Visual Commonsense Graphs: Reasoning about the Dynamic Context of a Still Image –Supplementary 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 12 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 1. 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 2 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 8 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 9 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 10. Popular places that are annotated include "office", "living room", "restaurant", "kitchen", and "party". Lastly, Figure 11 provides the length of event, place, and inference sentences.

B Qualitative Examples

We show more qualitative examples in Figure 3 and 4. 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)],

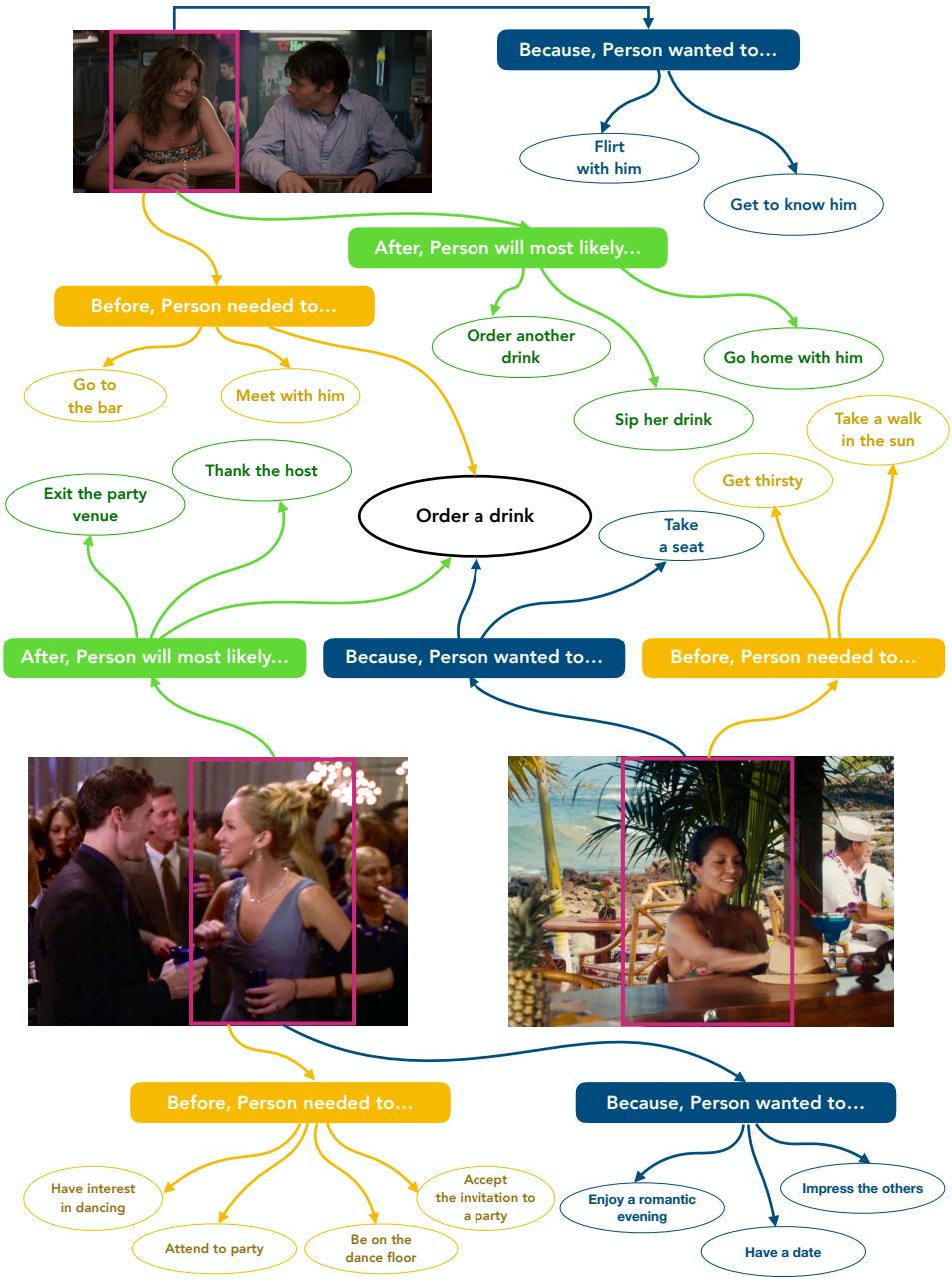


Fig. 1: Snapshot of our **Visual Commonsense Graphs**. Images from very distinct scenes are connected by the same inference sentence "order a drink".

