mathbf{y}, \mathbf{u})). \quad (4)$$

The hyperparameter γ is adjustable and allows us to fine-tune for different latency requirements.

Remark. In Shen et al. (2016) and Wieting et al. (2019)'s studies, the cost is directly defined on a translation candidate $\mathbf{u}$, and thus it is trivial to calculate the cost for a given $\mathbf{u}$. In comparision, $\mathrm{AL}_g(\mathbf{x}, \mathbf{u})$ depends not only on $\mathbf{u}$ but also on $g(i, \mathbf{u})$ specified by the read/write policy used in the our SiMT system. During the training, we access the SiMT model to incrementally compute the $g(i, \mathbf{u})$ for all i and then compute $\mathrm{AL}_g(\mathbf{x}, \mathbf{u})$ based on all $g(i, \mathbf{u})$ for each candidate $\mathbf{u}$ generated via decoding.

3.2 Generating n Candidates for Training SiMT

Generally, SiMT is trained by using cross-entropy loss, and its decoding does not consider multiple candidates. To calculate the objective function defined in eq. (3), our SiMT system generates a set of candidates $\mathcal{U}$ via decoding which also allows the

SiMT model to be exposed to the predictions and thereby mitigates exposure bias during the training. To this end, we try two different ways, including Beam search and Sampling search (Holtzman et al., 2020), to generate n-best candidates in SiMT. Beam search is a maximization-based decoding technique that optimizes output by favoring high-probability tokens. Sampling search (Holtzman et al., 2020) is a stochastic decoding approach that samples from the top-p portion of the probability distribution. In our experiments, we generate a set of 5-best candidates in the beam search and select 0.8 for top-p in the sampling search.

Moreover, to calculate the $\text{AL}_g(\mathbf{x}, \mathbf{u})$ of candidates defined in Eq. (1) which is dependent on the $g(i)$, we maintain both model score p_g as well as $g(i)$ (the number of waited source words before translating $\mathbf{u}_i$) at each time-step i. Specifically, during the decoding, the SiMT model uses the value of $g(i)$ to incrementally specify the source context and produce the next predictive distribution p_g. From this predictive distribution p_g, we select the top n-best or sample n partial candidates along with their respective $g(i)$ values.

Following Edunov et al. (2018); Wieting et al. (2019), we employ the two-step training paradigm to train SiMT to speed up the training process: we first train the SiMT model with the cross-entropy loss, and then we fine-tune the model by optimizing the bi-objectives (e.g., BLEU and AL) with the generated n-best candidates in our context consistency training. It is worth noting that n candidates are generated during the training, but the greedy search is only used in the inference.

4 Experiments

4.1 Dataset and System Settings

The proposed approach is evaluated on three widely used benchmarks, including IWSLT14 German→English (De-En), IWSLT15 Vietnamese→English (Vi-En) and WMT15 German → English (De-En). Experiments are conducted on SiMT systems including two policies: The fixed read/write system (**wait-k policy**); The adaptive read/write system (**wait-info policy**) (Zhang et al., 2022).

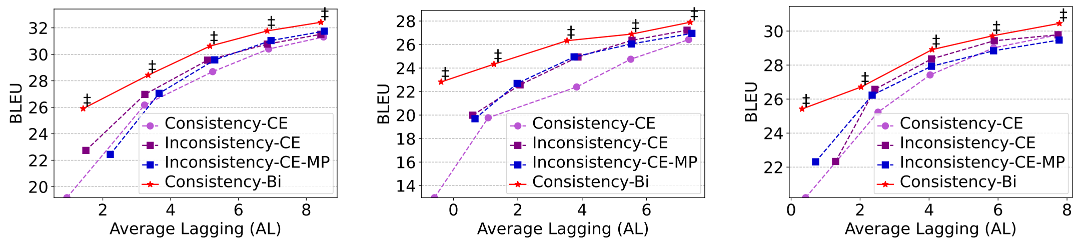
Baselines. The conventional training approach of SiMT systems is the context consistency training based on cross-entropy Ma et al. (2019), denoted **Consistency-CE**. In contrast, context inconsistency training, also based on cross-entropy, involves inconsistent context usage between training and inference stages, called **Inconsistency-CE**. Additionally, we implement a recently widely-used special case of context inconsistency training (Elbayad et al., 2020), termed **Inconsistency-CE-MP**, which involves sampling different values of k during the training.

