INK: Injecting kNN Knowledge in Nearest Neighbor Machine Translation

Wenhao Zhu[1], Jingjing Xu[2], Shujian Huang[1], Lingpeng Kong[3], Jiajun Chen[1]

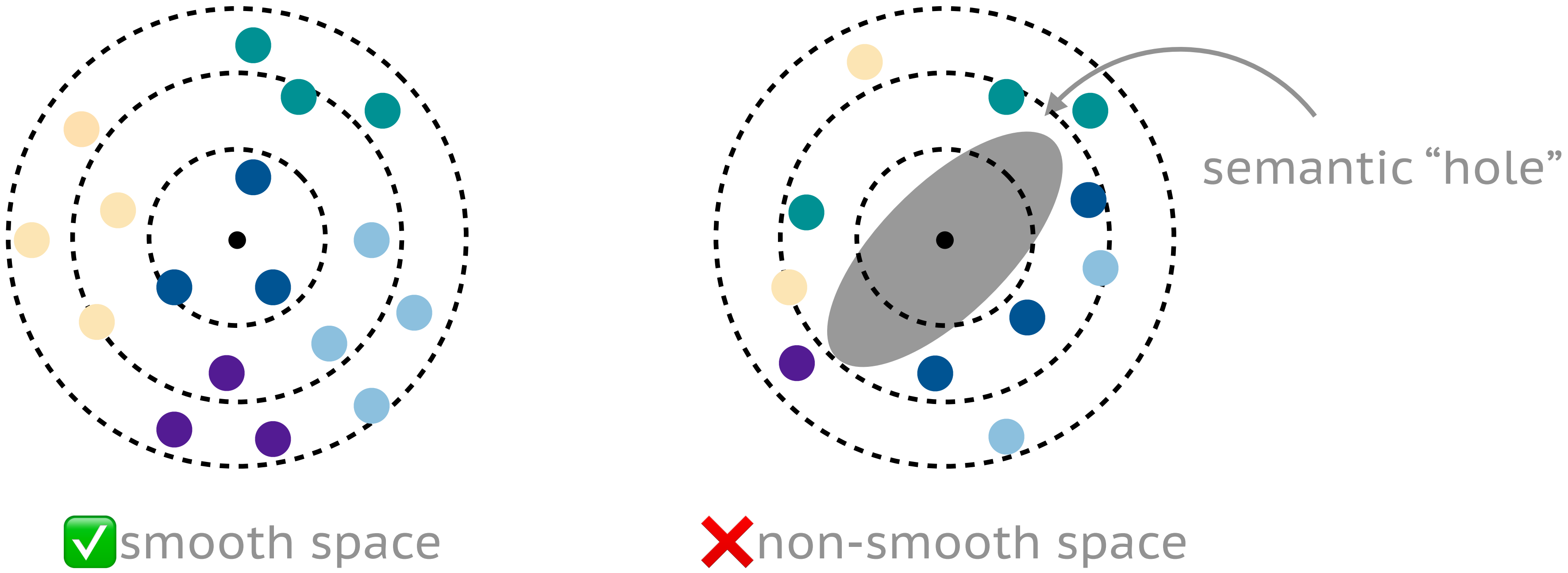
National Key Laboratory for Novel Software Technology, Nanjing University

Shanghai AI Laboratory   The University of Hong Kong

Neural Machine Translation

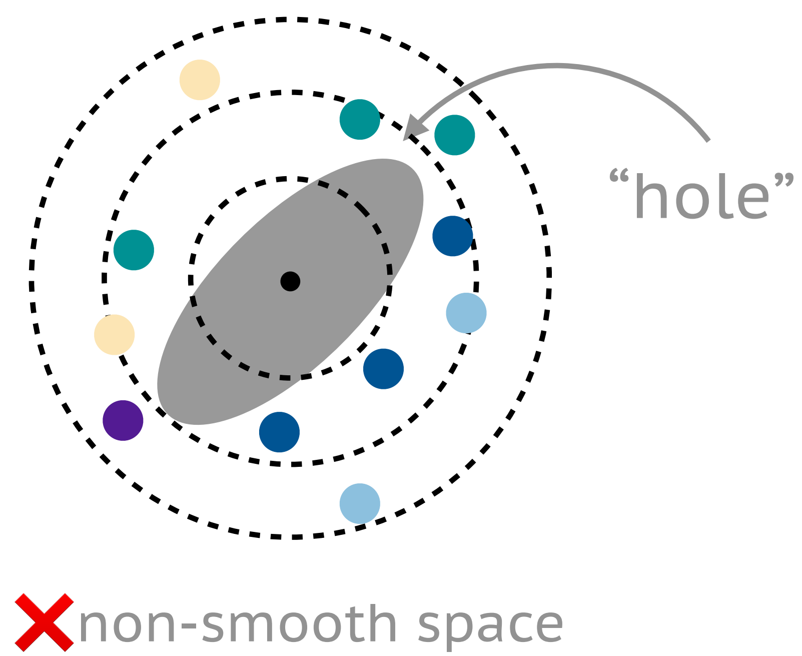
- NMT have achieved promising results in recent years.
- The target of NMT is to learn a generalized representation space to adapt to diverse scenarios.
- However, neural networks often induce a non-smooth representation space, limiting its generalization ability.

