Neural Machine Translation with Semantically Relevant Image Regions

Yuting Zhao[1], Mamoru Komachi[1], Tomoyuki Kajiwara[2], Chenhui Chu[3]

1. Tokyo Metropolitan University
2. Ehime University
3. Kyoto University

Background

Why multimodal machine translation ?

-- Semantics still poorly used in MT systems.

- A woman sitting on a very large rock smiling at the camera with trees in the background.
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem **sehr großen Felsen** und lächelt in die Kamera.
  - Felsen == stone (uncountable)
- Eine Frau sitzt vor Bäumen im Hintergrund auf einem **sehr großen Stein** und lächelt in die Kamera.
  - Stein == rock (individual stone)

MT system can’t learn everything from text only.

Related Work

*[Calixto and Liu, 2017]

Global visual feature

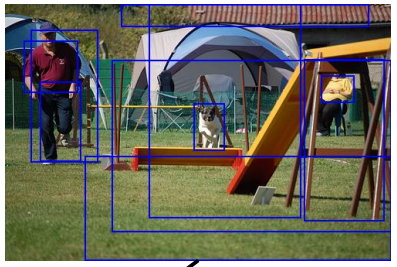
*[Calixto et al., 2017]

Conventional visual feature

OR

*[Zhao et al. , 2020]

Regional visual feature



man in red shirt watches dog on an agility course.

Source sentence

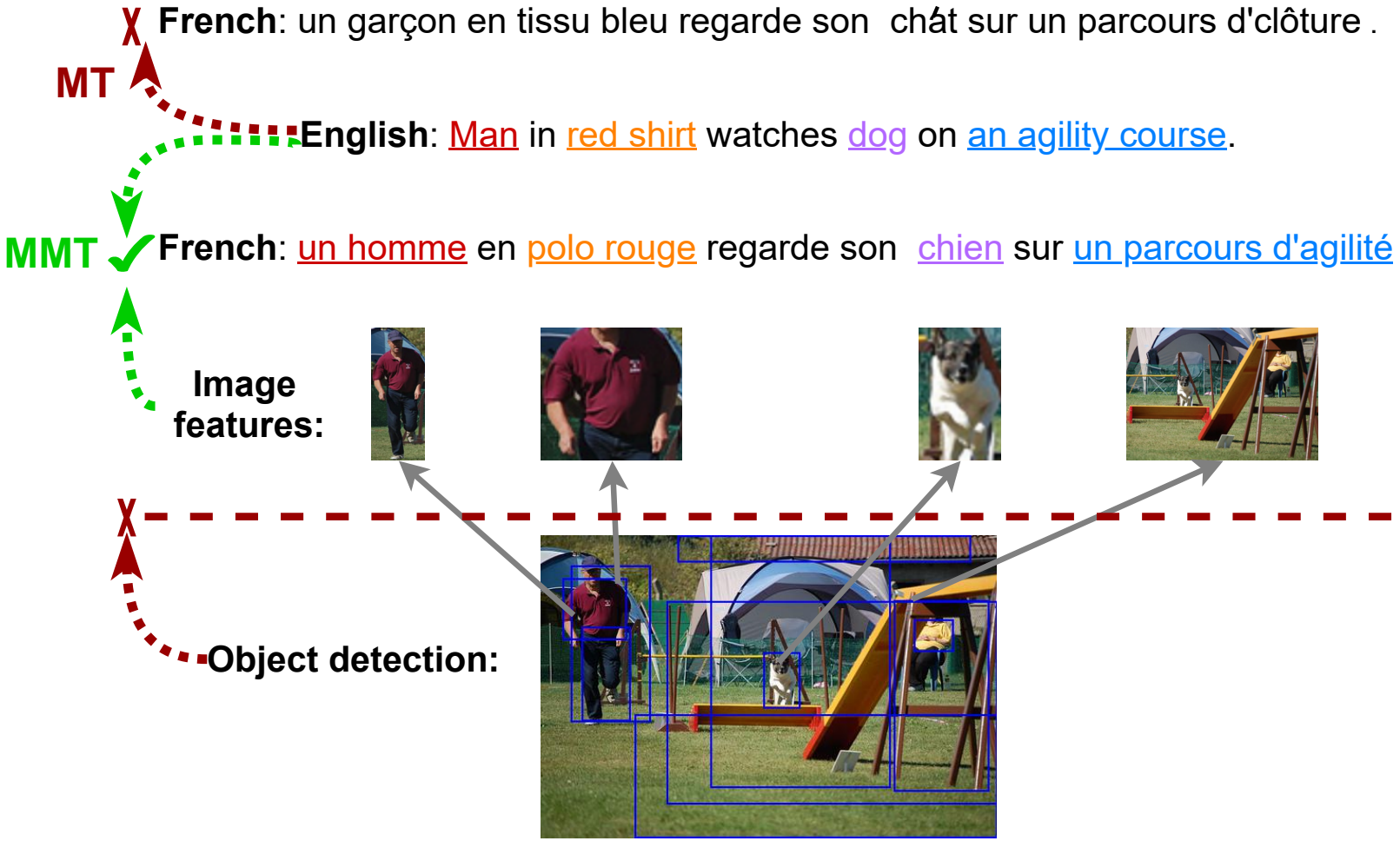
un homme en polo rouge regarde son chien sur un parcours d's agilité .

