

# VisualCOMET: Reasoning about the Dynamic Context of a Still Image

Jae Sung Park, Chandra Bhagavatula,  
Roosbeh Mottaghi, Ali Farhadi, Yejin Choi  
**ECCV 2020'**

Paul G. Allen School of Computer Science & Engineering, WA, USA  
Allen Institute for Artificial Intelligence, WA, USA

Presented by Dong Hui Im

ehdgnl101@korea.ac.kr

Data Intelligence Laboratory, Korea University

15th January 2021

How to learn **Dynamic story**  
underlying the visual scene?

# Previous work

---

## [Image captioning]

<Microsoft COCO Captions: Data Collection and Evaluation Server>



The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.



A horse carrying a large load of hay and two people sitting on it.

Automatic generation of captions for images.

# Previous work

---

## [Visual question answering]

<VQA: Visual Question Answering>



What color are her eyes?  
What is the mustache made of?



How many slices of pizza are there?  
Is this a vegetarian pizza?



Does it appear to be rainy?  
Does this person have 20/20 vision?

Given an image and a natural language question about the image,  
the task is to provide an accurate natural language answer.

# Previous work

## [Referring expressions]

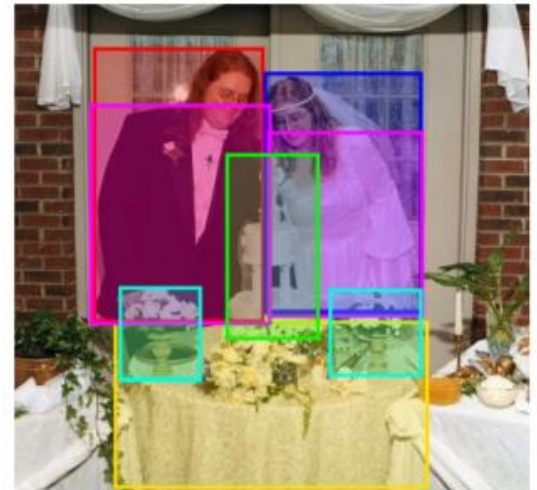
<Flickr30k Entities : Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models>



A man with pierced ears is wearing glasses and an orange hat.  
A man with glasses is wearing a beer can crotched hat.  
A man with gauges and glasses is wearing a Blitz hat.  
A man in an orange hat starring at something.  
A man wears an orange hat and glasses.



During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.  
A group of youths march down a street waving flags showing a color spectrum.  
Oriental people with rainbow flags walking down a city street.  
A group of people walk down a street waving rainbow flags.  
People are outside waving flags .



A couple in their wedding attire stand behind a table with a wedding cake and flowers.  
A bride and groom are standing in front of their wedding cake at their reception.  
A bride and groom smile as they view their wedding cake at a reception.  
A couple stands behind their wedding cake.  
Man and woman cutting wedding cake.

The content selection part determines which set of properties distinguish the intended target and the linguistic realization part defines how these properties are translated into natural language



## Related work

## <From Recognition to Cognition: Visual Commonsense Reasoning>



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.  
b) He just told a joke.  
c) He is feeling accusatory towards [person1].  
d) He is giving [person1] directions.

I chose **a)** because...






- a) [person1] has the pancakes in front of him.  
b) [person4] is taking everyone's order and asked for clarification.  
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.  
d) [person3] is delivering food to the table, and she might not know whose order is whose.



How did [person2] get the money that's in front of her?

- a) [person2] is selling things on the street.  
b) [person2] earned this money playing music.  
c) She may work jobs for the mafia.  
d) She won money playing poker.

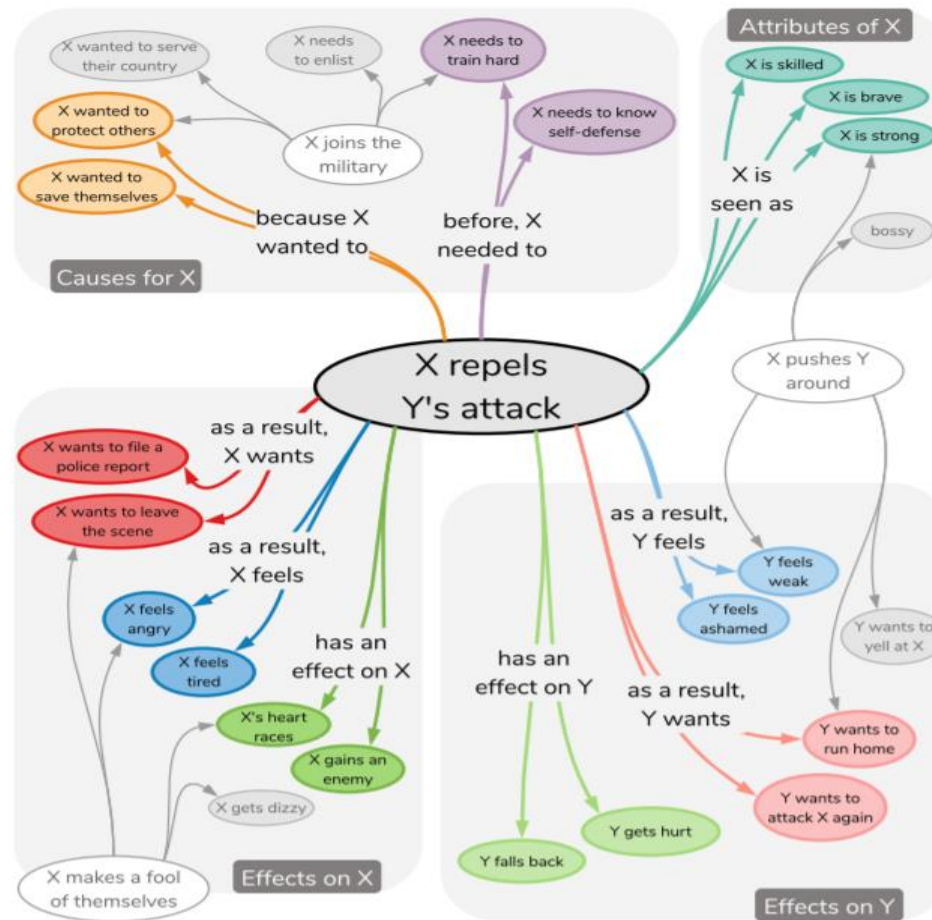
I chose **b)**  
because...