Ideally, all of the representations in a neighborhood should share the same target token.



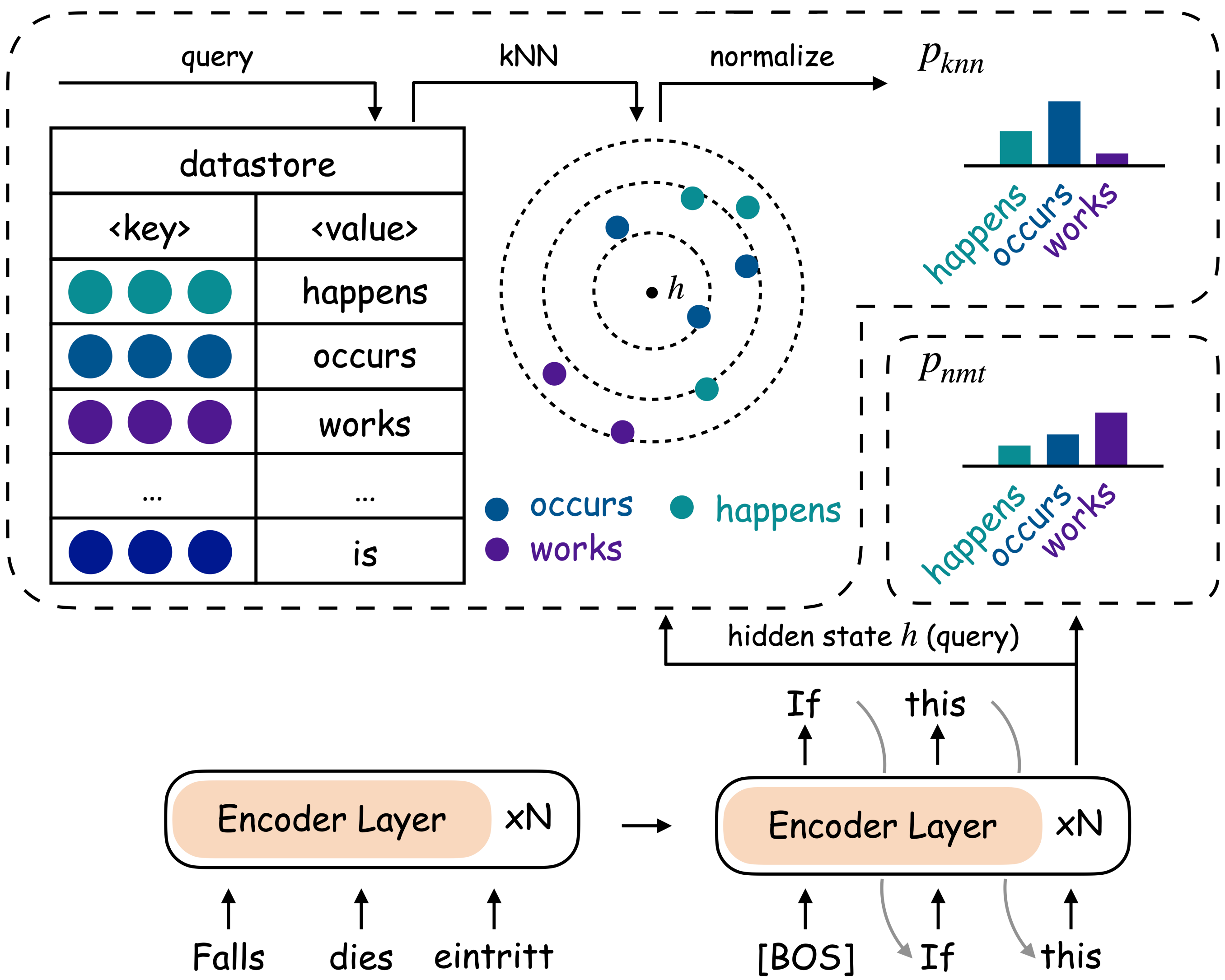
Non-smooth Representation Space of NMT Model

- non-smooth representation space
  - low-frequency tokens disperse sparsely.
  - many "holes" could be formed, where the semantic meaning can be poorly defined.
- As a result, when NMT is used to translate examples from an unseen domain, the performance drops sharply.



Previous Solution: kNN-MT

- kNN-MT (k-nearest neighbor machine translation)
  - saving representations and target tokens into a datastore
  - smoothing predictions with nearest neighbors



Drawbacks of kNN-MT

- Retrieving neighbors from a large datastore at each decoding step is time consuming
- Once the datastore is constructed, representations can not be easily updated