	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 1: Additional Statistics for **VisualCOMET**.

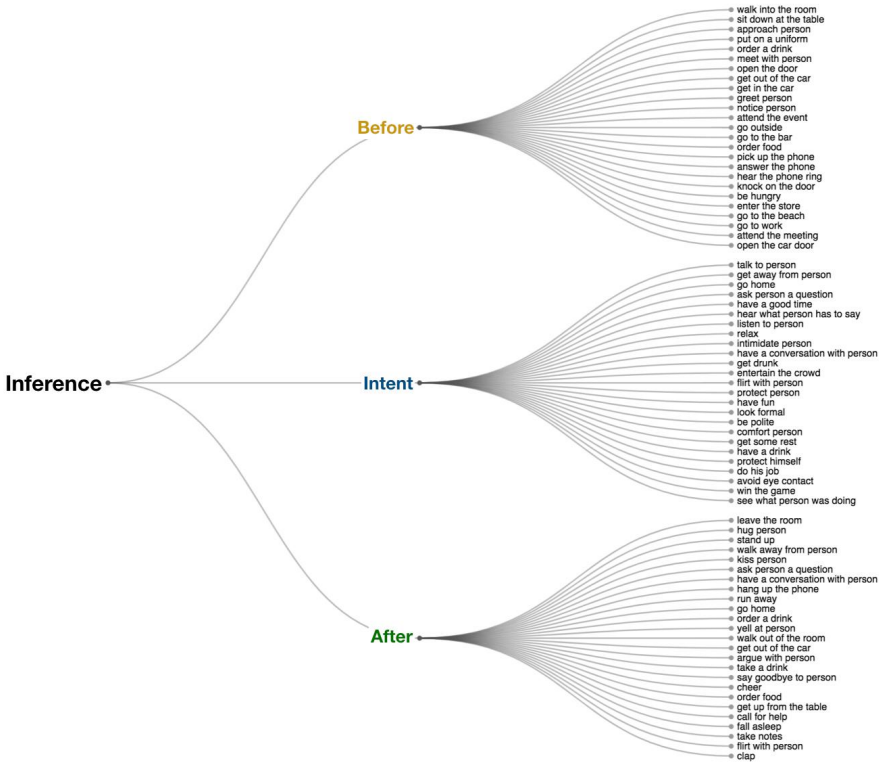


Fig. 2: Most frequent phrases mentioned per Inference Type

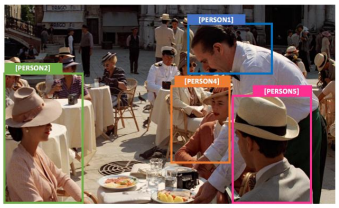
and Row 8 [Image + Event + Place + PG] in Table 2 of the main paper. We highlight obviously incorrect inference sentences as red, and plausible but not expected as orange.

Figure 3(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 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 3(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 4 and we see that Image + Text model generates more coherent and plausible inferences.

Inference vs Captioning Figure 5 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 5 (a). Top of Figure 5 (b) shows results from dense captioning model [3] that predicts the bounding boxes and associated captions. Bottom of the figure provides five captioning outputs using the strong baseline in [1]. 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 6 shows the template used for our two stage annotation pipeline. The first stage Figure 6(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 6(b), each event is then annotated with 2-4 *before* and *after* inferences.



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.

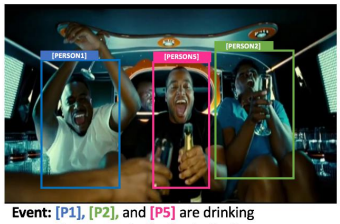
(b)

Fig. 3: **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.

Here, we assign two distinct workers to get the two inferences. In sum, each event is annotated with at least 10 inference sentences.

D Decoding Strategies

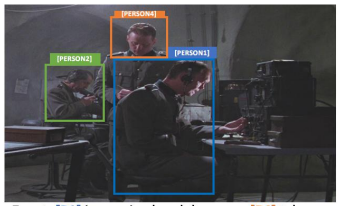
In the main paper, the inference sentences are generated using Nucleus Sampling [2], which is the state of the art decoding method to get more coherent and



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. 4: **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.

diverse sentences. Another option is to use beam search, which has shown to perform well in language metric but provides far less diverse sentences [9]. This is especially problematic for generating *multiple* inferences, where we want to avoid generating duplicating phrases within the inference set.