- a) She is playing guitar for money.  
b) [person2 ] is a professional musician in an orchestra.  
c) [person2 ] and [person1 ] are both holding instruments, and were probably busking for that money.  
d) [person1 ] is putting money in [person2 ]'s tip jar, while she plays music.

For cognition level, adopting the reasoning in the V+L task.

# Related work

<ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning>



Natural language commonsense graph using logical structure.

# Motivation

---



People can reason about the rich dynamic story underlying the visual scene.



# Contribution

---

1. They introduce a **new task of visual commonsense reasoning** for cognitive visual scene understanding, to reason about events before and after and people's intents at present.
2. They present the first large-scale **repository of Visual Commonsense Graphs** that contains more than 1M textual descriptions of commonsense inferences over 60K complex visual scenes.

# Contribution

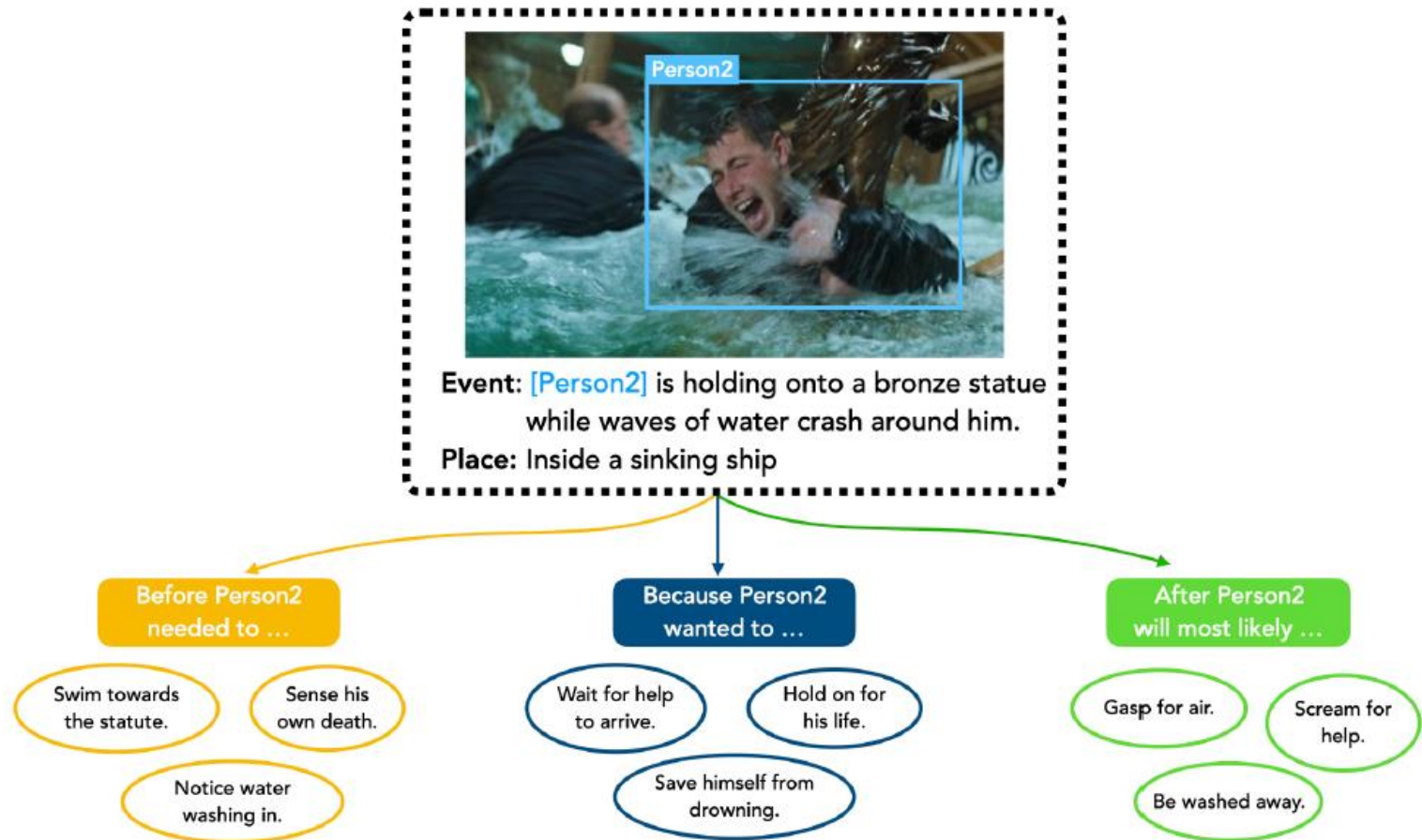
---

3. They extend the GPT-2 model to incorporate visual information and allow direct supervision for grounding people in images.

