Investigating Hallucinations in Simultaneous Machine Translation: Knowledge Distillation Solution and Components Analysis

Donglei Yu[1,2], **Xiaomian Kang**[1,2], **Yuchen Liu**[1,2], **Feifei Zhai**[1,2],
Nanchang Cheng [4], **Yu Zhou**[1,3*], **Chengqing Zong** [1,2]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
[4] Communication University of China
{yudonglei2021,kangxiaomian2014,yuchen.liu,yu.zhou,chengqing.zong}@ia.ac.cn

Abstract

Simultaneous Machine Translation (SiMT) generates target translation before receiving the whole source sentence and faces a serious hallucination problem. In contrast, traditional offline machine translation (OMT) models exhibit significantly fewer hallucinations. Motivated by this disparity, we propose Knowledge Distillation for SiMT (KD-SiMT), a simple yet effective method that utilizes the OMT model to mitigate hallucinations in SiMT. Experiments on Zh→En and De→En tasks demonstrate that KD-SiMT effectively reduces hallucinations and enhances the SiMT performance. Furthermore, we systematically investigate the deficiencies in SiMT models related to serious hallucinations and the effect of KD-SiMT. Specifically, we design targeted tasks and metrics to quantitatively evaluate the components in SiMT models from the perspectives of model structure and knowledge acquisition. Our analyses reveal that inaccurate source representations and imbalanced cross-attention are more likely to occur in SiMT models when generating hallucinations, while KD-SiMT alleviates these issues. Besides, we find that KD-SiMT equips SiMT models with sufficient faithfulness knowledge in training, thus reducing hallucinations.

1 Introduction

Simultaneous Machine Translation (SiMT) aims to generate target translation before receiving the whole source sentence, which acquires models to learn both translation ability and a read/write policy that decides between outputting a target word and waiting for a new source word (Ma et al., 2019; Elbayad et al., 2020b; Liu et al., 2021; Miao et al., 2021b; Wang et al., 2022).

However, this complex training objective hinders the SiMT models from learning translation ability, thus triggering serious hallucinations (Han et al., 2022), which means SiMT models generate fluency but unfaithful translation. This is extremely harmful to translation quality and disturbs user trust (Lee et al., 2018; Guerreiro et al., 2023).

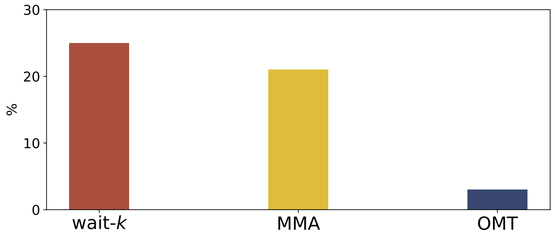


Figure 1: Percentage of hallucination sentences for each SiMT model and OMT model. We choose wait-*k* (Ma et al., 2019) and MMA (Ma et al., 2020) as representative SiMT models.

In contrast, with the focused training objective, traditional offline machine translation (OMT) models generate fewer hallucinations (Dale et al., 2022; Guerreiro et al., 2023). For a clear comparison, we conduct a manual analysis of hallucinations[1]. As shown in Figure 1, the percentage of hallucinations in SiMT is 20%, while in OMT is only about 3%.

This observed performance gap motivates our approach. In this paper, we propose Knowledge Distillation for SiMT (KD-SiMT), a simple yet effective method that utilizes OMT models to reduce hallucinations in SiMT models. KD-SiMT enhances the SiMT model by incorporating additional supervised signals derived from the hidden representations and output probabilities of the OMT model. Experiments on Zh→En and De→En SiMT tasks indicate that KD-SiMT significantly reduces hallucinations and improves translation quality.

Furthermore, we decouple SiMT models into

[*] Corresponding Author

[1] We train two typical SiMT models (Ma et al., 2019, 2020) and an OMT model with the same training set on Zh→En SiMT task. Then we randomly select 100 sentences under various latency levels to count the proportion of hallucination sentences in their translations manually. All models are based on Transformer (Vaswani et al., 2017).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7116–7131
April 29 - May 4, 2025

distinct components for individual analysis to identify deficiencies in SiMT models associated with hallucinations and to understand how KD-SiMT reduces these hallucinations. Specific tasks and metrics are designed for quantitative assessment. Our analyses involve examining the SiMT models from two primary perspectives: model structure and knowledge acquisition. From the perspective of model structure, we introduce the Auto-Encoder task and our defined Contribution Standard Deviation (CSD) to evaluate the encoder and decoder in SiMT models separately. Our evaluation reveals that *inaccurate source representations* and *imbalanced cross-attention assignment* are more likely to occur in SiMT models, and KD-SiMT effectively addresses these issues within the respective components. From the perspective of knowledge acquisition, we categorize the acquired knowledge into fluency knowledge and faithfulness knowledge. We find that SiMT models acquire adequate fluency knowledge but *limited faithfulness knowledge*, which is crucial for faithful translation. KD-SiMT can provide sufficient faithfulness knowledge for SiMT models during training. Our contributions can be summarized as follows:

- We propose KD-SiMT to reduce hallucinations in SiMT models with the aid of OMT models. Our proposed method is simple and compatible to various SiMT models. Experimental results on Zh→En and De→En SiMT tasks prove the effectiveness of KD-SiMT.

- We design specific tasks and metrics to quantitatively evaluate the components in SiMT models from the perspectives of model structure and knowledge acquisition. Our analyses identify deficiencies related to hallucinations in different components of SiMT models and demonstrate the enhancements KD-SiMT brings to these components.

2 Background

Prefix-to-Prefix Framework We consider a source sentence as $\mathbf{x} = (x_1, ..., x_M)$ and its corresponding target sentence $\mathbf{y} = (y_1, .., y_N)$. When the SiMT model generates y_n in the SiMT process, only a prefix of source sentence is available (Ma et al., 2019). We denote the source prefix as $\mathbf{x}_{\leq g_n}$, where g_n is the number of tokens in $\mathbf{x}_{\leq g_n}$. Therefore, the decoding probability of $\mathbf{y}$ is calculated as:

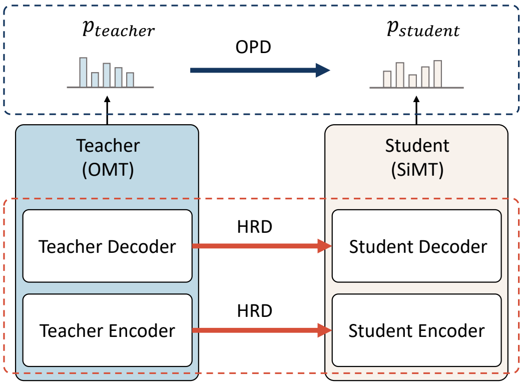


Figure 2: The overview of our proposed KD-SiMT. "HRD" means **H**idden **R**epresentation **D**istillation, and "OPD" means **O**utput **P**robability **D**istillation.

