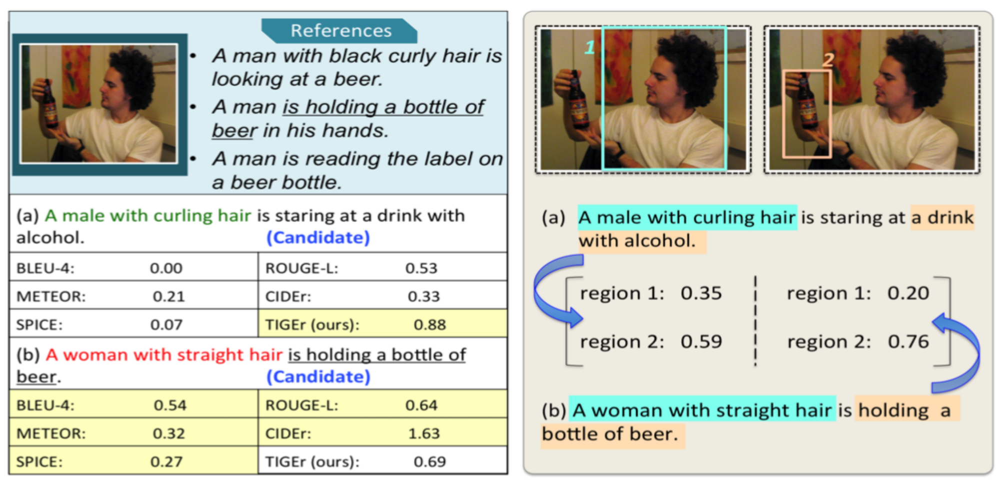
# TIGEr: Text-to-Image Grounding for Image Caption Evaluation

Ming Jiang[1], Qiuyuan Huang[2], Lei Zhang[2], Xin Wang[3], Pengchuan Zhang[2], Zhe Gan[2], Jana Diesner[1], Jianfeng Gao[2]

1. University of Illinois at Urbana-Champaign 2. Microsoft Research 3.University of California Santa Barbara

## Motivation and Contribution




- Metrics based on pure **text-level** comparison **loss image information** and face the challenge of **language ambiguity**.
- Propose a **novel automatic evaluation metric** called ***TIGEr***.
  - Consider both image content and human-generated references.
  - Measure the consistence with human attention distribution among image regions.