4. Their empirical results and human evaluations show that model trained **jointly with visual and textual** cues **outperform models with single modality** and can generate meaningful inferences from still images.

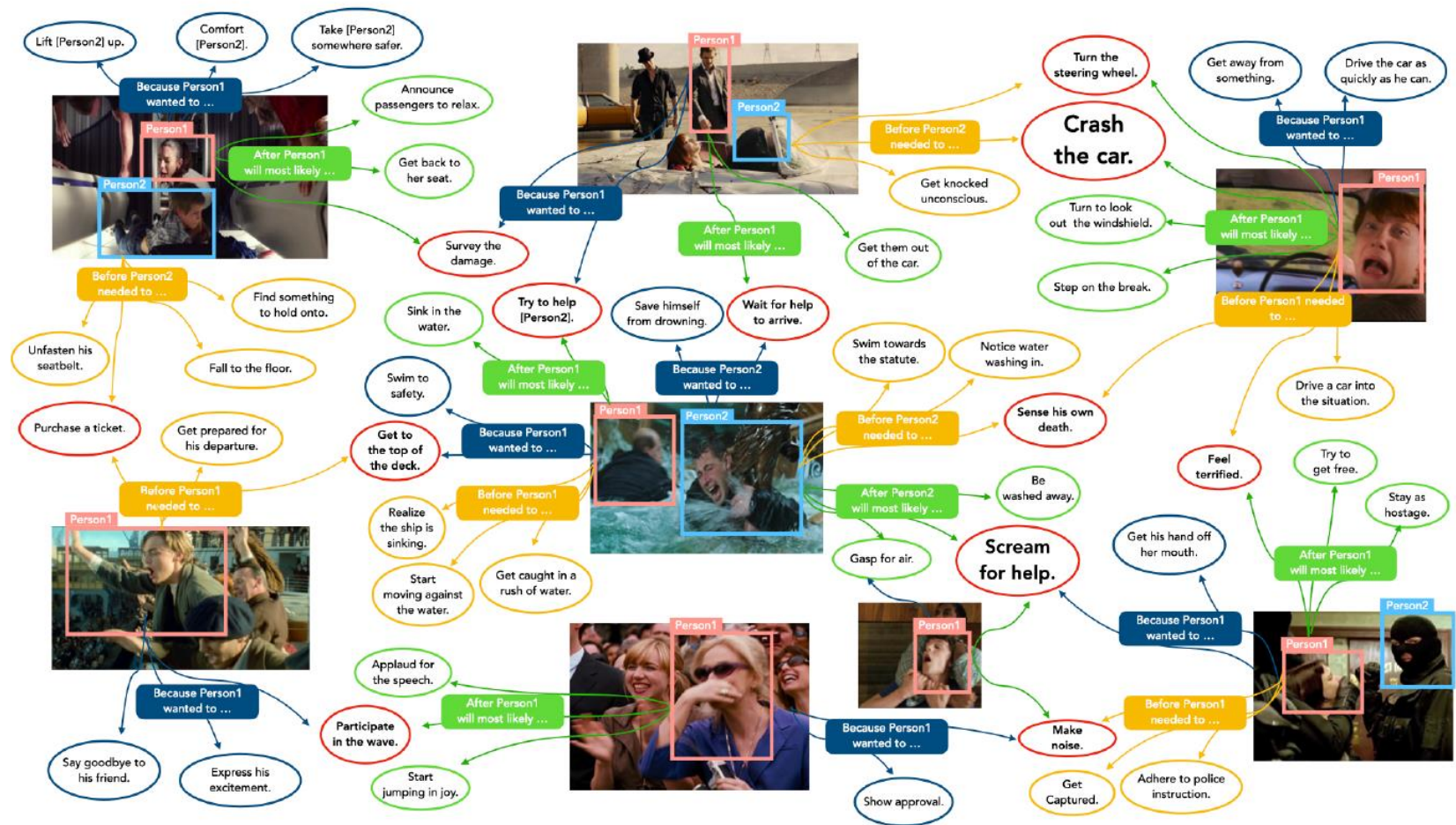
# Datasets

Publicly available at <https://visualcomet.xyz/dataset>



The data consists of event and intent at present, event before and after, and place.