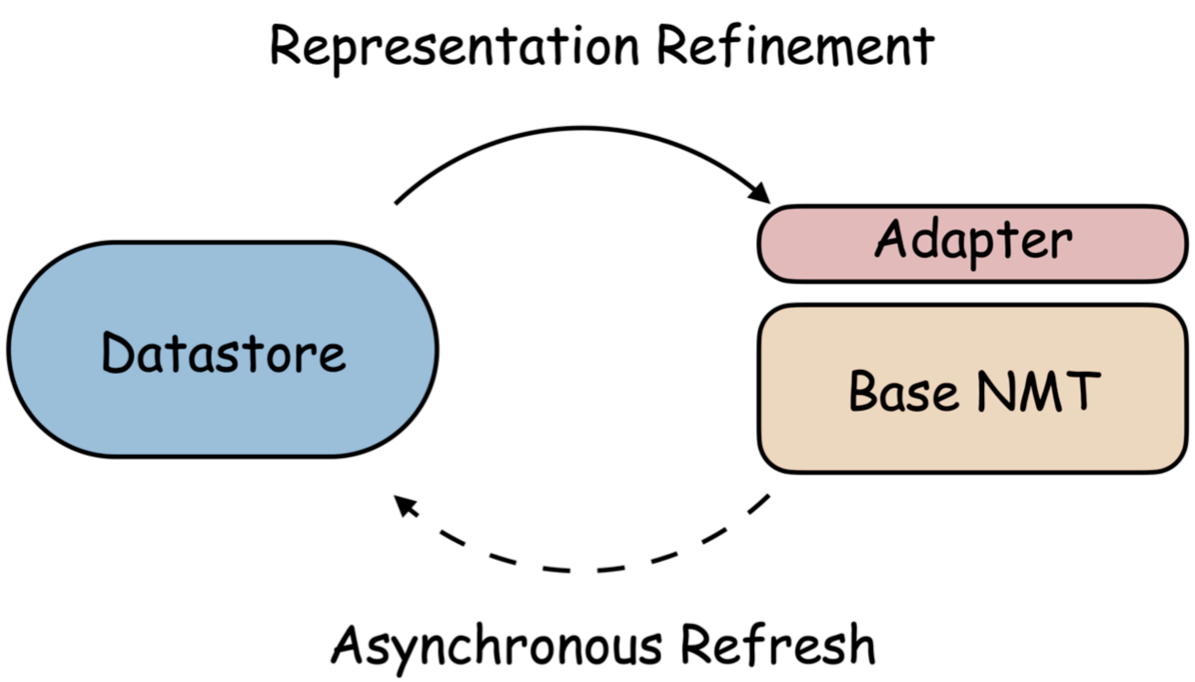
To overcome these drawbacks, we propose **INK** to INject kNN Knowledge into NMT.

Smoothing Representation Space with INK

- Overview of INK training loop
  - representation refinement

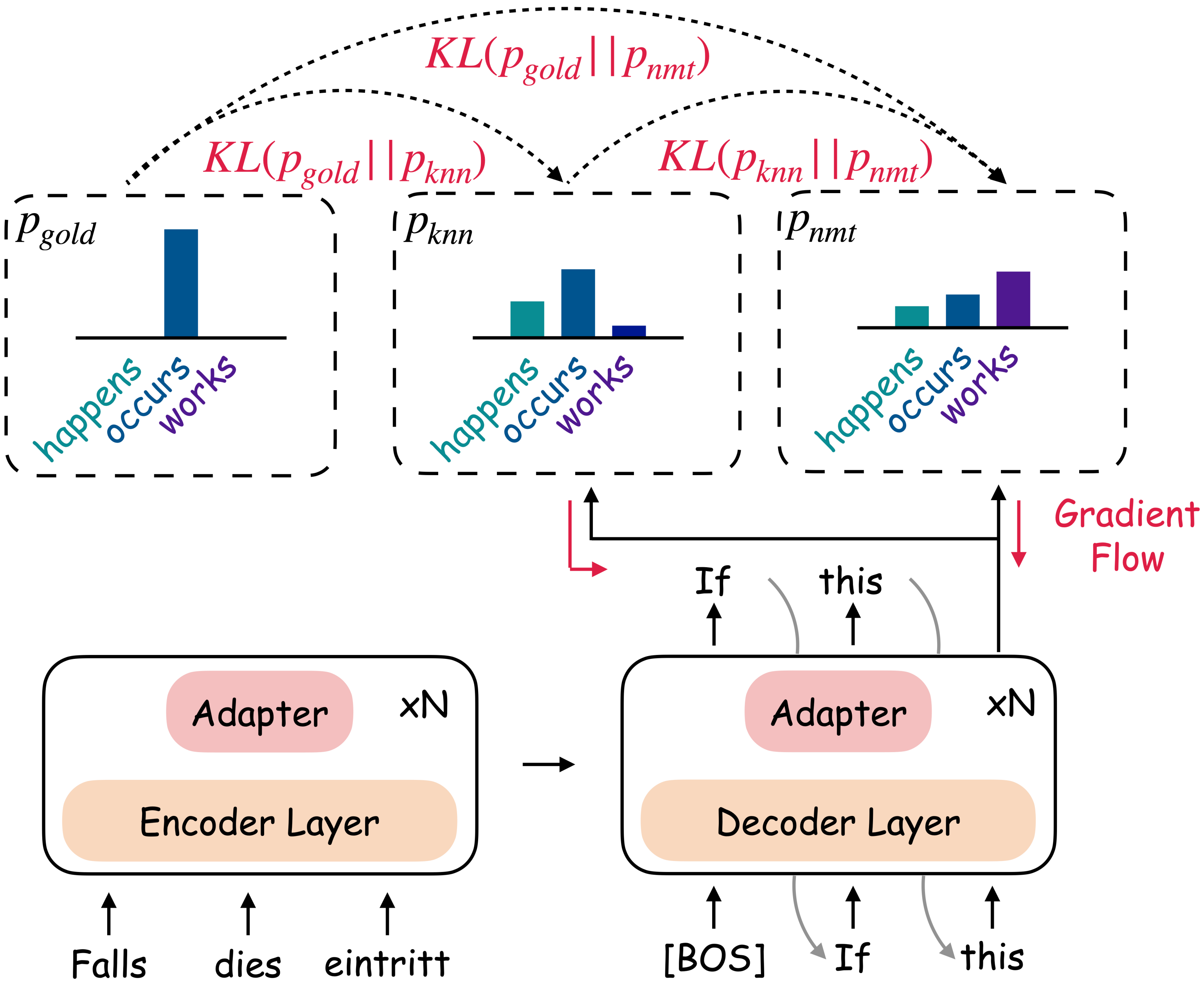
    extracting kNN knowledge to adjust representation
  - asynchronous refresh