Our Training Approaches. The proposed SiMT systems follow the standard evaluation paradigm (Ma et al., 2019) and report BLEU (Papineni et al., 2002) , ChrF (Popović, 2015) and COMET (Rei et al., 2020) scores for translation quality and Average Lagging (AL) (Ma et al., 2019) for latency mentioned in §3.1. The proposed context consistency training is based on bi-objectives, called **Consistency-Bi**, and we also implement the context consistency training based on BLEU as the uni-objective, called **Consistency-Uni** for further comparison. For generating n candidates, we implement Beam search in most cases, except the wait-k policy, for which we utilize the Sampling search strategy. The implementation of all systems is based on Transformer in the Fairseq Library (Ott et al., 2019). Appendix A provides detailed experimental settings.

4.2 Main Results

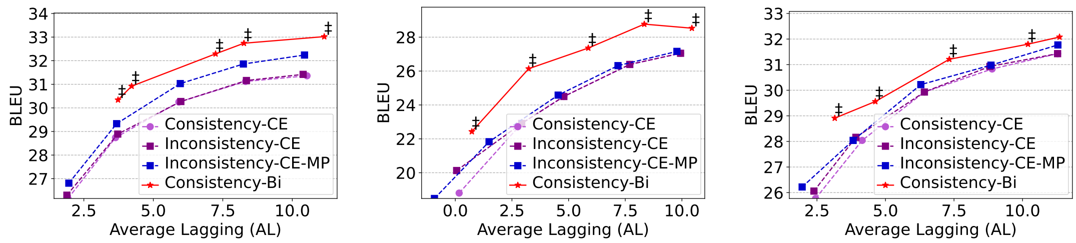
The results of BLEU scores are displayed in Figures 3 and 4. The results of ChrF scores are presented in Figures 5 and 6, while the results of COMET scores are shown in Figures 7 and 8. The experimental results for both metrics are comparable. Therefore, we will discuss using the BLEU scores as an example in this section. Within our proposed context consistency training approach (Consistency-Bi), all implemented SiMT systems (wait-k and wait-info) exhibit significant improvements in both translation quality and latency, as evidenced by an increase in BLEU scores and a decrease in AL across all the benchmarks. This reveals that our proposed methods not only yield substantial performance improvements but also demonstrate strong generalization capabilities.

In contrast to the original consistency training of the wait-k policy, our proposed Consistency-Bi achieves over 5 BLEU improvement at low latency (k=1) across all datasets. Specifically, our method improves 2.68 BLEU on the IWSLT14 De-En task, 4.39 BLEU on the IWSLT15 Vi-En task, and 1.91 on the WMT15 De-En task, respectively (average on all latency). Furthermore, compared with inconsistency training, the proposed method also



(a) IWSLT14 De-En (b) IWSLT15 Vi-En (c) WMT15 De-En

Figure 3: Translation quality (BLEU) v.s. latency (Average Lagging, AL) in Wait-k Policy. "‡/†" indicates significant difference ($p < 0.01/0.05$) from Consistency-CE. Specifically, we take the compare-mt Library (Neubig et al., 2019) to compute the significance testing results.



(a) IWSLT14 De-En (b) IWSLT15 Vi-En (c) WMT15 De-En

Figure 4: Translation quality (BLEU) v.s. latency (Average Lagging, AL) in Wait-info Policy.

demonstrates significant improvements, especially at low latency (k=1), achieving over 3 BLEU scores increases. This suggests that incorporating our proposed method enables a wait-k model trained consistently under the same wait-k inference setting to outperform an inconsistent one.