# Datasets

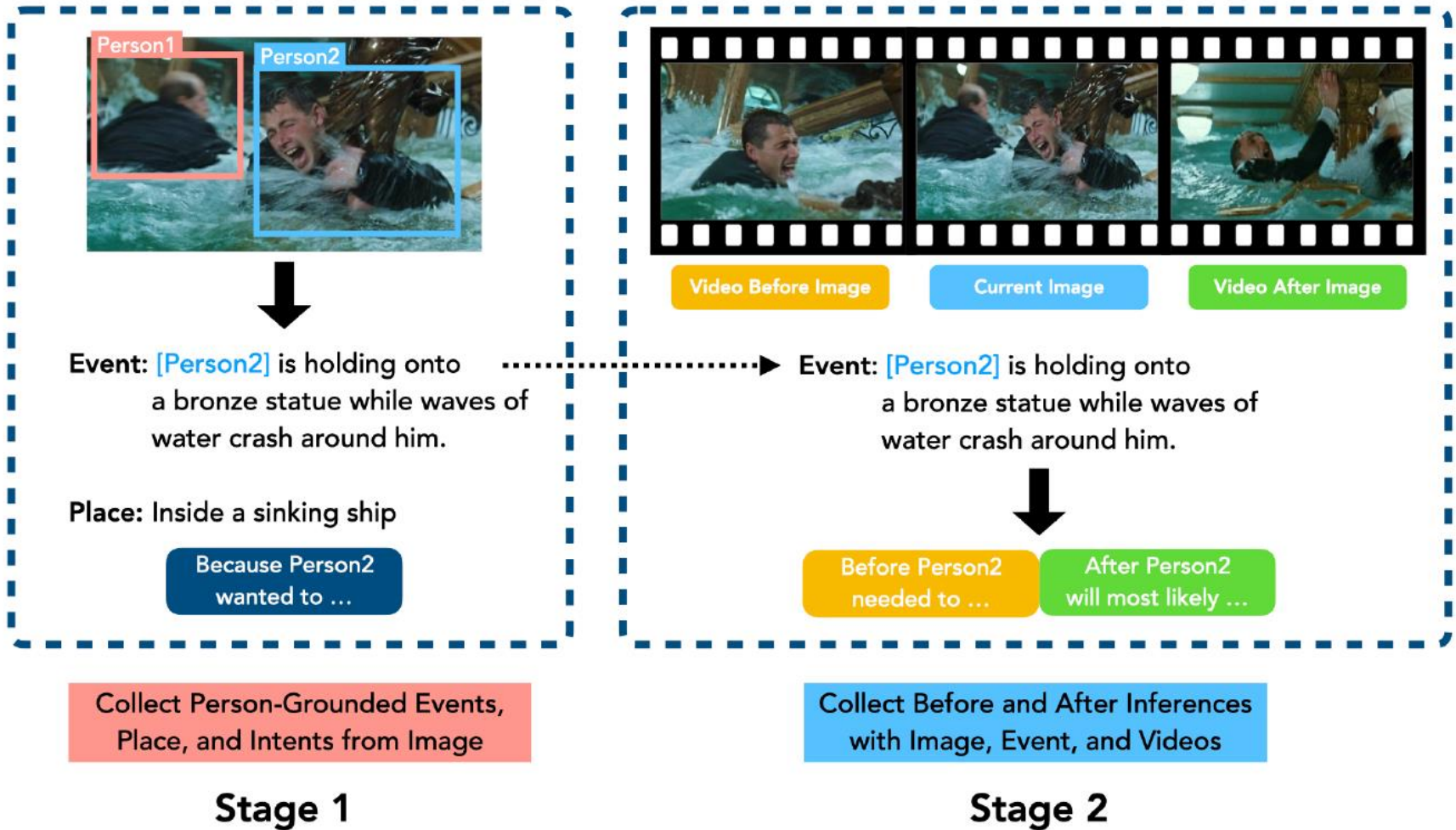


59,356 Image/Place  
139,377 Event at present

584,211 Inference on Event before  
586,418 Inference on Event After  
295,080 Inference on Intents at present