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{n=1}^{N} p(y_n \mid \mathbf{x}_{\leq g_n}, \mathbf{y}_{<n}) \tag{1}$$

SiMT Model Structure Existing SiMT models are commonly based on Transformer (Vaswani et al., 2017), which are formalized as:

$$\begin{aligned} e^l &= \text{Encoder}(e^{l-1}) \\ s^l &= \text{Decoder}(e^L, s^{l-1}), \quad l = 1, 2, ..., L \end{aligned} \tag{2}$$

where L is the number of layers in the encoder and decoder. It is noted that SiMT models typically employ unidirectional encoders to simulate streaming source inputs in the training process (Elbayad et al., 2020a; Ma et al., 2020; Zhang and Feng, 2022b).

3 Knowledge Distillation Solution

We propose KD-SiMT, which utilizes the OMT model as the teacher and the SiMT model as the student to reduce hallucinations in SiMT. Figure 2 provides an overview of our method. Further details are introduced below.

3.1 Hidden Representation Distillation

For the hidden representations, KD-SiMT introduces Hidden Representation Distillation (HRD) to make SiMT models learn the representations of encoder and decoder layers in OMT model. The cosine similarity is chosen as distillation loss:

$$\mathcal{L}_{\text{HRD}} = \sum_{l \in \{2,4,6\}} (2 - \cos(e_t^l, e_s^l) - \cos(s_t^l, s_s^l)) \tag{3}$$

where e_t^l,s_t^l are hidden representations of the l-th encoder and decoder layer in the teacher model, and e_s^l,s_s^l are counterparts in the student model. Notably, we focus on distillation on even layers to reduce computation overhead during training.

3.2 Output Probability Distillation

Hinton et al. (2015) proposed that knowledge could be transferred by allowing student models to learn from soft labels provided by teacher models. To transfer the acquired knowledge from the OMT model to the SiMT model, we use KL-divergence loss to perform the Output Probability Distillation (OPD) as follows:

$$\mathcal{L}_{\text{OPD}} = \sum_{i=n}^{N} p_t^n \log \frac{p_t^n}{p_s^n} \tag{4}$$

where N is the length of the target sequence, p_t^n and p_s^n describe the output probabilities of OMT and SiMT models at the n-th decoding step. Through OPD, SiMT models are encouraged to learn to generate more faithful translations.

3.3 Joint Training Framework

Due to the task gap between OMT and SiMT, using a pre-trained OMT as a teacher directly is adverse to KD (Zhang et al., 2021). To mitigate this challenge, we jointly train both teacher and student models. The embeddings of OMT and SiMT models are shared to reduce training overhead. Since the parameters of OMT model are not utilized in the inference process, KD-SiMT does not require additional computational resources for application.

The objective of the translation task is defined as cross-entropy loss for both teacher and student models, which are denoted as $\mathcal{L}_{\text{TCE}}$ and $\mathcal{L}_{\text{SCE}}$. The total loss is calculated in the following manner:

$$\mathcal{L} = \mathcal{L}_{\text{TCE}} + \mathcal{L}_{\text{SCE}} + \lambda_1 \mathcal{L}_{\text{HRD}} + \lambda_2 \mathcal{L}_{\text{OPD}} \tag{5}$$

where λ_1 and λ_2 are the super parameters.

4 Experiments

4.1 Settings

Implemented Model SiMT models are categorized into two groups based on their policies: ***fixed policies*** and ***adaptive policies*** (Zhang and Feng, 2022a,c). Following the setting in Ma et al. (2020), we validate the effectiveness of KD-SiMT on both fixed and adaptive policies:

- **wait-k** (Ma et al., 2019): A fixed policy that initially reads k tokens, followed by alternating between writing and reading one token.
- **multipath-wait-k (m-wait-k)** (Elbayad et al., 2020a): A fixed policy, which randomly samples different k during training and is similar to wait-k during inference.
- **MMA** (Ma et al., 2020): An adaptive policy that employs monotonic attention (Arivazhagan et al., 2019) to make read/write decisions.
- **ITST** (Zhang and Feng, 2022b): An adaptive policy that quantifies the information weight of each source token and makes the read/write decisions based on the received information.
- **HMT** (Zhang and Feng, 2023): An adaptive policy, which models the read/write decision-making process as the Hidden Markov Model.

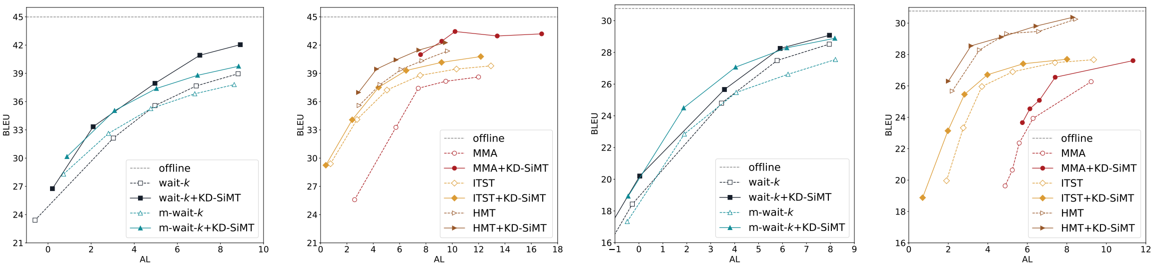
Most existing SiMT models are still based on the encoder-decoder Transformer architecture (Vaswani et al., 2017). Therefore, we believe that choosing the OMT model with a similar structure to SiMT models allows for a fairer comparison and a more accurate components analysis. We set $\lambda_1 = 0.1$ and $\lambda_2 = 1$ in our experiments. More details are provided in Appendix A.

Datasets For Zh→En SiMT task, we utilize LDC corpus (2.1M) for training, NIST 2008 for validation and NIST 2003,2004,2005,2006 for test. Byte-pair encoding (BPE) (Sennrich et al., 2016) is used in both Chinese and English, with a vocabulary size of 30k and 20k. For De→En SiMT task, we choose WMT15 (4.5M) as the training set, newstest 2013 as the validation set and newstest 2015 as the test set. A joint 32K vocabulary is applied.

Evaluation Metric We utilize BLEU (Papineni et al., 2002) to measure the translation quality, and Average Lagging (AL) (Ma et al., 2019) for latency. Besides, we use the Hallucination Rate (HR) (Chen et al., 2021) for analyzing hallucinations.

