VisualCOMET: Reasoning about the Dynamic Context of a Still Image

visualcomet.xyz

Jae Sung Park[1,2], Chandra Bhagavatula[2], Roozbeh Mottaghi[1,2], Ali Farhadi[1], Yejin Choi [1,2]

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

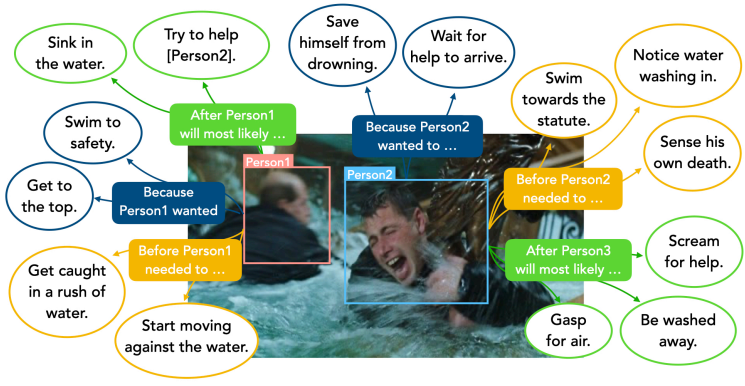


Fig. 1: Given a person in the image, **VisualCOMET** provides a *graph* of common sense inferences about 1) what needed to happen before, 2) intents of the people at present, and 3) what will happen next.

Abstract. Even from a single frame of a still image, people can reason about the dynamic story of the image *before*, *after*, and *beyond* the frame. For example, given an image of a man struggling to stay afloat in water, we can reason that the man fell into the water sometime in the past, the intent of that man at the moment is to stay alive, and he will need help in the near future or else he will get washed away. We propose **VisualCOMET**,[1] the novel framework of visual commonsense reasoning tasks to predict events that might have happened before, events that might happen next, and the intents of the people at present. To support research toward visual commonsense reasoning, we introduce the first large-scale repository of **Visual Commonsense Graphs** that consists of over **1.4 million** textual descriptions of visual commonsense inferences carefully annotated over a diverse set of 59,000 images, each paired with short video summaries of before and after. In addition, we provide person-grounding (i.e., co-reference links) between people appearing in the image and people mentioned in the textual commonsense descriptions, allowing for tighter integration between images and text. We establish strong baseline performances on this task and demonstrate that integration between visual and textual commonsense reasoning is the key and wins over non-integrative alternatives.

[1] **Visual Com**monsense **R**easoning in **T**ime.

1 Introduction

Given a still image, people can reason about the rich dynamic story underlying the visual scene that goes far beyond the frame of the image. For example, in Figure 1, given the image of a desperate man holding onto a statue in water, we can reason far beyond what are immediately visible in that still frame; sometime in the past, an accident might have happened and a ship he was on might have started sinking. Sometime in the future, he might continue struggling and eventually be washed away. In the current moment, his intent and motivation must be that he wants to save himself from drowning. This type of visual understanding requires a major leap from recognition-level understanding to cognitive-level understanding, going far beyond the scope of image classification, object detection, activity recognition, or image captioning. An image caption such as "a man in a black shirt swimming in water", for example, while technically correct, falls far short of understanding the dynamic situation captured in the image that requires reasoning about the context that spans before, after, and beyond the frame of this image. Key to this rich cognitive understanding of visual scenes is visual commonsense reasoning, which in turn, requires rich background knowledge about how the visual world works, and how the social world works.

In this paper, we propose **VisualCOMET**, a new framework of task formulations to reason about the rich visual context that goes beyond the immediately visible content of the image, ranging from events that might have happened before, to events that might happen next, and to the intents of the people at present. To support research toward visual commonsense reasoning, we introduce the first large-scale repository of **Visual Commonsense Graphs** that consists of **1.4 million** textual descriptions of visual commonsense inferences that are carefully annotated over a diverse set of about 59,000 people-centric images from VCR [52]. In addition, we provide person-grounding (i.e., co-reference links) between people appearing in the image and people mentioned in the textual commonsense descriptions, allowing for tighter integration between images and text. The resulting Visual Commonsense Graphs are rich, enabling a number of task formulations with varying levels of difficulties for future research.

We establish strong baseline performances on such tasks based on GPT-2 transformer architecture [32] to combine visual and textual information. Quantitative results and human evaluation show that integrating both the visual and textual commonsense reasoning is the key for enhanced performance. Furthermore, when the present eventual description is not available and only image is given, we find that the model trained to predict both events and inferential sentences performs better than the one trained to predict only inferences.

In summary, our contributions are as follows. (1) We 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) We 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. (3) We extend the GPT-2 model to incorporate visual information and allow direct supervision for grounding people in images. (4) 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.

2 Related Work

Visual Understanding with Language: Various tasks have been introduced for joint understanding of visual information and language, such as image captioning [8,47,36], visual question answering [1,17,26] and referring expressions [19,31,25]. These works, however, perform inference about only the current content of images and fall short of understanding the dynamic situation captured in the image, which is the main motivation of our work. There is also a recent body of work addressing representation learning using vision and language cues [41,24,39]. We propose a baseline for our task, which is inspired by these techniques.

Visual Commonsense Inference: Prior works have tried to incorporate commonsense knowledge in the context of visual understanding. [43] use human-generated abstract scenes made from clipart to learn common sense, but not on real images. [30] try to infer the motivation behind the actions of people from images. Visual Commonsense Reasoning (VCR) [52] tests if the model can answer questions with rationale using commonsense knowledge. While [52] includes rich visual common sense information, their question answering setup makes it difficult to have models to generate commonsense inferences. ATOMIC [35] provides a commonsense knowledge graph containing if-then inferential textual descriptions in generative setting; however, it relies on generic, textual events and does not consider visually contextualized information. In this work, we are interested in extending [52] and [35] for general visual commonsense by building a large-scale repository of visual commonsense graphs and models that can explicitly generate commonsense inferences for given images.

Visual Future Prediction: There is a large body of work on future prediction in different contexts such as future frame generation [33,38,51,48,27,46,6], prediction of the trajectories of people and objects [49,2,28], predicting human pose in future frames [13,50,7] and semantic future action recognition [21,55,40]. In contrast to all these approaches, we provide a compact description for the future events using language.

3 Task: Cognitive Image Understanding via Visual Commonsense Graphs

3.1 Definition of Visual Commonsense Graphs

The ultimate goal is to generate the entire visual commonsense graph illustrated in Figure 1 that requires reasoning about the dynamic story underlying the input image. This graph consists of four major components:

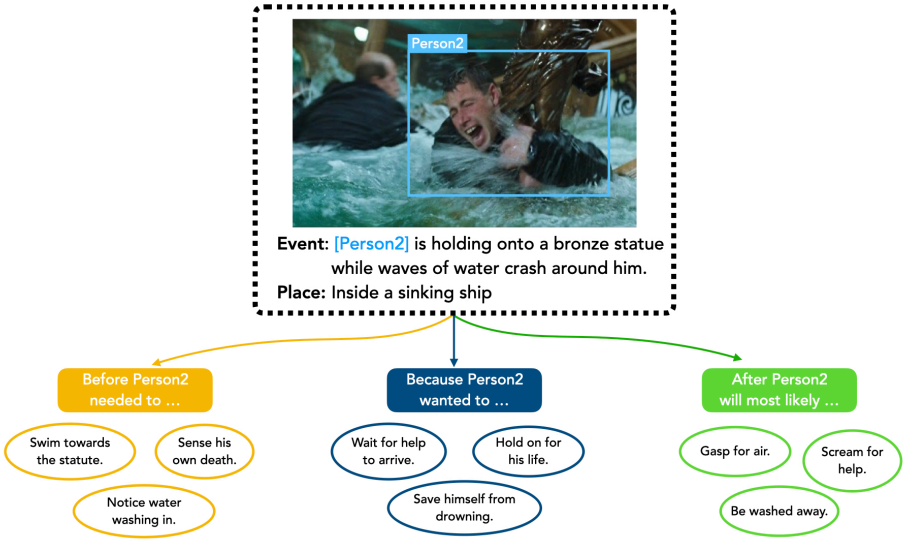


Fig. 2: **Task Overview:** Our proposed task is to generate commonsense inferences of **events before**, **events after** and **intents at present**, given an image, a description of an **event at present** in the image and a plausible scene / location of the image.

- (1) a set of textual descriptions of **events at present**,
- (2) a set of commonsense inferences on **events before**,
- (3) a set of commonsense inferences on **events after**, and
- (4) a set of commonsense inferences on people's **intents at present**.

The events before and after can broadly include any of the following: (a) *actions* people might take before and after (e.g., people jumping to the water), (b) *events* that might happen before and after (e.g., a ship sinking), and (c) *mental states* of people before and after (e.g., people scared and tired). Our design of the commonsense graph representation is inspired by ATOMIC [35], a text-only atlas of machine commonsense knowledge for *if-then* reasoning, but tailored specifically for cognitive understanding of visual scenes in images.

Location and Person Grounding: In addition, the current event descriptions are accompanied by additional textual descriptions of the **place** or the overall scene of the image, e.g., "at a bar" or "at a party". We also provide person-grounding (i.e., co-reference links) between people appearing in the image and people mentioned in the textual commonsense descriptions, allowing for tighter integration between images and text.

Dense Event Annotations with Visual Commonsense Reasoning: Generally speaking, the first component of the visual commonsense graph, "*events at present*", is analogous to dense image captioning in that it focuses on the

immediate visual content of the image, while components (2) - (4), *events before and after* and *intents at present*, correspond to visual commonsense reasoning.

Importantly, in an image that depicts a complex social scene involving multiple people engaged in different activities simultaneously, the inferences about before, after, and intents can be ambiguous as to which exact current event the inferences are based upon. Therefore, in our graph representation, we link up all the commonsense inferences to a specific event at present.

3.2 Definition of Tasks

Given the complete visual commonsense graph representing an image, we can consider multiple task formulations of varying degrees of difficulties. In this paper, we focus on two such tasks: (1) Given an image and one of the events at present, the task is to generate the rest of visual commonsense graph that is connected to the specific current event. (2) Given an image, the task is to generate the complete set of commonsense inferences from scratch.

4 Dataset Overview

We present the first large-scale dataset of Visual Commonsense Graphs for images with person grounding (i.e., multimodal co-reference chains). We collect a dataset of 1.4 million commonsense inferences over 59,356 images and 139,377 distinct events at present (Table 1). Figure 3 gives an overview of our Visual Commonsense Graphs including a diverse set of images, connected with the inference sentences [2].