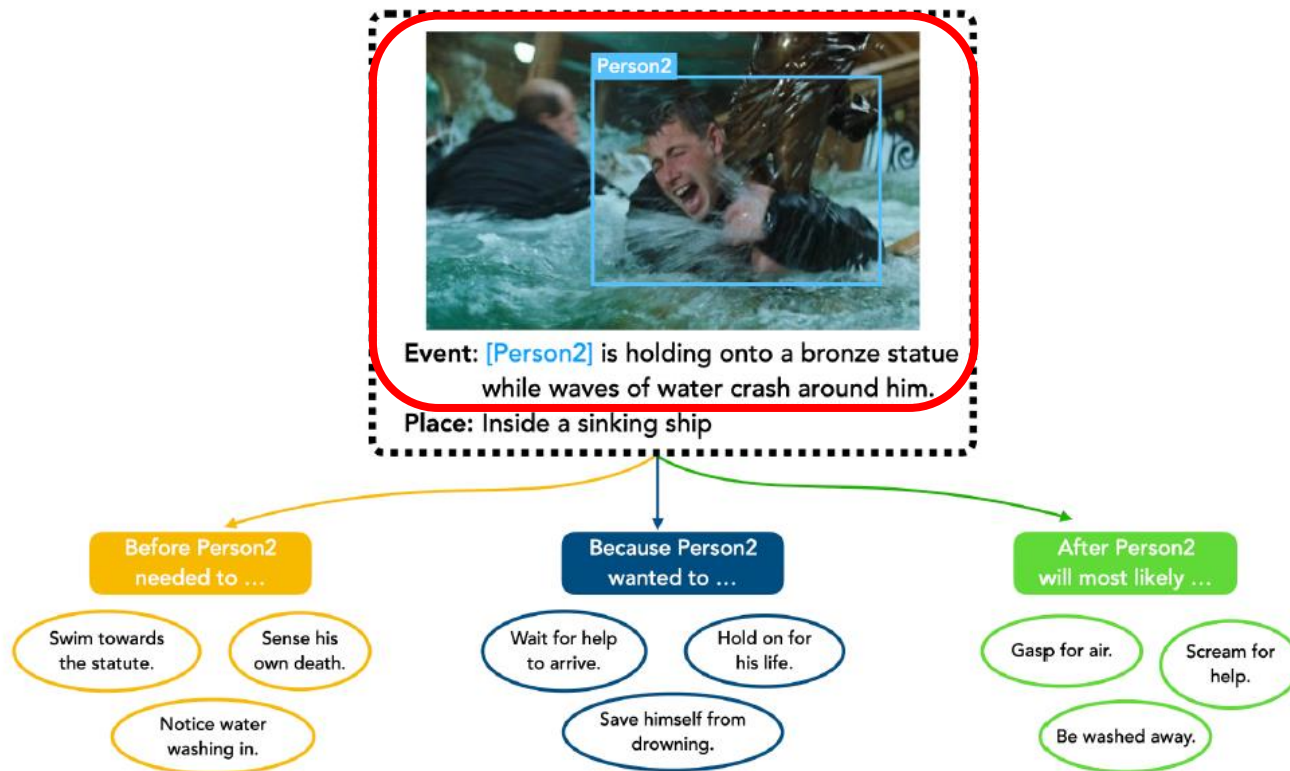


# Datasets



# Task

1. Given an image and one of the events at present and generate the rest of visual commonsense graph that is connected to the specific current event.



# Task

2. Given an image, the task is to generate the complete set of commonsense inferences from scratch.

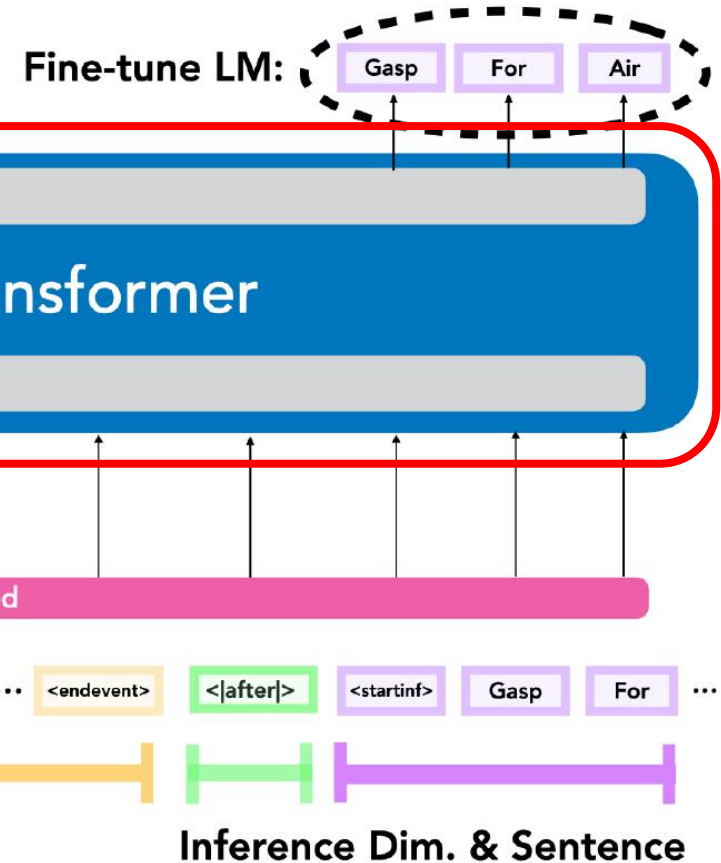


**Q & A**



# Model

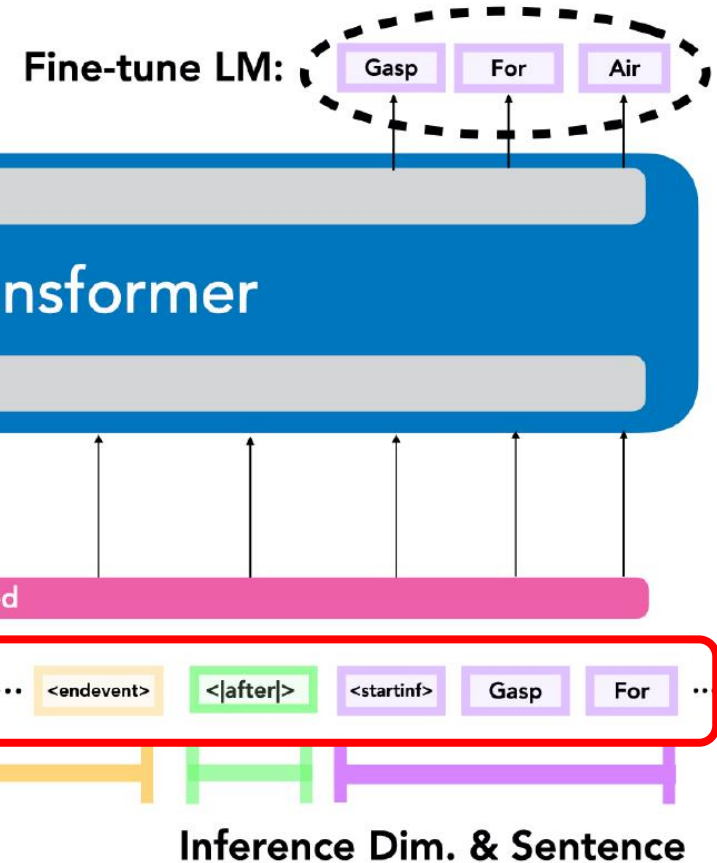
## [Overview]



Using GPT-2 single stream transformer model which has been shown to be more effective in vision and language task.