Table 2 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 [6], 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) [7]. 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 [6] 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 7 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 2: 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 3) and train the second model.

Table 3 shows the performance of two settings. We report the language metrics, CIDER [8], BLEU-4 [6], METEOR [4], and ROUGE [5], 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.

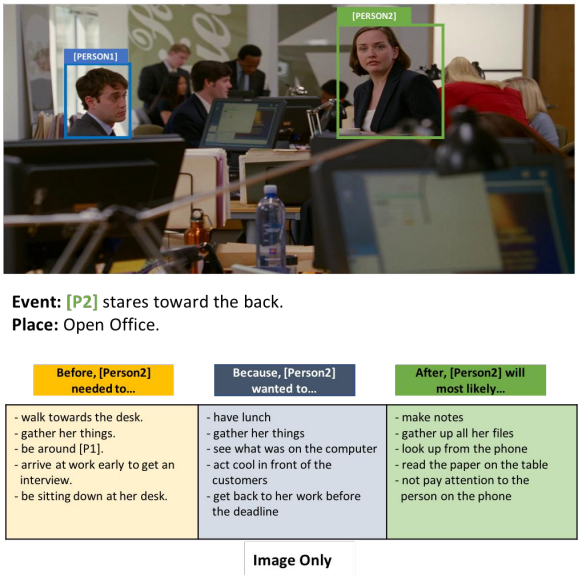
Inference using Generated Event Can the generated event be used as text input to generate the inferences? We use the generated event from Row3 in Table 3 as auxiliary text input and evaluate the quality of inferences. In Table 4 we show human evaluation using the same images and setup in Table 3 of the 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

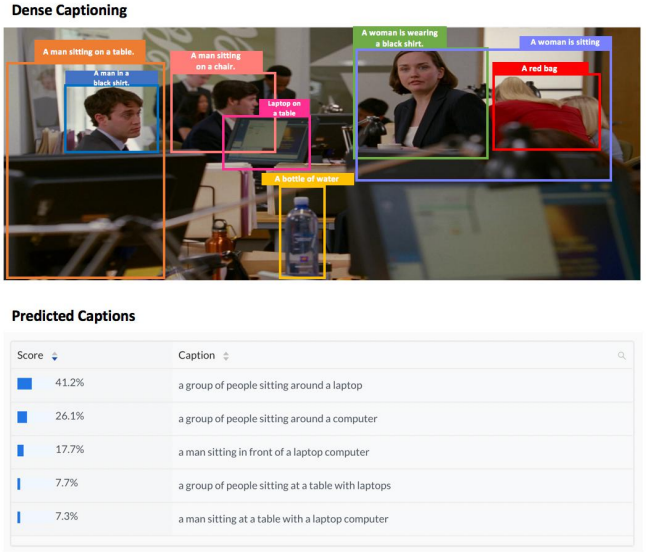
Table 3: Event + Place Generation Performance on Test Set. We report the following language metrics: CIDER (C), BLEU-4 (B-4), METEOR (M), and ROUGE (R). We additionally include vocab size and sentence length. See Section E for more details.

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 4: **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.

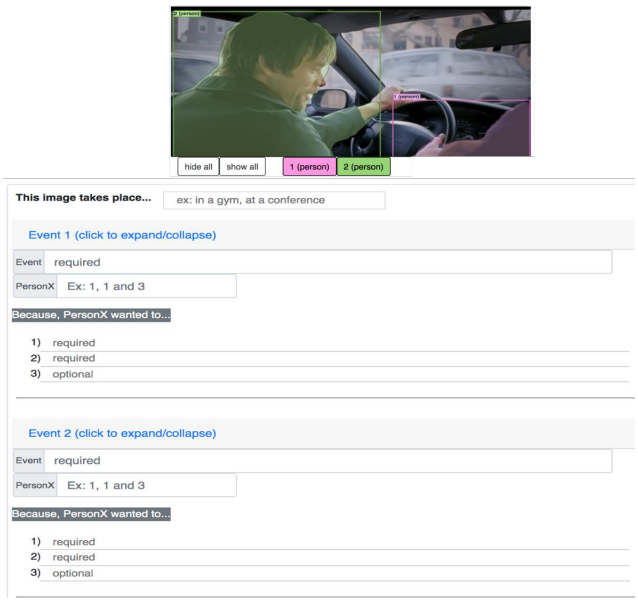


(a) Inference with Image Only Model (event and place are not taken as input, and shown just for visualization)