	Train	Dev	Test	Total
# Images/Places	47,595	5,973	5,968	**59,356**
# Events at Present	111,796	13,768	13,813	**139,377**
# Inferences on Events Before	467,025	58,773	58,413	584,211
# Inferences on Events After	469,430	58,665	58,323	586,418
# Inferences on Intents at Present	237,608	28,904	28,568	295,080
# Total Inferences	1,174,063	146,332	145,309	**1,465,704**

Table 1: **Statistics** of our Visual Commonsense Graph repository: there are in total 139,377 distinct Visual Commonsense Graphs over 59,356 images involving 1,465,704 commonsense inferences.

[2] Larger figure available in the Appendix.

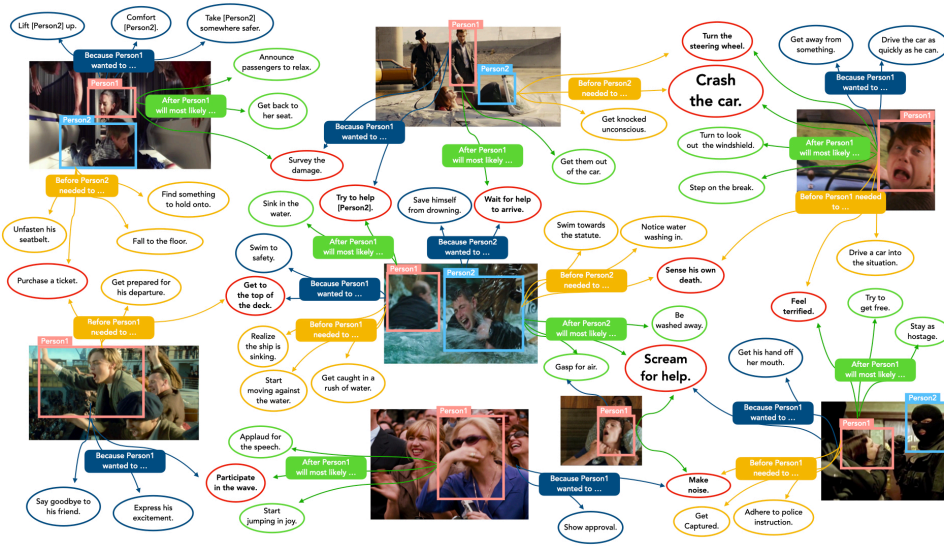


Fig. 3: Overview of our **Visual Commonsense Graphs**. We see that a diverse set of images are covered and connected with inference sentences. Red bubbles indicate if inference sentences shared by two or more images.

4.1 Source of Images

As the source of the images, we use the VCR [52] dataset that consists of images corresponding to complex visual scenes with multiple people and activities. The dataset also includes automatically detected object bounding boxes, and each person in the image uniquely identified with a referent tag (e.g. Person1 and Person2 in Fig 1).

4.2 Crowdsourcing Visual Commonsense Graphs

Annotating the entire commonsense graph solely from an image is a daunting task even for humans. We design a two-stage crowdsourcing pipeline to make the annotation task feasible and to obtain focused and consistent annotations. We run our annotation pipeline on Amazon Mechanical Turk (AMT) platform and maintain the ethical pay rate of at least $15/hr. This amounts to $4 per image on average. Figure 4 shows an overview of our annotation pipeline.

Stage 1: Grounded Event Descriptions with Locations and Intents

In the first stage, we show crowdworkers an image along with tags identifying each person in the image. Crowdworkers select a person and author a description for the event involving that person. One key concern during event annotation is to encourage crowdworkers to annotate informative, interesting events as opposed to low-level events like standing, sitting, looking, etc. While technically

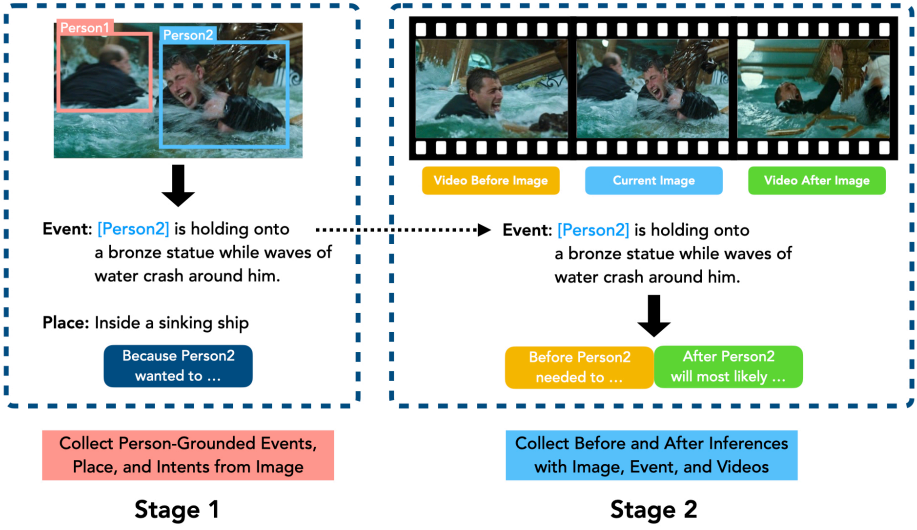


Fig. 4: **Annotation Pipeline:** Our two-stage crowdsourcing annotation pipeline used for collecting our high-quality Visual Commonsense Graphs.

correct, such descriptions do not contribute to higher-level understanding of the image. To obtain more meaningful events, we ask crowdworkers to write an event description and intent inferences at the same time. Finally, we ask crowdworkers to annotate the plausible location of the scene depicted in the image. In addition to priming workers, such location information provides more contextualized information for the task. The location information is not just a physical place, but can also include occasions, e.g., in a business meeting. At the end of stage 1, we collect (i) the location of an image, (ii) two to three events for each image, and (iii) two to four intents for each event of each image.

Stage 2: Collecting Before and After Inferences

In stage 2, we collect visual commonsense inferences of what might have happened *before* and what might happen *after* for each event description for each image annotated in stage 1 above. Images in our dataset were originally part of movie scenes. Based on the timestamp of the image being annotated, we show crowdworkers two short, fast-forwarded clips of events that happen before and after the image. This allows crowdworkers to author inferences that are more meaningful, rather than authoring correct but trivial inferences – e.g. "before, Person1 needed to be born", "after, Person1 will be dead", etc.

We assign two workers for each event and ask each to annotate between two and four *before* and *after* inferences. At the end of the two stages in our annotation pipeline, we have up to ten (2 intent, 4 before, 4 after) inferences for each pair of image and a textual description of event at present.

5 Our Approach

Our task assumes the following inputs for each image: a sequence of visual embeddings $\mathcal{V}$ representing the image and people detected in the image, grounded event description e, scene's location information p, and inference type r. Then, we wish to generate a set of possible inferences $H = \{s_1^r, s_2^r, ...s_{|H|}^r\}$.

5.1 Visual Features

The sequence of visual representations $\mathcal{V}$ consists of a representation of the whole image and an additional representations for each person detected in the image. We use Region of Interest (RoI) Align features [14] from Faster RCNN [34] as our visual embedding and pass it through a non-linear layer to obtain the final representation for an image or each detected person. The final sequence of representations $\mathcal{V} = \{v_0, v_1, ..v_k\}$ where k is the number of people detected.

As described in §4.2, we provide special tags identifying each person in the image (e.g. Person1 in Fig. 4) in our dataset. To use these tags, we introduce new person tokens, e.g. [Person1], in the vocabulary and create additional word embedding for these tokens. Then, we sum the visual representation for a person with the word embedding of the token referencing the person in text. This way, our model has visually grounded information about the image. We refer to this approach as "Person Grounding" (PG) input.

5.2 Text Representation

Transformer models used for language tasks [12,32] use special separator tokens to enable better understanding of the input structure. Since our task involves textual information of different kinds (event, place, and relation), we follow [5,53,4] to include special tokens for our language representation as well. Specifically, we append special token indicating the start and end of image (e.g. `s_img`, `e_img`), event, place, and inference fields. To generate inference statements, we use one of the three inference types (*before*, *intent*, *after*) as the start token, depending on the desired dimension.

5.3 Single Stream Vision-Language Transformer

We fix the model architecture as GPT-2 [32], a strong Transformer model [42] for natural language generation, conditioned on $\mathcal{V}, e, p$. Our model is a single stream transformer that encodes visual and language representations with a single transformer model, which has been shown to be more effective in vision and language tasks [9,54] compared to designing separate transformer models for each modality [24].

For each inference $s_h^r \in H$, our objective is to maximize $P(s_h^r | v, e, p, r)$. Suppose $s_h^r = \{w_{h1}^r, w_{h2}^r, ...w_{hl}^r\}$ is a sequence of l tokens. Then, we minimize the negative log-likelihood loss over inference instances in dataset:

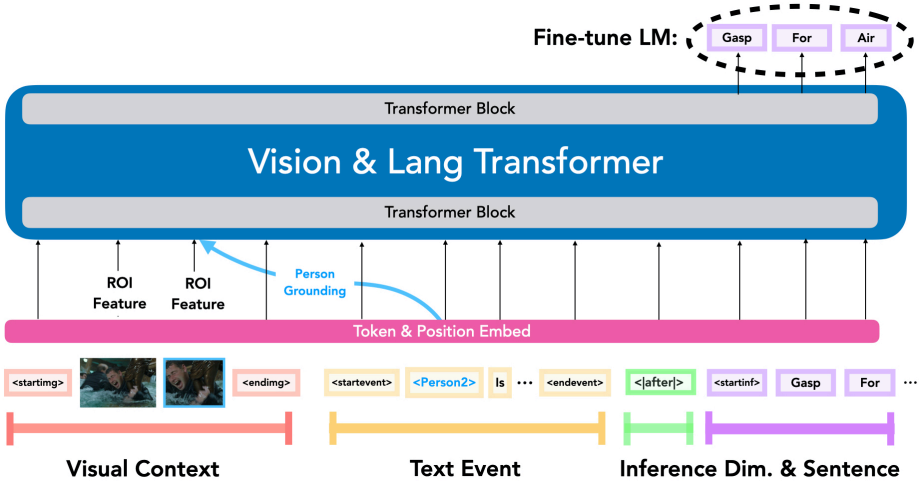


Fig. 5: **Model Overview.** Vision-Language Transformer for our approach. Our sequence of inputs uses special tokens indicating the start and end of image, event, place, and inference. We only show the start token in the figure for simplicity.