Target sentence

* [Calixto and Liu, 2017] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In EMNLP , pages 992–1003, 2017.
*[Calixto et al., 2017] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In ACL, pages 1913–1924, 2017.
*[Zhao et al. , 2020] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. Double attentionbased multimodal neural machine translation with semantic image regions. In EAMT , pages 105–114, 2020.

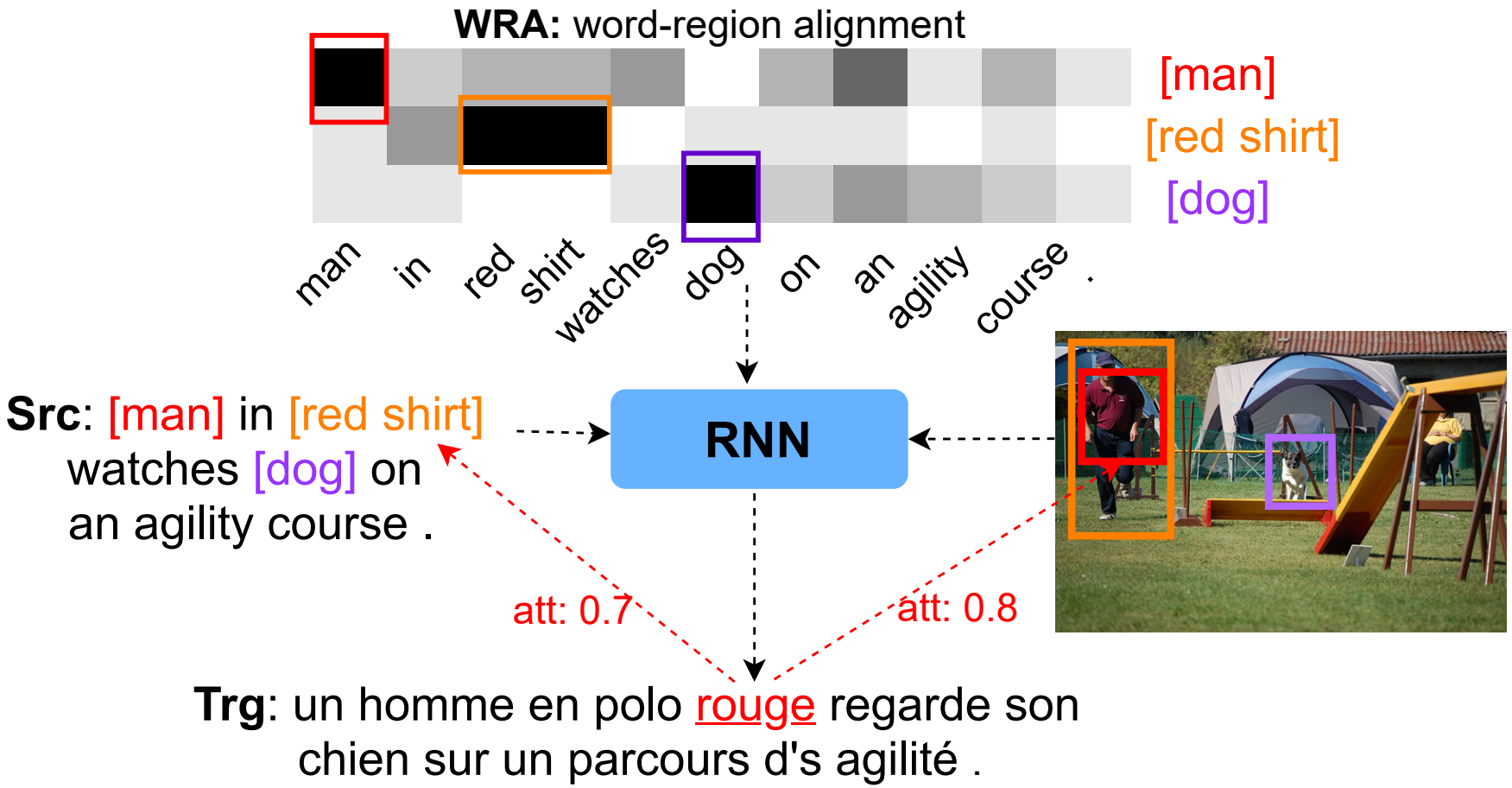
Motivation

How to focus on semantically relevant visual features?

