Data Augmentation by Concatenation for Low-Resource Translation: A Mystery and a Solution

Toan Q. Nguyen*, Kenton Murray†, David Chiang*

* University of Notre Dame

† Johns Hopkins University

Motivation

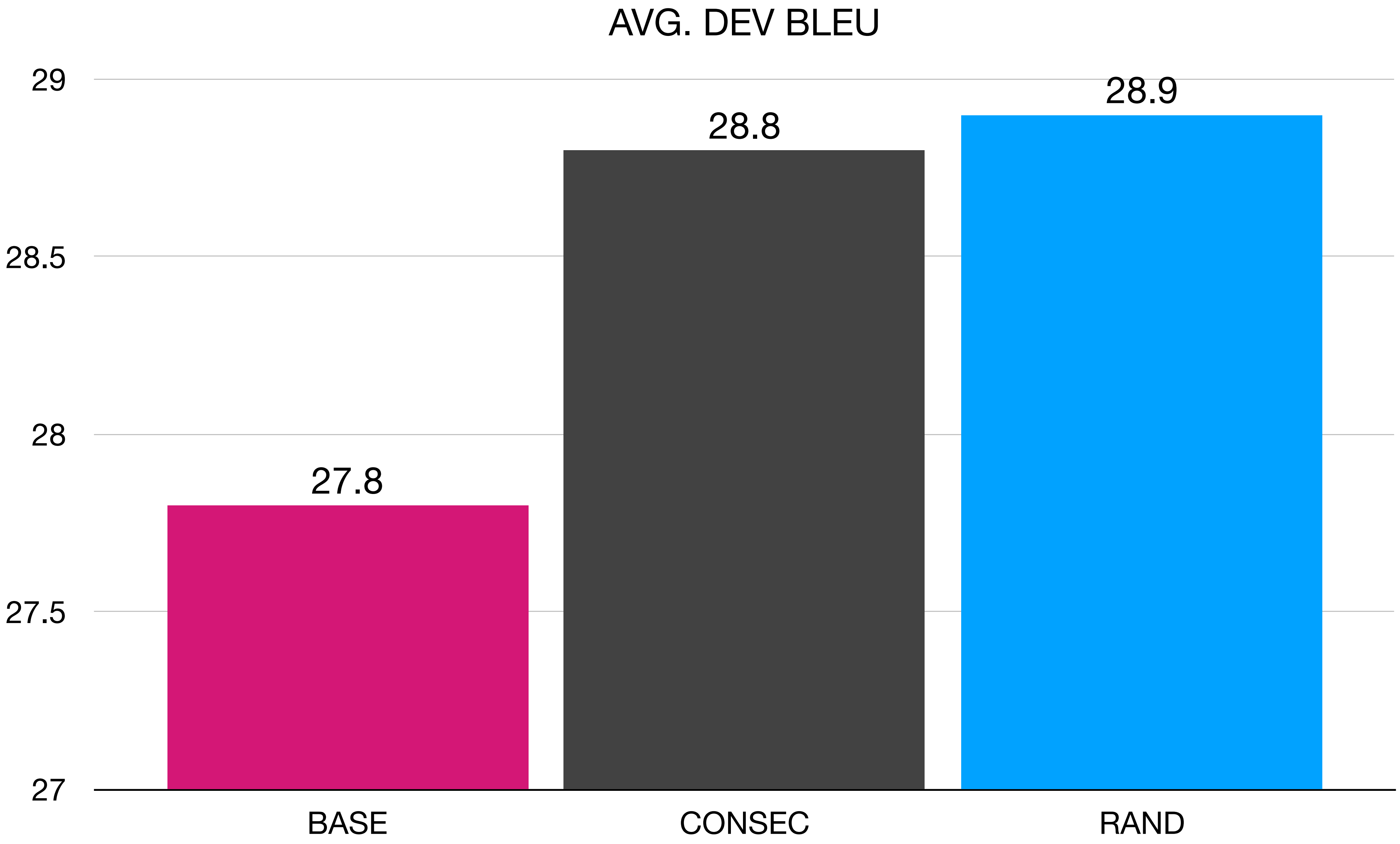
Concatenating ***consecutive pairs*** (during training) is a simple, **non-invasive** data augmentation method for NMT

Discourse context (from previous sentence) is often attributed for its improvement

	Source	Target
1837th pair	And I think back .	Và tôi nghĩ lại .
1838th pair	I think back to my father .	Tôi nghĩ lại về cha tôi .
Generated pair	And I think back . <EOS> I think back to my father .	Và tôi nghĩ lại . <EOS> Tôi nghĩ lại về cha tôi .

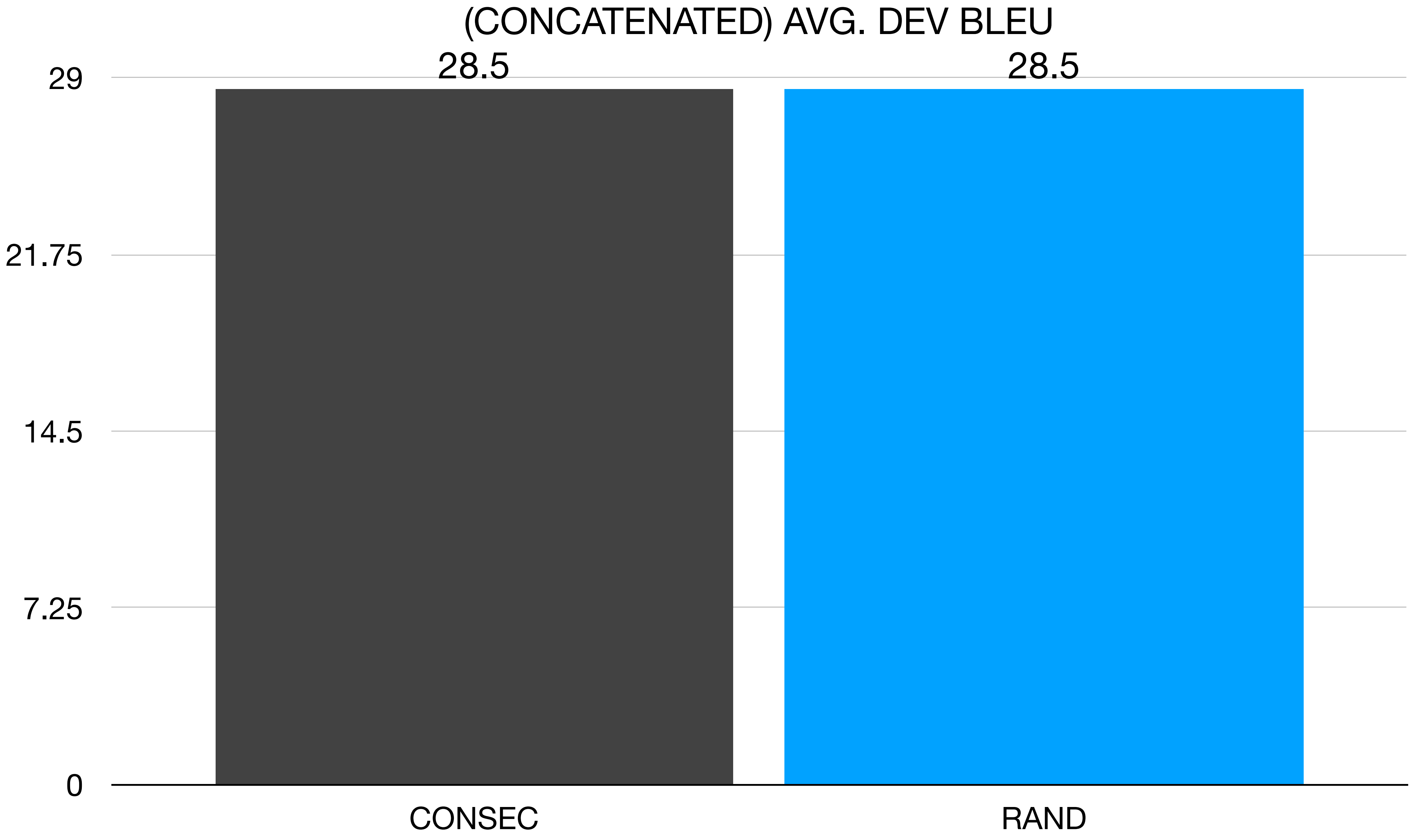
However, we found concatenating ***random pairs*** yields the same improvement (test on 4 low-resource language pairs)

	Source	Target
41864th pair	And this investment is actually Western-led .	Và sự đầu tư này chắc chắn do phương Tây dẫn đầu .
1838th pair	I think back to my father .	Tôi nghĩ lại về cha tôi .
Generated pair	And this investment is actually Western-led . <EOS> I think back to my father .	Và sự đầu tư này chắc chắn do phương Tây dẫn đầu . <EOS> Tôi nghĩ lại về cha tôi .