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

While our dataset provides events associated with each image, it is impractical to assume the availability of this information on new images. We experiment with a more general version of our model which does not take e and p as input. Nonetheless, we can supervise such models to generate e and p in the training phase. If we denote the *event at present* $\{e\} = \{w_1^e, w_2^e, ...w_n^e\}$ and place $\{p\} = \{w_1^p, w_2^p, ...w_m^p\}$ as a sequence of tokens, we apply the seq2seq loss on e, p (EP Loss in Section 6) as follows:

$$\begin{aligned}\mathcal{L} = &-\sum_{i=1}^{n} \log P(w_i^e | w_{<i}^e, v)) - \sum_{i=1}^{m} \log P(w_i^p | w_{<i}^p, e, v)) \\ &- \sum_{i=1}^{l} \log P(w_{hi}^r | w_{h<i}^r, r, e, p, v)\end{aligned} \tag{2}$$

6 Experiments and Results

6.1 Implementation Details

We use Adam optimizer [20] with a learning rate of 5e-5 and batch size of 64. Visual features for image and person embeddings use ResNet101 [15] backbone

pretrained on ImageNet [11]. We set the maximum number of visual features to 15. We use pre-trained GPT2-base model [32] as our model architecture with maximum total sequence length as 256. For decoding, we use nucleus sampling [16] with $p = 0.9$, which has shown to be effective generating text that is diverse and coherent. We have found beam search, which is a popular decoding scheme for generating multiple candidates, to be repetitive and produce uninteresting inferences. We report the effect of different decoding schemes in the supplementary material.

6.2 Experimental Setup

Baselines based on Different Inputs

In our experiments, we fix the same model architecture but ablate on the inputs available, e.g. place, event, and image. We also measure the effect of Person Grounding (PG) trick stated in Section 5.1. The models are trained with the same seq2seq objective in Eq. 1, and we mask out the visual and/or textual input based on the ablation of interest. We additionally experiment if learning to generate the event at present and place can improve the performance of generating the inferences using the objective in Eq. 2. For simplicity, we denote the loss on the two textual input as [+ EP. Loss]. Thus, we test two settings when generating the inferences: 1) one that uses event, place, and image, and 2) one that uses only image. We mark the two options in the Text Given column. [3]

Automatic Evaluation

Here, we describe the automatic evaluation measuring the quality of inference sentences. We first report the automatic metrics used in image captioning [8], such as BLEU-2 [29], METEOR [22], and CIDER [44] across the 5 inferences. Inspired by the metric in visual dialog [10], we also use perplexity score to rank the ground truth inferences and inferences from the different image. We append negatives such that there are 50 candidates to choose from, rank each candidate using perplexity score, and get the average accuracy of retrieved ground truth inferences (Acc@50 in Table 2). Note that perplexity is not necessarily the perfect measure to rank the sentences, but good language models should still be able to filter out inferences that do not match the content in image and event at present. Lastly, we measure the diversity of sentences, so that we do not reward the model for being conservative and making the same predictions. We report the number of inference sentences that are unique within the generated sentences divided by the total number of sentences (Unique in Table 2), and the number of generated sentences that are not in the training data divided by the total number of sentences (Novel in Table 2). To capture the semantic diversity, we replace the predicted person tags with the same tag when calculating the above diversity scores.

[3] We have tried running inferences on predicted events at present, but have gotten worse results than using no events. We report the results on predicted events in the supplemental.

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

Table 2: **Ablation Results.** Ablations of our baseline model on the Validation set. We use nucleus sampling with $p = 0.9$ to generate 5 sentences for all models. Automatic metrics used are BLEU-2 (B-2) [29], METEOR (M) [22], and CIDER (C) [44]. Acc@50 is the accuracy of correctly retrieved inference sentences with 50 candidates to choose from. Unique is the number of inference sentences that are unique within the generated sentences, divided by the total number of sentences. Novel refers to the number of generated sentences that are not in the training data, divided by the total number of sentences. Text Given is when model is given any textual input during test time to generate the inferences. We bold the models based on the following order: 1) Best Text only model, 2) Best Image + Text model given visual and text input, 3) Best Image only model, and 4) Best Image + Text model given just visual input.

6.3 Results

Table 2 shows our experimental results testing multiple training schemes and input modalities. We make the following observations: 1) Adding PG trick gives a boost for model over all metrics. 2) Model trained with both visual and textual (Image + Event + Place + PG) modalities outperform models trained with only one of modality (Event + Place; Image + PG) in every metric, including retrieval accuracy and diversity scores. This indicates that the task needs visual information to get higher quality inferences. 3) Adding place information helps in general. 4) Models with access to textual event and place information during test time, generate higher quality sentences than the same models without them (Text Given Yes vs No). This is not surprising as our dataset was collected with workers looking at the event, and the event already gives a strong signal understanding the content in the image. 5) Lastly, adding the EP Loss boosts the

Modalities	B-2	M	C	Human Before	Human Intent	Human After	Human Avg
With Text Input.							
Event + Place	11.00	9.65	15.12	54.9	52.6	42.9	50.1
Image + Event + Place + PG	**12.71**	**11.13**	**17.36**	**63.36**	**63.5**	**56.0**	**61.0**
Without Text Input.							
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	**9.33**	**10.12**	**10.82**	**42.9**	**36.8**	**34.8**	**38.2**
GT	-	-	-	83.8	84.5	76.0	81.4

Table 3: **Generated Inference Results.** BLEU-2 (B-2) [29], METEOR (M) [22], CIDER (C), [44] and Human scores for the generated inferences on the Test split. We select 200 random images and generate 5 sentences for each of the three inference type (3000 sentences total). Then, we assign three annotators to determine if each inference sentence is correct, and take the majority vote. The models are chosen based on their best performance on the validation set when visual and/or textual modalities are available (bolded models in Table 2).

performance if only the image content is available in the test time. This indicates that training the model to recognize events at present helps the performance, when the model has to generate inferences directly from image.

Human Evaluation

While the numbers in automatic evaluation give favorable results to our Image + Text model, they are not sufficient enough to evaluate the quality of generated inferences. We choose the best performing model when only image, text, or both inputs are available (model trained with no input and bolded models in Table 2). We take 200 random images and the generated inferences, and ask the humans to evaluate their quality based on just the image content. Even for models that use ground truth inferences, we do not show the events to the workers and make them rely on image to make the decision. Specifically, we ask three different workers to evaluate if each inference is likely (1) or unlikely (0) to happen based on the image. We then take the majority out of three and calculate the average across all the inferences.

Table 3 shows automatic metrics and human evaluation scores on the test split. We notice a similar pattern based on our automatic metric results: Image + Text model outperforms the Text only model (61.0 vs 50.1 on average) when text input is given in test time, and Image + Text model outperforms Image only model when text input is not given (38.2 vs 34.4 on average). We see that Text only model performs better than the Image + Text model without text input in test time, as the event sentence already describes the relevant details in the image and is a strong signal itself. Note that there is still a 20 point gap between our best model and ground truth inferences, meaning there is more room to improve our best model.

6.4 Qualitative Examples

Figure 9 presents some qualitative examples comparing the outputs of the various systems with the human annotated ground truth inferences. Overall, models that integrate information from both the visual and textual modalities generate more consistent and better contextualized predictions than models that only use either visual or textual information.

Specifically, the first example (on the top) illustrates that in the absence of the event description, a model that solely relies on visual information generates incorrect predictions like "order a drink at a bar", "dance and have fun" etc. – none of which are reasonable in the context of the event description. Similarly, a model that solely relies on the textual description, but not the visual information, generates "get off of the stage" and even predicts "her job as a scientist". This inference could be true in the absence of the visual features, but the image clearly shows that the person is in the audience, and not the one giving a presentation, nor she is portrayed as a scientist.

This pattern continues in the bottom example. [Person2] clearly looks worried but the Text only model predicts that he wants to "alleviate his boredom", and does not incorporate this visual detail. Image only model again hallucinates wrong objects like "have grabbed the wire". On the other hand, Image + Text model has the appropriate balance between the two models by stating there is possibly a criminal nearby as Person 2 is making an urgent call, and still predicts relevant visual details in the image. Thus, we see that both visual and textual features contribute to generating coherent inferences.

7 Conclusion

We present **VisualCOMET**, a novel framework of visual commonsense reasoning tasks to predict events that might have happened *before*, events that that might happen *after*, and the intents of people at *present*. To support research in this direction, we introduce the first large-scale dataset of Visual Commonsense Graphs consisting of 1.4 million textual descriptions of visual commonsense inferences carefully annotated over a diverse set of 59,000 images.

We present experiments with comprehensive baselines on this task, evaluating on two settings: 1) Generating inferences with textual input (event and place) and images, and 2) Directly generating inferences from images. For both setups, we show that integration between visual and textual commonsense reasoning is crucial to achieve the best performance.

Acknowledgements

This research was supported in part by NSF (IIS1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031), and gifts from Allen Institute for Artificial Intelligence.

Event: [Person2] is looking forward, watching the presentation closely.
Place: At a presentation

	Before, [Person2] needed to...	Because, [Person2] wanted to...	After, [Person2] will most likely...
Text Only	- walk towards the desk. - hear the words. - go to the business event. - arrive at the political event. - attend the event.	- show that she knows the target area. - solve a problem. - do her job as a scientist. - hear all the details at the conference. - appreciate the ideas of the speakers	- turn towards others. - wish the others luck. - ask questions - clap at the end of the presentation - get off of the stage
Image Only	- decide what she wanted to watch. - witness a good performer - order a drink at the bar - arrive at the party - get dressed up for the night.	- smoke her cigarette. - dance in a flirtatious manner. - see what other people are looking at. - hear the story while they wait. - observe the performance.	- have loads of fun - dance and have fun - eat her meal at the table - wipe her hands with a napkin - get into an argument with the man
Image + Text	- enter the presentation. - purchase a ticket. - put on her blouse. - enjoy the presentation. - attend the event.	- judge the presentation given by the presenter. - witness the event. - see what will be the next step. - enjoy the presentation. - appreciate the vision of the presenter.	- watch someone speak. - nod in agreement - ask questions. - clap at the end of the presentation. - get excited
Ground Truth	- watch a presentation - wait for everyone to come - listen to the lecture - watch the product demonstration	- think about what is being said. - tell the group her opinion.	- watch the screens - become very angry - talk with the others about the product - go over to examine it.

Event: [Person2] is holding a phone to his ear as he takes an urgent call
Place: Outside.

	Before, [Person2] needed to...	Because, [Person2] wanted to...	After, [Person2] will most likely...
Text Only	- wait outside his house for the call. - see what time the call was taking. - *hear the phone ring.* - *hear the phone ring.* - *hear the phone ring.*	- make sure it was there - hear the call. - respond to the call. - hear the information better. - alleviate his boredom.	- hang up the phone. - yell for help. - have a conversation on the phone. - end the call. - put his phone back in his pocket.
Image Only	- have grabbed the wire. - gather the team. - see something happening to the left. - hear the phone ring. - get inside the building.	- make someone angry. - fantasize. - check for trouble in the background. - hear what the person is saying. - be heard.	- say goodbye - fear for his life - start shooting someone - yell at the person in front of him - let go of the device
Image + Text	- take out his phone. - enter the crime scene. - receive a call. - *hear the phone ring.* - *hear the phone ring.*	- make sure his voice is heard - tape the call. - stop the criminal from going further. - hear the person on the phone. - get information.	- take pictures - slam the phone - lay down on a table - unlock the front door - put the phone in his pocket
Ground Truth	- be told some important information. - hear his phone ringing. - see that an important person was calling. - listen for someone on the line.	- find out what the emergency is - plug his ear so he can hear the call better	- go somewhere he can hear the call better. - ask the caller to speak up - stand up and scan for the person calling - question the caller as to his intentions.