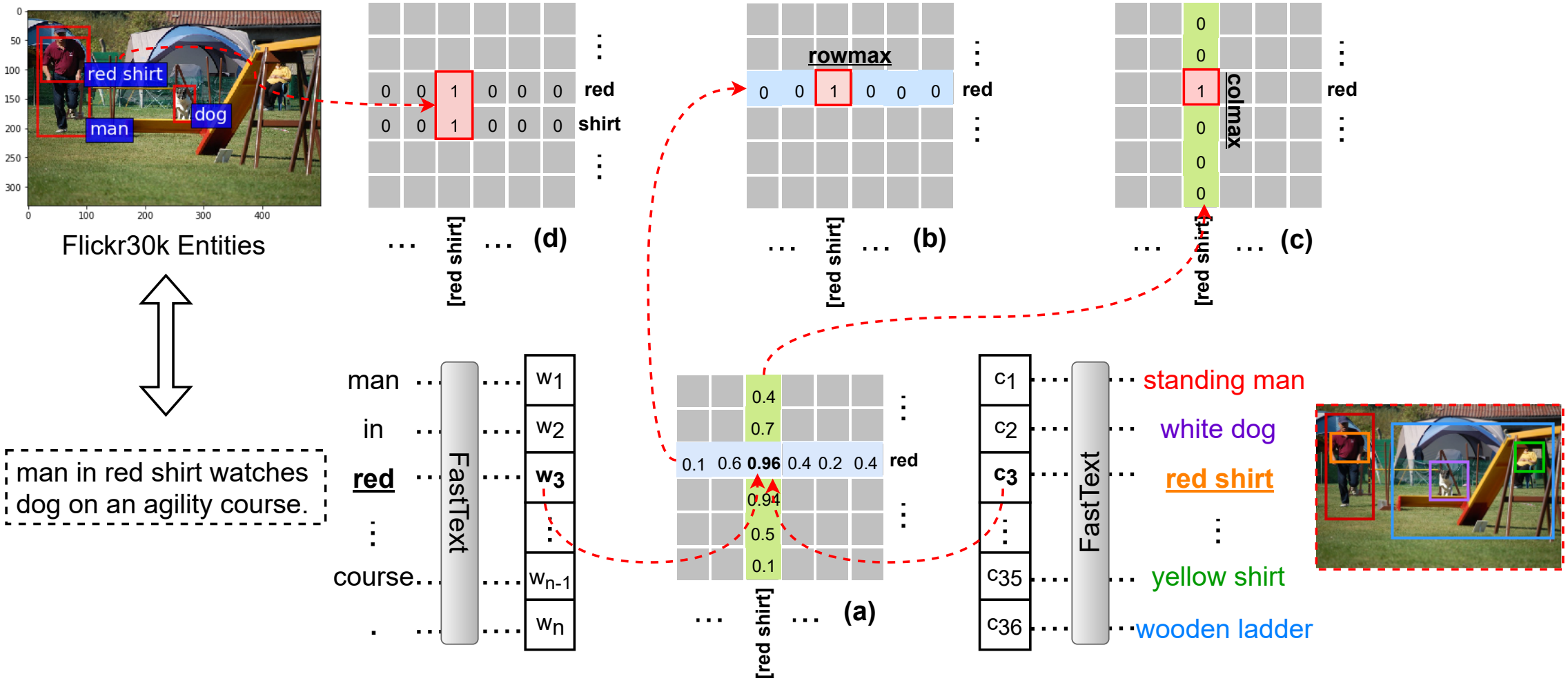


Overview

MNMT-WRA

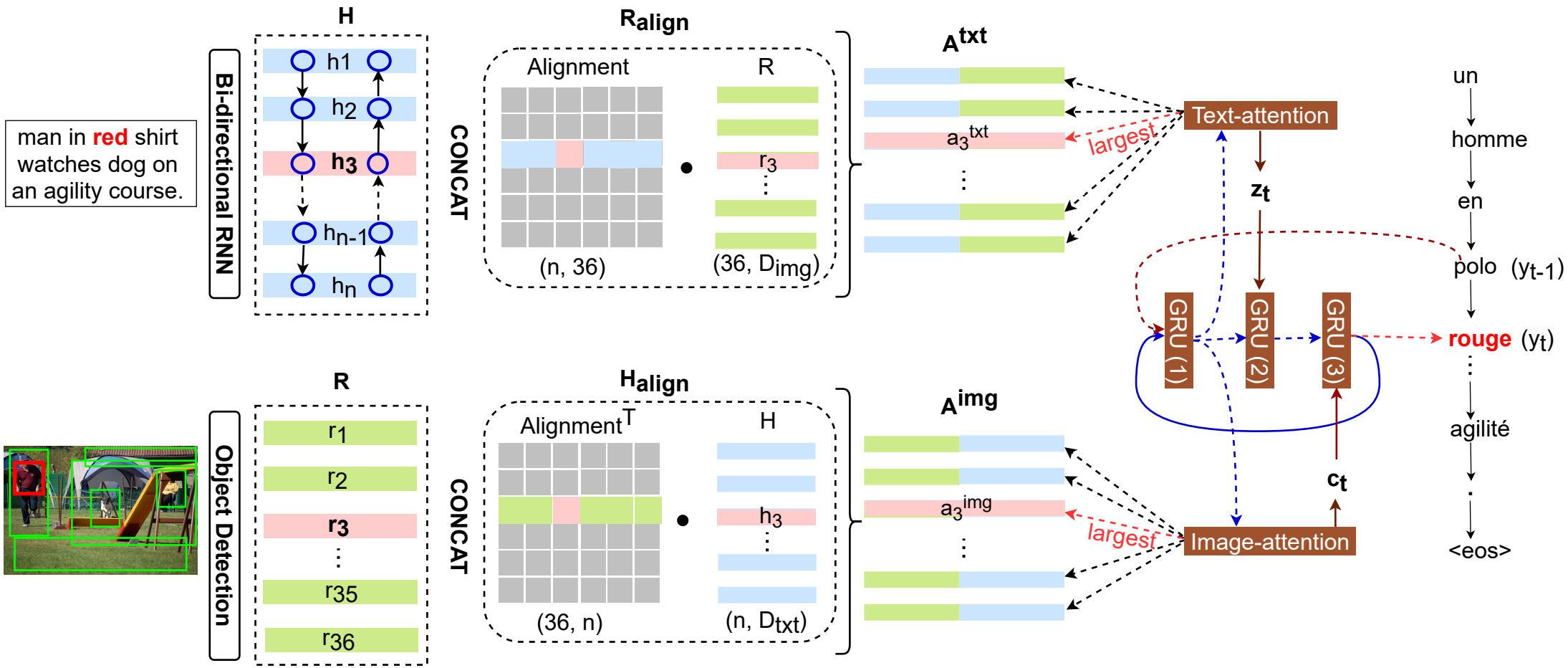


WRA Generation



The WRA. (a): Soft alignment (SA). (b/c): Hard alignment (HA). (d): Entity alignment (EA).