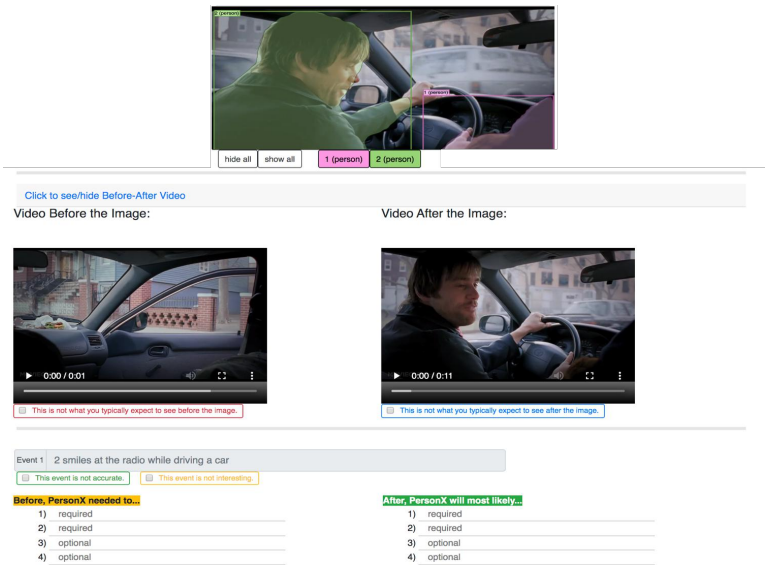


(b) Results from Dense Captioning [3] and Bottom-up and Top-down image captioning model [1]

Fig. 5: **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. 6: **Our Two-Stage Annotation Pipeline**. See Section C for more details.

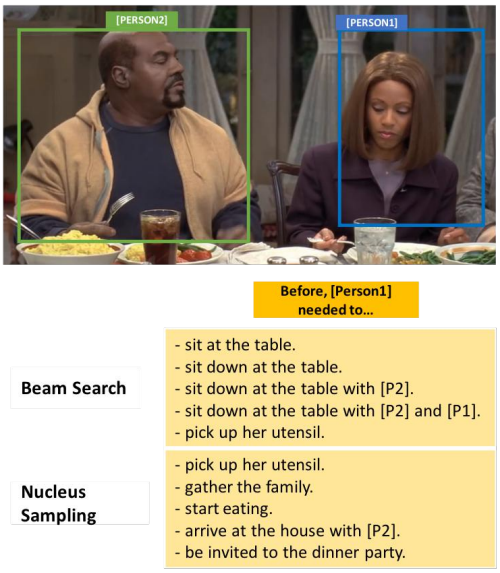
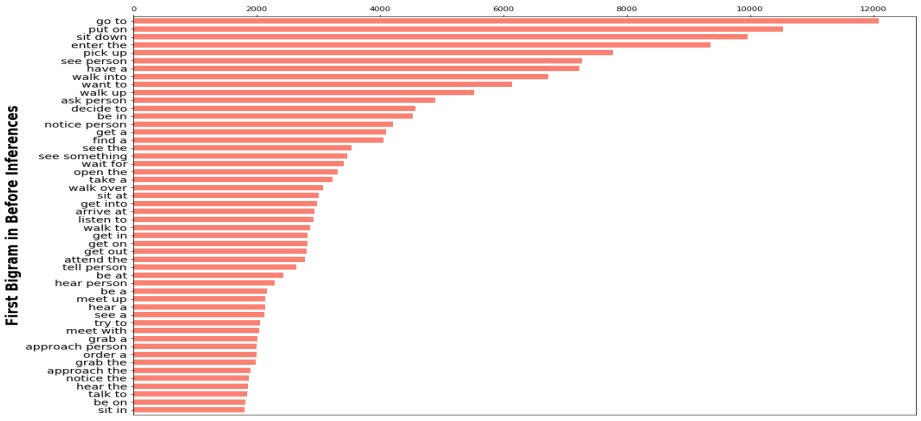
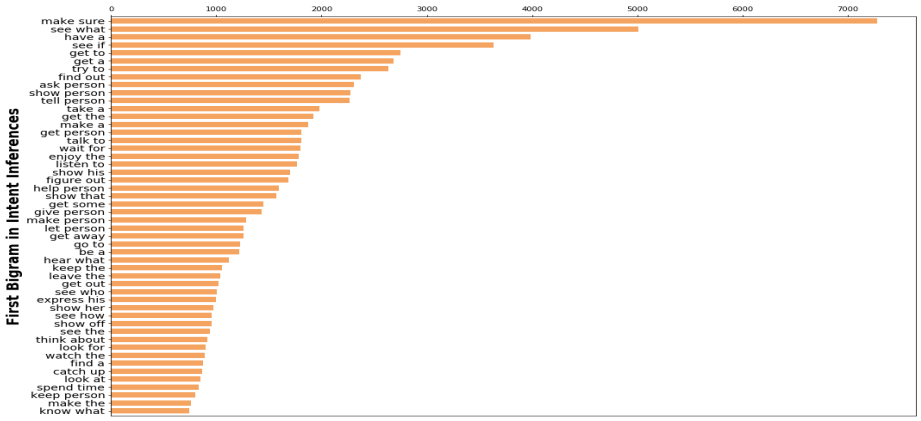