    using updated representation to refresh the datastore



Smoothing Representation Space with INK

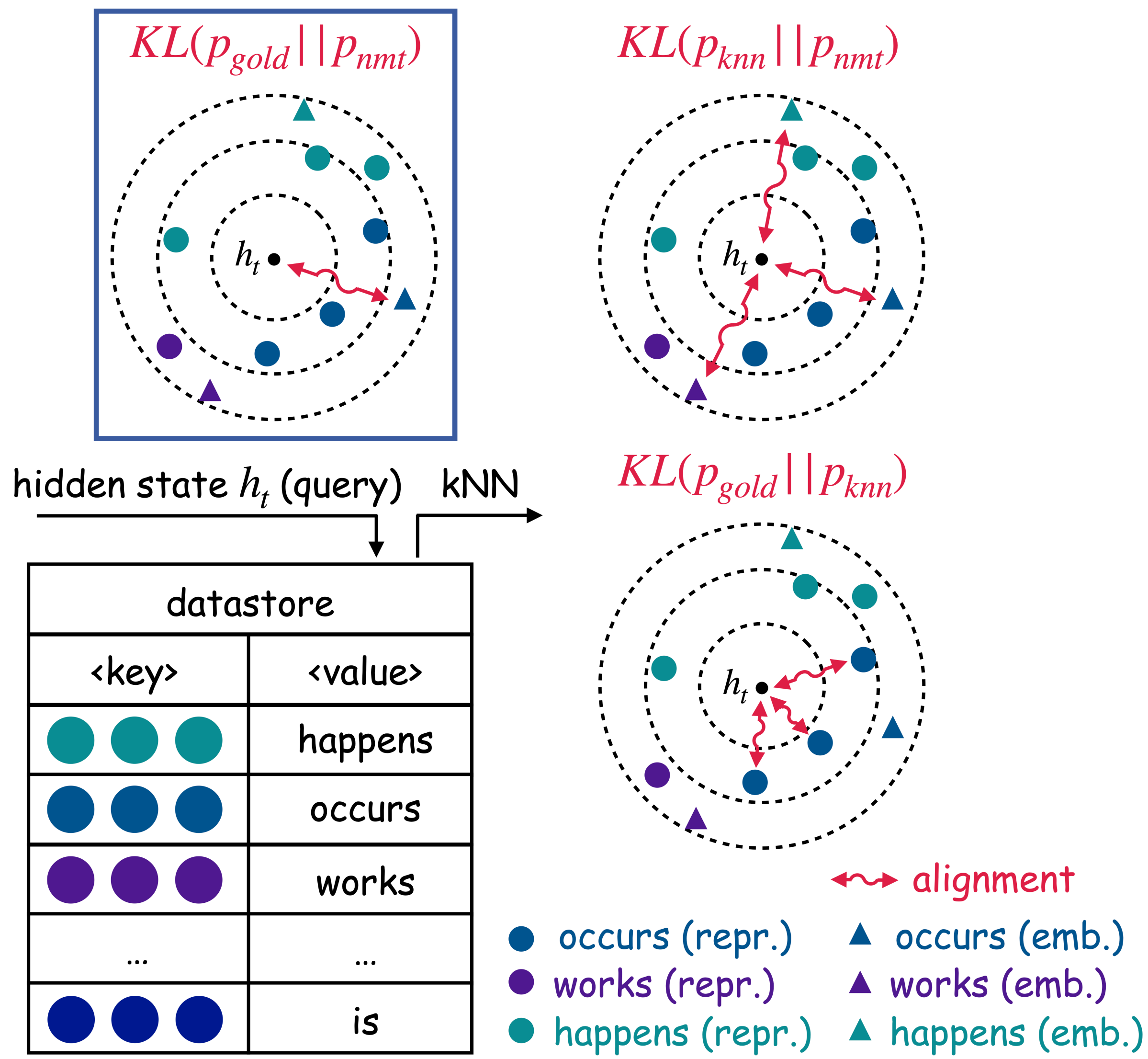
- We adjust the representation by aligning three kinds of representations with KL-divergence.