WRA Integration



The proposed MNMT-WRA.

Datasets

Multi30K + MSCOCO:

Train: 30k training images.

Dev:1,014 validation images.

Test:

- Test2016: 1,000 testing images.
- Test2017: 1,000 testing images.
- Mscoco:461 testing images.

Tasks:

English->German (En-De)

English->French (En-Fr)

Baselines

NMT: En-De and En-Fr textual part of Multi30k.

- a 2-layer bidirectional GRU encoder and 2-layer cGRU decoder.

MNMT: MNMT with global visual features extracted by ResNet-50.

- a 2-layer bidirectional GRU encoder and 3-layer deepGRU decoder.
- Visual feature dimension: 2048

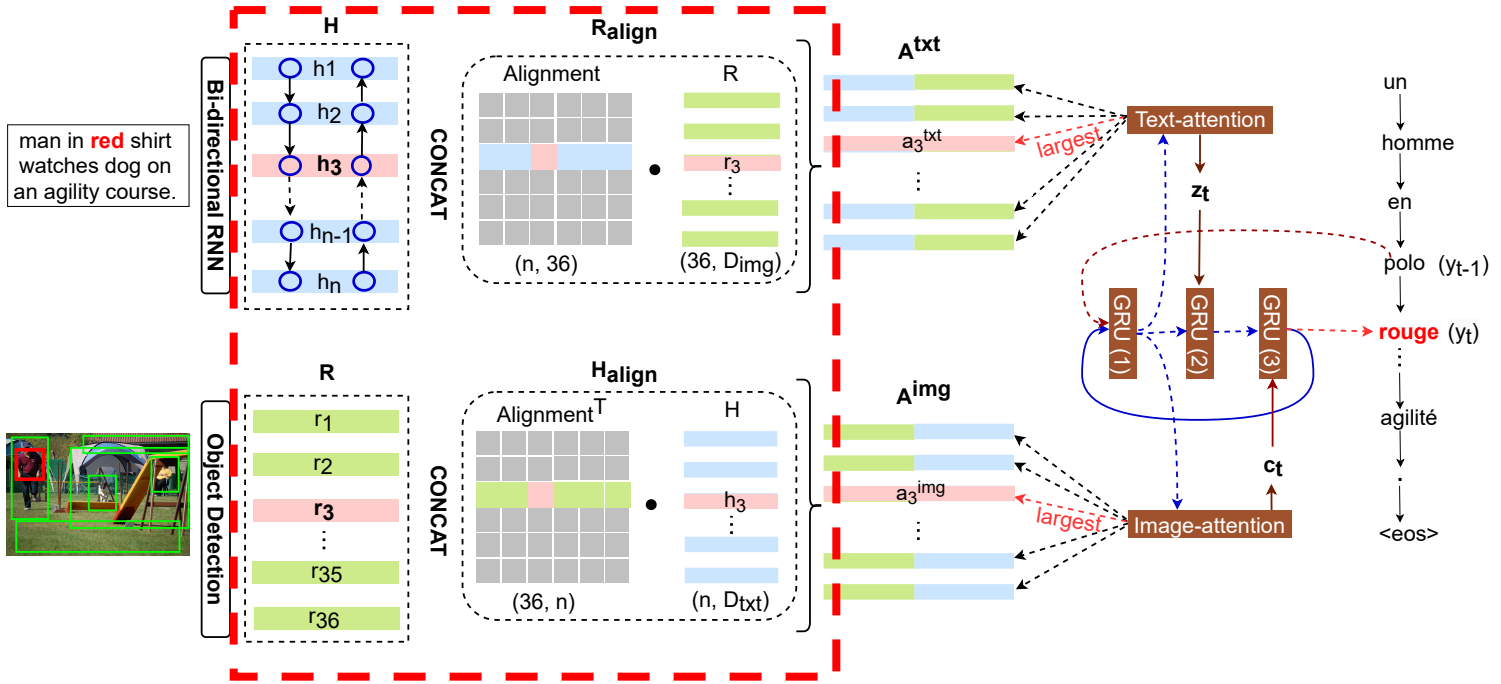
RAMNMT: Double attention-based MNMT with regional visual features extracted by Faster R-CNN.