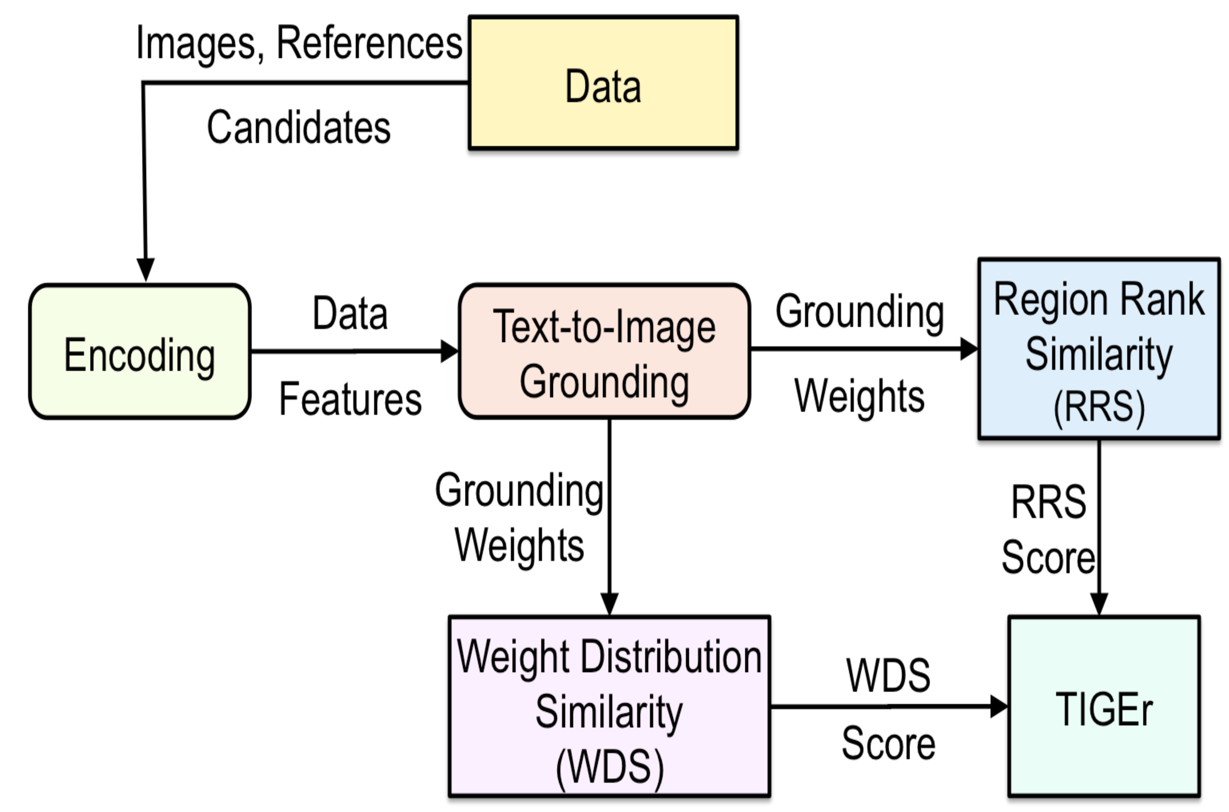
## TIGE Framework

**Data Encoding**
- Region-level & Word-level embedding vectors

**Text-to-Image Grounding**
- Grounding a caption into each image region.

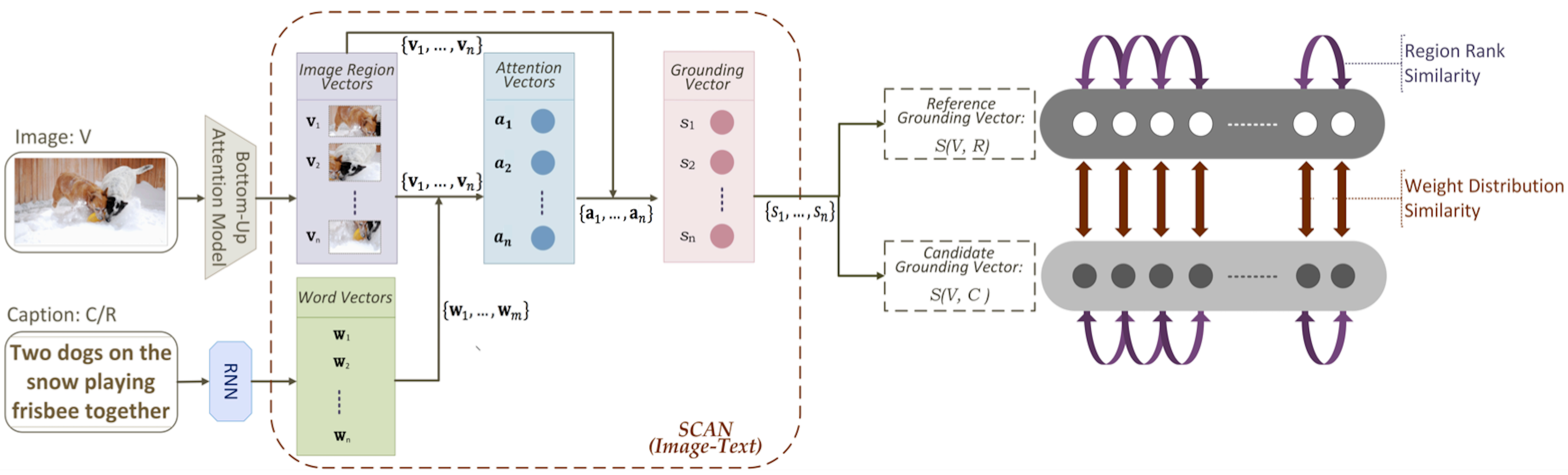
**[(Reference vs. Candidate) | Image]**
- RRS: how similar is the order of image regions based on grounding weights?
- WDS: how similar is the attention distributed by a caption among image regions?