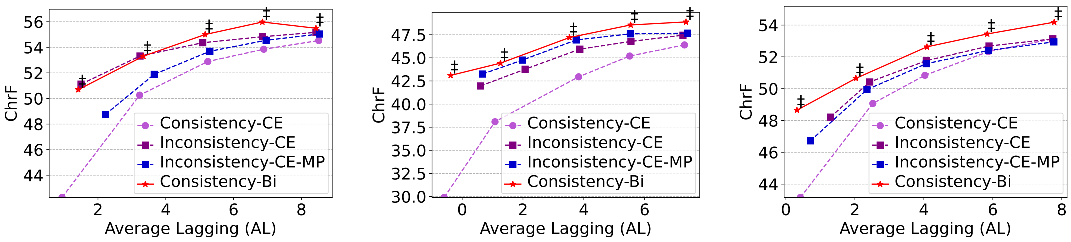
To evaluate whether our method could achieve improvements with advanced adaptive SiMT systems, we apply our proposed training method to wait-info policy (Zhang et al., 2022). The results are depicted in Figure 4. Similarly, in comparison to the three baseline training methods, we observe a significant enhancement in translation quality across all latencies. However, in IWSLT15 Vi-En and WMT15 De-En tasks, Inconsistency-CE and Inconsistency-CE-MP are not significantly better than Consistency-CE. This can be attributed to the advanced policy, which makes more informed read/write decisions based on information.

4.3 Ablation Study

Ablation Studies on Consistency-Bi and Consistency-Uni. To validate the effectiveness of Consistency-Bi, we perform the ablation studies on Consistency-Bi (Both BLEU and AL) and Consistency-Uni (BLEU only) in Figure 9. The experiments reveal that compared with Consistency-Uni, Consistency-Bi not only results in lower latency but also yields superior translation quality, especially in low latency scenarios (k=1), except for k=3, where Consistency-Uni is slightly better than Consistency-Bi. It is largely attributed to the latency as part of the training objectives.

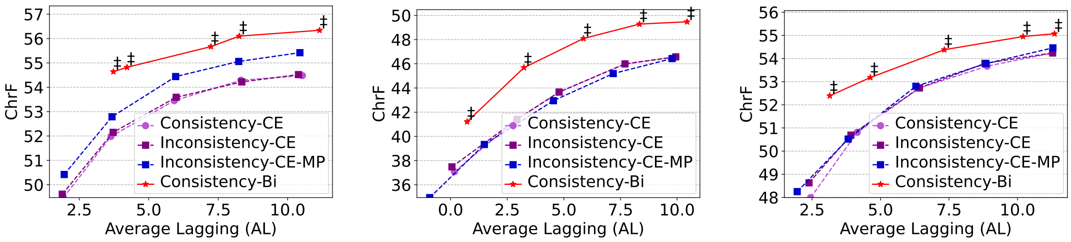
Ablation studies on n-best candidates generations. We conduct the ablation studies on two types of n-best generation methods (Beam search and Sampling search) under both wait-k and wait-info policies, as depicted in Figure 10. The results reveal that under the wait-k policy, the performance of Consistency-Bi using sampling search is slightly superior to that using beam search. Conversely, under the wait-info policy, employing beam search yields slightly better results compared to sampling search. These findings suggest the choice of generation method is not notably sensitive.

Variation in hyperparameter γ. Fine-tuning hyperparameter γ defined in (4) aims to achieve a better trade-off between BLEU and latency in our proposed Consistency-Bi. As illustrated in Table 5, as γ increases, AL decreases while the BLEU scoresimproves, reaching its peak at γ = 0.4. This indicates that our proposed method can simultaneously optimize two objectives and achieve a value that is relatively optimally balanced



(a) IWSLT14 De-En (b) IWSLT15 Vi-En (c) WMT15 De-En

Figure 5: Translation quality (ChrF) v.s. latency (Average Lagging, AL) in Wait-k Policy.