- Regional feature amount: 36
- Visual feature dimension: 2048

Experimental Settings (1/3)

MNMT-WRA:

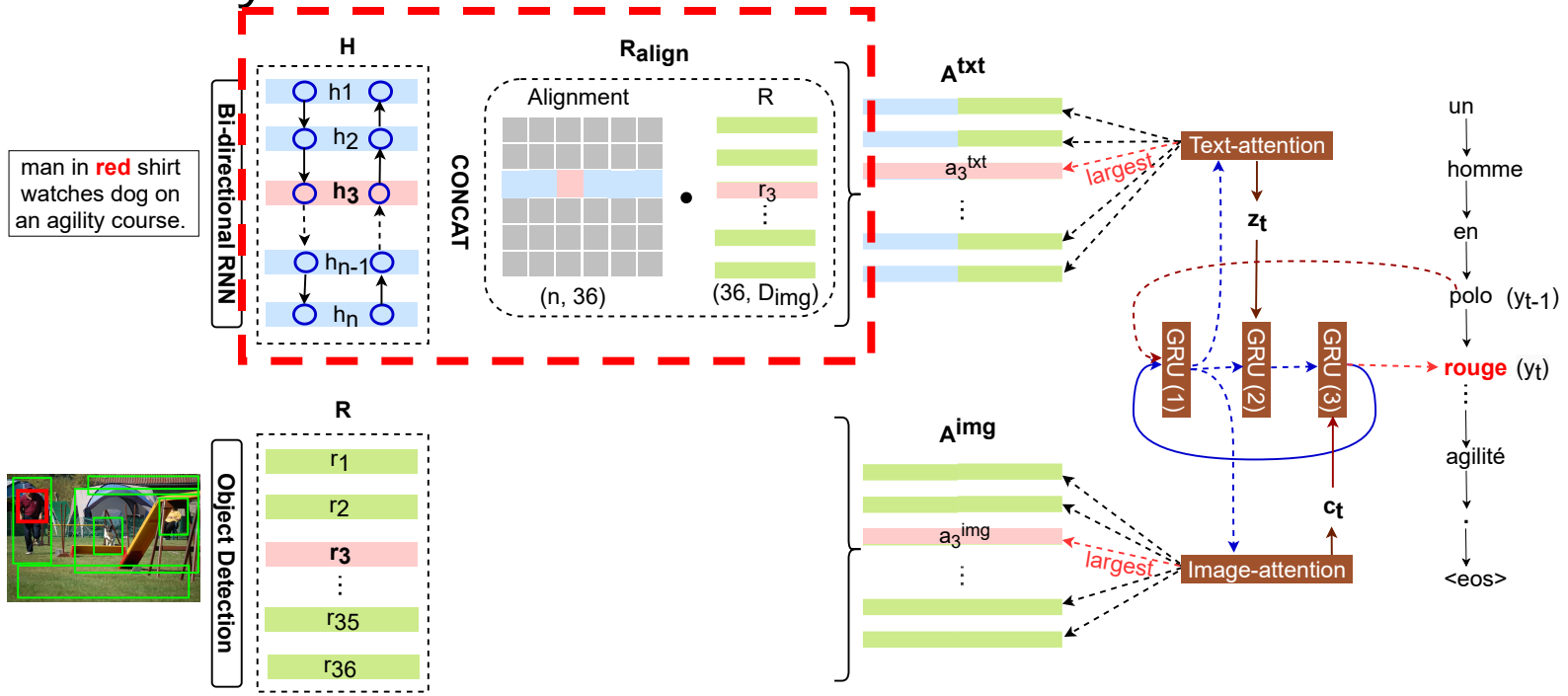
- **MNMT-WRA (C+HA/SA):** WRA (HA/SA) integrated in the text and image sides crossly.



Experimental Settings (2/3)

MNMT-WRA:

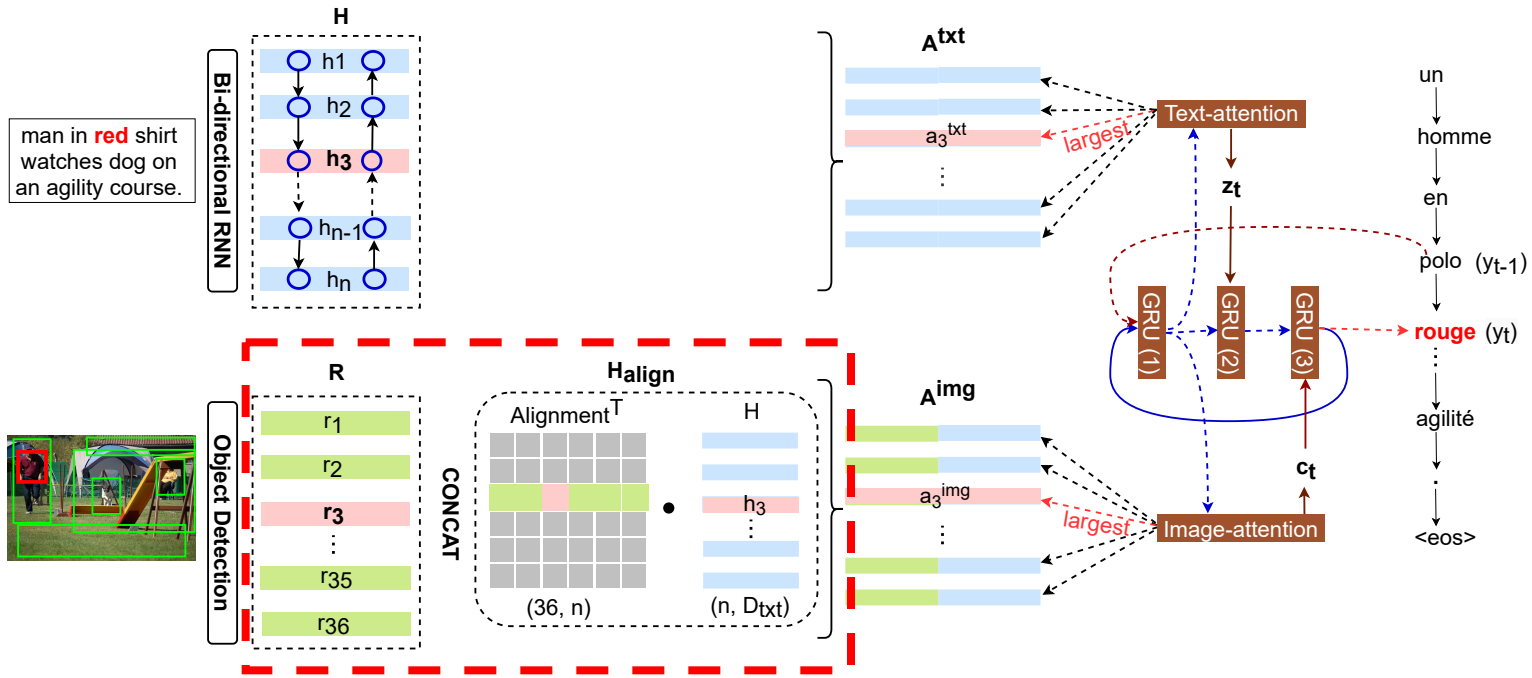
- **MNMT-WRA (T+HA/SA):** WRA (HA/SA) integrated in the text side only.