# Model

## [Overview]



Additional special tokens to indicating  
the start and end of image, event, place, and inference type

# Model

---

[EP Loss]

$$\mathcal{L} = - \sum_{i=1}^l \log P(w_{hi}^r | w_{h<i}^r, r, e, p, v)$$

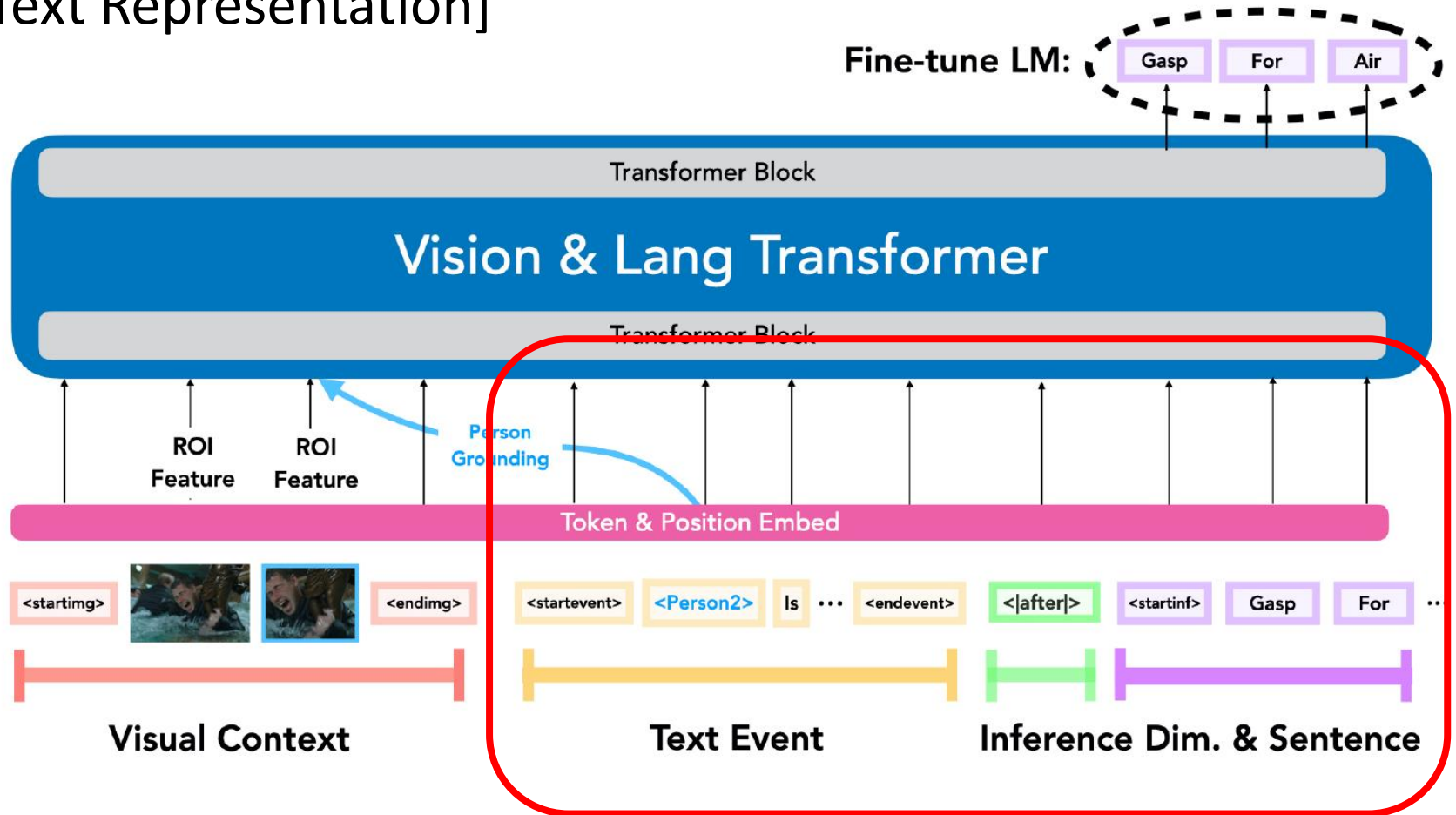
Base Loss function

$$\begin{aligned} \mathcal{L} = & - \sum_{i=1}^n \log P(w_i^e | w_{<i}^e, v)) - \sum_{i=1}^m \log P(w_i^p | w_{<i}^p, e, v)) \\ & - \sum_{i=1}^l \log P(w_{hi}^r | w_{h<i}^r, r, e, p, v) \end{aligned}$$

EP Loss function

# Model

[Text Representation]