Smoothing Representation with INK

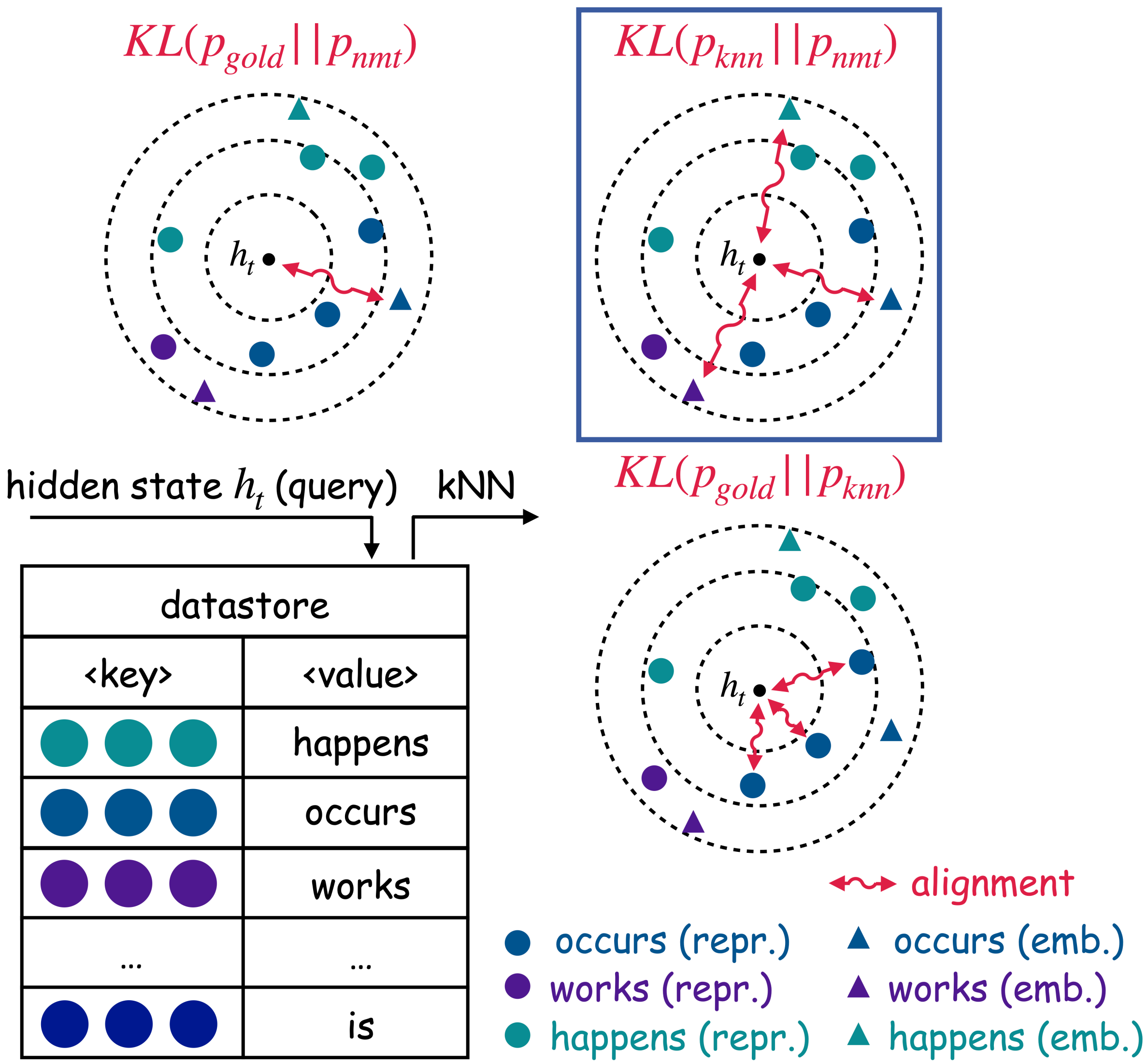
- Aligning contextualized representations and token embeddings.

$$\mathcal{L}_t^a = D_{\mathrm{KL}}[\, p_{\mathrm{gold}}(y|X, Y_{<t}) \parallel p_{\mathrm{nmt}}(y|X, Y_{<t}$$
$$= -\log \frac{\sum_{(w,v)\in\mathcal{E}} \mathbb{1}(v = y_t)\kappa(h_t, w)}{\sum_{(w,v)\in\mathcal{E}} \kappa(h_t, w)}$$