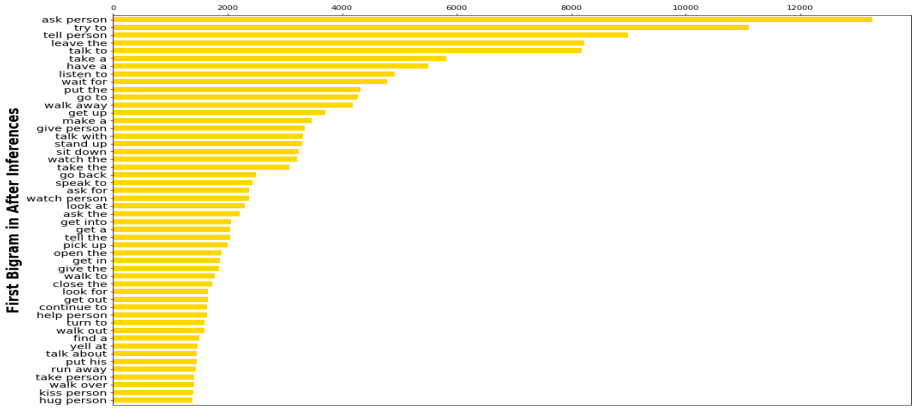
Fig. 7: Comparison between beam search and Nucleus Sampling from the same model. We see that beam search repeats the phrase "sit down at the table", while Nucleus Sampling gets more diverse and richer sentences.