AVG. DEV BLEU
29
28.5
28
27.5
27
27.8
28.8
28.9
BASE
CONSEC
RAND

What if we provide models with more discourse contexts?

New dev set: each sentence is the concatenation of two consecutive dev sentences



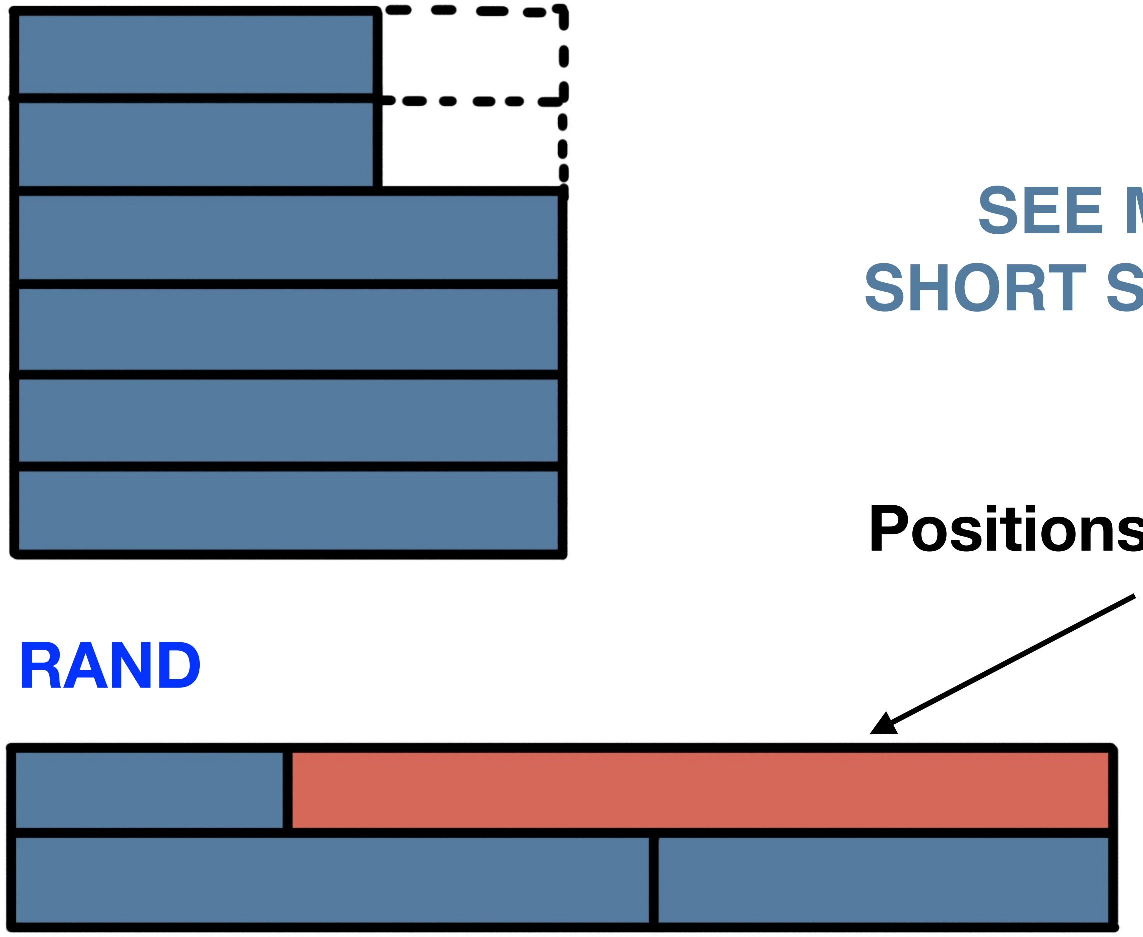
Hypotheses

If it’s not discourse, then what’s the reason?

We have three hypotheses:

- Position Diversity
- Context Diversity
- Length Diversity

Position Diversity



BASELINE

SEE MOSTLY
SHORT SENTENCES

Positions are shifted

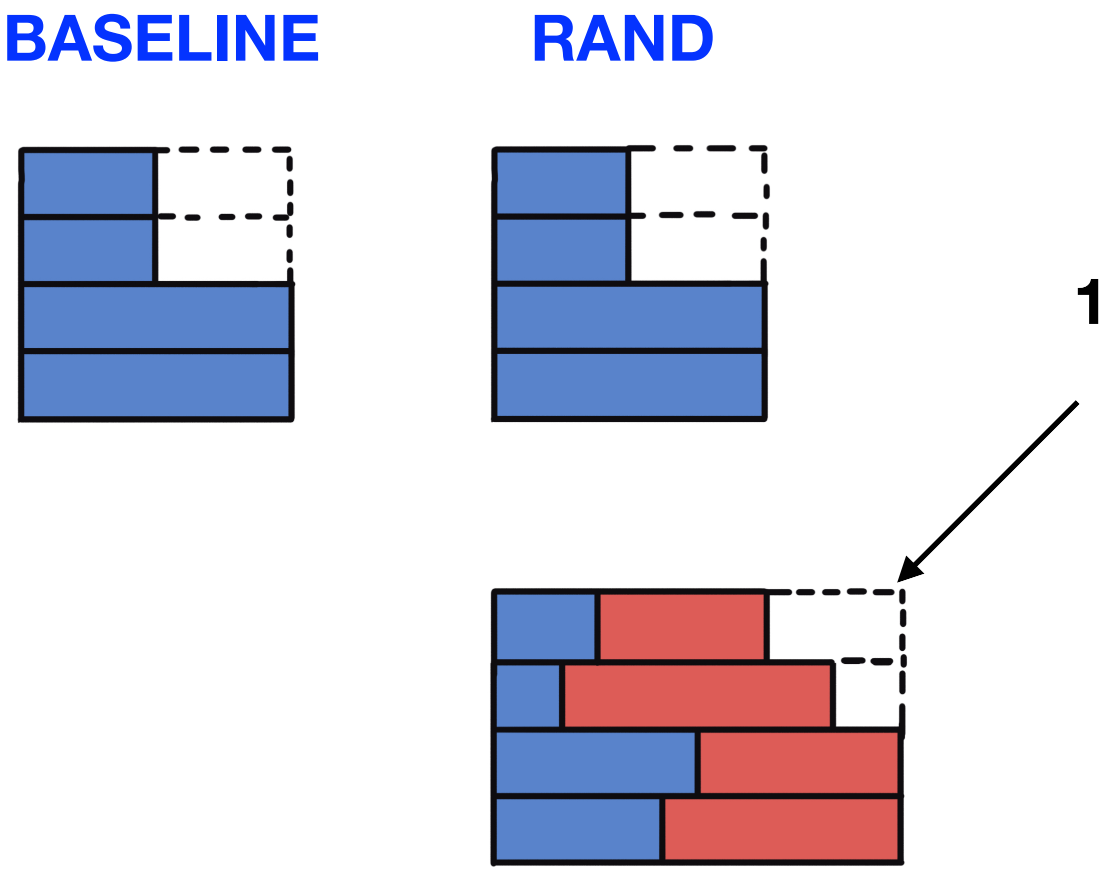
RAND

SEE MORE LONG
SENTENCES

Position Diversity

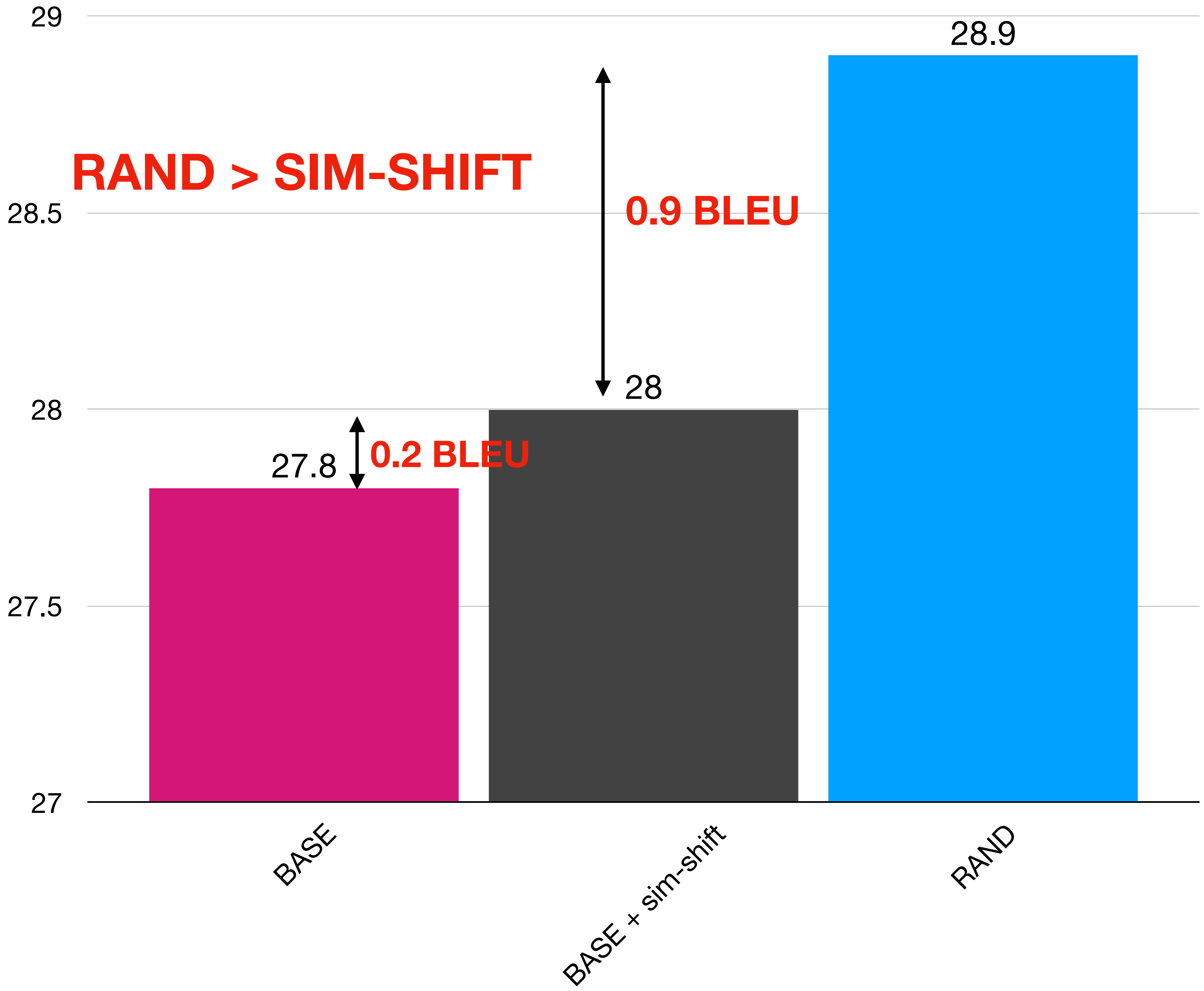
If improvement comes from position shifting, we should be able to reproduce it

sim-shift: randomly shift sentences by d sampled from train lengths at rate 1/3 (this makes it the same as concatenation)



1/3 train sentences are shifted

POSITION SHIFTING AVG. DEV BLEU



RAND > SIM-SHIFT

0.9 BLEU

0.2 BLEU

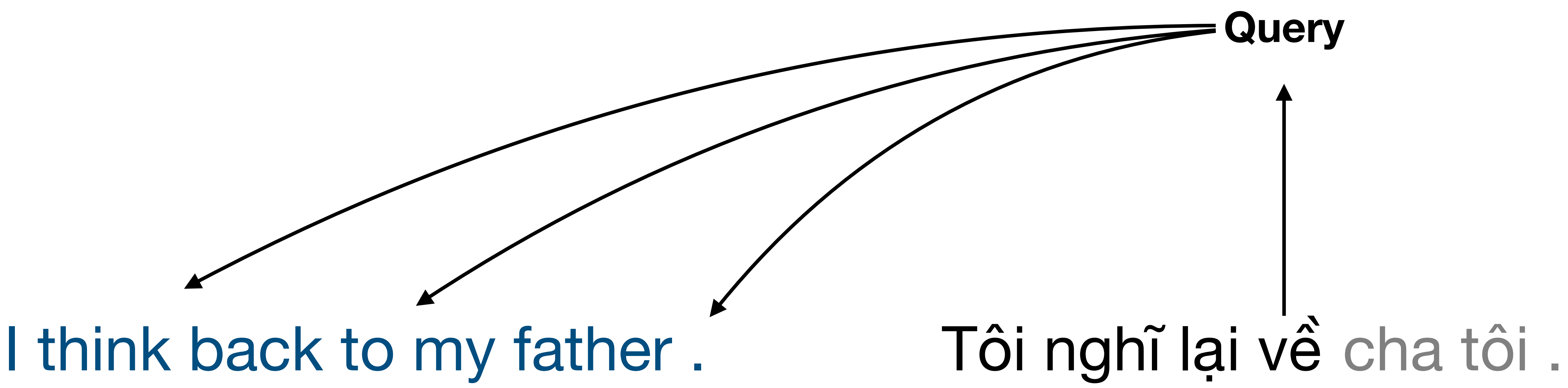
Context Diversity

I think back to my father . → Tôi nghĩ lại về cha tôi .

Context Diversity

I think back to my father . → Tôi nghĩ lại về cha tôi .