Smoothing Representation with INK

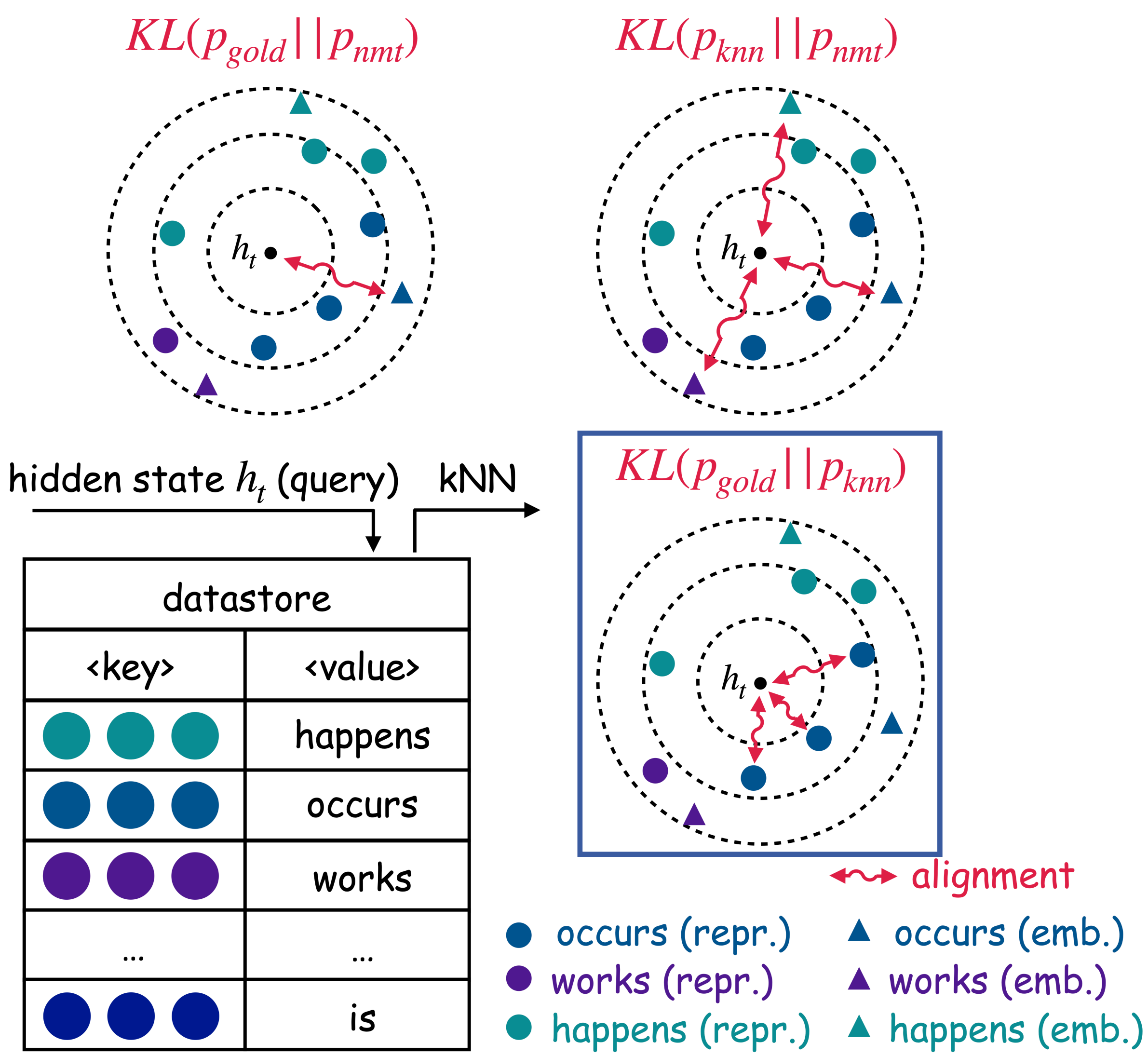
- Aligning contextualized representations and kNN token embeddings.



$$\mathcal{L}_t^i = D_{\mathrm{KL}}[\, p_{\mathrm{knn}}(y|X, Y_{<t}) \parallel p_{\mathrm{nmt}}(y|X, Y_{<t})\,]$$
$$= -\sum_{\bar{y}\in\mathcal{Y}} p_{\mathrm{knn}}(\bar{y}) \cdot \log \frac{\sum_{(w,v)\in\mathcal{E}} \mathbb{1}(v=\bar{y})\kappa(h_t, w)}{\sum_{(w,v)\in\mathcal{E}} \kappa(h_t, w) \cdot p_{\mathrm{knn}}(\bar{y})}$$

Smoothing Representation with INK

- Aligning contextualized representations of the same target token.



$$
\begin{aligned}
\mathcal{L}_t^r &= D_{\mathrm{KL}}[\, p_{\mathrm{gold}}(y|X, Y_{<t}) \parallel p_{\mathrm{knn}}(y|X, Y_{<t}) \,] \\
&= -\log \frac{\sum_{(\hat{h},\hat{y}) \in \mathcal{N}_k} \mathbb{1}(\hat{y} = y_t)\kappa(h_t, \hat{h})}{\sum_{(\hat{h},\hat{y}) \in \mathcal{N}_k} \kappa(h_t, \hat{h})}
\end{aligned}
$$

Smoothing Representation with INK

- Overall Training Procedure
  - optimizing adapter with the combined learning objective

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(X,Y)\in\mathcal{B}} \sum_{t} (\mathcal{L}_t^a + \alpha\mathcal{L}_t^i + \beta\mathcal{L}_t^r)$$

  - refreshing datastore asynchronously
  - runing the loop until convergence

- During inference, we only need to load the off-the-shelf NMT model and tuned adaptation parameters.

Experiment Setting

- NMT Model
  - winner model of WMT'19 news translation task
- Target Domains
  - Medical, Law, IT, Koran
- Baselines
  - V-kNN, A-kNN, R-kNN: different implementation of kNN-MT
  - Adapter: adjusting representations without kNN knowledge
  - kNN-KD: using kNN knowledge to train a NMT from scratch.