Fig. 6: **Qualitative Results.** Qualitative Examples comparing our best Text only, Image only, and Image + Text only model. Red highlights inference statements that are incorrect. Orange highlights if the sentences are plausible, but not expected. We see that our Image + Text model gives more consistent and contextualized predictions than the baseline models.

Supplemental Material

We provide detailed statistics about the **VisualCOMET** dataset including its language diversity, and qualitative examples of inferences made by various model variants. We also show results from additional experiments for varying decoding schemes and performance for event description and place generation.

Figure 18 shows a snapshot of our **Visual Commonsense Graphs**. The three images show very distinct scenes, but the graph allows us to reason that the *intent* of the person sitting at a shack (bottom right image), the *before* event for the woman at an indoor bar (top left image), and the likely *after* event for the woman in the ballroom (bottom left) are identical – to "order a drink". Each image is associated with several inferences of the three types: (i) intents at present, (ii) events before, and (iii) events after.

A Dataset Statistics

Additional statistics of the dataset are provided in Table 4. On average, there are 2.12 *Intent*, 4.30 *Before*, and 4.31 *After* Inferences for each event. Each image has 2.34 events on average (place is always annotated once for each image). Figure 8 shows a breakdown of most frequent phrases per each inference type. *Before* and *After* inferences tend to focus on action statements, specifically activities involving entering or leaving the place. *Intent* inferences mostly involve various interactions with another person and also include person's mental states, such as "have a good time", "be polite", and "look formal".

We also provide more detailed distribution of the sentences. Figure 14 shows the number of occurrences of starting bigram (first two words) for each inference type. As we see, the distribution is vastly different based on the inference type, and there is no overlapping bigram among the top 5 phrases. Figure 15 shows the a) noun and b) verb distributions of the event sentences. We omit person in noun, and linking verbs in verb distributions for visualization purposes. We show histogram of unique place phrases in Figure 16. Popular places that are annotated include "office", "living room", "restaurant", "kitchen", and "party". Lastly, Figure 17 provides the length of event, place, and inference sentences.

B Qualitative Examples

We show more qualitative examples in Figure 9 and 10. Following Figure 6 of the main paper, we use the best performing model when Text only, Image only, and Image + Text input are given. Specifically, the models are Row 3 [Event + Place], Last Row [Image + Event + Place + PG + EP Loss (No Text Given)], and Row 8 [Image + Event + Place + PG] in Table 2 of the main paper. We highlight obviously incorrect inference sentences as red, and plausible but not expected as orange.

Figure 9(a) shows Person1 [P1] serving food and "putting a platter on the table". While the event and place information does not mention that [P1] is a

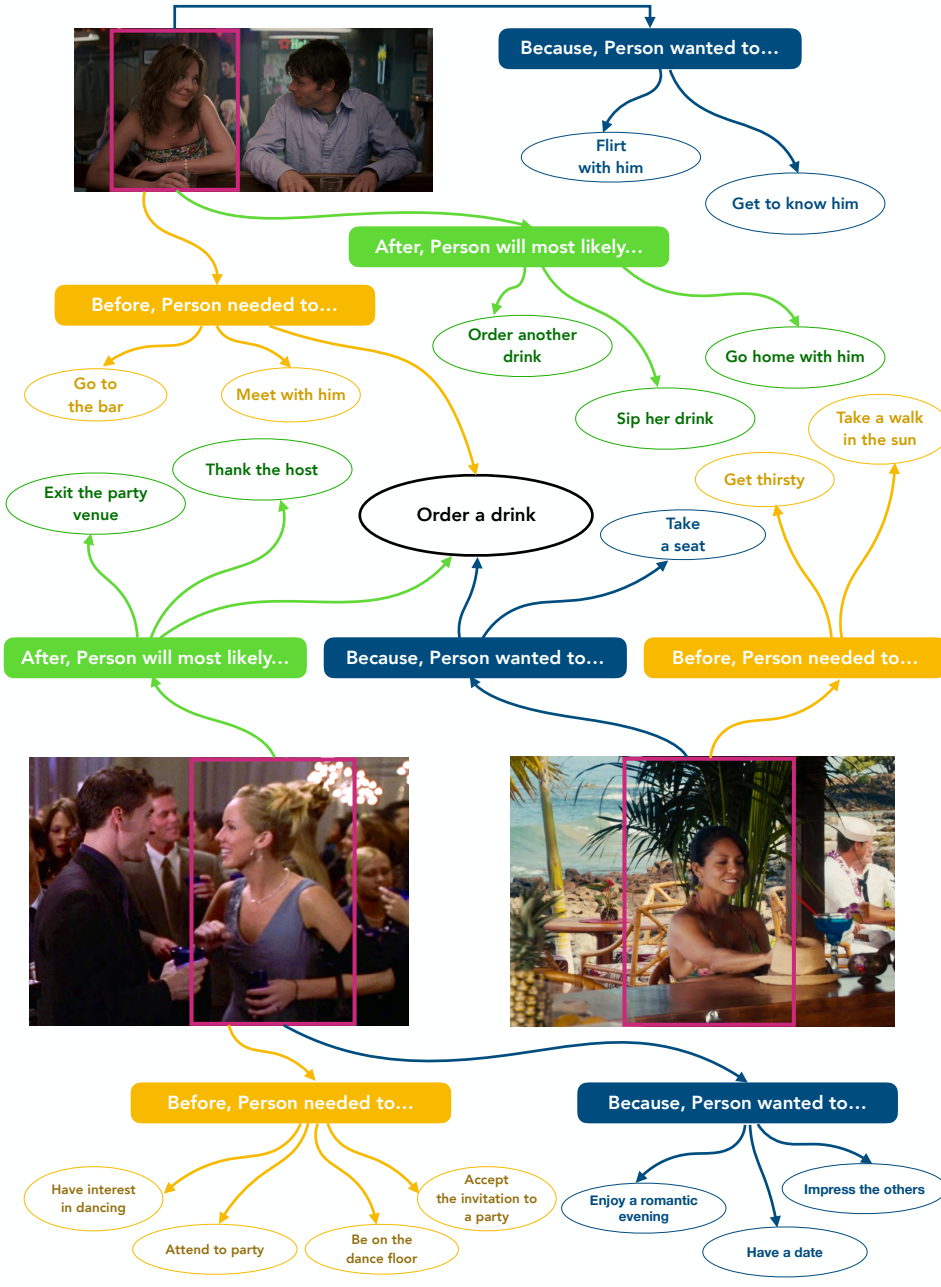


Fig. 7: Snapshot of our **Visual Commonsense Graphs**. Images from very distinct scenes are connected by the same inference sentence "order a drink".

	Avg Count
# of *Intent* Inference per Event	2.12
# of *Before* Inference per Event	4.30
# of *After* Inference per Event	4.31
# of Event per Image	2.34
# of Unique Persons Mentioned in Event	1.51
# of Unique Persons Mentioned in Inference	0.27
# of Words in Event	9.93
# of Words in Place	3.44
# of Words in Inference	4.8

Table 4: Additional Statistics for **VisualCOMET**.

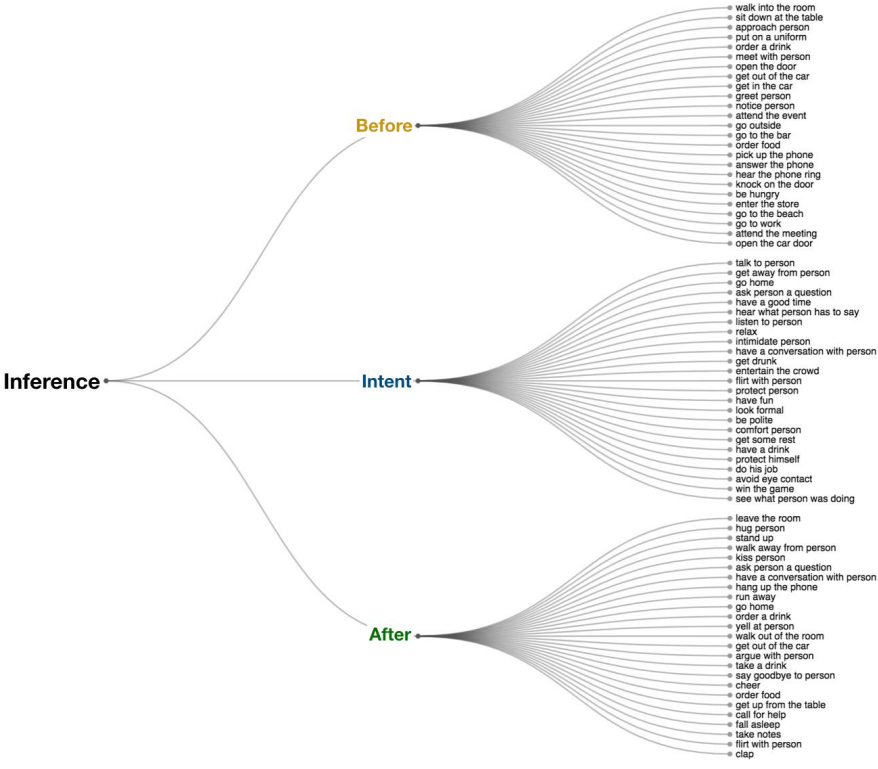


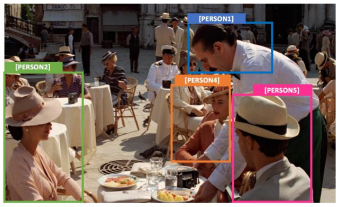
Fig. 8: Most frequent phrases mentioned per Inference Type

waiter, our Image + Text model uses the visual information to correctly infer that he needed to "be hired as a waiter at a formal event". The model also generates inferences that involve other relevant people (e.g. "serve [P2], [P4], [P5]"). Text only model fails to infer that [P1] is a waiter and sees him as the one joining the meal. For example, the model generates "ask [P2] for a menu" and "sip the water" in the *after* inferences. Image only model can generate inferences involving other people and recognize that the place is a restaurant; however, it fails to get the detail that [P1] is the one serving the food. Figure 9(b) shows an example focusing on person's mental state. While the image takes place at an outdoor party, it is unlikely that Person2 [P2] will dance, based on the event "is alone and feeling awkward" and her passive body language. We see that Image only and Text only models fail to incorporate this information and generate typical activities at a party, such as "dancing" or "drinking". Image + Text model makes inferences that suggests [P2] is not having fun and even predicts that she might "return to her car and drive away" or "yell at the people" after the event. Additional examples are shown in Figure 10 and we see that Image + Text model generates more coherent and plausible inferences.