(a) IWSLT14 De-En (b) IWSLT15 Vi-En (c) WMT15 De-En

Figure 6: Translation quality (ChrF) v.s. latency (Average Lagging, AL) in Wait-info Policy.

between BLEU and AL, which can effectively enhance both translation quality and latency.

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6
BLEU	23.5	23.37	23.08	23.56	**24.21**	21.09	17.74
AL	1.68	1.62	1.53	1.14	0.16	-1.48	**-2.93**

Table 5: Ablation studies on various γ in wait-1 training with wait-1 inference of Consistency-Bi.

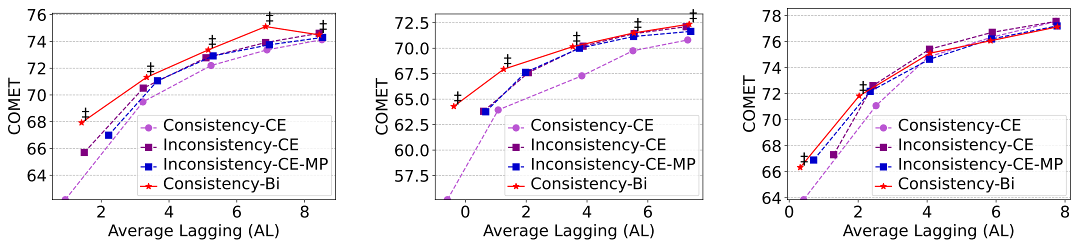
4.4 Analysis

Counterintuitive Phenomenon Mitigation. To explore whether the counterintuitive phenomenon described in §2.1 is alleviated, we conduct experiments using models trained with wait-k' but tested with wait-k , as illustrated in Figure 11. Figure 11(a) presents the results of the original training method. Optimal results for inference with k are generally achieved when k'=9, except for k=3, where k'=5 yields the best. In contrast, our proposed training method demonstrates that the best results tested with wait-k closely match with the diagonal line as depicted in Figure 11(b). Specifically, when inference with k=1 and 9, the best results match the models trained with the same value of k'. For k=3, 5, and 7, although the best results come from different models, the differences are not significant. These findings suggest that our method exhibits improved consistency between training and inference compared with the origin.

Correlation between training loss and translation quality. We analyze the correlation between BLEU scoresand training loss, similar to the analysis described in §2.2. The results shown in Figure 12 demonstrate that, compared with Consistency-CE, proposed Consistency-Bi exhibits a strong correlation between training loss and translation quality, even when using a small k.

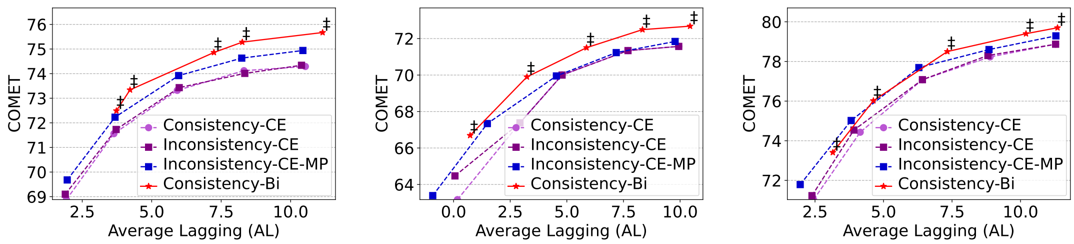
Exposure Bias. To assess whether our method successfully mitigates exposure bias discussed in §2.2, we conduct wait-1 decoding experiments using both Consistency-CE and Consistency-Bi under the prefix-constrained decoding setting (Wuebker et al., 2016). The detailed experimental settings are as described in §2.2. Figure 13 reveals that as the number of gold prefixes decreases, the performance of Consistency-Bi improves, while the performance of Consistency-CE deteriorates. This suggests that the proposed method effectively mitigates exposure bias, enhancing the model's performance when relying on prediction rather than on gold prefixes.