Experiment Results

- We explore the following research questions:
  - RQ1: Can we smooth the representation space via small adapter and drop datastore aside during inference?
  - RQ2: How much improvement can be brought by using kNN knowledge to adjust the representation distribution?
  - RQ3: Will together using adapter and datastore bring further improvement?

Main Results

- INK system achieves the best performance by smoothing the representation space.

Systems	Mem.	Medical		Law		IT		Koran		Avg.	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
Off-the-shelf NMT	-	46.87	40.00	57.52	45.47	39.22	38.39	-1.32	16.26	35.57	35.03
*k*NN-KD	-	56.20	56.37	68.60	60.65	-1.57	1.48	-13.05	19.60	27.55	34.53
NMT + Datastore Augmentation											
V-*k*NN	$\times$1.7	53.46	54.27	66.03	61.34	51.72	45.56	0.73	20.61	42.98	45.45
A-*k*NN	$\times$1.7	57.45	56.21	69.59	63.13	56.89	47.37	4.68	20.44	47.15	46.79
R-*k*NN†	$\times$1.7	58.05	54.16	69.10	60.90	54.60	45.61	3.99	20.04	46.44	45.18
R-*k*NN	$\times$43.8	57.70	57.12	70.10	**63.74**	57.65	48.50	5.28	20.81	47.68	47.54
NMT + Representation Refinement											
Adapter	$\times$1.0	60.14	56.88	70.87	60.64	66.86	48.21	4.23	21.68	50.53	46.85
INK (ours)	$\times$1.0	**61.64***	**57.75***	**71.13**	61.90*	**68.45***	**49.12***	**8.84***	**23.06***	**52.52**	**47.85**

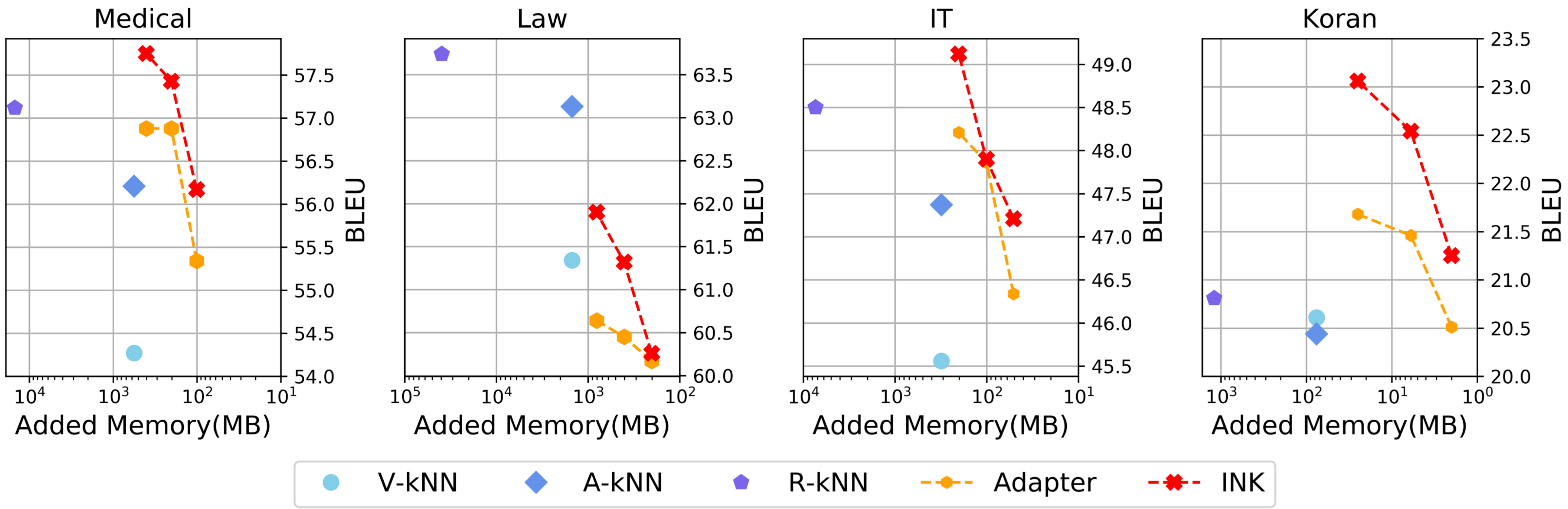
Main Results (Cont.)

- Representation refinement according to kNN knowledge brings larger performance improvement.