(a) Before

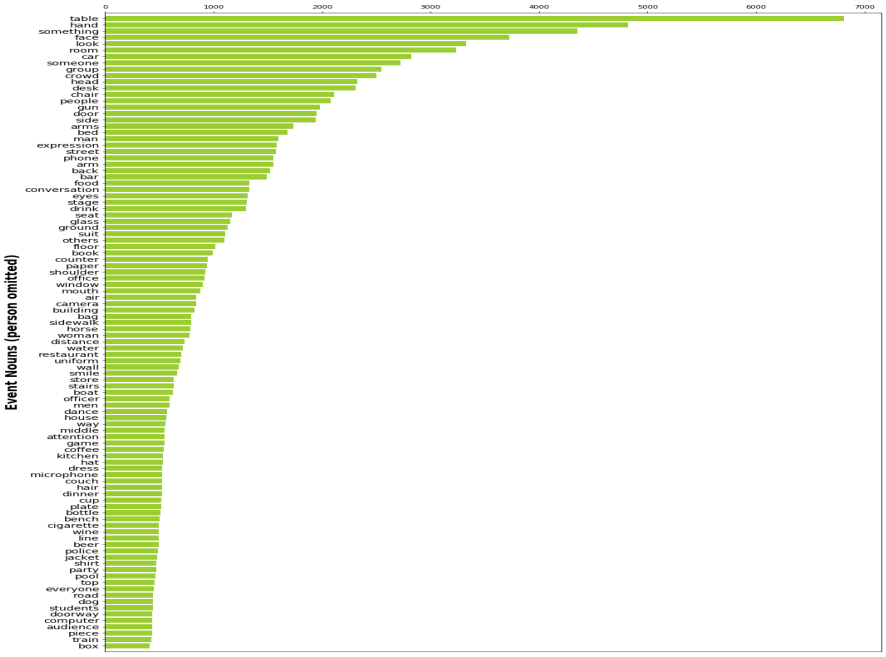


(b) Intent

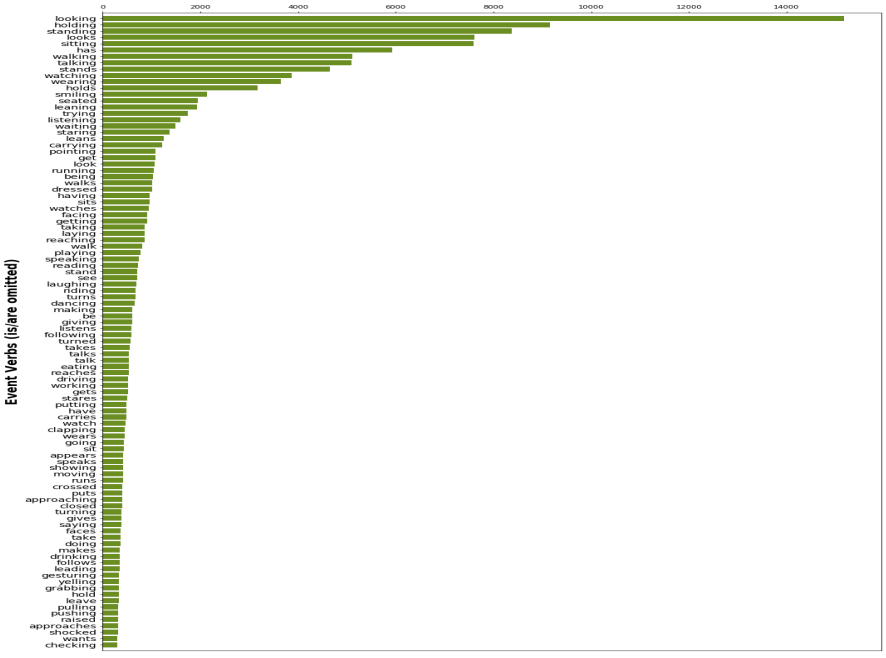


(c) After

Fig. 8: 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. 9: Most Frequent Noun & Verbs in Event Sentences