Training Efficiency. We demonstrate the training efficiency of our proposed method and the Consistency-CE Baseline in Table 6. All experiments were performed on the NVIDIA GeForce



(a) IWSLT14 De-En (b) IWSLT15 Vi-En (c) WMT15 De-En

Figure 7: Translation quality (COMET) v.s. latency (Average Lagging, AL) in Wait-k Policy.



(a) IWSLT14 De-En (b) IWSLT15 Vi-En (c) WMT15 De-En

Figure 8: Translation quality (COMET) v.s. latency (Average Lagging, AL) in Wait-info Policy.

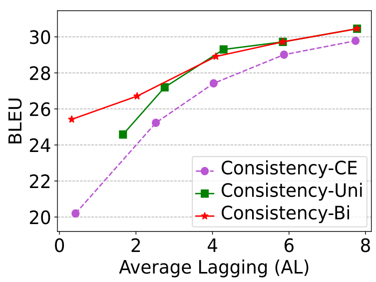
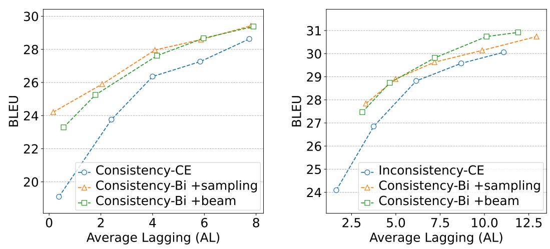


Figure 9: Ablation studies between Consistency-Bi and Consistency-Uni on WMT15 De-En test set of wait-k.

	Init	Wait-1	Wait-3	Wait-5	Wait-7	Wait-9	Total
Consistency-CE	-	5.06	5.05	5.09	5.04	5.06	25.30
Consistency-Bi	5.06	1.60	1.76	1.59	1.54	1.62	13.17

Table 6: The total GPU time (GPU-Hours) required for training each model in the wait-k system on the WMT15 De-En task.

RTX 4090. Initially, we train the SiMT models utilizing the standard cross-entropy loss, followed by fine-tuning during the context consistency training phase. This stage necessitates training on a limited number of updates, approximately 2000 steps, thereby reducing GPU time consumption. In our scenario, "GPU-Hours" represents the cumulative number of hours spent by all GPUs used for training. For example, in Table 6, when the fine-tuning for the initial model takes the Consistency-



(a) Wait-k Policy (b) Wait-info Policy

Figure 10: Ablation studies on n-best candidates generations (Beam search and Sampling search) on the valid set of WMT15 De-En.

Bi and wait-5 policy, the total training time "GPU-Hours" is 1.59: fine-tuning of the initial model takes 0.397 hours on the four NVIDIA GeForce RTX 4090 GPUs, resulting in a total training time of 1.59 GPU-Hours.

Case Studies. In addition, We provide an example under the wait-3 policy in Table 7 to validate the effectiveness of proposed method.

5 Related Work

Existing SiMT sudies can be mainly categorized into two types (i.e., fixed or adaptive policy) according the READ/WRITE policy.

As the fixed policy, Dalvi et al. (2018) introduced STATIC-RW, and Ma et al. (2019)

Ground Truth Target	"and she 's an optometrist in st. petersburg , and she plays with optics ."
Consistency-CE(Wait-3) Hypothesis	"and it 's an optical optical in st. petersburg , and it 's playing with optics ."
Inconsistency-CE(Wait-9) Hypothesis	"and it 's an optimist in st. petersburg , and it plays with optics ."
Consistency-Bi(Wait-3) Hypothesis	"and she 's an optic woman st. petersburg , and she plays with optics ."

Table 7: Translation examples of inference with $k = 3$: Consistency-CE model outputs a repeated word "optical," leading to translation oscillation. In the Inconsistency-CE output, such errors are absent, resulting in a more coherent translation. Moreover, both Consistency-CE and Inconsistency-CE still exhibit a gender error, translating "she" as "it." Consistency-Bi further corrects this error.