Systems	Mem.	Medical COMET	Medical BLEU	Law COMET	Law BLEU	IT COMET	IT BLEU	Koran COMET	Koran BLEU	Avg. COMET	Avg. BLEU
Off-the-shelf NMT	-	46.87	40.00	57.52	45.47	39.22	38.39	-1.32	16.26	35.57	35.03
*k*NN-KD	-	56.20	56.37	68.60	60.65	-1.57	1.48	-13.05	19.60	27.55	34.53
NMT + Datastore Augmentation											
V-*k*NN	×1.7	53.46	54.27	66.03	61.34	51.72	45.56	0.73	20.61	42.98	45.45
A-*k*NN	×1.7	57.45	56.21	69.59	63.13	56.89	47.37	4.68	20.44	47.15	46.79
R-*k*NN†	×1.7	58.05	54.16	69.10	60.90	54.60	45.61	3.99	20.04	46.44	45.18
R-*k*NN	×43.8	57.70	57.12	70.10	**63.74**	57.65	48.50	5.28	20.81	47.68	47.54
NMT + Representation Refinement											
Adapter	×1.0	60.14	56.88	70.87	60.64	66.86	48.21	4.23	21.68	50.53	46.85
INK (ours)	×1.0	**61.64***	**57.75***	**71.13**	61.90*	**68.45***	**49.12***	**8.84***	**23.06***	**52.52**	**47.85**

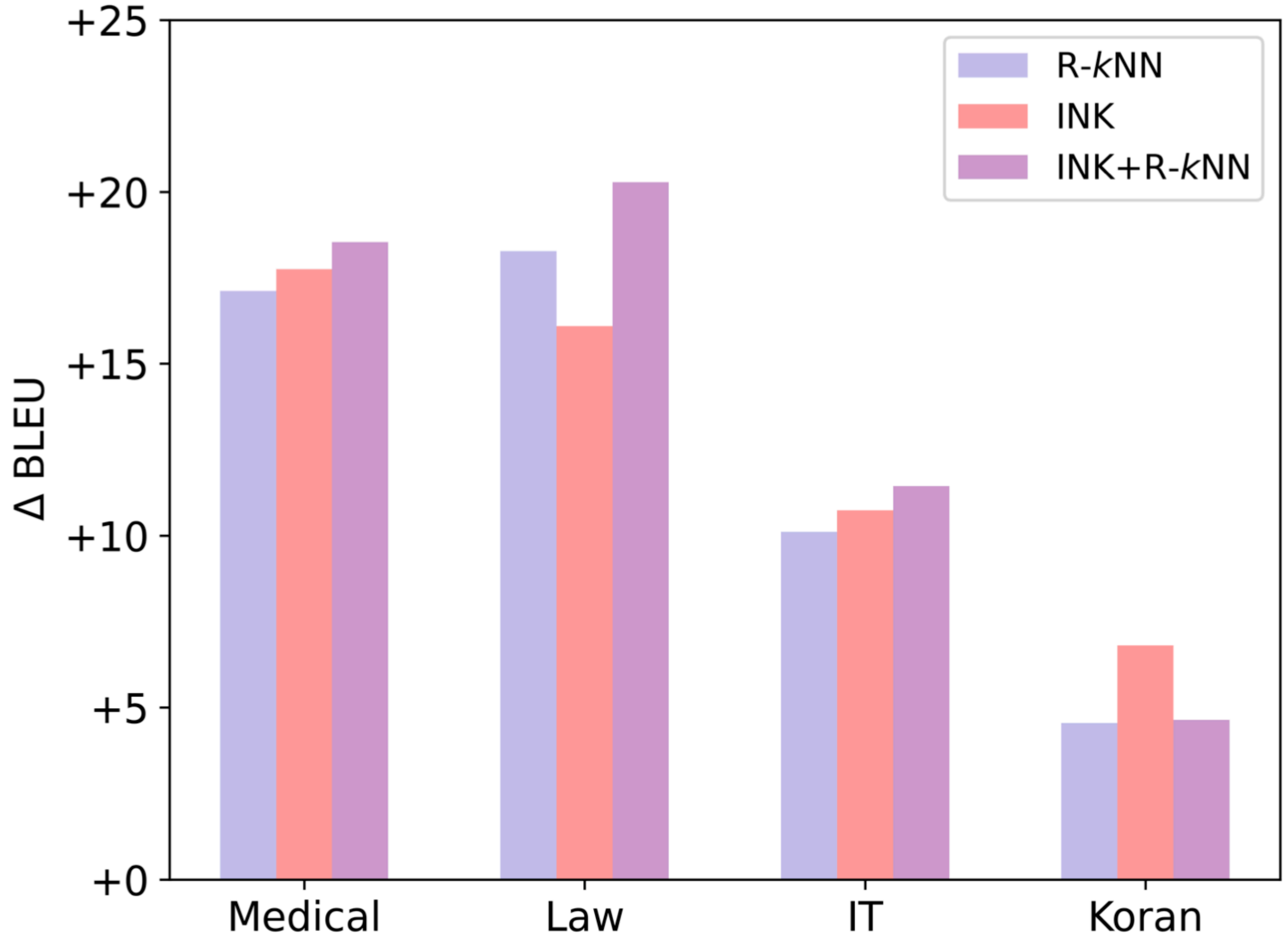
Main Results (Cont.)

- Representation refinement according to kNN knowledge brings larger performance improvement.



Main Results (Cont.)

- Jointly applying adapter and datastore can further smooth predictions.



Conclusion

- We propose a novel training framework INK, to iteratively refine the representation space of the NMT model according to kNN knowledge.
  - INK system achieves an average gain of 1.99 COMET and 1.0 BLEU.
  - Compared with kNN-MT baselines, our INK achieves better translation performance with 0.02× memory space and 1.9× inference speed up.