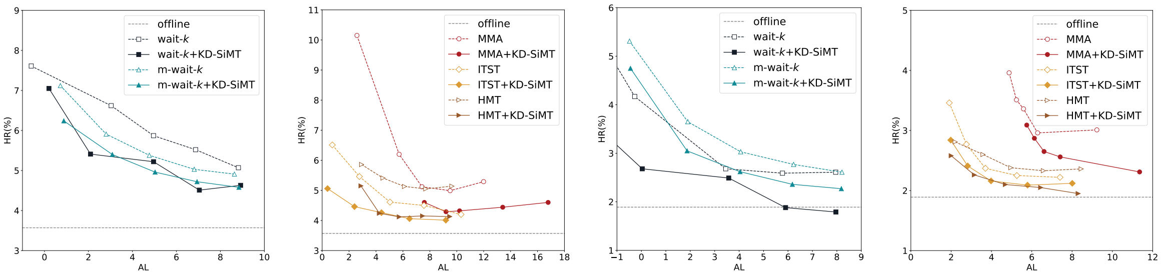
4.2 Main Results

Translation Quality We present the translation quality under various latency levels of different SiMT models in Figure 3. For fixed policies, KD-SiMT enhances translation quality across all latency settings. In Zh→En task, KD-SiMT yields an average improvement of 2.65 BLEU for wait-k and 2.25 BLEU for m-wait-k. Similarly, in De→En task, the improvements are 0.99 BLEU and 1.57 BLEU respectively. For adaptive policies, SiMT models with KD-SiMT also achieve higher BLEU



(a) Zh→En, fixed policies (b) Zh→En, adaptive policies (c) De→En, fixed policies (d) De→En, adaptive policies

Figure 3: Translation quality against latency of different SiMT models with/without KD-SiMT on Zh→En and De→En SiMT tasks.



(a) Zh→En, fixed policies (b) Zh→En, adaptive policies (c) De→En, fixed policies (d) De→En, adaptive policies

Figure 4: Hallucination Rate (%, ↓) against latency of different SiMT models with/without KD-SiMT on Zh→En and De→En SiMT tasks.

scores, especially for MMA. Additional numerical results are provided in Appendix B.

Hallucination Rate The HR of different SiMT models are shown in Figure 4. With the application of KD-SiMT, HR notably decreases across all models under various latency levels, especially for wait-k (0.78 on Zh→En, 1.04 on De→En) and MMA (0.76 on Zh→En, 0.67 on De→En). These results indicate that KD-SiMT effectively reduces hallucinations in both fixed and adaptive policies.

5 Components Analysis for SiMT Models

To further investigate hallucinations in SiMT, we decouple the SiMT model into components and analyze their performances on Zh→En SiMT task. From the perspective of model structure, we separately evaluate the encoder and decoder in the SiMT model. From the perspective of knowledge acquisition, we categorize the knowledge obtained in the SiMT model into fluency knowledge and fluent knowledge for individual analysis. For each component, specific tasks and metrics are designed. With these metrics, we can identify the specific deficiencies in each component when the SiMT model generates hallucinations (Sec.5.1) and analyze the impacts of KD-SiMT on these components (Sec.5.2). Besides, ablation studies and case studies are also conducted to further validate the effectiveness of KD-SiMT in reducing hallucinations of SiMT models (Sec.5.3 and 5.4).

5.1 Deficiencies in Each Component

We assess each component in SiMT models in the following. To investigate the association between the performance of these components and hallucinations, we divide samples into hallucination samples and non-hallucination samples for SiMT models[2], evaluating each group separately.

5.1.1 Source Representations in Encoder

Motivation The source representations from the encoder contain semantic information of source inputs, and inaccurate representations can trigger hallucinations (Weng et al., 2020). Hence, it is valuable to assess the quality of source representations in SiMT encoders when generating hallucinations.

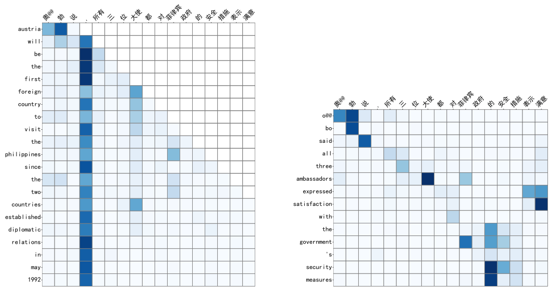
[2] We categorize samples according to Hallucination Rates (HR) (Chen et al., 2021). Those with HR exceeding 20% are classified as hallucination samples, while the rest are considered non-hallucination samples.

Model		BLEU$_{\text{AE}}$ ↑			CSD ↓			$S_{fluency}$ ↑			S_{faith} ↑		
		All	H.	N.H.	All	H.	N.H.	All	H.	N.H.	All	H.	N.H.
SiMT	wait-k (Ma et al., 2019)	91.61	81.92	95.05	7.19	9.97	6.50	-138.64	-167.61	-131.39	-107.58	-215.05	-80.71
	m-wait-k (Elbayad et al., 2020a)	89.97	80.83	94.62	8.48	10.64	7.94	-149.06	-165.64	-144.92	-99.80	-187.20	-77.95
	MMA (Ma et al., 2020)	86.67	74.82	93.05	8.29	11.97	7.37	-140.36	-151.52	-137.57	-108.52	-214.45	-82.04
	ITST (Zhang and Feng, 2022b)	90.84	79.74	96.60	8.26	10.37	7.73	-152.95	-158.75	-151.53	-111.52	-213.03	-86.14
	HMT (Zhang and Feng, 2023)	89.10	77.73	91.94	7.97	11.34	7.21	-145.83	-154.52	-142.39	-102.64	-151.18	-90.33
OMT(Vaswani et al., 2017)		96.03	–	–	5.67	–	–	-149.95	–	–	-86.65	–	–

Table 1: Evaluation of existing SiMT models and OMT model on Zh→En SiMT task about the source representations, cross-attention assignment and acquired knowledge. "All" means the evaluation results on the complete test samples. "H." and "N.H." mean the corresponding evaluation results on hallucination samples and non-hallucination samples. "BLEU$_{\text{AE}}$" means the BLEU on AE task. Note that all values of "CSD" in the table are in units of $\times 10^{-3}$.

Source	在 参加 比赛 的 八十 个 国家 中 , 南韩 是 第二十二 个 、 北韩 是 第二十三 个 入场 的 国家 。
Reference	among the 80 nations participating in competitions , south korea was the 22nd **and north korea the 23rd nation** to enter the stadium .
AE	在 参加 比赛 的 八十 个 国家 中 , 南韩 是 第二十二 个 入场 的 (missing: 北韩 是 第二十三 个 入场 的) 国家 。
SiMT	south korea is the twenty-second **and the 23rd place** to compete in 80 countries .

Table 2: Example illustrating the quality source representations when hallucinations happen. The "**reconstruction errors**" in AE, "**hallucinations**" in SiMT, and corresponding "**correct tokens**" for AE and SiMT are marked.