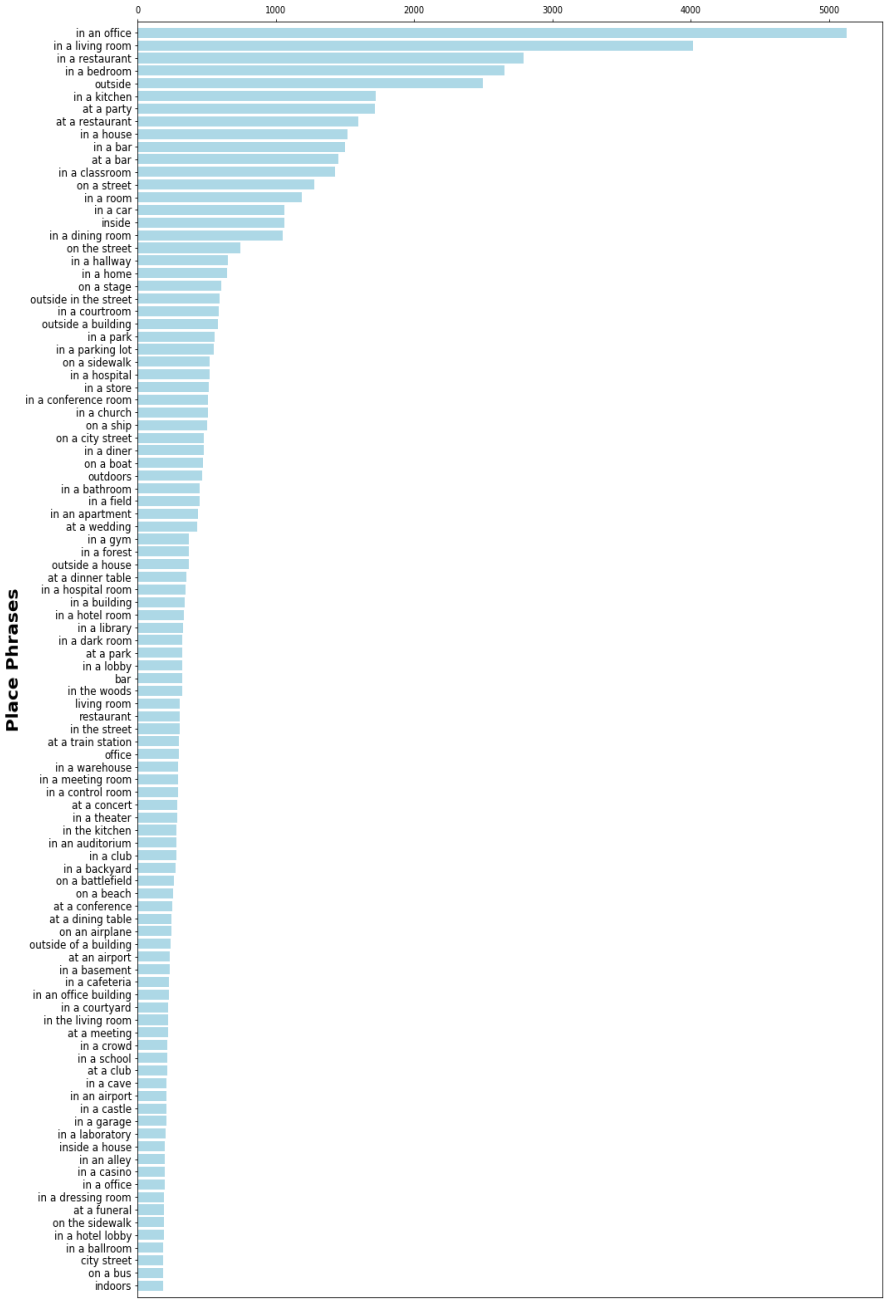
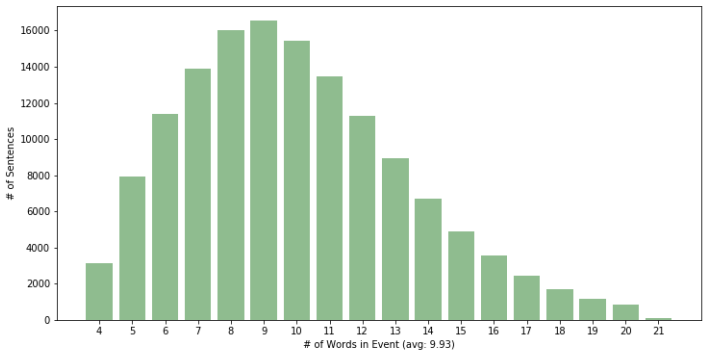
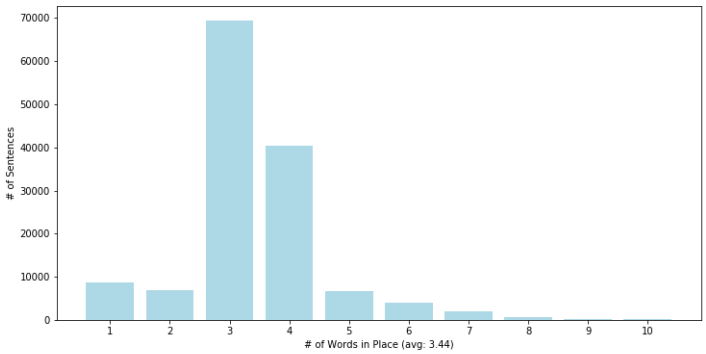