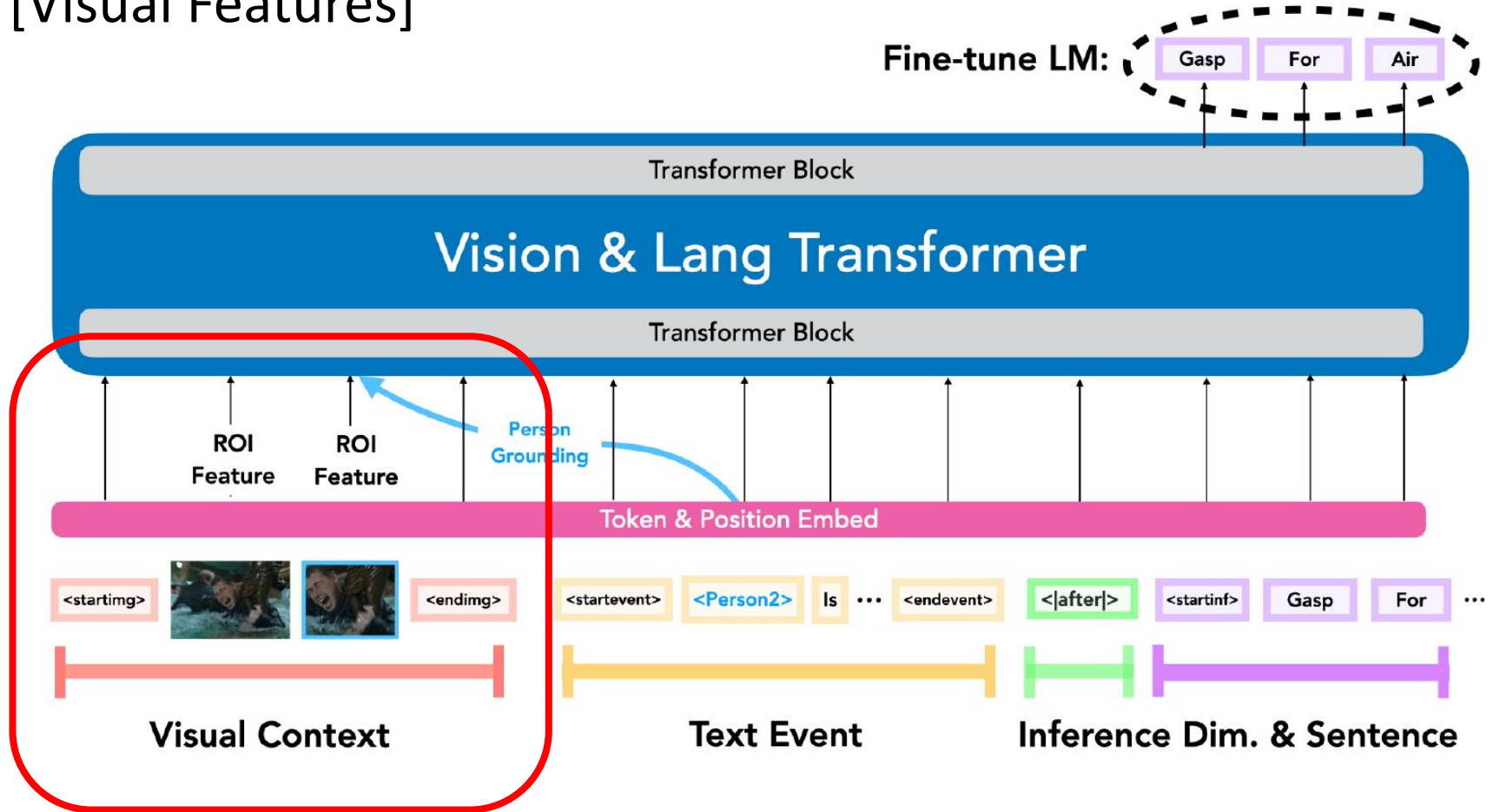


Transformer models used for language tasks use special separator tokens to enable better understanding of the input structure.



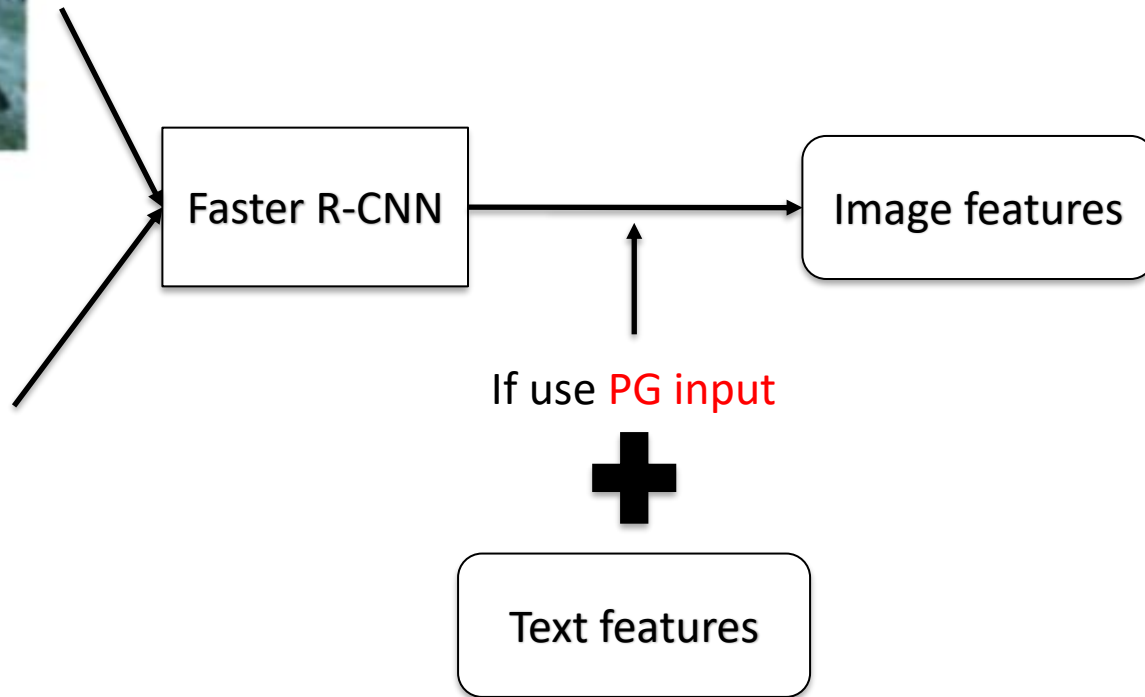
# Model

[Visual Features]