(a) SiMT (b) OMT

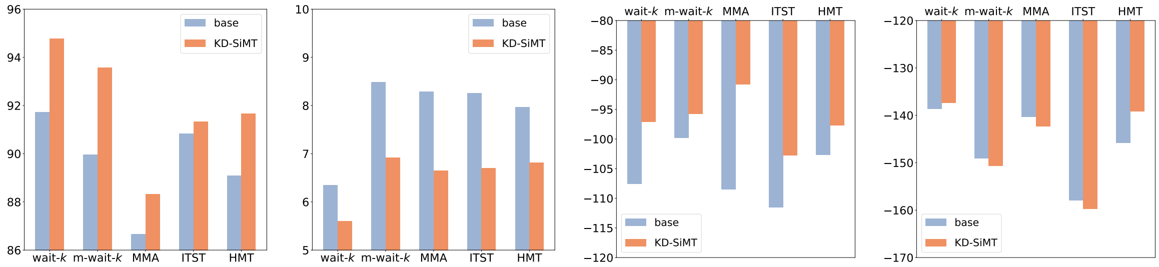
Figure 5: Attention visualization for SiMT and OMT on hallucination case. We choose wait-k as a representative of SiMT models, and the attention weights are from the top decoder layer of both models.

Evaluation Due to poor interpretability, directly assessing the semantic information contained in the source representations is challenging. Therefore, we employ the **Auto-Encoder** (AE) task to verify whether these representations accurately convey the information of source tokens. Specifically, we extract the source representations from a well-trained SiMT model, and then train an additional decoder to reconstruct the source inputs from the source representations. The perfect reconstruction of source sentences indicates high-quality source representations. Conversely, incomplete or incorrect reconstruction suggests that semantic information is inaccurate in these source representations. We argue that the quality of source representations is correlated with hallucinations in SiMT. Taking Table 2 as an example, when reconstruction errors arise, the corresponding part in the translation from the SiMT model also exhibits hallucinations. Therefore, we employ BLEU (Papineni et al., 2002) to assess the AE performance of various SiMT encoders, thus evaluating the source representations.

Results The results (BLEU$_{\text{AE}}$) are presented in Table 1. It is evident that reconstructing the original source sentences from the source representations of SiMT encoders is more challenging than from those of the OMT encoder. This indicates a more severe information inaccuracy of the source representations in SiMT models. Besides, the AE performance on hallucination samples is inferior, while on non-hallucination samples, it is notably better, even closely approximating OMT. This suggests that SiMT models produce worse source representations when generating hallucinations.

5.1.2 Cross-attention Assignment in Decoder

Motivation The cross-attention mechanism is widely applied in decoders to choose suitable information from source input for generating tokens at each step. Existing studies (Lee et al., 2018; Yan et al., 2022) indicate through qualitative analysis that imbalanced attention assignments in translation models will exacerbate the hallucinations. As illustrated in Figure 5, during the SiMT model generating hallucination tokens, the cross-attention is imbalanced, consistently assigning higher attention scores to some meaningless tokens and failing to select appropriate information. In contrast, information from each token is equally selected in the OMT model and faithful translation is produced.



(a) BLEU$_{\text{AE}}$ (b) CSD ($\times 10^{-3}$) (c) S_{faith} (d) $S_{fluency}$

Figure 6: Comparison of different components in SiMT models with/without KD-SiMT.

Evaluation We introduce a metric to quantitatively evaluate the imbalance in cross-attention, thus assessing the performance of various SiMT decoders. For each source token x_m ($1 \leq m \leq M$), we denote its contribution score as c_m, which is calculated as follows:

$$c_m = \frac{\max_n\{c_{mn}\}}{\sum_{m'=1}^{M} \max_n\{c_{m'n}\}}, 1 \leq n \leq N \tag{6}$$

where c_{mn} is the attention score on x_m when generating the n-th target token y_n. When cross-attention is imbalanced, attention scores are concentrated on a few source tokens, causing significant contribution score disparities among them, while balanced attention maintains uniformity among them. Consequently, we define the Contribution Standard Deviation (CSD) as the metric, which is computed as follows:

$$\text{CSD} = \sqrt{\frac{\sum_{m=1}^{M}(c_m - \bar{c})^2}{M}} \tag{7}$$

where $\bar{c}$ is the mean of c_m. A larger CSD means more imbalanced attention assignment.

Results The results (CSD) in Table 1 demonstrate that cross-attention in SiMT models is generally more imbalanced compared to the OMT model. Similarly, CSD values on hallucination samples are obviously higher, which means the imbalance in cross-attention is more acute when SiMT models generate hallucinations. These results indicate that imbalanced cross-attention is a deficiency in SiMT models associated with hallucinations.

5.1.3 Knowledge Acquisition Analysis

Motivation During training, models extract knowledge necessary for tasks from the provided training data. In translation tasks, such as OMT or SiMT, the knowledge that models need to learn can be categorized into two types:

- ***fluency knowledge***: Models need to learn how to generate fluent and natural sentences without syntax errors. This type of knowledge is also crucial for Language Modeling (LM).
- ***faithfulness knowledge***: The target sentences produced by translation models should keep semantic consistency with their corresponding source sentences.

To that end, hallucinations in SiMT may stem from acquiring sufficient fluency knowledge but inadequate faithfulness knowledge. To verify this view, we separate these two types of knowledge and conduct a quantitative assessment respectively.

Evaluation Existing studies (Hinton et al., 2015) show that the knowledge gained in models is reflected in output probabilities, which we select as our evaluation metric. For fluency knowledge, we initially train a language model on the same corpus used for the translation task. In this way, the language model acquires the same fluency knowledge as translation models, while not including any faithfulness knowledge. We denote the sentence generation probabilities from the language model as p_{lm}, which can effectively score the fluency of the output from translation models, thereby serving as a suitable measure for evaluating fluency knowledge. To evaluate faithfulness knowledge, we utilize p_{mt}, the probability that a translation model accurately decodes the reference. Under these grounds, we define $S_{fluency}$ and S_{faith} to assess fluency knowledge and faithfulness knowledge, which are calculated as follows:

$$\begin{aligned} S_{fluency} &= \log^{p_{lm}} \\ S_{faith} &= \log^{p_{mt}} \end{aligned} \tag{8}$$

For both $S_{fluency}$ and S_{faith}, larger values indicate more acquired knowledge.

Results The results in Table 1 indicate that $S_{fluency}$ of SiMT models and OMT model are comparable in both hallucination and non-hallucination samples. This suggests that SiMT models possess adequate fluency knowledge, even when generating hallucinations. However, all SiMT models gain lower S_{faith} than the OMT model, with a more pronounced gap in hallucination samples. In contrast, on non-hallucination samples, SiMT models achieve S_{faith} comparable to the OMT model. This reveals that SiMT models lack the corresponding faithfulness knowledge for correctly translating the sentences in hallucination samples.