**TIGEr**
- Average value of RRS and WDS




## TIGEr Workflow

- Encoding images and texts by a pre-trained Bottom-Up Attention and a RNN model.
- Grounding texts and images by a pre-trained SCAN model.
- Calculating RRS based on Normalized Discounted Cumulative Gain (NDCG).
- Measuring WDS based on KL Divergence.




## Metric Performance

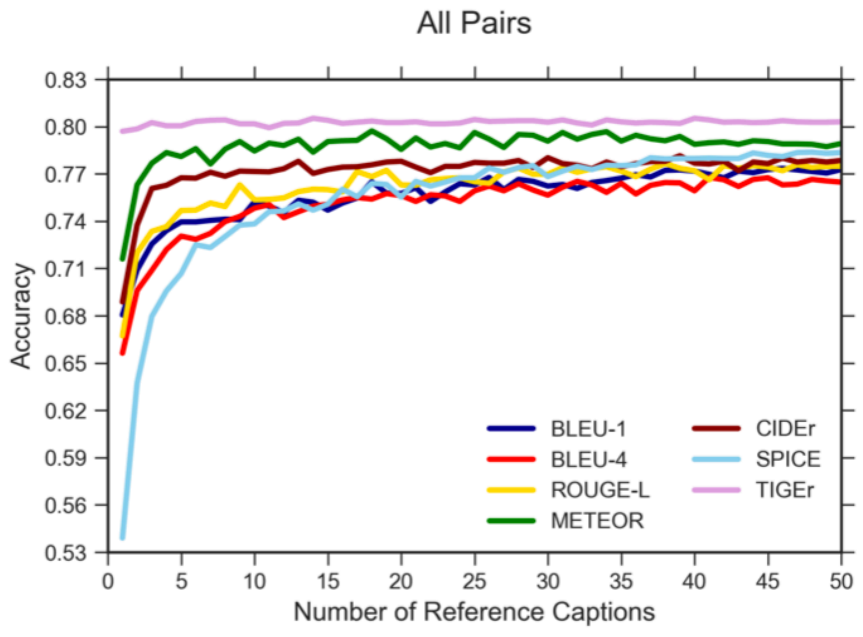
- TIGEr achieved a noticeable improvement in the assessment of caption quality on three benchmark datasets.
- Identifying irrelevant human-written captions in HI is relatively easy for all metrics, while judging the quality of two correct human-annotated captions in HC is more difficult than other comparison groups.
- Given the change of reference sizes, TIGEr achieves a higher judgment accuracy and more stable performance.