Context Diversity

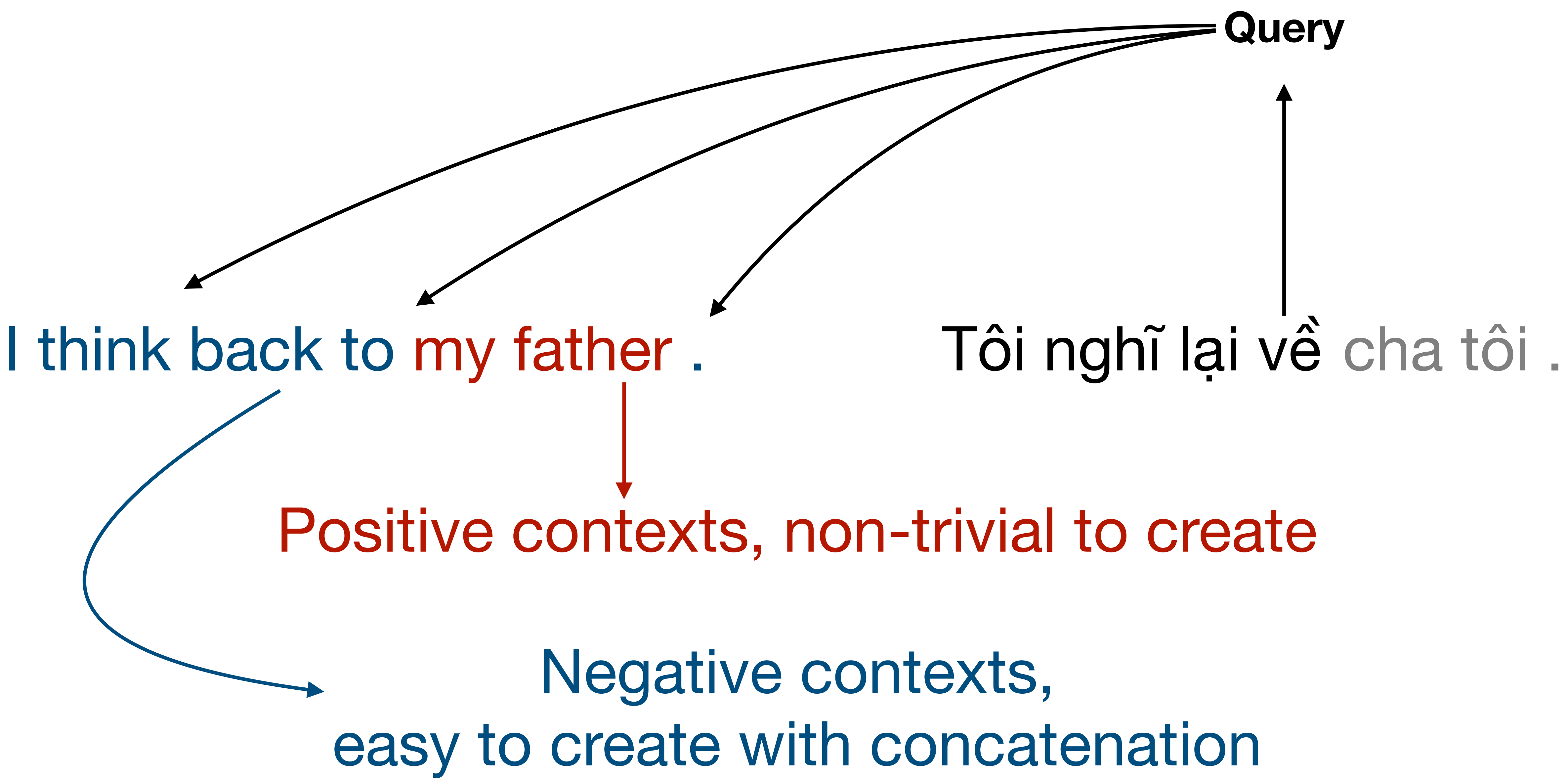


Query

I think back to my father .

Tôi nghĩ lại về cha tôi .

Context Diversity



Query

I think back to my father .

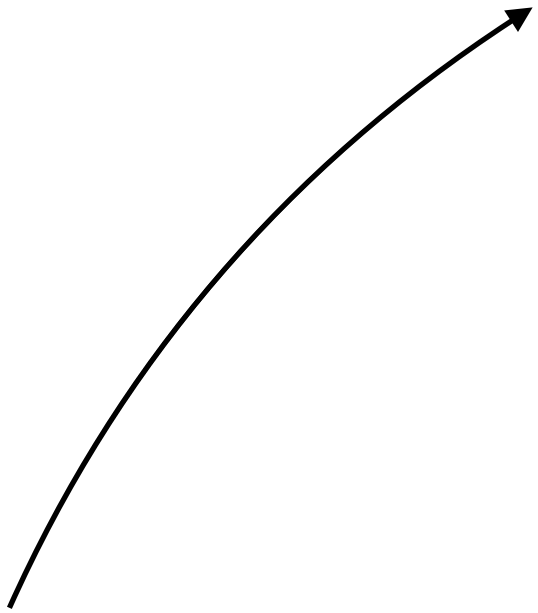
Tôi nghĩ lại về cha tôi .

Positive contexts, non-trivial to create

Negative contexts,
easy to create with concatenation

Context Diversity

“And this investment… <EOS>” acts as negative context

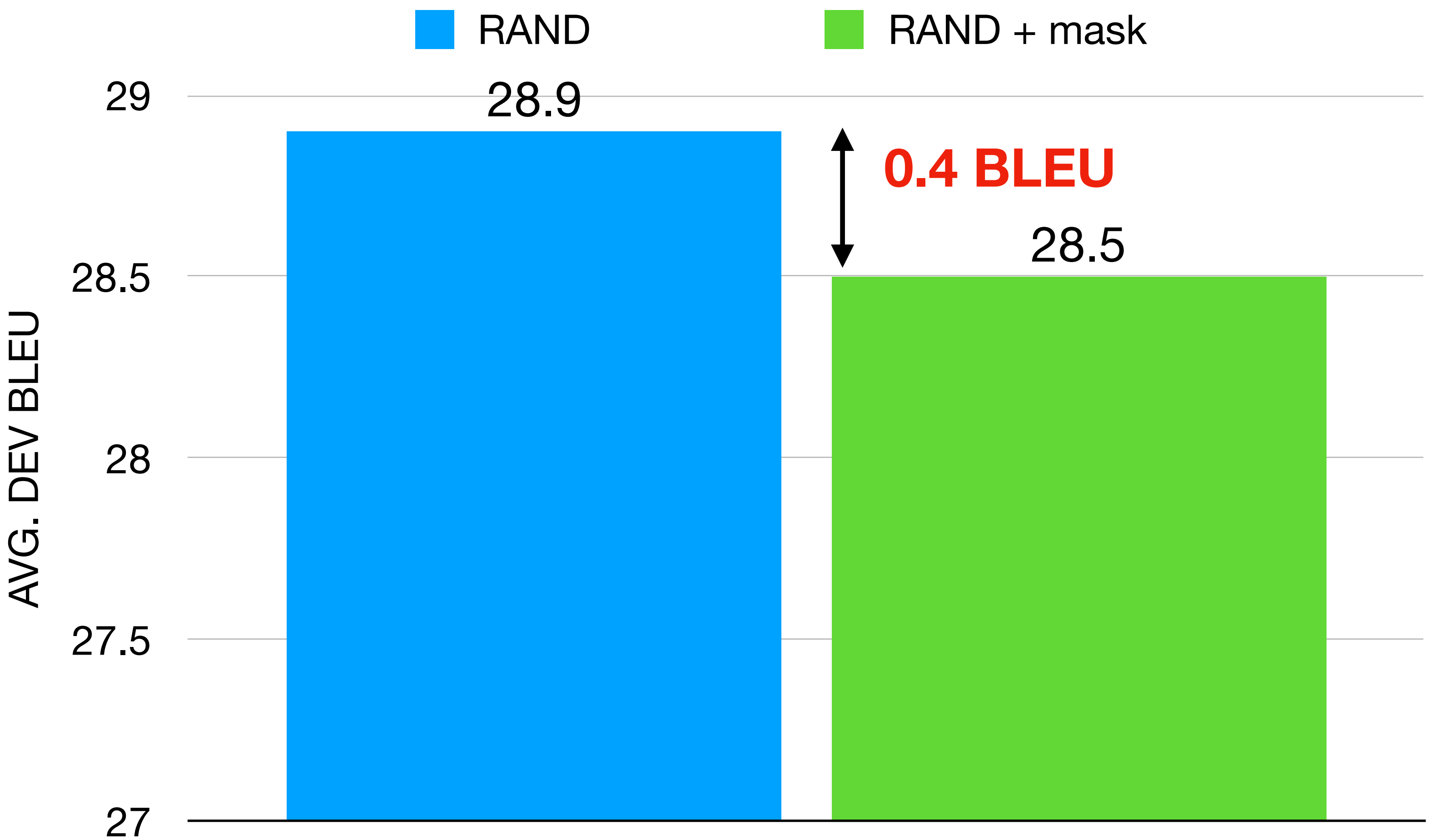


And this investment is actually Western-led . <EOS> I think back to my father .

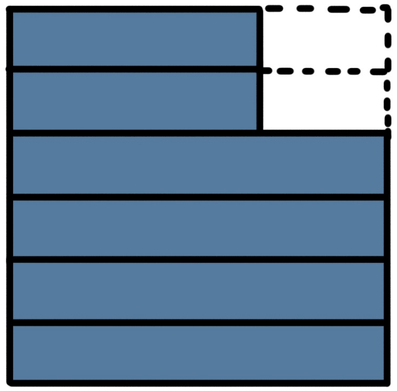
Và sự đầu tư này chắc chắn do phương Tây dẫn đầu . <EOS> Tôi nghĩ lại về cha tôi .

