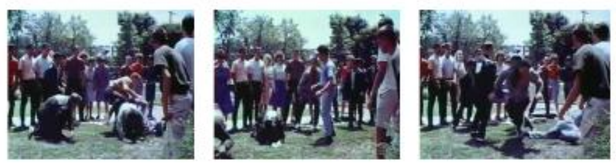
Ground truth: Two women are shopping in a store.
Two girls are shopping.
Baseline model: A man is doing a monkey in a store.
Multi-task model: A woman is shopping in a store.



Ground truth: Two men are fighting.
A group of boys are fighting.
Baseline model: A group of men are dancing.
Multi-task model: Two men are fighting.

(a)

Ground truth: A woman slices a shrimp tail.
A girl is cutting a fish tale.
Baseline model: A person is cutting the something.
Multi-task model: A woman is cutting a piece of meat.

Ground truth: Two men are talking aggressively.
The boy is talking.
Baseline model: A man is crying.
Multi-task model: A man is talking.

(b)

Ground truth: A monkey and a deer are fighting.
A gazelle is fighting with a baboon.
Baseline model: A man is walking on the ground.
Multi-task model: A monkey is walking.

Ground truth: A dog climbs into a dryer.
A dog is in a washing machine.
Baseline model: A man is playing.
Multi-task model: A man is playing with a toy.

(c)

Figure 5: Examples of generated video captions on the YouTube2Text dataset: (a) complex examples where the multi-task model performs better than the baseline; (b) ambiguous examples (i.e., ground truth itself confusing) where multi-task model still correctly predicts one of the possible categories (c) complex examples where both models perform poorly.

	Relevance	Coherence
Not Distinguishable	70.7%	92.6%
SotA Baseline Wins	12.3%	1.7%
Multi-Task Wins (M-to-M)	**17.0%**	**5.7%**

Table 5: Human evaluation on YouTube2Text video captioning.

	Relevance	Coherence
Not Distinguishable	84.6%	98.3%
SotA Baseline Wins	6.7%	0.7%
Multi-Task Wins (M-to-1)	**8.7%**	**1.0%**

Table 6: Human evaluation on entailment generation.

the multi-task models are always better than the strongest baseline for both video captioning and entailment generation, on both relevance and coherence, and with similar improvements (2-7%) as the automatic metrics (shown in Table 1).

5.5 Analysis

Fig. 5 shows video captioning generation results on the YouTube2Text dataset where our final M-to-M multi-task model is compared with our strongest attention-based baseline model for three categories of videos: (a) complex examples where the multi-task model performs better than the baseline; (b) ambiguous examples (i.e., ground truth itself confusing) where multi-task model still correctly predicts one of the possible categories (c) complex examples where both models perform poorly. Overall, we find that the multi-task model generates captions that are better at both temporal action prediction and logical entailment (i.e., correct subset of full video premise) w.r.t. the ground truth captions. The supplementary also provides ablation examples of improvements by the 1-to-M video prediction based multi-task model alone, as well as by the M-to-1 entailment based multi-task model alone (over the baseline).

On analyzing the cases where the baseline is better than the final M-to-M multi-task model, we find that these are often scenarios where the multi-task model's caption is also correct but the baseline caption is a bit more specific, e.g., "a man is holding a gun" vs "a man is shooting a gun".

Finally, Table 7 presents output examples of our entailment generation multi-task model (Sec. 5.3), showing how the model accurately learns to produce logically implied subsets of the premise.

Given Premise	Generated Entailment
a man on stilts is playing a tuba for money on the boardwalk	a man is playing an instrument
a child that is dressed as spiderman is ringing the doorbell	a child is dressed as a superhero
several young people sit at a table playing poker	people are playing a game
a woman in a dress with two children	a woman is wearing a dress
a blue and silver monster truck making a huge jump over crushed cars	a truck is being driven

Table 7: Examples of our multi-task model's generated entailment hypotheses given a premise.

6 Conclusion

We presented a multimodal, multi-task learning approach to improve video captioning by incorporating temporally and logically directed knowledge via video prediction and entailment generation tasks. We achieve the best reported results (and rank) on three datasets, based on multiple automatic and human evaluations. We also show mutual multi-task improvements on the new entailment generation task. In future work, we are applying our entailment-based multi-task paradigm