5.2 Impact of KD-SiMT

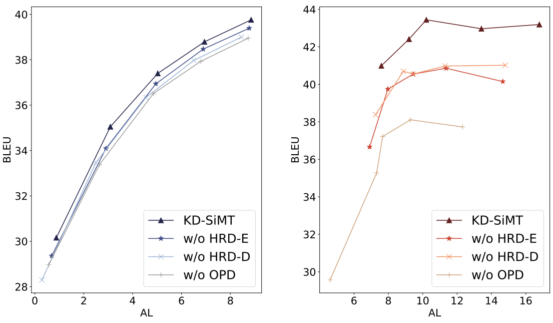
To validate the impact of KD-SiMT on these components, we compare the performance of components in SiMT models with and without KD-SiMT. The results are shown in Figure 6.

For the source representations, the results (BLEU_{AE}) are presented in Figure 6(a). All SiMT encoders with KD-SiMT achieve higher AE performance, which means that KD-SiMT improves the quality of source representations in SiMT models.

For the cross-attention assignments, the results (CSD) in Figure 6(b) demonstrate that CSD of SiMT models with KD-SiMT are lower than those without KD-SiMT. This proves that KD-SiMT makes the attention assignment in SiMT decoders more balanced, rather than only focusing on a small portion of source information.

For acquired knowledge, the results in Figure 6(c) show that all SiMT models with KD-SiMT archive higher S_{faith}. This indicates that SiMT models with KD-SiMT can learn more faithfulness knowledge. In contrast, $S_{fluency}$ (Figure 6(d)) of SiMT models with or without KD-SiMT are even the same, which means there is no extra fluency knowledge from the OMT model. These results further reveal SiMT models' preference for fluency knowledge and lack of faithfulness knowledge.

Based on these analyses, it is evident that SiMT models exhibit serious deficiencies in different components when hallucinations happen, including inaccurate source representations, imbalanced cross-attention assignment, and insufficient faithfulness knowledge. Concurrently, KD-SiMT enhances these components, thus improving the translation quality and reducing hallucinations.



(a) m-wait-k

(b) MMA

Figure 7: Effect of distillation modules about translation quality against latency on Zh→En task. "w/o HRD-E" and "w/o HRD-D" respectively mean without the hidden representation distillation in encoder and decoder. "w/o OPD" means without output probability distillation.

Model	BLEU_{AE}	CSD	S_{faith}	HR
m-wait-k + KD-SiMT	93.58	6.92	-95.76	5.40
-w/o HRD-E	91.16	6.90	-96.93	5.74
-w/o HRD-D	93.34	7.47	-95.70	5.63
-w/o OPD	93.49	5.79	-98.40	5.79
MMA + KD-SiMT	88.33	6.65	-90.77	4.29
-w/o HRD-E	87.87	6.77	-91.41	4.81
-w/o HRD-D	89.12	8.48	-89.62	4.53
-w/o OPD	88.05	6.40	-101.89	4.56

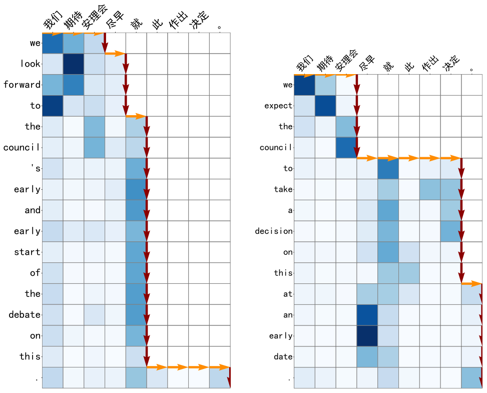
Table 3: Effect of each knowledge distillation module in ability improvements on Zh→En task. Note that all values of "CSD" in the table are in units of $\times 10^{-3}$.

5.3 Ablation Study

To explore the effect of each distillation module, we conduct ablation studies on Zh→En SiMT task using m-wait-k and MMA, representing fixed and adaptive policies respectively. The results are presented in Figure 7 and Table 3.

In Figure 7, it is evident that models with all distillation modules consistently exhibit the highest performance across various latency levels, indicating the effectiveness of each distillation module. For m-wait-k, each module has a roughly equivalent impact on model performance. For MMA, the influence of OPD is most significant. The effect of these distillation modules on hallucinations is shown in Table 3. We can see that all modules alleviate hallucinations in SiMT.

Table 3 also presents the effect of each distillation module on the abilities of different components in SiMT models. Models without HRD in the encoder ("w/o HRD-E") exhibit inferior performance on the AE task. Conversely, models without



(a) MMA (b) MMA with KD-SiMT

Figure 8: Attention Visualization for MMA and MMA with KD-SiMT on hallucination case.

HRD in the decoder ("w/o HRD-D") exhibit the highest CSD, while models without OPD acquire the least faithfulness knowledge. Therefore, each distillation module contributes to improving specific components in SiMT models, aligning with our expectations. Besides, HR scores in Table 3 also reveal that all distillation modules contribute to reducing hallucinations in SiMT models.

5.4 Case Study

Figure 8 displays the cross-attention visualization for MMA with and without KD-SiMT. MMA exhibits imbalanced cross-attention, focusing too much on the token "就" (means "on") and leading to hallucinations. In contrast, MMA with KD-SiMT demonstrates a more balanced cross-attention. Intriguingly, we observe that MMA with KD-SiMT performs a more effective read/write policy to decide when to continue translating, avoiding the generation of tokens lacking supporting evidence from the current source inputs. Specifically, as depicted in Figure 8, when "就" is fed, MMA with KD-SiMT chooses to wait for the subsequent source tokens until it encounters the source token "决定" (means "decision"), while MMA continues to generate text, eventually producing hallucination tokens "early start of the debate". This phenomenon may offer insights into how imbalanced cross-attention disrupts SiMT models and triggers hallucinations, which we intend to explore in future research. More cases can be found in Appendix C.

5.5 Manual Analysis

To further evaluate the effect of KD-SiMT on reducing hallucinations in SiMT, we conduct a manual evaluation on Zh→En SiMT task. Specifically, we randomly sample 100 sentences respectively in the outputs of wait-k and MMA and calculate the percentage of sentences exhibiting hallucinations. Figure 9 presents the results, indicating an average reduction of 8.8% for wait-k and 14.0% for MMA. Consequently, we can conclude that KD-SiMT effectively mitigates the hallucination issue across all latency levels in both fixed and adaptive policies.

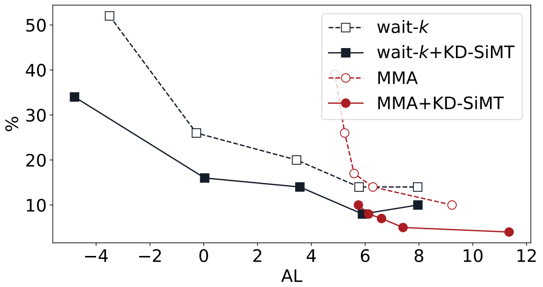


Figure 9: Percentage of hallucination sentences in SiMT models with/without KD-SiMT on Zh→En.