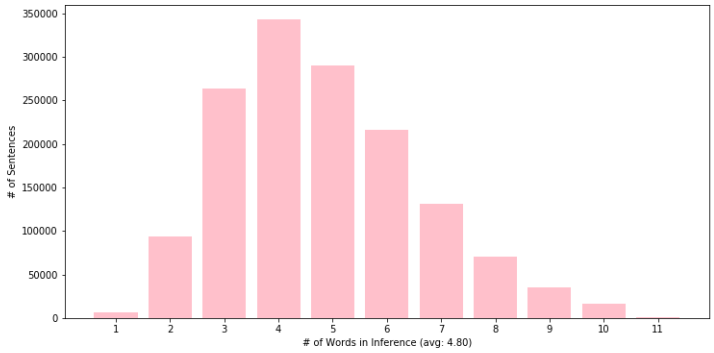
Fig. 10: Place Phrases



(a) Number of Words in Event



(b) Number of Words in Place



(c) Number of Words in Inference

Fig. 11: Sentence Length

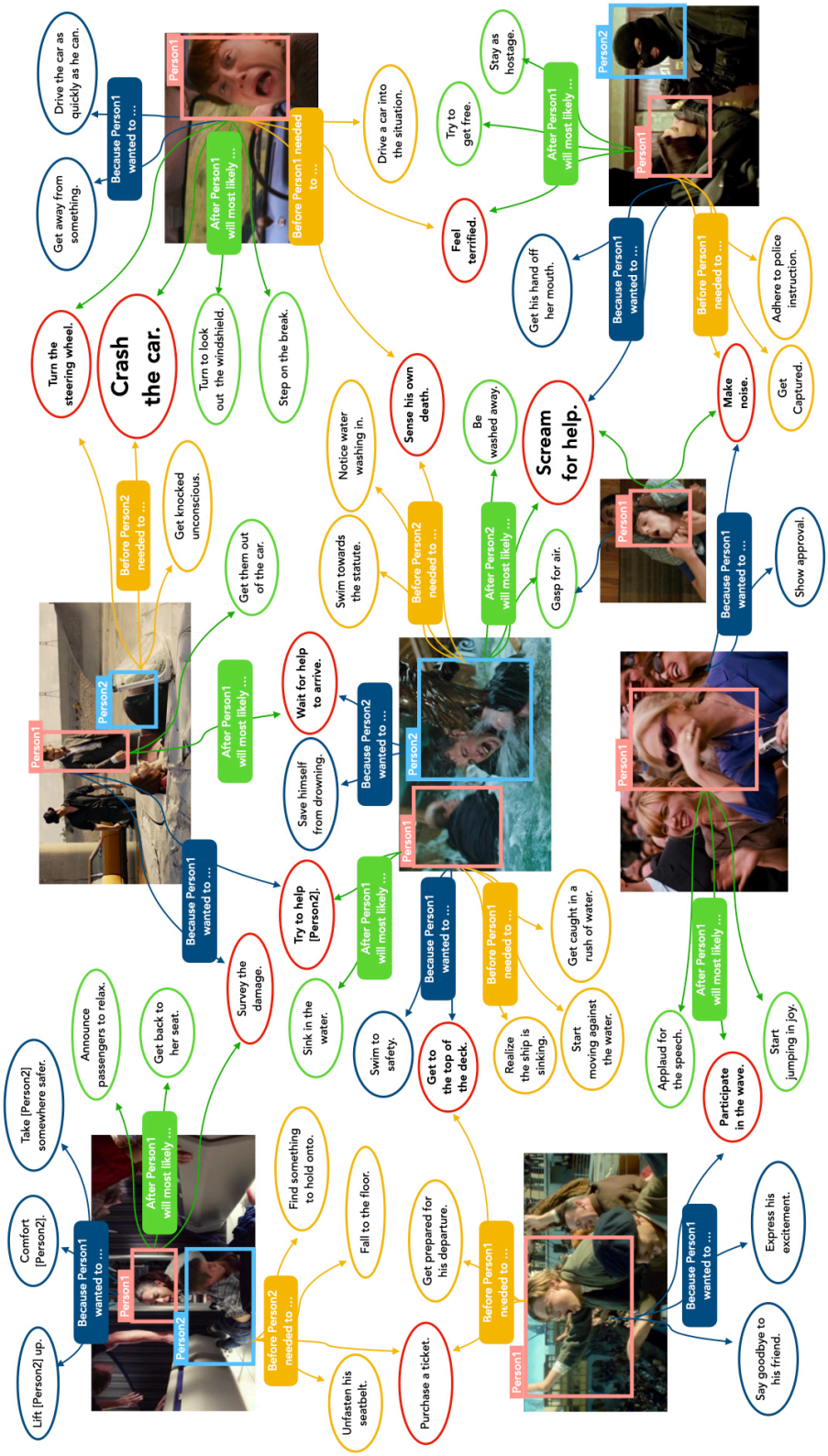


Fig. 12: Overview of our Visual Commonsense Graphs

References

1. 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) 4, 10
2. Holtzman, A., Buys, J., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv (2019) 5
3. 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) 4, 10
4. 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) 8
5. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (2004) 8
6. 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) 6, 7, 8
7. 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) 7
8. 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) 8
9. 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) 6