Experimental Settings (3/3)

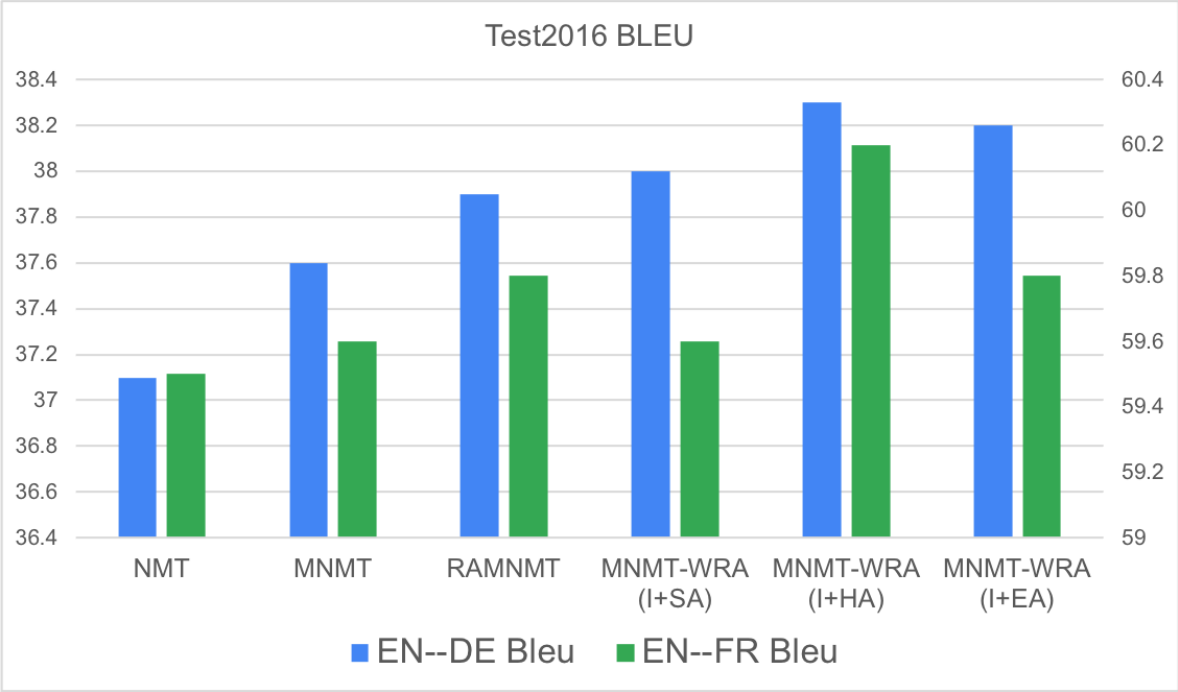
MNMT-WRA:

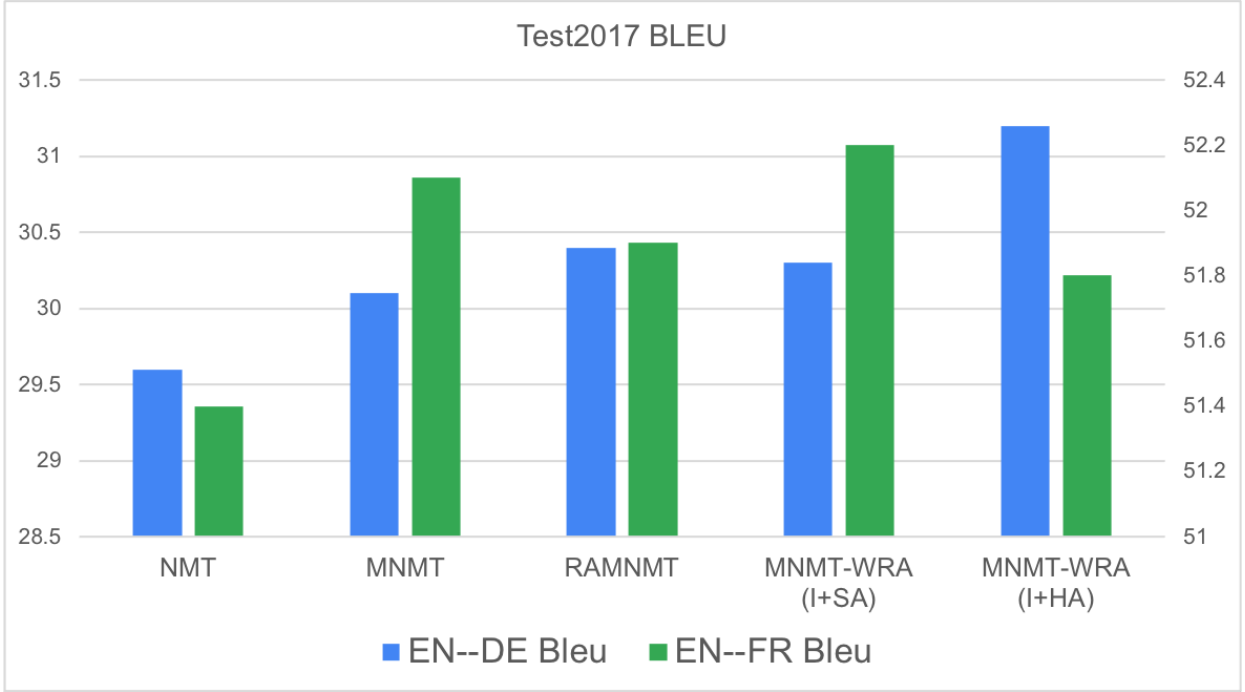
- **MNMT-WRA (I+HA/SA/EA):** WRA (HA/SA/EA) integrated in the image side only.