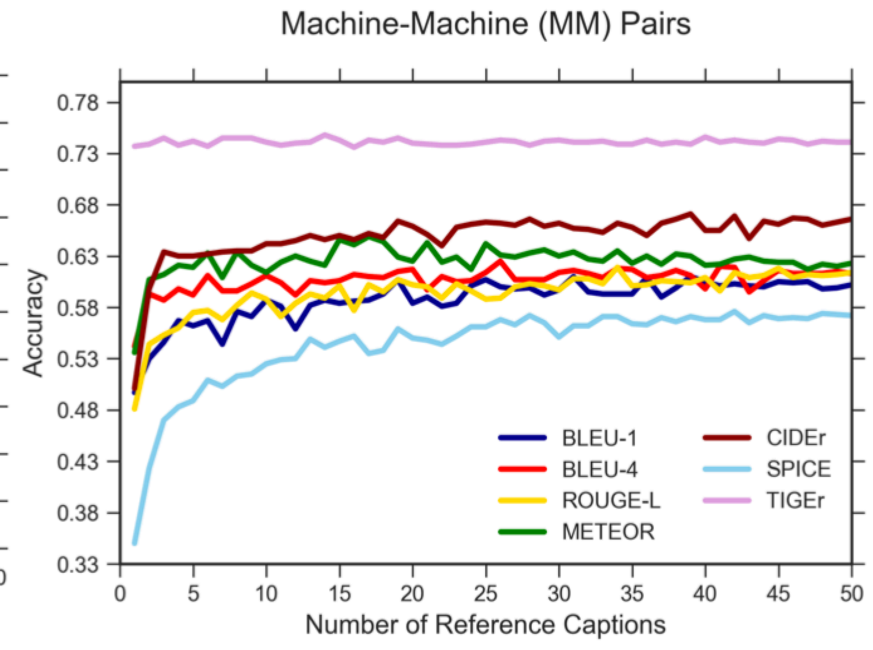
| | Composite | | Flickr8k | |
|---|---|---|---|---|
| | $\tau$ | $\rho$ | $\tau$ | $\rho$ |
| BLEU-1 | 0.280 | 0.353 | 0.323 | 0.404 |
| BLEU-4 | 0.205 | 0.352 | 0.138 | 0.387 |
| ROUGE-L | 0.307 | 0.383 | 0.323 | 0.404 |
| METEOR | 0.379 | 0.469 | 0.418 | 0.519 |
| CIDEr | 0.378 | 0.472 | 0.439 | 0.542 |
| SPICE | 0.419 | 0.514 | 0.449 | 0.596 |
| **Ours** | | | | |
| RRS | 0.388 | 0.479 | 0.418 | 0.521 |
| WDS | 0.433 | 0.526 | 0.464 | 0.572 |
| TIGEr | **0.454** | **0.553** | **0.493** | **0.606** |

*Caption-level correlation between metrics and human grading scores in Composite and Flickr 8K dataset by using Kendall tau and Spearman rho. All p-values <0.01.*

| | HC | HI | HM | MM | All |
|---|---|---|---|---|---|
| BLEU-1 | 51.20 | 95.70 | 91.20 | 58.20 | 74.08 |
| BLEU-4 | 53.00 | 92.40 | 86.70 | 59.40 | 72.88 |
| ROUGE-L | 51.50 | 94.50 | 92.50 | 57.70 | 74.05 |
| METEOR | **56.70** | 97.60 | **94.20** | 63.40 | 77.98 |
| CIDEr | 53.00 | 98.00 | 91.50 | 64.50 | 76.75 |
| SPICE | 52.60 | 93.90 | 83.60 | 48.10 | 69.55 |
| TIGEr (ours) | 56.00 | **99.80** | 92.80 | **74.20** | **80.70** |

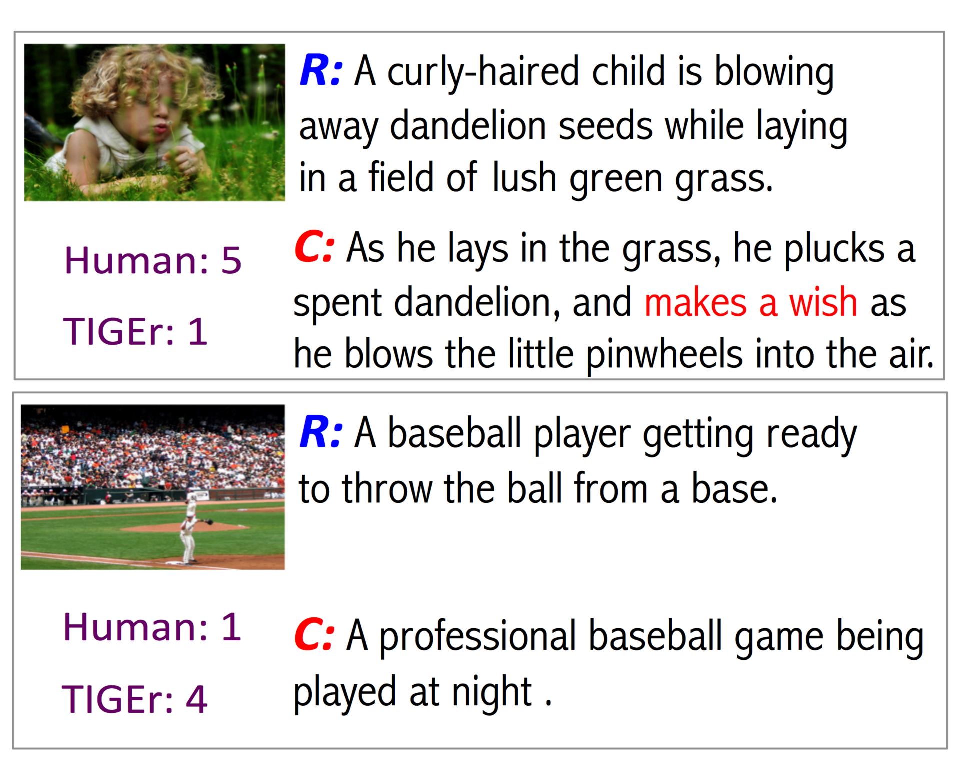
*Accuracy of metrics at matching human judgments on PASCAL-50S with 5 reference captions. The highest accuracy per pair type is shown in bold font. HC: human-human correct, HI: human-human incorrect, HM: human-machine, MM: machine-machine, ALL: all pairs.*