Inference vs Captioning Figure 11 shows an example highlighting the main difference between our task and other visual captioning models. For fair comparison with image captioning models, we show the inference sentences using Image only model in Figure 11 (a). Top of Figure 11 (b) shows results from dense captioning model [18] that predicts the bounding boxes and associated captions. Bottom of the figure provides five captioning outputs using the strong baseline in [3]. We see that captioning models are mostly correct, such as the phrase "A woman is wearing a black shirt" and caption "a group of people sitting around a laptop". The descriptions, however, miss the detail of people working in the office. On the other hand, our Image only model can go beyond the simple details of sitting down at the desk and generate more contextualized information in office environment, such as "arrive at work early to get an interview", "see what was on the computer", and "gather up all her files". Using our visual commonsense graphs, we see that we can infer more salient and detailed information in still images that captioning tasks fail to provide.

C Annotation Template

Figure 12 shows the template used for our two stage annotation pipeline. The first stage Figure 12(a) involves writing at least two events and place per image. Then, each event is given optional choice of writing 2-3 *intent* inferences. Note only one worker is assigned for each image in the first stage. In the second stage Figure 12(b), each event is then annotated with 2-4 *before* and *after* inferences. Here, we assign two distinct workers to get the two inferences. In sum, each event is annotated with at least 10 inference sentences.



Event: [P1] is putting a platter on the table.
Place: At an Outdoor Restaurant.

	Before, [PersonX] needed to...	Because, [PersonX] wanted to...	After, [PersonX] will most likely...
Text Only	- buy groceries. - gather the other chefs for dinner. - enter the restaurant. - get up from the table. - put food on the platter.	- have dessert. - tend to the patrons. - see what everyone is doing. - ensure the food is taken care of - be friendly.	- chat with [P2]. - sip the water. - place the plate of food on the table. - get up and walk over to his table. - ask [P2] for a menu.
Image Only	- have drinks. - gather in a banquet hall. - order some food from the waitress. - arrive at the table. - be seated by a server.	- keep [P5] from interrupting [P2] and [P4]'s dinner. - greet [P2], [P4] and [P5]. - look at [P4] and [P2]. - hear what [P2], [P4], and [P5] had to say. - be friendly with [P2], [P4], and [P5].	- finish eating. - dance on the benches. - eat his meal. - enjoy the food. - put food on a plate.
Image + Text	- wait for everyone to sit down. - gather the others. - place the plates in a container - receive an order for platter. - be hired as a waiter at a formal event.	- have [P2], [P4], and [P5] to eat - greet [P2], [P4], and [P5]. - serve [P2], [P4], and [P5]. - get [P2], [P4], and [P5]'s attention. - serve [P2], [P4], and [P5] their meal.	- take drinks. - greet the person. - place the plates in a bowl. - get back to his work duties. - go back to the kitchen to get more food.
Ground Truth	- become a waiter. - grab a plate of food. - approach the table. - take the order. - get the food from the kitchen.	- wait on [P2], [P4], and [P5]. - do his job well.	- stand up straight. - leave the table. - ask if anything else is needed. - take more orders. - bring the check.

(a)

Event: [P2] is alone and feeling awkward.
Place: An Outdoor Party.

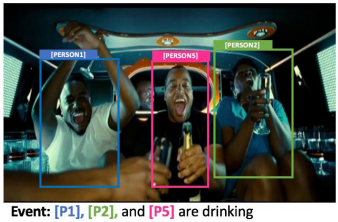
	Before, [PersonX] needed to...	Because, [PersonX] wanted to...	After, [PersonX] will most likely...
Text Only	- walk onto the lawn. - gather with the crowd. - go to the event. - hear someone's wrong. - get drunk.	- have others try and help her feel better. - gather his thoughts. - not get invited to a date. - act like he is not in the mood. - be alone.	- talk with others. - dance for the dancers - respond to other's question. - hear some good news. - be overcome with emotion.
Image Only	- walk towards the dance floor. - gather up the perfume supplies. - be asked for a favor. - attend a party. - stand up.	- make herself feel special. - dance and have fun. - get her hair done. - enjoy the party. - get away from the guy in a dress.	- make conversation. - dance on the patio. - watch as she is very nervous. - wipe her hands with a towel. - put her gloves on.
Image + Text	- walk onto the sidewalk. - greet the people. - put on her make up. - arrive at the event. - get dressed up for the event.	- have everyone join her at the party. - greet her friend. - ask people to be more careful. - end the date early. - be alone.	- make faces at someone walking by. - greet the other. - return to her car and drive away. - yell at the people at the party. - be awkward around people.
Ground Truth	- be stood up by her date. - arrived at the party alone. - plan to meet up with a date at the party. - find P2's date not at the garden party.	- go to a party. - meet new friends.	- look for someone she knows. - smile as her date finally shows up. - get through the crowd to the food table. - eat finger sandwiches alone.

(b)

Fig. 9: **Qualitative Results.** Qualitative Examples comparing our best Text only, Image only, and Image + Text model. Red highlights inference statements that are incorrect. Orange highlights if the sentences are plausible, but not expected. [PersonX] in the inference type refers to the subject of the event.

D Decoding Strategies

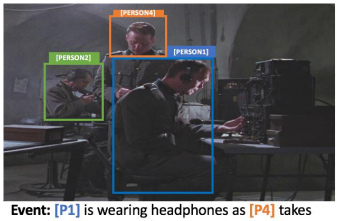
In the main paper, the inference sentences are generated using Nucleus Sampling [16], which is the state of the art decoding method to get more coherent and diverse sentences. Another option is to use beam search, which has shown to perform well in language metric but provides far less diverse sentences [45]. This



Event: [P1], [P2], and [P5] are drinking champagne in the back of a limo.
Place: In a limo.

	Before, [PersonX] needed to...	Because, [PersonX] wanted to...	After, [PersonX] will most likely...
Text Only	- show [P2] their appointment. - receive the champagne. - go on a date with [P2]. - get up on the limo. - get into the limo.	- have champagne. - sip on the champagne. - do something nice for a tip. - enjoy a night out at the bar. - get drunk.	- finish their champagne. - dance and have loads of fun. - drink more champagne. - enjoy the limo ride. - get shot.
Image Only	- have someone capture them. - gather the others. - board the spaceship with [P1], [P2], and [P5]. - receive information from [P5] and [P1]. - be sharing their space.	- keep someone from hanging on his own. - fantasize over the TV show. - do their job with the camera. - impress the crowd with his skill. - get information.	- make [P2] follow through with the plan. - dance with [P1], [P2], and [P5]. - watch [P1] and [P5]'s performance. - enjoy the show with [P2] and [P5]. - play the movie with [P5] and [P1].
Image + Text	- walk onto the stage. - tire of the festivities. - be driving to a new location. - attend a party. - get into the limo.	- have champagne. - sate their addiction. - relax. - enjoy the limo ride. - get drunk.	- keep partying. - dance in the champagne. - look out at the air. - receive an award. - get hit by a car.
Ground Truth	- rent the limo for an event. - pick everyone up in the limo. - plan a guy's night out. - pop open the cork.	- get a buzz on the way to a bachelor party. - go out on the town in style.	- have drinks in the limo. - tip the driver at the end of the night. - go out to a bar. - get drunk and spill in the limo.

(a)



Event: [P1] is wearing headphones as [P4] takes notes next to him.
Place: In a command center.

	Before, [PersonX] needed to...	Because, [PersonX] wanted to...	After, [PersonX] will most likely...
Text Only	- turn his computer on. - hear the bad translation about an item. - listen to a report. - read an important message from the headset. - use his headphones to monitor the data.	- keep [P4] informed - hear the information [P4] knew. - listen to a different stream. - hear more about what [P4] is saying - be able to take notes on the computer screens.	- talk into his headset. - wipe his hands. - watch for signs of movement. - converse with [P2]. - ask [P4] about the device he is holding.
Image Only	- make [P4] feel uncomfortable. - gather with [P4]. - listen to what [P4] has to say. - arrive at the meeting with [P4]. - be impressed by [P4].	- show [P4] that he is not his enemy. - solve a case with [P4]. - remain silent to get away from [P4]. - hear what [P4] has to say. - get along with [P4].	- show [P4] his empty plate. - sip from his mug as [P4] chugs it. - order [P4] to do the dirty work for him. - argue with [P4]. - be shocked by what [P4] says.
Image + Text	- turn towards [P4]. - hear the plan [P4] gave him. - listen to the report from [P4]. - unplug the headphones. - put his headphones on.	- listen to [P4]'s orders. - hear the information [P4] gives him. - do his job as a crew member - hear his orders. - hear what [P4] has to say.	- finish listening to what [P4] says. - yell at the subordinates for being slow. - look up from the page. - read the notes [P4] is holding. - put his headphones on.
Ground Truth	- put the headphones on his ears - intercept the code from the . enemy. - have some headphones. - be around [P4].	- intercept messages from the enemy. - decipher the code used by the enemy.	- tell everyone the message he received. - translate the code. - take off their headphones. - ask [P4] what they are writing.

(b)

Fig. 10: **Qualitative Results.** Qualitative Examples comparing our best Text only, Image only, and Image + Text model. Red highlights inference statements that are incorrect. Orange highlights if the sentences are plausible, but not expected. [PersonX] in the inference type refers to the subject of the event.

is especially problematic for generating *multiple* inferences, where we want to avoid generating duplicating phrases within the inference set.

Table 5 shows the comparison between the two decoding schemes and generate 5 sentences for each inference. We use the models from Row 3, 8, 10, and 12 in Table 2 of the main paper. We report BLEU-2 [29], and diversity metrics, such as proportion of unique inferences (UI), and ratio of unique unigrams/bigrams to number of words within the set of 5 sentences (DIV1/2-S) [37]. In language

metric, we see that the model performance is consistent regardless of the decoding strategy: Image + Text model (Image + Event + Place + PG) outperforms other Text only and Image only baselines for Nucleus Sampling and beam search. Image + Text model also gets the most number of unique sentences for the both decoding schemes. While BLEU-2 [29] scores are higher using beam search, we see that the diversity scores are much worse. Specifically, UI drops by half, and DIV1/2-S scores also suffer for the best performing model. We also see that Nucleus Sampling gets similar DIV1/2-S to the ground truth across all models, while there is around 30 and 20 point gap respectively for beam search methods. Note that getting the highest DIV1/2-S does not necessarily indicate having the highest diversity if these scores above a certain threshold. For instance, the model trained with No Input gets the highest DIV1-S and even higher than ground truth sentences, while UI is close to 0.

