A. IntentQA dataset

A.1. More Details about the Annotation Process

In order to ensure that all of the QA pairs we select are relevant to intent, we hired workers on Amazon Mechanical Turk for manual annotation. We design four questions to help with the annotation:

Q1: According to the given video and QA, whether the annotated ‘Action’ represents a physical action, such as an obvious body movement (e.g., grab, throw) or static body pose (e.g., keep, stay)?

Q2: Whether the annotated ‘Action’ actually occurs in the video?

Q3: Whether the annotated ‘Action’ is performed by humans?

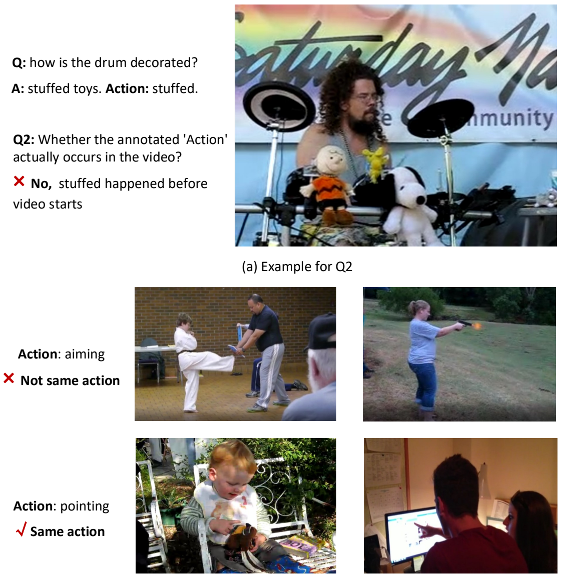
Q4: The annotated ‘Action’ labels are the same in these two videos, but are the corresponding physical action similar with each other?

Q1, Q2 and Q3 ensure that the action we control is a physical action that can be observed in the video, completed by a person, rather than mental behavior, animal behavior, or other actions mentioned but not occurring in the video, etc. This ensures that the social intents retained in the selected QAs are triggered by specific observable actions. An example of Q2 is given in Fig. 6.a, where ‘stuffed toys’ is something that has happened before the video starts, so it cannot be considered an action that occurs in the video. Q4 ensures that when we compare two samples, we determine that it is the same actions triggering the intents that are compared, not just the same words for different actions. An example of Q4 is given in Fig. 6.b, where the two actions of ‘aiming’ shown in the figure are considered to be very different, not one action, even though the words describing them are both ‘aiming’. The ‘pointing’ shown in the figure obviously describes the same action. We hope that the actions referred to in our selected QAs for comparison are physically similar, but the underlying intents are different only due to the different contexts in the videos.

A.2. Dataset Statistics

As shown in Fig. 7, we have collected statistics on the number of questions per video (see Table 1). The total number of videos is 4303. The question numbers of the majority of videos lie between 1 and 7. The distribution is relatively balanced, with the peak at around 3-4 questions.

As shown in Fig. 9, we have collected data on the length of questions and answers. Most answers are between 1-5 words, with a median of 3 words. The lengths of answers for different types of questions are similar, and there is no significant difference in the lengths of answers for any particular type of question. Most questions are between 5-22 words, with the most common number of words being 11-14. *TP&TN* questions have a significantly longer average length than *CW* and *CH* questions due to the need to describe two events for each action. *CH* questions have the shortest average length, while *CW* questions have a more balanced length distribution.


Q: how is the drum decorated?
A: stuffed toys. Action: stuffed.
Q2: Whether the annotated 'Action' actually occurs in the video?
✕ No, stuffed happened before video starts
Action: aiming
✕ Not same action
Action: pointing
√ Same action

(a) Example for Q2

(b) Example for Q4

Figure 6: Example of Q2 and Q4 used in human annotation. Q2 ensures that the action mentioned in the video is happening in the video, not an action that has already happened or is going to happen outside the video. Q4 ensures that the same word actually represents the same action, not totally different physical actions of a polysemy.