6 Related Works

Simultaneous Machine Translation According to the used read/write policies, existing SiMT methods can be divided into fixed policy and adaptive policy. For fixed policy, Ma et al. (2019) proposed wait-k, which reads k tokens before starting translation. Elbayad et al. (2020a) proposed multipath wait-k, which enhances wait-k by randomly sampling k during training. Zhang et al. (2021) utilized knowledge distillation to guide the SiMT encoders mapping accurate source representations. Zhang and Feng (2021) used the mix-of-experts structure to realize the universal SiMT model across different latency levels. For adaptive policy, Gu et al. (2017) used reinforcement learning to make read/write decisions. Ma et al. (2020) utilized multi-head monotonic attention to guide the SiMT model in learning an adaptive policy. Miao et al. (2021b) proposed a generative framework with a latent variable to dynamically decide between "read" and "write". Zhang and Feng (2022b) added a module measuring the received information and proposed an information-based policy. Zhang and Feng (2023) used Hidden Markov Model to realize an adaptive policy based on the hidden state probability. Nevertheless, the hallucination problem in existing SiMT models is still serious.

Hallucinations in Machine Translation For OMT model (Vaswani et al., 2017; Ma et al., 2023; Liang et al., 2024; Zhang et al., 2025), there are fewer hallucinations when the inputs are in-domain and are not disturbed (Dale et al., 2022; Guerreiro et al., 2023). Consequently, existing studies about hallucinations in OMT mainly focus on out-of-domain and noised scenarios. Weng et al. (2020) used multitask learning to enhance the faithfulness of OMT. Miao et al. (2021a) pointed out that one reason for hallucinations is the overconfidence of the language model mechanism in the decoder. Yan et al. (2022) found that the deficient encoder and vulnerable cross-attentions are to blame. Ji et al. (2023) summarized the causes of hallucinations from perspectives of data and model architecture. In SiMT, few researches are proposed. Chen et al. (2021) found that using training corpus with fewer reordering could alleviate hallucinations. Guo et al. (2023) used reinforcement learning to create tailored references for SiMT to reduce hallucinations. Wang et al. (2023) proposed two-stage beam search to create monotonic reference translations. However, to our best knowledge, there is still no research that systematically investigates hallucinations in SiMT.

7 Conclusion

In this paper, we focus on investigating serious hallucinations in SiMT. We propose KD-SiMT to reduce hallucinations in SiMT models. Experiments show that KD-SiMT is effective and achieves significant improvements. Furthermore, we conduct component analysis to explore the deficiencies in SiMT models related to hallucinations and the effect of KD-SiMT on the SiMT models. From the perspective of model structure, we find that KD-SiMT enhances the source representations and cross-attention assignments in SiMT models. From the perspective of knowledge acquisition, our analyses suggest that KD-SiMT equips SiMT models with sufficient faithfulness knowledge during training, thus improving translation quality.

Limitations

In this paper, we propose KD-SiMT to reduce hallucinations in SiMT models. We also conduct a quantitative analysis to investigate the performances of each component in SiMT models when generating hallucinations. However, there is still room for improvement. For example, the efficiency of KD-SiMT might be further improved. Due to the task gap between SiMT and OMT, the effectiveness of KD-SiMT could also be affected. It is worth exploring a more efficient and effective training paradigm for SiMT models to reduce hallucinations. Besides, our analyses can be further refined, which may lead to insightful conclusions. For instance, when hallucinations occur, fixed policies and adaptive policies may exhibit distinct characteristics, and the representations across different layers may also follow different patterns. We will leave these for our future work.

Acknowledgments

The research work has been supported by the Natural Science Foundation of China under Grant No. 62206286.

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.

Junkun Chen, Renjie Zheng, Atsuhito Kita, Mingbo Ma, and Liang Huang. 2021. Improving simultaneous translation by incorporating pseudo-references with fewer reorderings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5857–5864.

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020a. Efficient wait-k models for simultaneous machine translation.

Maha Elbayad, Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Antoine Caubrière, Benjamin Lecouteux, Yannick Estève, and Laurent Besacier. 2020b. ON-TRAC consortium for end-to-end and simultaneous speech translation challenge tasks at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 35–43, Online. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK 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.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Shoutao Guo, Shaolei Zhang, and Yang Feng. 2023. Simultaneous machine translation with tailored reference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3070–3084, Singapore. Association for Computational Linguistics.

HyoJung Han, Marine Carpuat, Jordan Boyd-Graber, UMIACS CS, and LCS iSchool. 2022. Simqa: Detecting simultaneous mt errors through word-by-word question answering.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7077–7088.

Dan Liu, Mengge Du, Xiaoxi Li, Yuchen Hu, and Lirong Dai. 2021. The USTC-NELSLIP systems for simultaneous speech translation task at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 30–38, Bangkok, Thailand (online). Association for Computational Linguistics.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *International Conference on Document Analysis and Recognition*, pages 484–501. Springer.

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 *International Conference on Learning Representations*.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021a. Prevent the language model from being overconfident in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3456–3468.

Yishu Miao, Phil Blunsom, and Lucia Specia. 2021b. A generative framework for simultaneous machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6697–6706.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022. The HW-TSC's simultaneous speech translation system for IWSLT 2022 evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 247–254, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Shushu Wang, Jing Wu, Kai Fan, Wei Luo, Jun Xiao, and Zhongqiang Huang. 2023. Better simultaneous translation with monotonic knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2334–2349, Toronto, Canada. Association for Computational Linguistics.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2022. Probing causes of hallucinations in neural machine translations. *arXiv preprint arXiv:2206.12529*.

Shaolei Zhang and Yang Feng. 2021. Universal simultaneous machine translation with mixture-of-experts wait-k policy. *arXiv preprint arXiv:2109.05238*.

Shaolei Zhang and Yang Feng. 2022a. Gaussian multi-head attention for simultaneous machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3019–3030.

Shaolei Zhang and Yang Feng. 2022b. 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 and Yang Feng. 2022c. Modeling dual read/write paths for simultaneous machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2461–2477, Dublin, Ireland. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2023. Hidden markov transformer for simultaneous machine translation. In *The Eleventh International Conference on Learning Representations*.

Shaolei Zhang, Yang Feng, and Liangyou Li. 2021. Future-guided incremental transformer for simultaneous translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14428–14436.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025. From chaotic ocr words to coherent document: A fine-to-coarse zoom-out network for complex-layout document image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10877–10890.