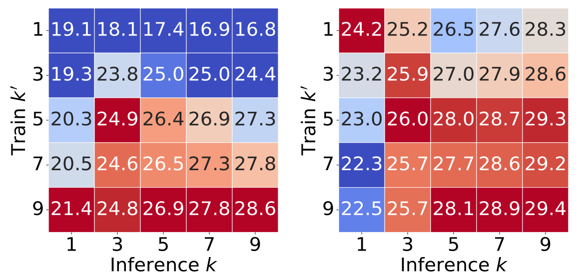


Figure 11: BLEU scores comparison between the original and proposed training methods using wait-k' during training and wait-k during inference on the WMT15 De-En valid set. The diagonal line indicates consistency between training k' and inference k.

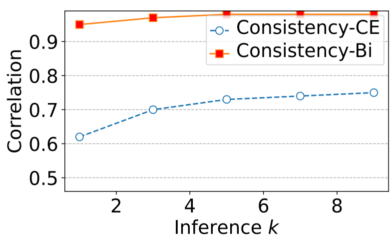


Figure 12: Comparison of correlation between BLEU scoresand training loss (cross-entropy loss for Consistency-CE and bi-objectives loss for Consistency-Bi) on training subset of WMT15 De-En task.

proposed the wait-k policy. Building upon this, Elbayad et al. (2020) enhanced the wait-k policy by introducing the practice of sampling different values of k during training. Additionally, Han et al. (2020) incorporated meta-learning into the wait-k policy, and Zhang et al. (2021) proposed future-guided training for the wait-k policy.

Alternatively, many notable works develop an adaptive policy for SiMT (Zheng et al., 2019; Zhang et al., 2020; Wilken et al., 2020; Miao et al., 2021; Zhang and Feng, 2022; Zhang et al., 2022). For instance, Zheng et al. (2020) propose the adaptive policy through a heuristic ensemble of multiple wait-k models. Other studies (Zheng

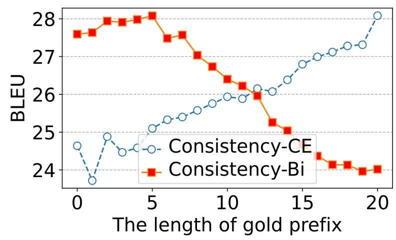


Figure 13: BLEU scores comparison between original Consistency-CE model and ours proposed Consistency-Bi model for wait-1 decoding under the prefix-constrained decoding setting.

et al., 2019; Arivazhagan et al., 2019; Ma et al., 2020; Zhang and Zhang, 2020; Zhang et al., 2020) resort to an adaptive policy controller to determine the READ/WRITE action and then integrate the controller into the SiMT model.

The above studies overlook the counterintuitive phenomenon about the context usage between training and inference, and our work thereby provides comprehensive analysis on this phenomenon and propose an effective approach to address this phenomenon, which is general enough to be applied into both policies.

6 Conclusion

This paper pays attention to a counterintuitive phenomenon in the context of usage between training and inference in SiMT. Subsequently, we conduct a comprehensive analysis and make the noteworthy discovery that this phenomenon primarily stems from the weak correlation between translation quality and training loss as well as exposure bias between training and inference. Based on our findings, we propose a context consistency training method that incorporates both translation quality and latency as bi-objectives and alleviates the exposure bias issue during the training. Experiments verify the effectiveness of the proposed approach, making the context-consistent SiMT successful for the first time.

Limitations

Our context consistency training approach necessitates a search for an appropriate hyperparameter, denoted as γ shown in Table 5, to strike a balance between translation quality and latency. Further research is required to establish an efficient method for this purpose.

Acknowledgements

We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions. The work was supported by the National Natural Science Foundation of China under Grant 62276077, Guangdong Basic and Applied Basic Research Foundation (2024A1515011205), and Shenzhen College Stability Support Plan under Grants GXWD20220811170358002 and GXWD20220817123150002.