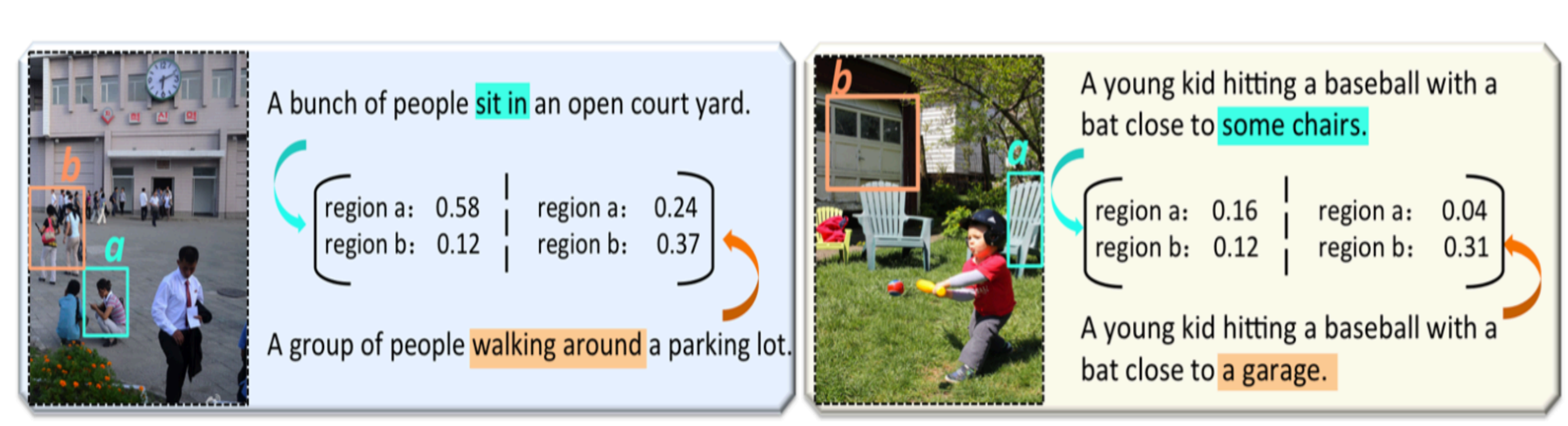




## Analysis

- Image region has a higher grounding weight with the corresponding caption than other unrelated regions.
- Text-to-image grounding is more challengeable at action-level compared to object-level.
- Reference captions may not fully cover visual information and TIGEr can measure a caption quality by considering the semantic information of image contents.
- Human interpretation inspired by the image is hard to be judged by an automatic evaluation metric.








## Related Resource

**REO-Relevance, Extraness, Omission: A Fine-grained Evaluation for Image Captioning. In *EMNLP-IJCNLP'19*.**
- A fine-grained evaluation on description adequacy
- Candidate vs. Image or (Image + References)

**Github Link:**
- https://github.com/SeleenaJM/CapEval