# Model

[Visual Features]



# Code & Setting

---

[Code] Publicly available

<https://github.com/jamespark3922/visual-comet>

[Setting]

Parameters

- Optimizer : Adam
- Learning rate :  $5e-5$
- Batch size : 64

Visual feature embedding

- ResNet101 backbone pretrained on ImageNet
- Maximum # of visual features : 15

Pre-trained GPT2-base model

- Maximum total sequence length as 256
- Nucleus sampling with  $p = 0.9$

# Experiments

[Validation set]

Modalities	Text Given	B-2	M	C	Acc@50	Unique	Novel
Place	Yes	5.87	6.25	4.69	14.55	6.84	47.57
Event	Yes	10.99	9.58	14.81	31.95	39.56	47.19
Event + Place	Yes	11.46	9.82	15.73	33.06	41.39	47.61
Image + Place	Yes	7.42	7.33	6.69	20.39	27.70	46.50
Image + Event	Yes	12.52	10.73	16.49	37.00	42.83	47.40
Image + Event + Place	Yes	12.78	10.87	17.12	38.25	43.83	48.15
Image + Event + Place + EP Loss	Yes	11.15	10.02	13.60	33.23	42.13	51.25
Image + Event + Place + PG	Yes	13.50	11.55	18.27	38.72	44.49	49.03
Image + Event + Place + PG + EP Loss	Yes	12.10	10.74	15.00	34.07	42.33	51.73
No Input	No	3.76	5.23	2.07	6.87	0.00	33.33
Image	No	6.79	7.13	5.63	18.22	26.38	46.80
Image + PG	No	8.20	8.44	7.61	21.5	29.09	45.53
Image + Event + Place	No	6.97	7.55	6.01	16.81	24.75	45.27
Image + Event + Place + EP Loss	No	7.06	7.77	6.37	20.02	31.60	50.77
Image + Event + Place + PG	No	8.80	9.19	8.77	17.35	27.42	47.37
Image + Event + Place + PG + EP Loss	No	10.21	10.66	11.86	22.7	33.90	49.84
GT	-	-	-	-	-	74.34	54.98



# Experiments

[Validation set]

Modalities	Text Given	B-2	M	C	Acc@50	Unique	Novel
Place	Yes	5.87	6.25	4.69	14.55	6.84	47.57
Event	Yes	10.99	9.58	14.81	31.95	39.56	47.19
<b>Event + Place</b>	Yes	11.46	9.82	15.73	33.06	41.39	47.61
Image + Place	Yes	7.42	7.33	6.69	20.39	27.70	46.50
Image + Event	Yes	12.52	10.73	16.49	37.00	42.83	47.40
Image + Event + Place	Yes	12.78	10.87	17.12	38.25	43.83	48.15
Image + Event + Place + EP Loss	Yes	11.15	10.02	13.60	33.23	42.13	51.25
<b>Image + Event + Place + PG</b>	Yes	<b>13.50</b>	<b>11.55</b>	<b>18.27</b>	<b>38.72</b>	<b>44.49</b>	49.03
Image + Event + Place + PG + EP Loss	Yes	12.10	10.74	15.00	34.07	42.33	<b>51.73</b>
No Input	No	3.76	5.23	2.07	6.87	0.00	33.33
Image	No	6.79	7.13	5.63	18.22	26.38	46.80
<b>Image + PG</b>	No	8.20	8.44	7.61	21.5	29.09	45.53
Image + Event + Place	No	6.97	7.55	6.01	16.81	24.75	45.27
Image + Event + Place + EP Loss	No	7.06	7.77	6.37	20.02	31.60	50.77
Image + Event + Place + PG	No	8.80	9.19	8.77	17.35	27.42	47.37
<b>Image + Event + Place + PG + EP Loss</b>	No	<b>10.21</b>	<b>10.66</b>	<b>11.86</b>	<b>22.7</b>	<b>33.90</b>	<b>49.84</b>
GT	-	-	-	-	-	74.34	54.98

# Experiments

[Test set]

Modalities	B-2	M	C	Human Before	Human Intent	Human After	Human Avg
<i>With Text Input.</i>							
Event + Place	11.00	9.65	15.12	54.9	52.6	42.9	50.1
Image + Event + Place + PG	<b>12.71</b>	<b>11.13</b>	<b>17.36</b>	<b>63.36</b>	<b>63.5</b>	<b>56.0</b>	<b>61.0</b>
<i>Without Text Input.</i>							
No Input	3.57	5.20	1.89	5.3	4.9	3.5	4.6
Image + PG	7.82	8.17	7.30	38.2	34.8	30.3	34.4
Image + Event + Place + PG + EP Loss	<b>9.33</b>	<b>10.12</b>	<b>10.82</b>	<b>42.9</b>	<b>36.8</b>	<b>34.8</b>	<b>38.2</b>
GT	-	-	-	83.8	84.5	76.0	81.4

- 1) Adding PG trick gives a boost for model.
- 2) Trained with visual and textual modalities is better than one modality.
- 3) Place information is helpful
- 4) EP-loss boost a model performance if only the image contents available

# Conclusions

---

1. They present VisualCOMET, a visual commonsense reasoning task that consists of before, after and intent.
2. They introduce large-scale dataset for visual commonsense Graphs.
3. They present baselines for this task with 2 setting : Text given and only image.
4. They show that integration between visual and textual commonsense reasoning is crucial to achieve the best performance.

**Q & A**