Hyper-parameter	
encoder layers	6
encoder attention heads	8
encoder embed dim	512
encoder ffn embed dim	1024
decoder layers	6
decoder attention heads	8
decoder embed dim	512
decoder ffn embed dim	1024
dropout	0.1
optimizer	adam
adam-β	(0.9, 0.98)
clip-norm	1e-7
lr	5e-4
lr scheduler	inverse sqrt
warmup-updates	4000
warmup-init-lr	1e-7
weight decay	0.0001
label-smoothing	0.1
max tokens	4096

Table 4: Hyper-parameters of our experiments.

A Hyper-parameters

The hyper-parameters of our experiments are shown in Table 4.

B Numerical Results

Table 5, 6, 7, 8 9 show the numerical results on Zh→En SiMT task. Table 10 shows the numerical results on De→En SiMT task. Table 11 shows the numerical results of hallucination rate on different SiMT models. Table 12 shows the numerical results of manual evaluation on hallucinations. Additionally, we also provide the COMET scores of the Zh→En task in Table 13.

C Hallucination Cases

Table 14 shows the additional hallucination cases in SiMT models.

	03		04		05		06		AVERAGE	
	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU
wait1	-0.53	23.23	-0.49	25.49	-0.74	22.75	-0.65	22.19	-0.60	23.42
+KD	0.26	26.76	0.20	27.92	0.12	25.95	0.24	26.40	0.20	26.76
wait3	3.17	32.58	3.20	34.06	2.74	30.38	3.02	31.50	3.03	32.13
+KD	2.46	34.40	2.02	34.10	2.02	32.88	1.88	31.96	2.09	33.34
wait5	5.08	35.53	5.12	37.76	4.81	34.19	4.84	34.82	4.96	35.58
+KD	5.11	38.22	5.17	39.97	4.71	37.50	4.86	36.07	4.96	37.94
wait7	7.08	38.19	7.03	39.84	6.64	35.85	6.75	36.78	6.88	37.67
+KD	7.37	42.89	7.08	41.34	6.86	38.84	6.87	40.62	7.05	40.92
wait9	9.06	39.22	8.96	41.17	8.68	37.04	8.58	38.40	8.82	38.96
+KD	9.20	43.35	9.05	43.32	8.81	40.21	8.63	41.30	8.92	42.05

Table 5: The numerical results of waitk on Zh→En SiMT task.

	03		04		05		06		AVERAGE	
	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU
m-wait1	0.81	29.19	0.87	30.06	0.42	26.43	0.80	27.54	0.72	28.31
+KD	0.91	30.94	0.91	30.88	0.78	29.25	0.92	29.57	0.88	30.16
m-wait3	3.01	33.33	3.03	34.33	2.46	30.50	2.72	32.33	2.80	32.62
+KD	3.22	36.14	3.20	36.00	2.94	33.79	2.99	34.26	3.09	35.05
m-wait5	4.92	35.72	5.02	37.43	4.49	32.77	4.66	35.11	4.77	35.26
+KD	5.11	38.20	5.18	38.60	4.91	35.70	4.93	37.08	5.03	37.40
m-wait7	7.00	37.53	7.01	39.09	6.58	34.44	6.65	36.26	6.81	36.83
+KD	7.13	39.94	7.08	40.06	6.80	36.54	6.75	38.62	6.94	38.79
m-wait9	8.84	38.41	8.86	39.94	8.40	35.43	8.46	37.43	8.64	37.80
+KD	9.02	40.48	8.97	40.54	8.81	38.21	8.58	39.79	8.85	39.76

Table 6: The numerical results of m-waitk on Zh→En SiMT task.

	03		04		05		06		AVERAGE	
	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU
MMA-0.3	2.08	24.47	2.34	26.75	2.46	24.72	3.47	26.40	2.59	25.59
+KD	7.79	42.24	7.47	42.07	7.82	39.71	7.29	39.94	7.59	40.99
MMA-0.25	6.16	33.78	5.36	35.05	5.55	32.34	5.82	31.89	5.72	33.27
+KD	9.54	43.57	9.04	43.52	9.56	42.13	8.67	40.43	9.20	42.41
MMA-0.2	7.64	38.23	7.40	38.94	7.62	36.42	7.03	36.07	7.42	37.42
+KD	10.39	44.26	10.20	44.22	10.72	42.98	9.53	42.31	10.21	43.44
MMA-0.15	9.66	39.02	9.45	39.91	9.89	37.47	9.03	36.23	9.51	38.16
+KD	13.87	43.87	13.21	44.13	14.17	42.27	12.41	41.60	13.42	42.97
MMA-0.1	12.19	39.60	12.02	40.36	12.51	37.22	11.25	37.27	11.99	38.61
+KD	17.01	44.34	16.69	44.01	17.93	42.88	15.52	41.54	16.79	43.19

Table 7: The numerical results of MMA on Zh→En SiMT task. MMA-λ means the super parameter λ used in MMA model, which is used to control latency.

	03		04		05		06		AVERAGE	
	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU
ITST-0.2	0.92	29.69	0.59	30.32	0.55	28.42	0.96	29.18	0.76	29.40
+KD	0.56	29.33	0.20	29.62	0.26	28.03	0.59	30.00	0.40	29.25
ITST-0.3	2.90	34.65	2.64	34.63	2.72	33.02	2.82	34.11	2.77	34.10
+KD	2.57	35.06	2.31	34.76	2.34	32.58	2.39	33.87	2.40	34.07
ITST-0.4	5.24	37.99	4.91	37.77	5.18	35.96	4.83	37.19	5.04	37.23
+KD	4.59	38.50	4.24	38.06	4.54	35.90	4.26	37.61	4.41	37.52
ITST-0.5	7.82	39.59	7.56	39.36	7.87	37.28	7.02	38.91	7.57	38.79
+KD	6.77	40.42	6.48	40.22	6.801	37.46	6.23	39.10	6.49	39.30
ITST-0.6	10.72	40.37	10.25	40.35	10.91	37.79	9.41	39.32	10.32	39.46
+KD	9.48	41.17	9.11	41.21	9.60	38.24	8.55	40.04	9.19	40.17
ITST-0.7	13.07	40.37	12.93	40.44	13.83	38.26	11.91	40.10	12.94	39.79
+KD	12.28	41.89	12.17	41.73	12.75	38.58	11.46	40.93	12.17	40.78

Table 8: The numerical results of ITST on Zh→En SiMT task. ITST-δ means the super parameter δ used in inference, which is used to control latency.