Figure 13 qualitatively shows the problem of using beam search over sampling methods. Beam search is prone to repeating the same phrases across the set, such as "sit down at the table", which are correct but not desirable for our task. On the other hand, Nucleus Sampling captures correct inference statements but also diverse and rich in content. This suggests that sampling based decoding scheme is far preferable to beam search, when generating multiple candidates.

Modalities	**BLEU-2** ↑	**UI**↑	**DIV1-S**	**DIV2-S**
Nucleus Sampling				
No Input	4.88	0.00	89.30	75.20
Event + Place	10.49	47.42	82.89	75.22
Image + PG.	7.84	35.62	83.70	75.99
Image + Event + Place + PG.	**11.76**	**51.99**	80.36	74.89
Beam Search				
No Input	7.36	0.00	54.00	48.70
Event + Place	18.97	23.64	56.10	54.50
Image + PG.	13.21	8.79	53.91	52.75
Image + Event + Place + PG.	**19.81**	**26.49**	54.70	53.92
GT	-	83.08	86.13	75.63

Table 5: Generating Inferences using Beam Search vs Nucleus Sampling on the Test set.

E Event and Place Generation

We report the performance of event and place generation given an image. We try two training schemes with the same model architecture used for generating

inferences: 1) train only on event and place, and 2) train on event, place, and inference. The second model is the same model [Image + Event + Place + PG + EP Loss] in Table 2 of the main paper. Note that there are around 10 times more inference sentences than events, meaning the second setup has access to 10 times more data. For fair comparison between the two models, we randomly sample 10% of the data (Row 2 in Table 6) and train the second model.

Table 6 shows the performance of two settings. We report the language metrics, CIDER [44], BLEU-4 [29], METEOR [22], and ROUGE [23], vocab size, and sentence length. Overall, we see that the two models perform similarly when the same amount of data are given. CIDER is higher for the first model, while the rest of language metrics are lower. When we use the entire data (All) for the second setup, we see that the improvement is significant for both language metrics and vocab size.

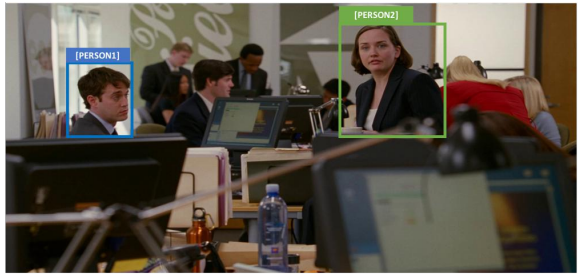
Inference using Generated Event Can the generated event be used as text input to generate the inferences? We use the generated event from Row3 in Table 6 as auxiliary text input and evaluate the quality of inferences. In Table 7 we show human evaluation using the same images and setup in Table 3 of the main paper. Under the section *With Generated Text Input*, we see that the Image + Text model performs better than Text only model, when generated event and place is given as input. However, the scores are lower than the best model without text input (36.0 vs 38.2). Note that this does not indicate that event and place information are not useful. As mentioned in the main paper, the model trained to generate event, place, and inference [Image + Event + Place + PG + EP Loss] performs the best when image is only given as input.

Training Scheme	**C**	**B-4**	**M**	**R**	**Vocab**	**Sent Len**
Image → Event + Place	17.61	1.85	11.78	22.62	1632	9.61
Image → Event + Place + Inference (10%)	15.69	2.35	12.01	23.34	1618	10.10
Image → Event + Place + Inference (All)	22.97	3.47	13.21	25.23	2578	9.71
GT					3799	9.98

Table 6: Event + Place Generation Performance on Test Set. We report the following language metrics: CIDER (C), BLEU-4 (B-4), METEOR (M), and ROUGE (R). We additionally include vocab size and sentence length. See Section E for more details.

Modalities	Human Before	Human Intent	Human After	Human Avg
With Generated Text Input				
Event + Place	34.6	35.8	29.5	33.3
Image + Event + Place + PG.	**38.9**	**37.5**	**31.7**	**36.0**
Image + Event + Place + PG + EP Loss.	37.2	32.9	30.4	33.5
With GT Text Input.				
Event + Place	54.9	52.6	42.9	50.1
Image + Event + Place + PG	**63.36**	**63.5**	**56.0**	**61.0**
Without Text Input.				
No Input	5.3	4.9	3.5	4.6
Image + PG	38.2	34.8	30.3	34.4
Image + Event + Place + PG + EP Loss	**42.9**	**36.8**	**34.8**	**38.2**

Table 7: **Generated Inference Results.** Human score for the generated inferences on the Test split. We select 200 random images and generate 5 sentences for each of the three inference type (3000 sentences total). Then, we assign three annotators to determine if each inference sentence is correct, and take the majority vote. Refer to Table 2 and Section 6.2 for model details. We see that the best model using generated event and place as input provides a worse performance than the best model without the text input.



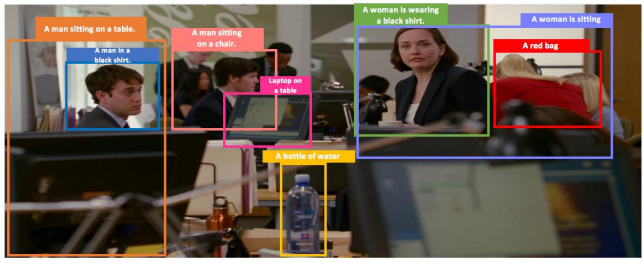
Event: [P2] stares toward the back.
Place: Open Office.

Before, [Person2] needed to...	Because, [Person2] wanted to...	After, [Person2] will most likely...
- walk towards the desk. - gather her things. - be around [P1]. - arrive at work early to get an interview. - be sitting down at her desk.	- have lunch - gather her things - see what was on the computer - act cool in front of the customers - get back to her work before the deadline	- make notes - gather up all her files - look up from the phone - read the paper on the table - not pay attention to the person on the phone

Image Only

(a) Inference with Image Only Model (event and place are not taken as input, and shown just for visualization)

Dense Captioning

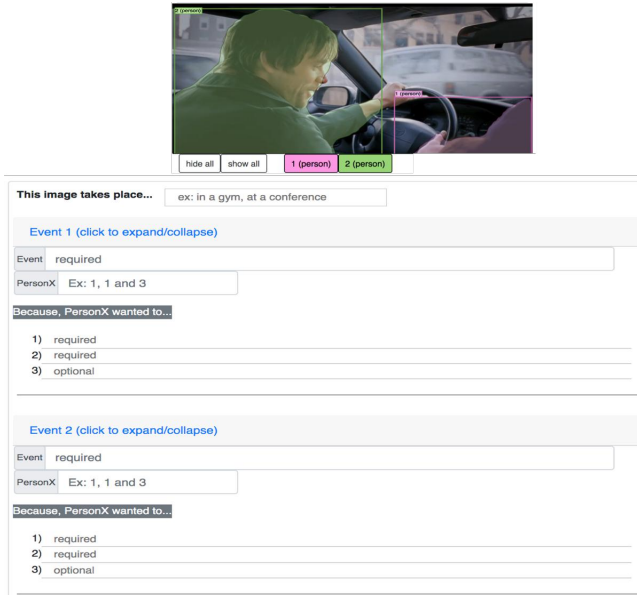


Predicted Captions

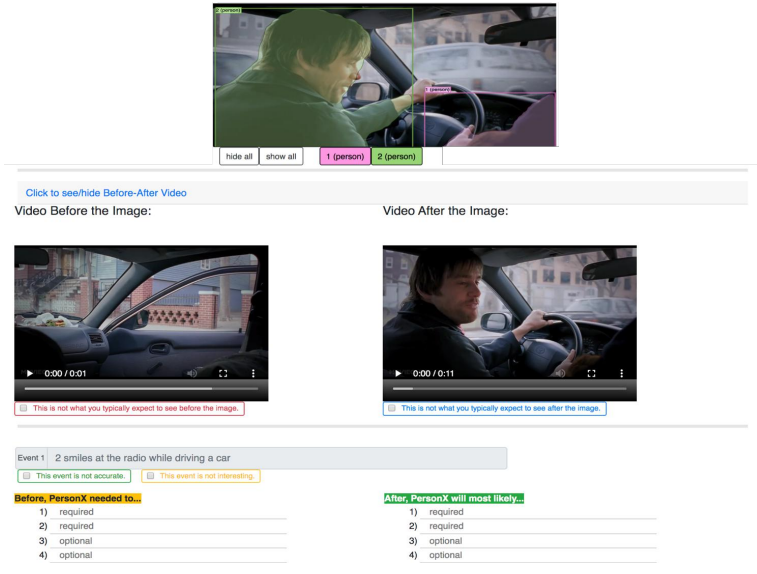
Score	Caption
41.2%	a group of people sitting around a laptop
26.1%	a group of people sitting around a computer
17.7%	a man sitting in front of a laptop computer
7.7%	a group of people sitting at a table with laptops
7.3%	a man sitting at a table with a laptop computer

(b) Results from Dense Captioning [18] and Bottom-up and Top-down image captioning model [3]

Fig. 11: **Difference between Inference and Captioning**. We see that our task (a) generates sentences that are more diverse and rich in content than the captioning models (b).



(a) We annotate event, place, and intent inferences in the First Annotation Stage.



(b) We annotate before and after inferences in the Second Annotation Stage.

Fig. 12: **Our Two-Stage Annotation Pipeline**. See Section C for more details.