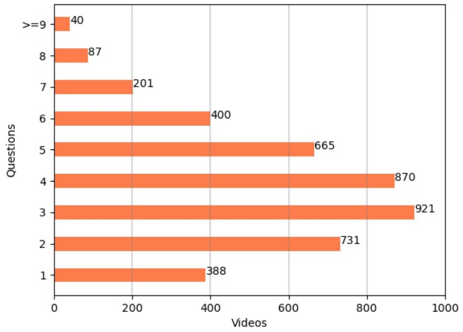
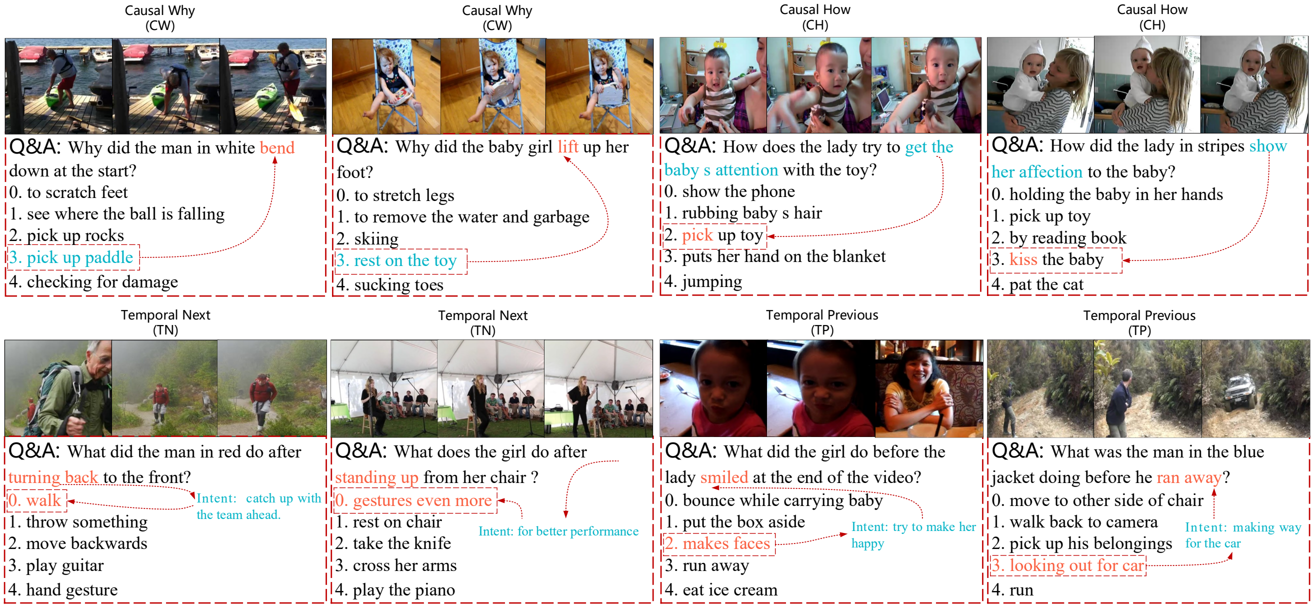


Figure 7: Distribution of question per video (Section 3).

As shown in Fig. 10, we have collected data on the distribution of action’s Lemmatized Verbs in the dataset. The distribution of action’s Lemmatized Verbs shows two extremes: some very common actions have a large number of available comparison samples, such as ‘hold’, ‘move’, ‘look’, ‘put’, and ‘point’, and these corresponding intentions in these different samples are also more diverse. However, many other actions have fewer comparison samples,


Causal Why (CW)

Q&A: Why did the man in white bend down at the start?
0. to scratch feet
1. see where the ball is falling
2. pick up rocks
3. pick up paddle
4. checking for damage

Causal Why (CW)

Q&A: Why did the baby girl lift up her foot?
0. to stretch legs
1. to remove the water and garbage
2. skiing
3. rest on the toy
4. sucking toes

Causal How (CH)

Q&A: How does the lady try to get the baby s attention with the toy?
0. show the phone
1. rubbing baby s hair
2. pick up toy
3. puts her hand on the blanket
4. jumping

Causal How (CH)

Q&A: How did the lady in stripes show her affection to the baby?
0. holding the baby in her hands
1. pick up toy
2. by reading book
3. kiss the baby
4. pat the cat

Temporal Next (TN)

Q&A: What did the man in red do after turning back to the front?
0. walk
1. throw something
2. move backwards
3. play guitar
4. hand gesture

Intent: catch up with the team ahead.