	03		04		05		06		AVERAGE	
	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU	AL	BLEU
HMT-(2,4)	3.12	35.65	3.17	37.07	2.55	33.95	2.90	35.70	2.93	35.59
+KD	3.13	37.73	2.90	38.04	2.59	34.71	2.88	37.38	2.88	36.97
HMT-(3,6)	4.74	38.22	4.55	38.98	4.48	36.75	4.33	37.28	4.52	37.81
+KD	4.55	40.13	4.24	40.33	4.07	37.58	4.12	39.77	4.25	39.45
HMT-(5,6)	6.42	40.21	6.11	40.34	6.05	38.17	5.90	38.95	6.12	39.42
+KD	6.00	41.04	5.83	41.70	5.63	38.60	5.47	40.36	5.73	40.43
HMT-(7,6)	7.93	41.22	7.77	41.48	7.71	38.62	7.36	40.00	7.70	40.33
+KD	7.75	42.36	7.53	42.20	7.35	40.12	7.19	41.18	7.46	41.47
HMT-(9,8)	10.01	41.92	9.70	41.87	9.68	40.25	9.17	41.45	9.64	41.37
+KD	9.71	43.00	9.56	43.30	9.45	40.62	9.18	42.11	9.48	42.26

Table 9: The numerical results of HMT on Zh→En SiMT task. HMT-(L, K) means the super parameter (L, K) used in inference, which is used to control latency.

	k	1	3	5	7	9
waitk	AL	-3.5	-0.274	3.452	5.774	7.953
	BLEU	10.06	18.43	24.8	27.49	28.52
waitk+KD	AL	-4.803	0.036	3.574	5.901	7.961
	BLEU	7.9	20.2	25.66	28.25	29.08
m-waitk	AL	-0.4852	1.899	4.065	6.23	8.21
	BLEU	17.33	22.85	25.47	26.62	27.55
m-waitk+KD	AL	-0.447	1.87	4.04	6.18	8.185
	BLEU	18.93	24.5	27.07	28.29	28.89
	λ	0.1	0.2	0.3	0.4	0.5
MMA	AL	9.23	6.29	5.59	5.24	4.876
	BLEU	26.27	23.93	22.36	20.64	19.63
MMA + KD	AL	11.35	7.41	6.61	6.13	5.75
	BLEU	27.61	26.55	25.08	24.54	23.66
	δ	0.2	0.3	0.4	0.5	0.6
ITST	AL	1.91	2.76	3.7	5.26	7.4
	BLEU	19.96	23.33	25.96	26.9	27.48
ITST+KD	AL	0.7	1.98	2.82	3.98	5.79
	BLEU	18.88	23.14	25.46	26.71	27.41
	(L, K)	(2,4)	(3,6)	(5,6)	(7,6)	(9,8)
HMT	AL	2.20	3.58	4.96	6.58	8.45
	BLEU	25.67	28.29	29.33	29.47	30.25
HMT+KD	AL	2.00	3.15	4.69	6.44	8.30
	BLEU	26.3	28.55	29.11	29.81	30.37

Table 10: Numerical results on De→En SiMT task.

Model	Zh→En						De→En					
	k=1	k=3	k=5	k=7	k=9	Avg. Δ	k=1	k=3	k=5	k=7	k=9	Avg. Δ
wait-k	7.61	6.62	5.87	5.52	5.07		6.90	4.17	2.68	2.59	2.61	
+KD	7.05	5.41	5.22	4.51	4.63	−**0.77 (12.7%)**	4.94	2.68	2.49	1.88	1.79	−**1.03 (26.1%)**
m-wait-k	7.12	5.91	5.38	5.03	4.91		5.31	3.65	3.03	2.77	2.61	
+KD	6.24	5.40	4.96	4.72	4.58	−**0.43 (7.2%)**	4.75	3.05	2.62	2.36	2.27	−**0.47 (13.7%)**
	λ=0.3	λ=0.25	λ=0.2	λ=0.15	λ=0.1	Avg. Δ	λ=0.3	λ=0.25	λ=0.2	λ=0.15	λ=0.1	Avg. Δ
MMA	10.15	6.20	5.12	4.99	5.29		3.96	3.51	3.36	2.96	3.01	
+KD	4.60	4.29	4.32	4.44	4.60	−**1.90 (25.0%)**	3.09	2.87	2.65	2.56	2.31	−**0.62 (18.3%)**
	δ=0.2	δ=0.3	δ=0.4	δ=0.5	δ=0.6	Avg. Δ	δ=0.2	δ=0.3	δ=0.4	δ=0.5	δ=0.6	Avg. Δ
ITST	6.51	5.46	4.61	4.50	4.20		3.46	2.77	2.37	2.25	2.22	
+KD	5.06	4.46	4.26	4.06	4.01	−**0.68 (12.4%)**	2.84	2.41	2.16	2.09	2.12	−**0.29 (10.4%)**

Table 11: Effect of KD-SiMT on Hallucination Rate (%, ↓) of SiMT models.

	k	1	3	5	7	9
waitk	AL	-0.60	3.03	4.96	6.88	8.82
	PHS	52	26	20	14	14
waitk+KD	AL	0.20	2.10	4.96	7.05	8.92
	PHS	34	16	14	8	10
	λ	0.3	0.25	0.2	0.15	0.1
MMA	AL	2.59	5.72	7.42	9.51	11.99
	PHS	39	26	17	14	10
MMA+KD	AL	7.59	9.20	10.21	13.42	16.79
	PHS	10	8	7	5	4

Table 12: Proportion of hallucination sentences (PHS,%) in Zh→En SiMT task.

wait-k	AL	-0.65	3.02	4.84	6.75	8.58
	COMET	67.59	74.35	76.19	77.63	78.48
+KD	AL	0.24	1.88	4.86	6.87	8.63
	COMET	70.36	74.49	77.50	78.67	79.28
MMA	AL	3.47	5.82	7.03	9.03	11.25
	COMET	68.81	75.11	76.95	77.69	78.18
+KD	AL	7.29	8.67	9.53	12.41	15.52
	COMET	78.65	79.68	79.99	79.80	79.78

Table 13: COMET scores on Zh→En task.

Source	据 了解 , 做 小时工 已经 成 了 一些 正在 求职 的 毕业生 积累 工作 经验 、 锻炼 工作 技能 、 提高 沟通 技巧 、 增长 社会 阅历 的 重要 途径 。
Reference	it was learned that working at hourly positions has become an important approach for some job - seeking graduates to accumulate work experience , practice work skills , enhance communication skills , and increase exposure to society .
MMA	it is understood that being small workers has become **a result of** some of the graduates who are seeking jobs **in the past** , a key way to improve their skills for work , communication skills , and a history of social experience .
+KD	it is learned that part-time workers have become important channels for graduates seeking jobs to accumulate work experience , exercise work skills , enhance communication skills , and increase social experience .
Source	埃及 是 世界 四 大 文明 古国 之一 , 它 悠久 的 历史 吸引 着 各 地 的 游客 。
Reference	for generations , danish people have been pursuing human rights , democracy , and freedom beyond politics .
m-waitk	egypt is **the only country** in the world with the largest ancient civilization , the ancient civilization that has attracted many visitors .
+KD	egypt is one of the four ancient civilizations of the world . its long history has attracted tourists from all over the world .

Table 14: The instructions about KD reducing hallucinations in SiMT. The "**hallucinations**" are marked.