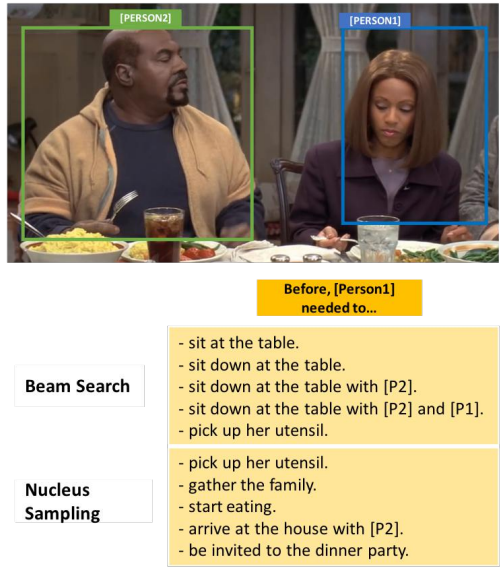
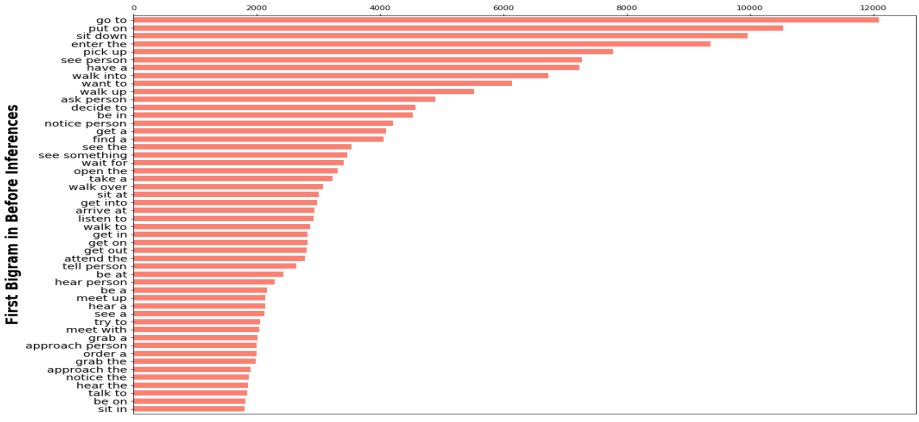
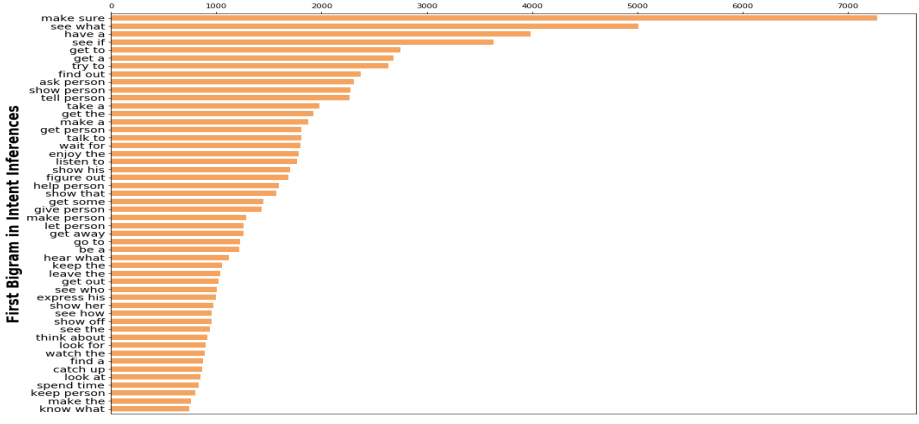


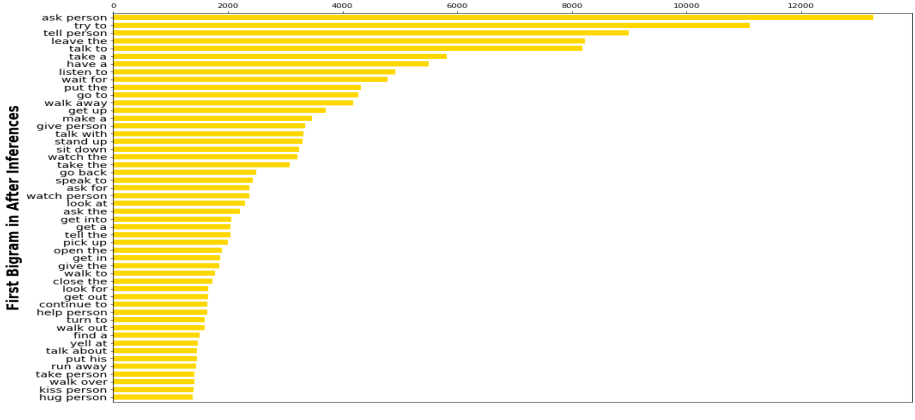
Fig. 13: Comparison between beam search and Nucleus Sampling from the same model. We see that beam search repeats the phrase "sit down at the table", while Nucleus Sampling gets more diverse and richer sentences.



(a) Before

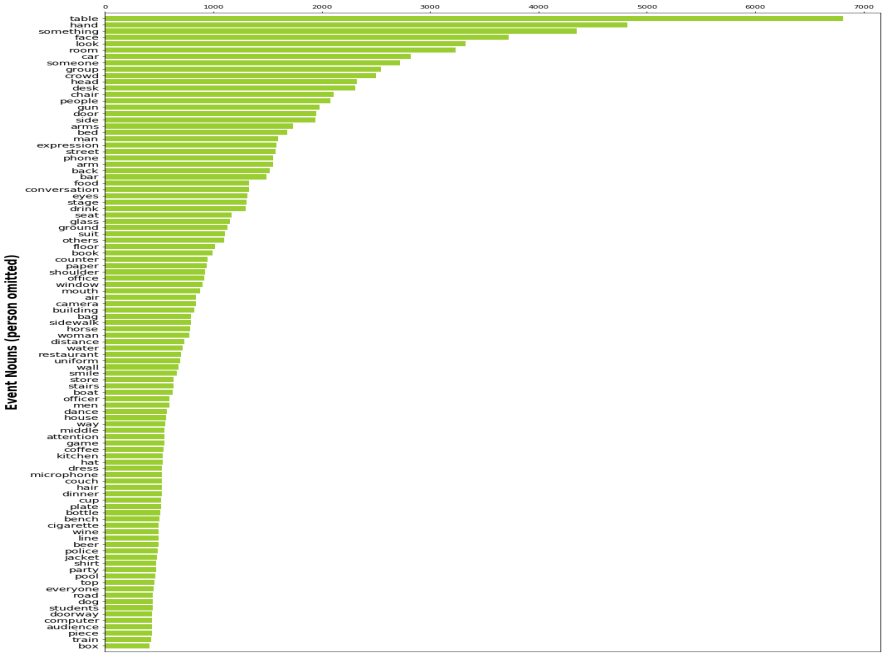


(b) Intent

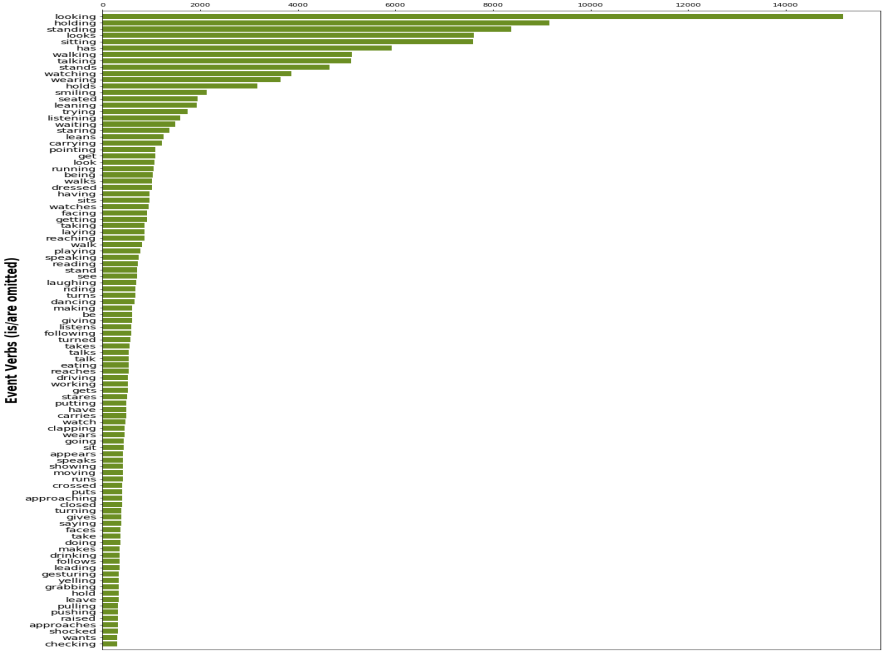


(c) After

Fig. 14: Most Frequent Starting bigram in a) Before, b) Intent, and c) After Inferences.



(a) Nouns in Event Sentences



(b) Verb Phrases in Event Sentences

Fig. 15: Most Frequent Noun & Verbs in Event Sentences

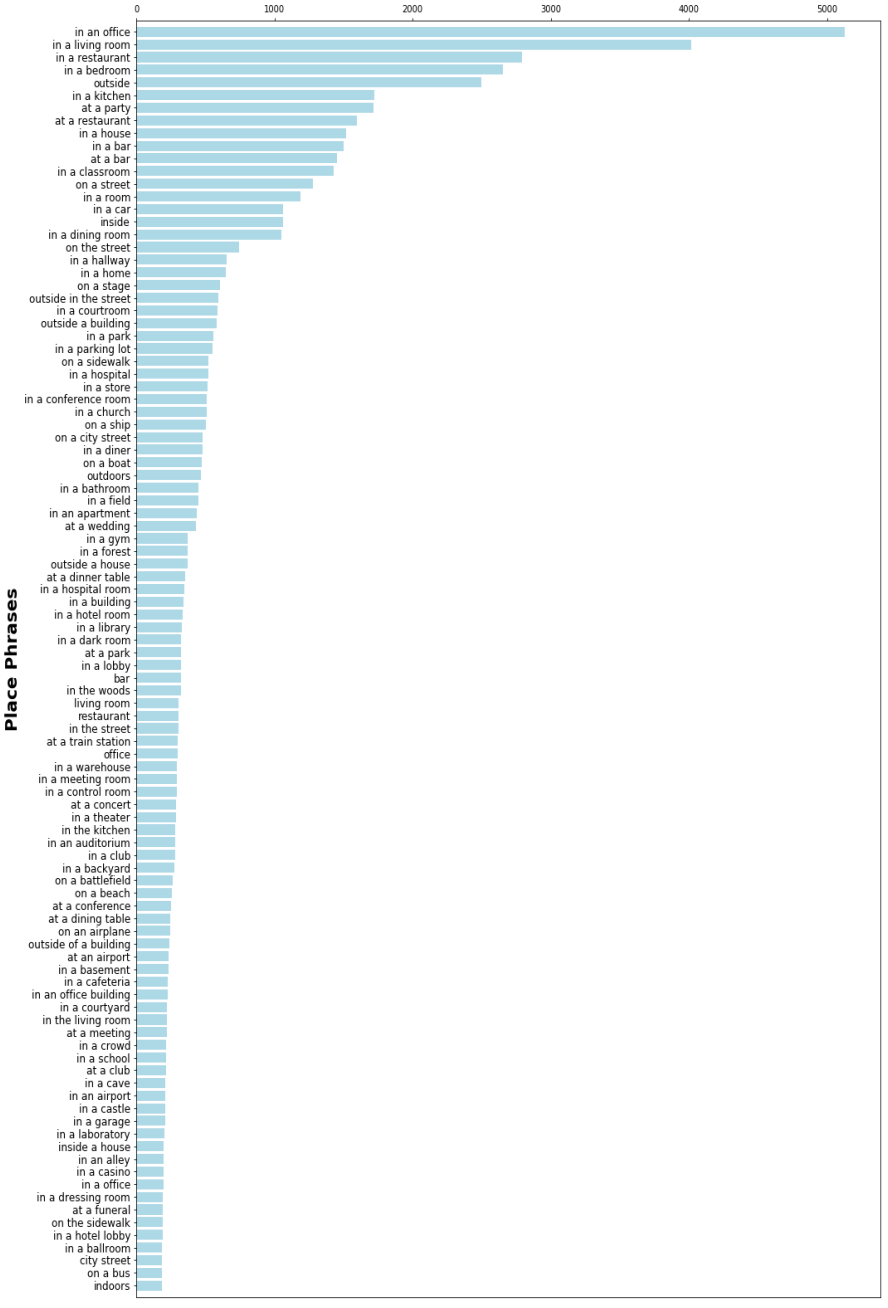
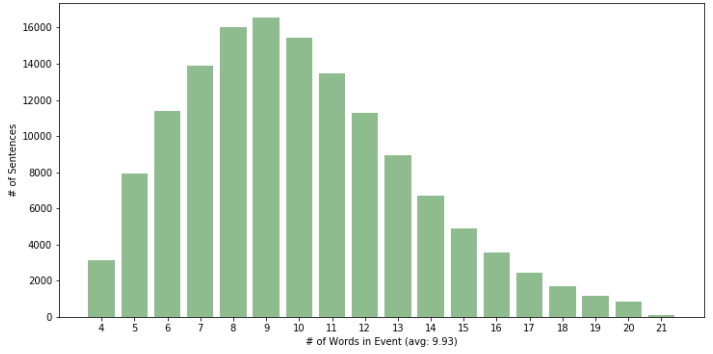
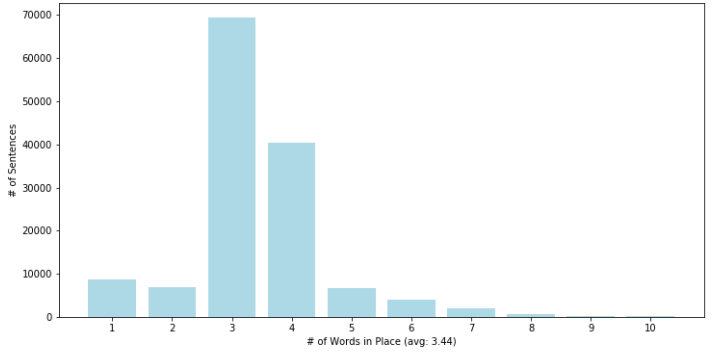