Context Diversity

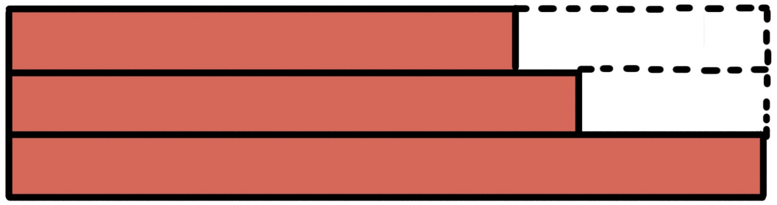
We **mask** attentions such that the two sentences in a concatenated one cannot see each other (same for cross attention)



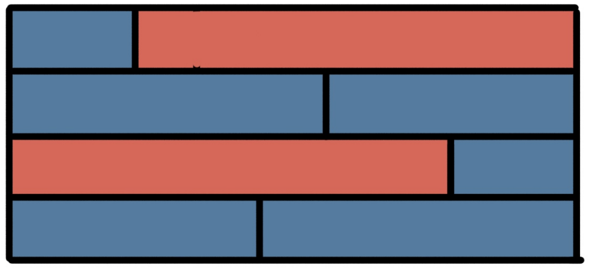
Length Diversity



SHORT SENTENCES, SIMPLE WORDS



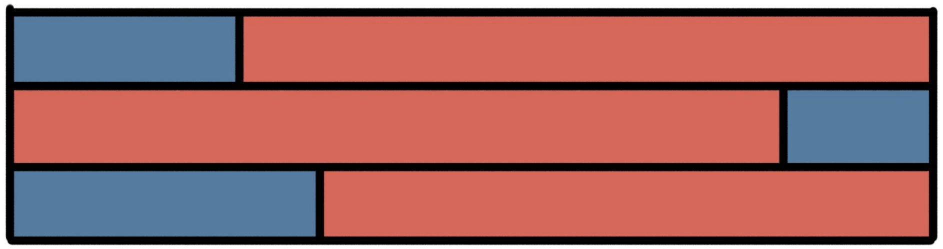
LONG SENTENCES, COMPLICATED WORDS

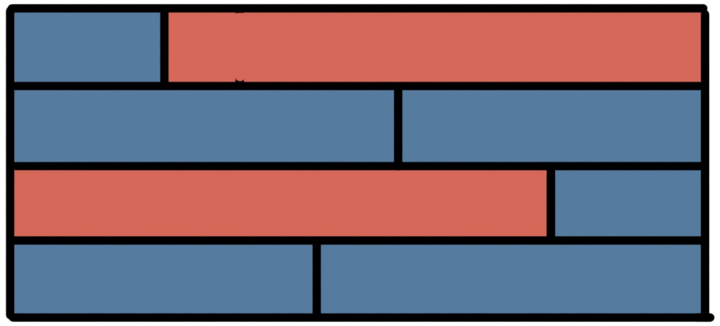


MIXED LONG & SHORT, FEATURE DIVERSITY

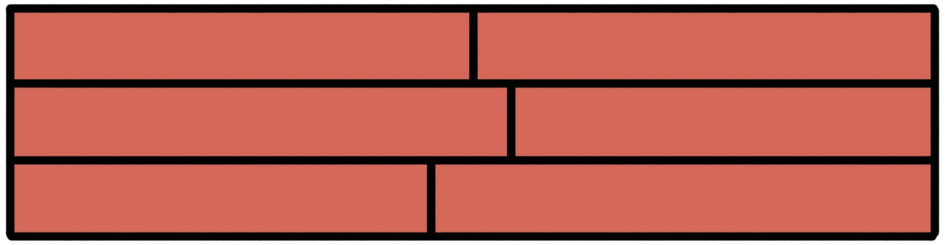
Length Diversity