Results (1/2)

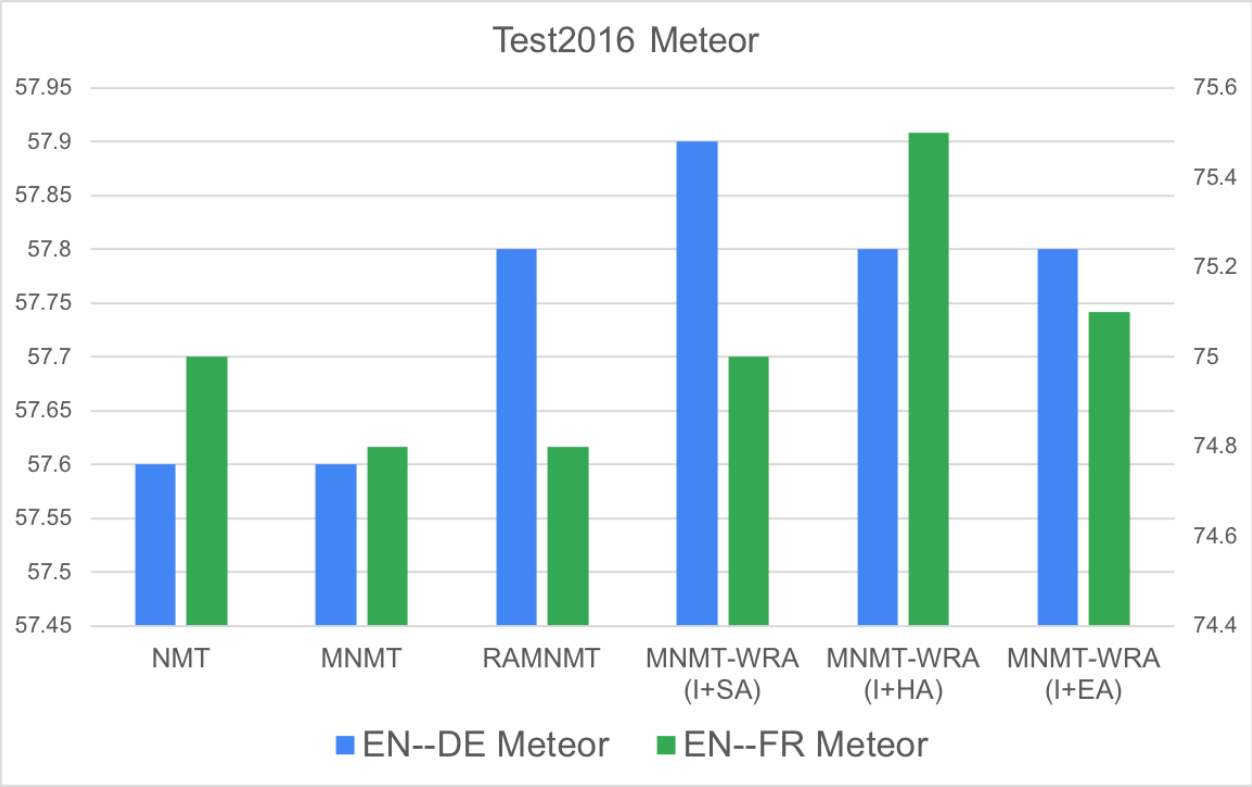
Evaluation: BLEU

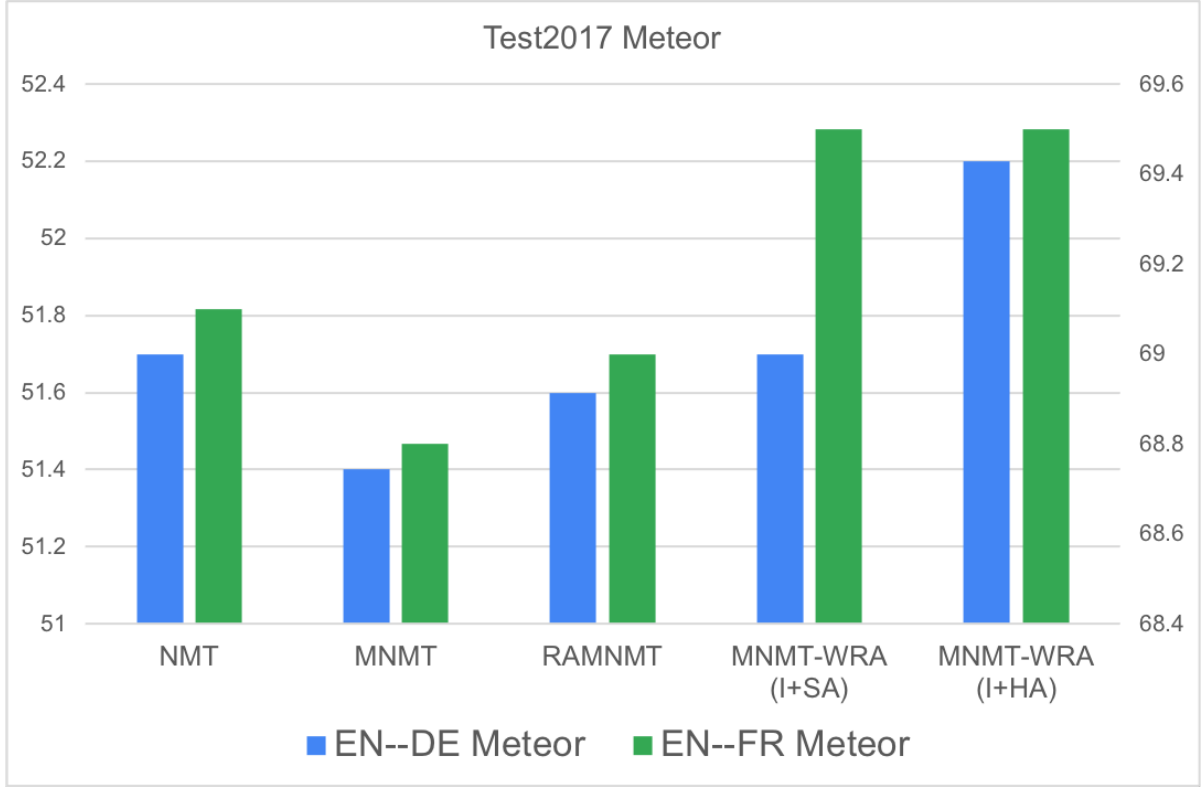




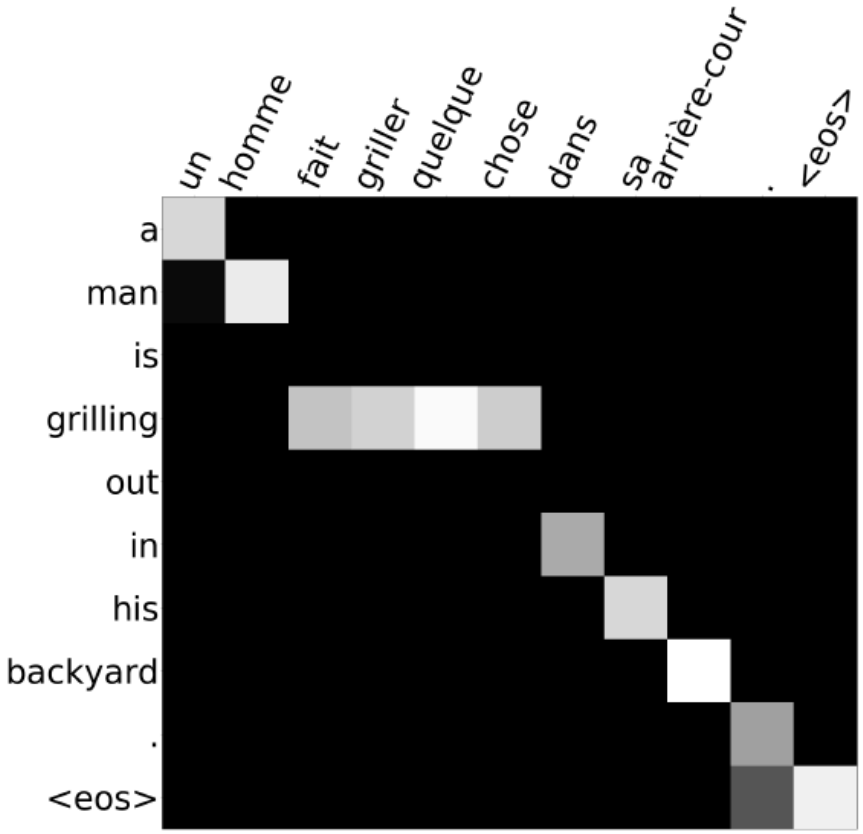
Results (2/2)

Evaluation: METEOR





Improved Example



English: a man is grilling out in his **backyard**.
Reference: un homme fait un barbecue dans son **arrière-cour**.
RAMNMT: un homme fait griller quelque chose dans sa **cour** (yard).
MNMT-WRA: un homme fait griller quelque chose dans sa **arrière-cour** (backyard).

Conclusion

(1) We propose WRA to guide the model to translate certain words based on certain image regions.

(2) The proposed MNMT-WRA model outperforms competitive baselines.

(3) The analysis demonstrates that MNMT-WRA utilizes visual information effectively.