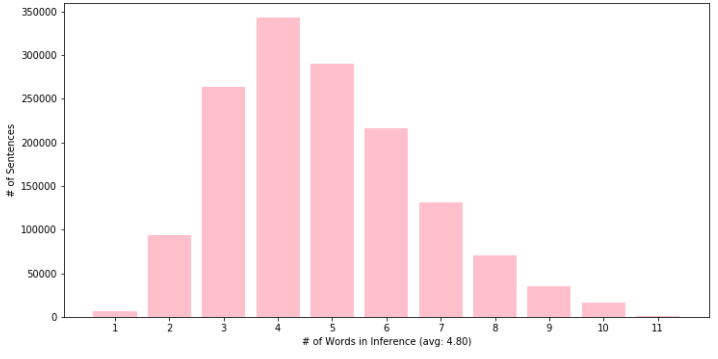
Fig. 16: Place Phrases



(a) Number of Words in Event



(b) Number of Words in Place



(c) Number of Words in Inference

Fig. 17: Sentence Length

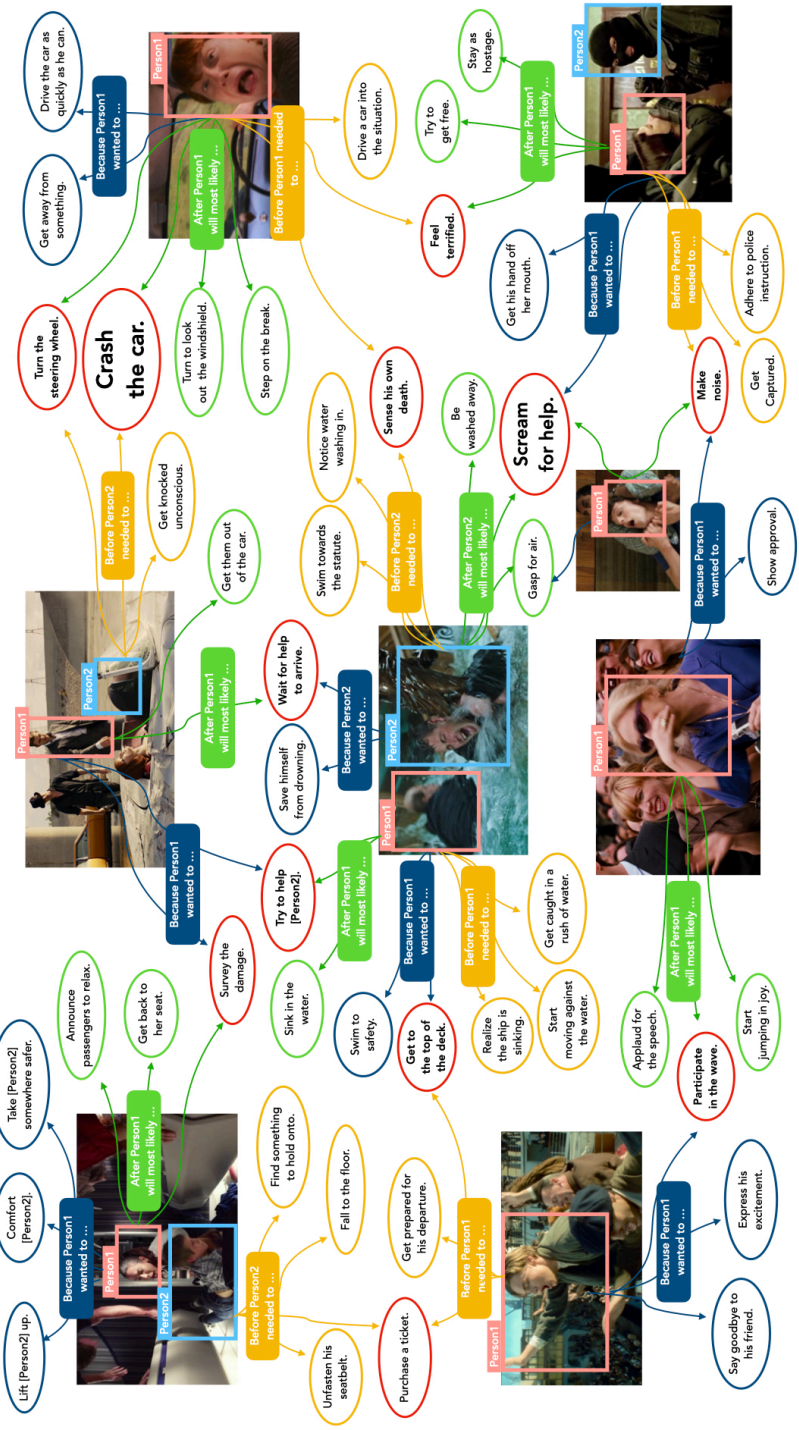


Fig. 18: Overview of our Visual Commonsense Graphs

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision **123**, 4–31 (2015) 3
2. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: CVPR (2016) 3
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 18, 24
4. Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., tau Yih, W., Choi, Y.: Abductive commonsense reasoning. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=Byg1v1HKDB 8
5. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: COMET: Commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4762–4779. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1470, https://www.aclweb.org/anthology/P19-1470 8
6. Castrejón, L., Ballas, N., Courville, A.C.: Improved vrnns for video prediction. In: ICCV (2019) 3
7. Chao, Y.W., Yang, J., Price, B.L., Cohen, S., Deng, J.: Forecasting human dynamics from static images. In: CVPR (2017) 3
8. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv (2015) 3, 10
9. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv (2019) 8
10. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M.F., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017) 10
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 10
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv (2018) 8
13. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV (2015) 3
14. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) 8
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 9
16. Holtzman, A., Buys, J., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv (2019) 10, 19
17. Johnson, J.E., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017) 3
18. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4565–4574 (2015) 18, 24

19. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 3
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv (2014) 9
21. Lan, T., Chen, T.C., Savarese, S.: A hierarchical representation for future action prediction. In: ECCV (2014) 3
22. Lavie, M.D.A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2014) 10, 11, 12, 22
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (2004) 22
24. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019) 3, 8
25. Mao, J., Huang, J., Toshev, A., Camburu, O., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016) 3
26. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019) 3
27. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2016) 3
28. Mottaghi, R., Rastegari, M., Gupta, A., Farhadi, A.: "what happens if..." learning to predict the effect of forces in images. In: ECCV (2016) 3
29. Papineni, K., Roukos, S., Ward, T., jing Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2002) 10, 11, 12, 20, 21, 22
30. Pirsiavash, H., Vondrick, C., Torralba, A.: Inferring the why in images. arXiv (2014) 3
31. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV (2015) 3
32. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8) (2019) 2, 8, 10
33. Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. arXiv (2014) 3
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015) 8
35. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2019) 3, 4
36. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2018) 3
37. Shetty, R., Rohrbach, M., Hendricks, L.A., Fritz, M., Schiele, B.: Speaking the same language: Matching machine to human captions by adversarial training. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) 20
38. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: ICML (2015) 3

39. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: ICLR (2020) 3
40. Sun, C., Shrivastava, A., Vondrick, C., Sukthankar, R., Murphy, K., Schmid, C.: Relational action forecasting. In: CVPR (2019) 3
41. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP (2019) 3
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS) (2017) 8
43. Vedantam, R., Lin, X., Batra, T., Zitnick, C.L., Parikh, D.: Learning common sense through visual abstraction. In: ICCV (2015) 3
44. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 10, 11, 12, 22
45. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q.H., Lee, S., Crandall, D.J., Batra, D.: Diverse beam search: Decoding diverse solutions from neural sequence models. ArXiv **abs/1610.02424** (2016) 19
46. Villegas, R., Pathak, A., Kannan, H., Erhan, D., Le, Q.V., Lee, H.: High fidelity video prediction with large stochastic recurrent neural networks. In: NeurIPS (2019) 3
47. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015) 3
48. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NeurIPS (2016) 3
49. Walker, J., Gupta, A., Hebert, M.: Patch to the future: Unsupervised visual prediction. In: CVPR (2014) 3
50. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV (2017) 3
51. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: NeurIPS (2016) 3
52. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019) 2, 3, 6
53. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y.: Defending against neural fake news. In: Advances in Neural Information Processing Systems (NIPS) (2019) 8
54. Zhou, L., Hamid, P., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and question answering. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2020) 8
55. Zhou, Y., Berg, T.L.: Temporal perception and prediction in ego-centric video. In: ICCV (2015) 3