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*, Heidelberg, Germany.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *ArXiv preprint*, abs/1606.02012.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient wait-k models for simultaneous machine translation. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1461–1465. ISCA.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

Shoutao Guo, Shaolei Zhang, and Yang Feng. 2023. Glancing future for simultaneous machine translation. *ArXiv preprint*, abs/2309.06179.

Hou Jeung Han, Mohd Abbas Zaidi, Sathish Reddy Indurthi, Nikhil Kumar Lakumarapu, Beomseok Lee, and Sangha Kim. 2020. End-to-end simultaneous translation system for IWSLT2020 using modality agnostic meta-learning. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 62–68, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language*

Translation: Evaluation Campaign, pages 76–79, Da Nang, Vietnam.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020. Monotonic multihead attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yishu Miao, Phil Blunsom, and Lucia Specia. 2021. A generative framework for simultaneous machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6697–6706, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Patrick Wilken, Tamer Alkhouli, Evgeny Matusov, and Pavel Golik. 2020. Neural simultaneous speech translation using alignment-based chunking. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 237–246, Online. Association for Computational Linguistics.

Joern Wuebker, Spence Green, John DeNero, Saša Hasan, and Minh-Thang Luong. 2016. Models and inference for prefix-constrained machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Berlin, Germany. Association for Computational Linguistics.

Ruiqing Zhang and Chuanqiang Zhang. 2020. Dynamic sentence boundary detection for simultaneous translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 1–9, Seattle, Washington. Association for Computational Linguistics.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 2280–2289, Online. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2021. Universal simultaneous machine translation with mixture-of-experts wait-k policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7306–7317, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2022. Information-transport-based policy for simultaneous translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 992–1013, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shaolei Zhang, Yang Feng, and Liangyou Li. 2021. Future-guided incremental transformer for simultaneous translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14428–14436. AAAI Press.

Shaolei Zhang, Shoutao Guo, and Yang Feng. 2022. Wait-info policy: Balancing source and target at information level for simultaneous machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2249–2263, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020. Simultaneous translation policies: From fixed to adaptive. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2847–2853, Online. Association for Computational Linguistics.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354, Hong Kong, China. Association for Computational Linguistics.

A Detailed Experimental Settings

We conduct experiments on the following datasets, which are the widely-used SiMT benchmarks.

IWSLT14 German→English (De→En) (Cettolo et al., 2013) we train on 160K pairs, develop on 7K held-out pairs, and test on TED dev2010+tst2010-2013 (6,750 pairs). Following the previous setting (Elbayad et al., 2020), all data is tokenized and lower-cased and we segment sequences using byte pair encoding (Sennrich et al., 2016) with 10K merge operations. The resulting vocabularies are of 8.8K and 6.6K types in German and English respectively.

IWSLT15[4] Vietnamese → English (Vi→En) (Luong and Manning, 2015) we train on 133K pairs, develop on TED tst2012 (1,553 pairs), and test on TED tst2013 (1,268 pairs). The corpus is simply tokenized by SentencePiece (Kudo and Richardson, 2018), resulting in 16K and 8K word vocabularies in English and Vietnamese respectively.

WMT15[5] German → English (De→En) (Callison-Burch et al., 2009) is a parallel corpus with 4.5M training pairs. We use newstest2013 (3003 pairs) as the dev set and newstest2015 (2169 pairs) as the test set. The corpus is simply tokenized by SentencePiece resulting in 32k shared word vocabularies.

The implementation of all systems is based on Transformer (Vaswani et al., 2017) and adapted from Fairseq Library (Ott et al., 2019). Following Ma et al. (2019); Elbayad et al. (2020), we apply Transformer-Small (4 heads) for IWSLT15 Vi-En and IWSLT14 De-En, Transformer-Base (8 heads) for WMT15 De-En. To avoid the recalculation of the encoder hidden states when a new source token is read, unidirectional encoder (Elbayad et al., 2020) is proposed to make each source token only attend to its previous words.

[4] `nlp.stanford.edu/projects/nmt/`
[5] `www.statmt.org/wmt15/translation-task`