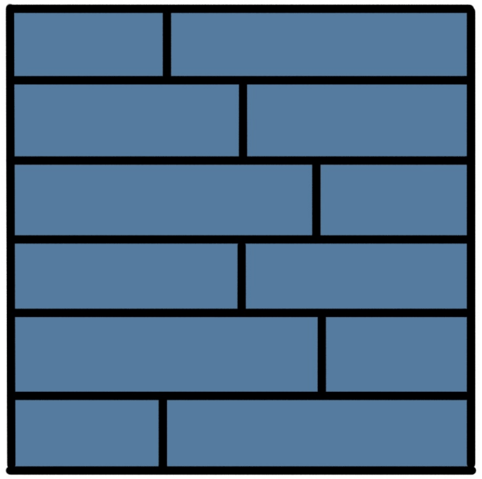
RAND





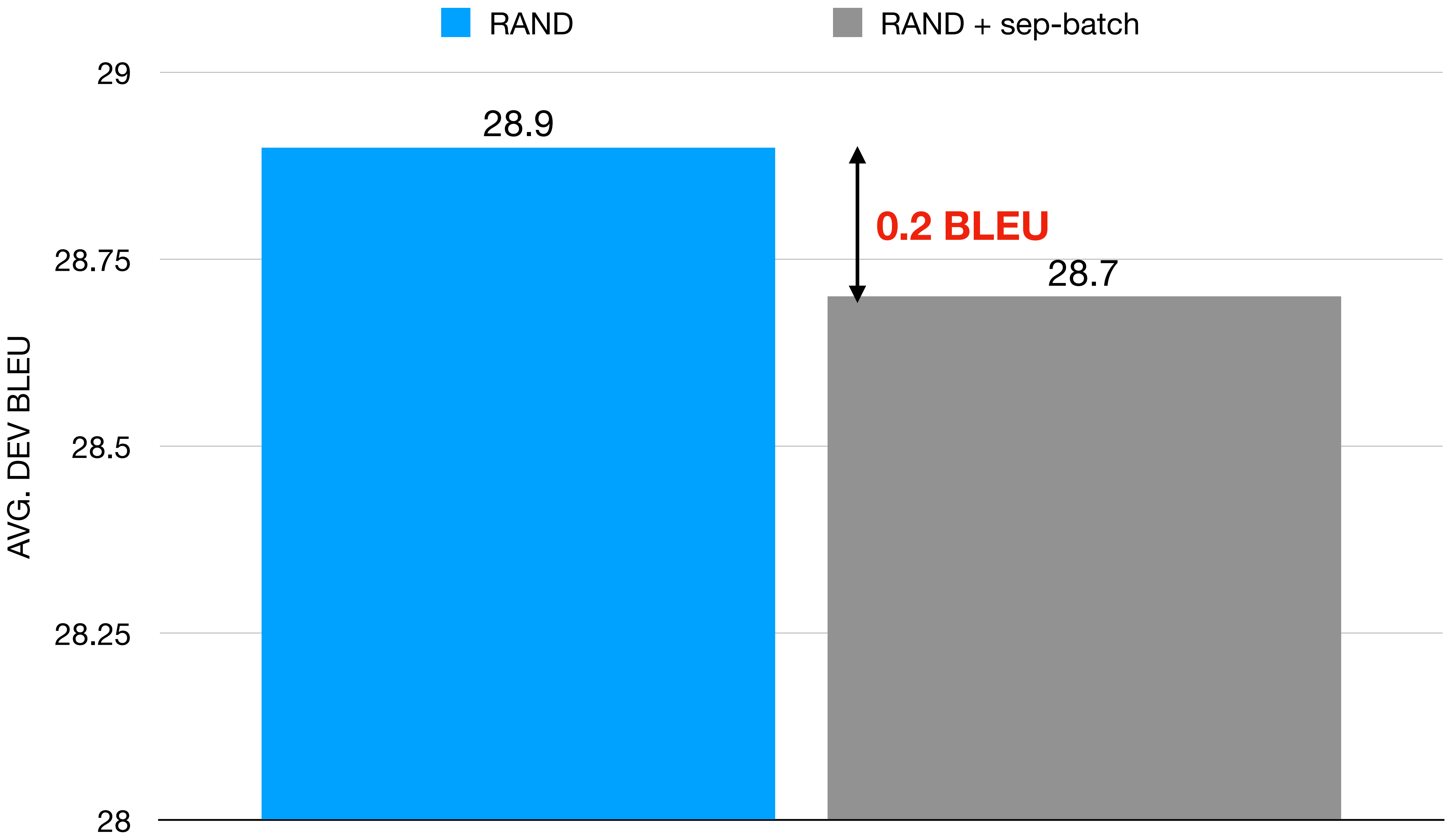
RAND + SEP-BATCH





SORT → CONCAT
REMOVE LENGTH
DIVERSITY

Length Diversity



Putting together

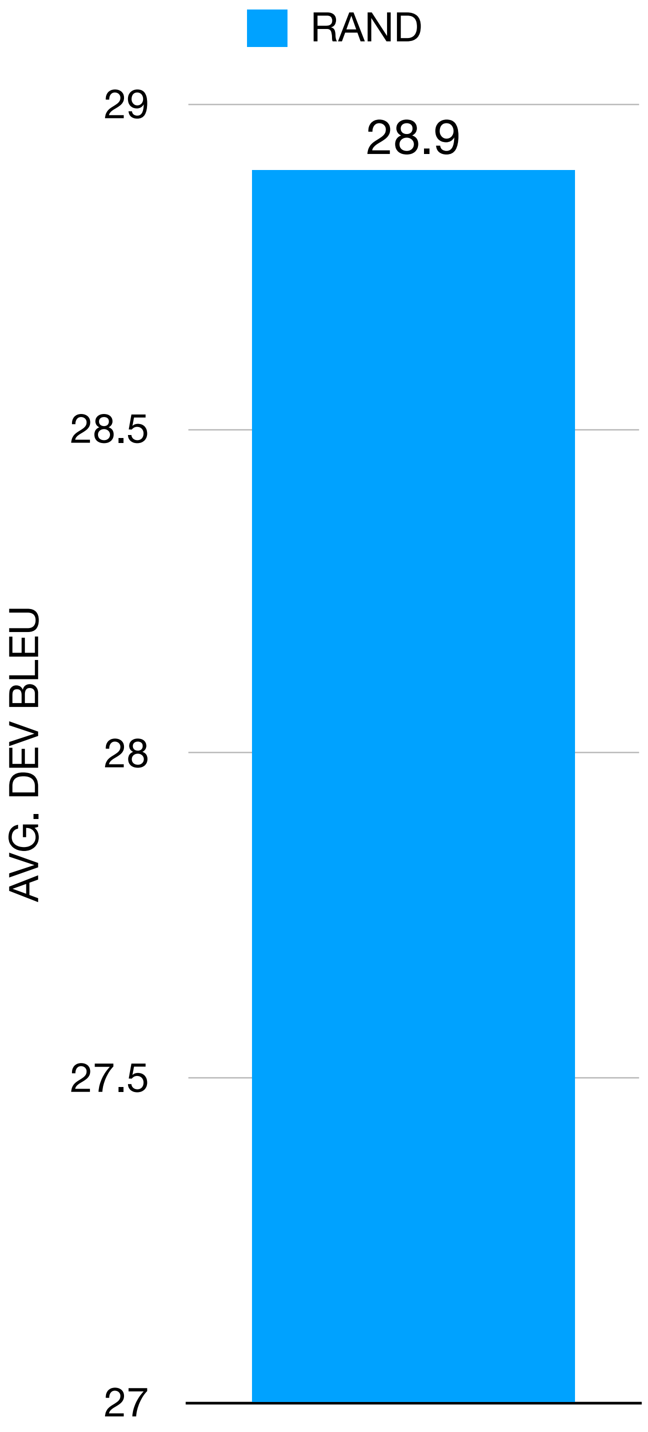
Apply both **mask** and **sep-batch** to concatenation

Reset the positions of the second sentence in a concatenated example

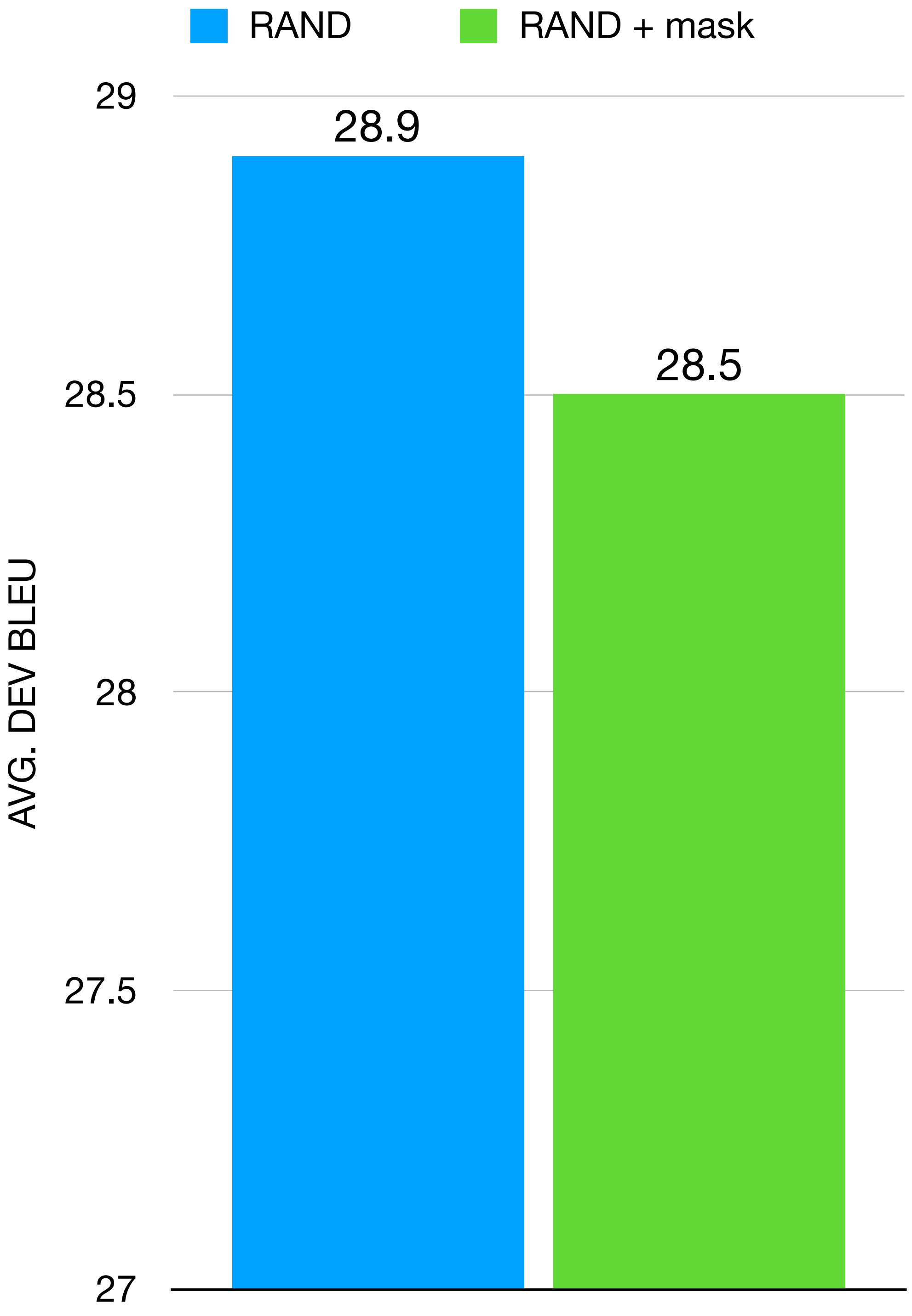
0 1 2 3 4 5 6 7
And this investment is actually Western-led . <EOS>

0 1 2 3 4 5 6
I think back to my father .