Temporal Next (TN)

Q&A: What does the girl do after standing up from her chair ?
0. gestures even more
1. rest on chair
2. take the knife
3. cross her arms
4. play the piano

Intent: for better performance

Temporal Previous (TP)

Q&A: What did the girl do before the lady smiled at the end of the video?
0. bounce while carrying baby
1. put the box aside
2. makes faces
3. run away
4. eat ice cream

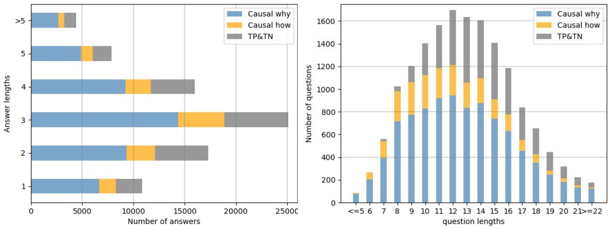
Intent: try to make her happy

Temporal Previous (TP)

Q&A: What was the man in the blue jacket doing before he ran away?
0. move to other side of chair
1. walk back to camera
2. pick up his belongings
3. looking out for car
4. run

Intent: making way for the car

Figure 8: More examples about the dataset. The red box frames the correct answer. Actions are colored in red while intents are colored in blue.

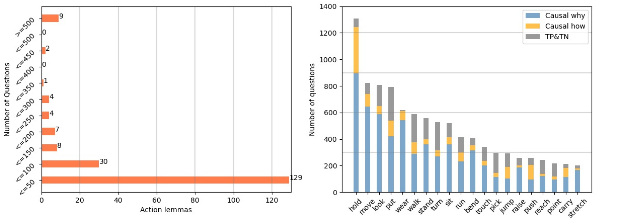


(a) Distribution of answers

(b) Distribution of questions

Figure 9: (a) Distribution of answer length. We showed the number of answers for each length and the proportion of questions belonging to each type, and it can be seen that answers with length 3 are the most numerous. (b) Distribution of question length. We showed the number of questions for each length and their type proportion, and it can be seen that the most questions are concentrated in 11-14 words.

with most actions having fewer than fifty comparison video samples, and thus the corresponding intentions are much fewer. For some actions, there are even some formulas or rules, such as 'smile/clap' corresponding to 'expressing a happy/appreciative emotion' and 'sit' corresponding to 'for rest'. It is unreasonable to require a balanced distribution of action's Lemmatized Verbs, as it reflects the diversity of intentions behind the actions, which is naturally unbalanced.



(a) Distribution of action lemmas

(a) Distribution of Top20 action lemmas

Figure 10: (a) Distribution of action's Lemmatized Verbs. It shows that most action's Lemmatized Verbs have fewer than 50 questions, while some common action's Lemmatized Verbs have a large number of questions, with 9 action's Lemmatized Verbs having more than 500 questions. (b) Distribution of the top 20 action's Lemmatized Verbs with the most questions. It shows that 'hold' has the most questions, with more than 1300 questions.

A.3. More Examples of Datasets

As shown in Fig. 8, We provide more examples of four types of questions with different actions. The 'causl why' type of questions ask the intention according to the action, the 'causal how' questions ask the action according to the intention, and the 'TP&TN' questions connect two actions with the intention.

B. Experimental Details and Hyperparameter Settings.

For both the multiple-choice and open-ended QA, we use the answer accuracy as the metric. We utilize VGT [64] to build video region graph by sparsely sampling 8 clips from each video, with each clip containing 4 frames, and each frame containing 10 regions detected by Faster RCNN [48]. We adopt the default number of Video Graph layers of VGT and the default hyperparameters of Dynamic Graph Transformer (DGT). We use Adam optimizer with the initial learning rate set to 3×10^{-5} under a cosine annealing schedule. We trained for 30 epochs with a batchsize of 64. The penalty coefficient λ is empirically set to 0.85. In 'The top-k nodes from the cross-modal graph', k is empirically set to 3. The margin in the triplet loss function formula (Eq. (12)) is empirically set to 2.2.