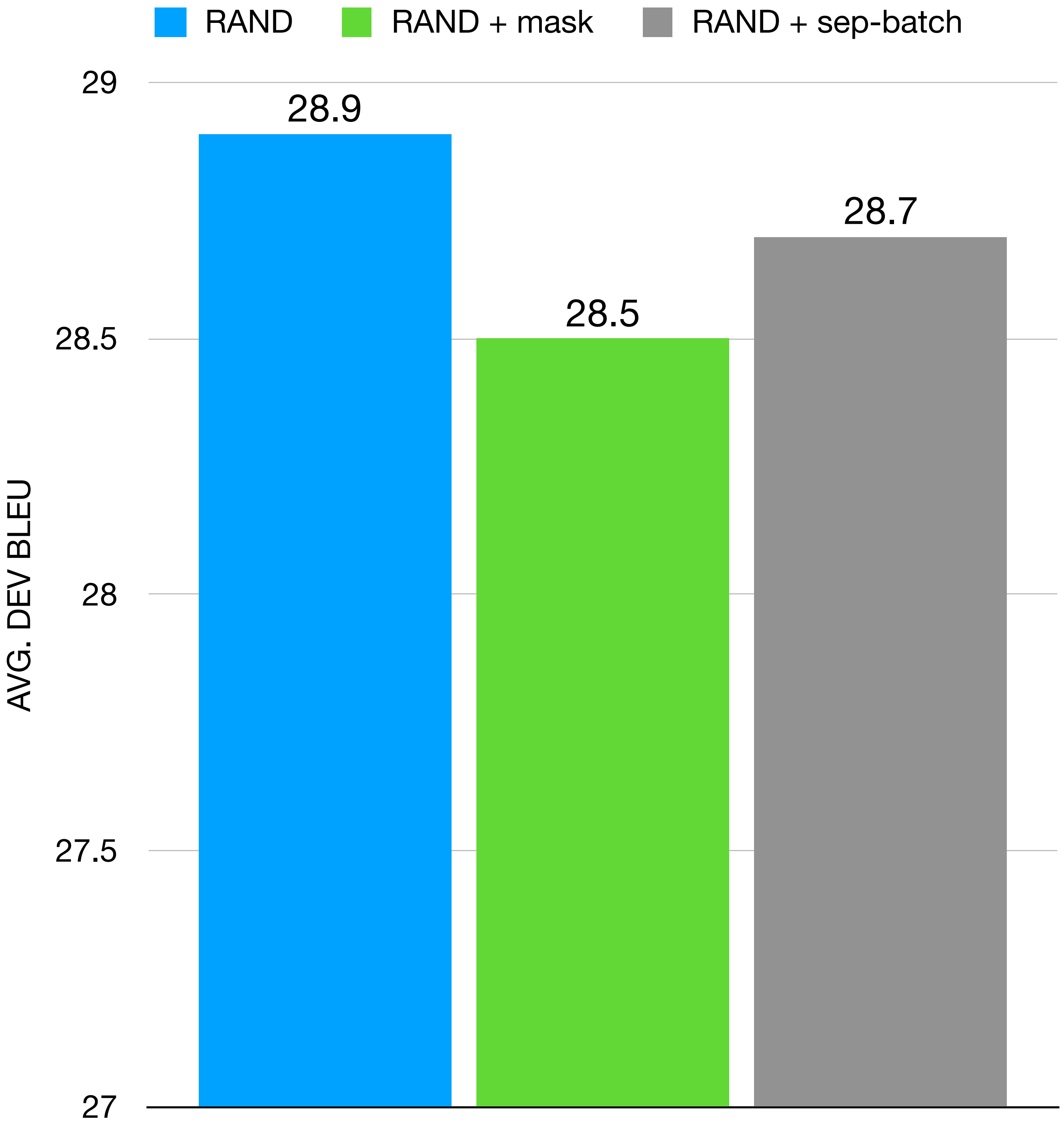
Putting together



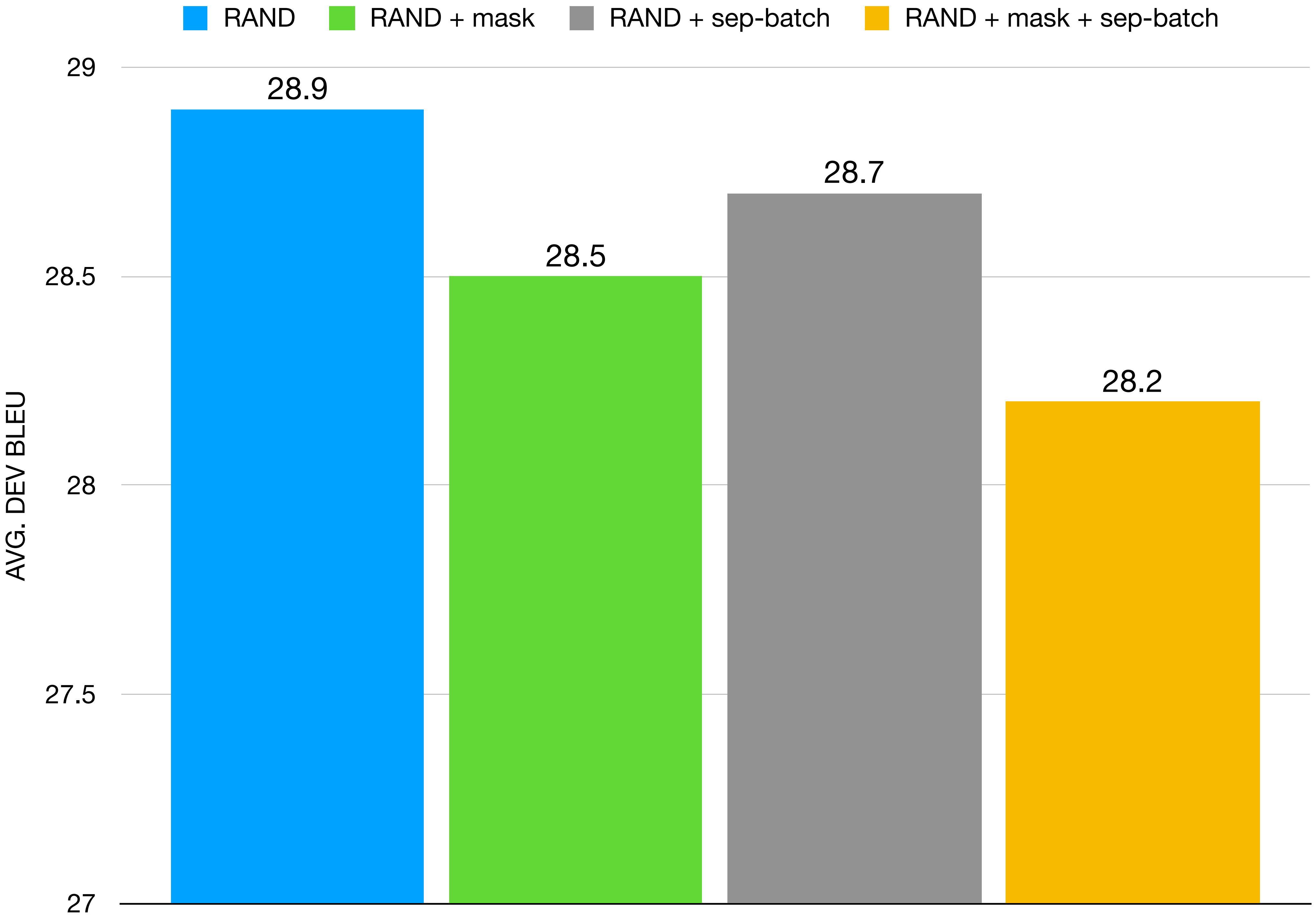
Putting together



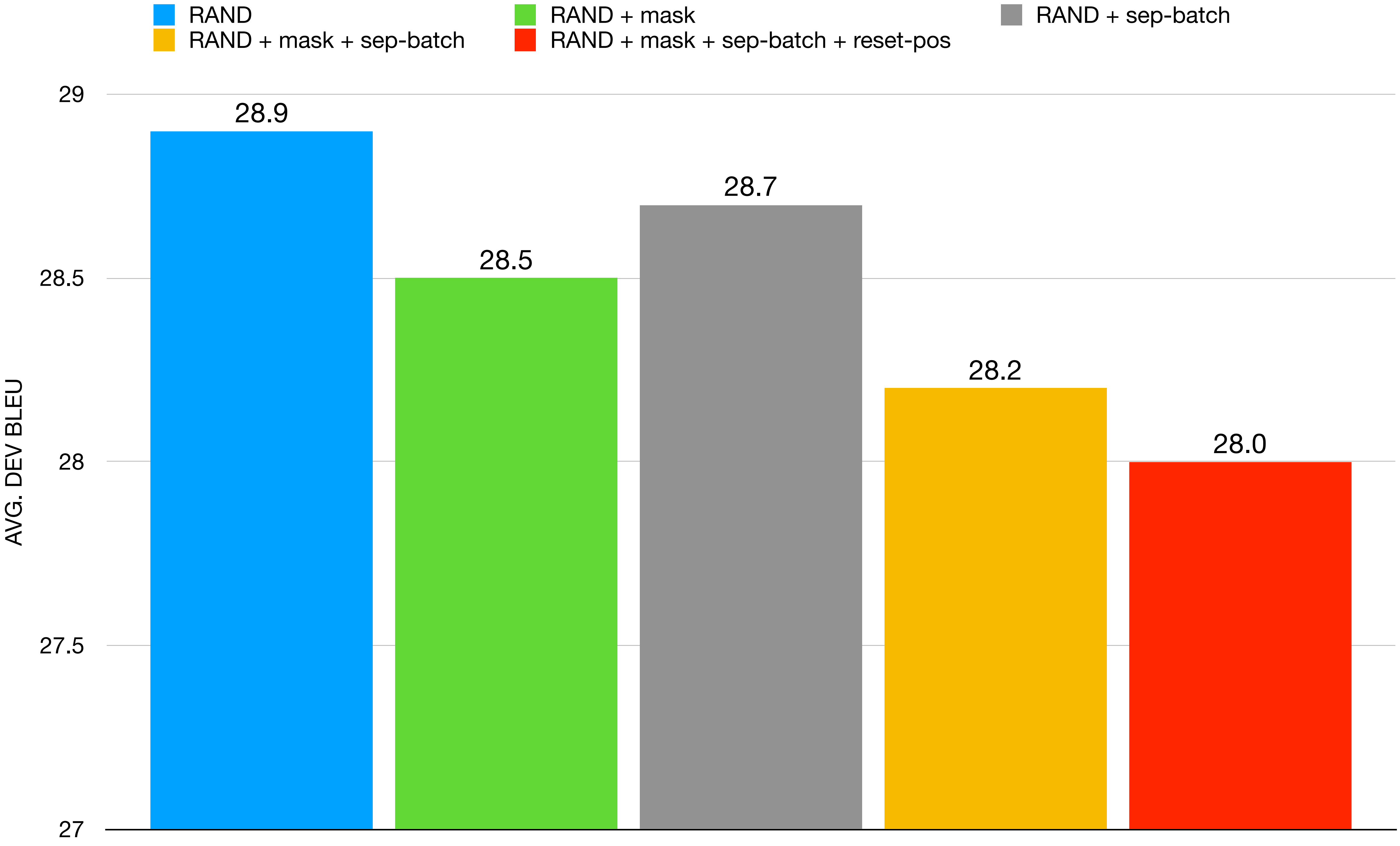
Putting together



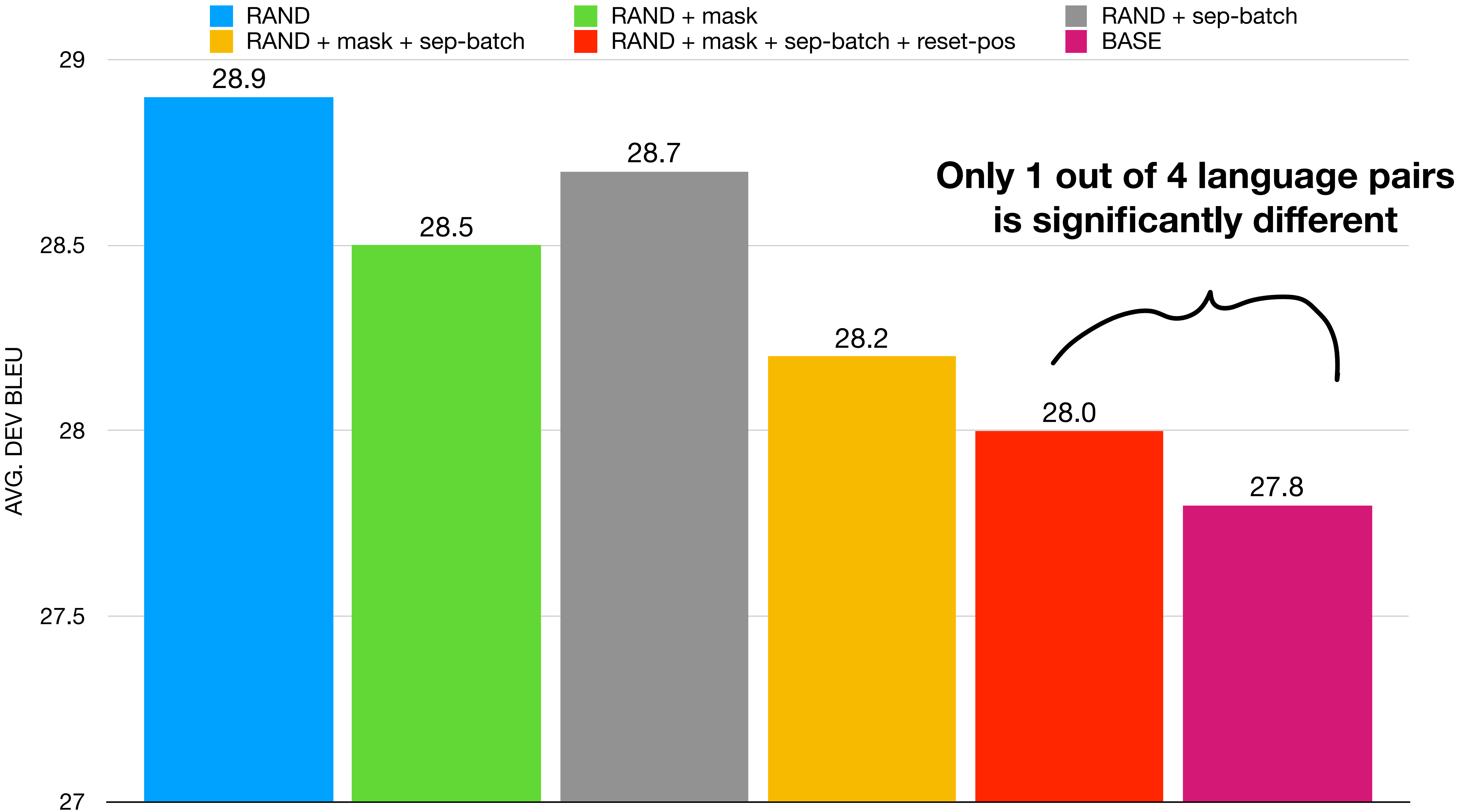
Putting together



Putting together



Putting together



Only 1 out of 4 language pairs is significantly different

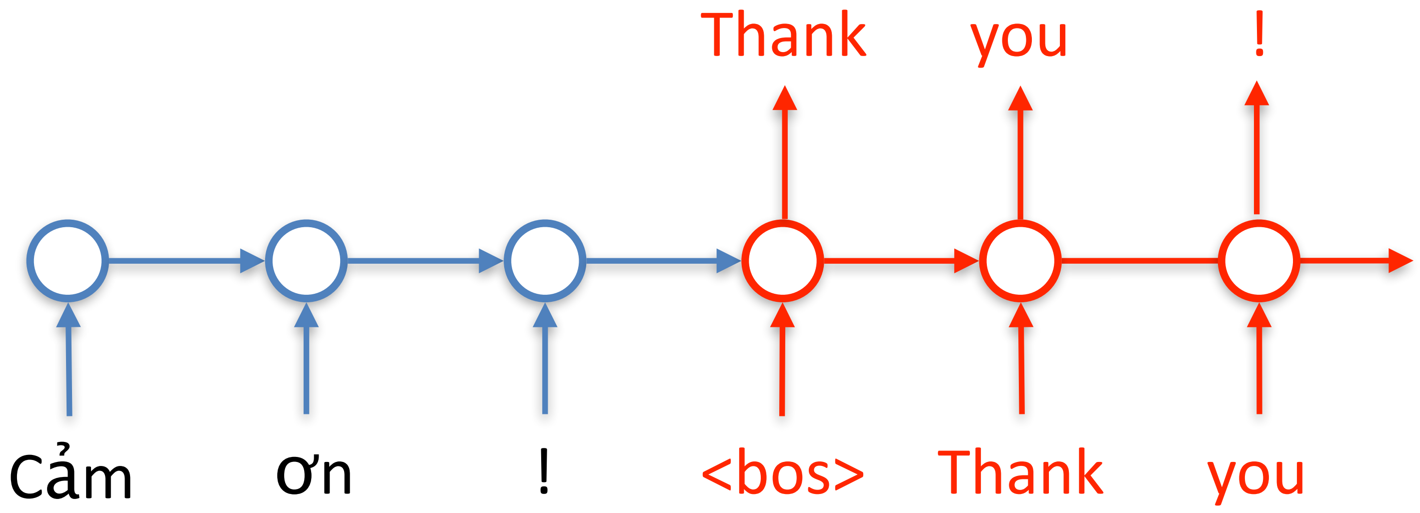
Conclusion

Concatenation is a simple yet effective & non-invasive data augmentation method for low-resource NMT

Its improvement **doesn't come from discourse context**

But from: ***position diversity***, ***context diversity***, and ***length diversity***

Details in our paper "Data Augmentation by Concatenation for Low-Resource Translation: A Mystery and a Solution" (Nguyen et al, 2021)


Thank
you
!
Cảm
ơn
!
<bos>
Thank
you