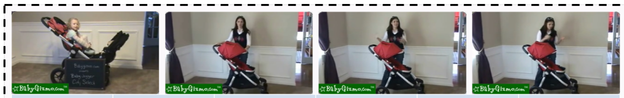
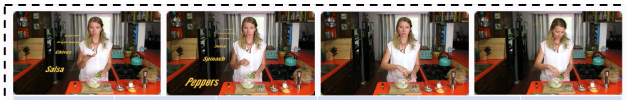
Supplementary for Hierarchical Modular Network for Video Captioning



Object#0	woman	lady	female	feminie	wife
	0.992	0.889	0.861	0.704	0.682
Object#1	girl	daughter	female	schoolgirl	girlfriend
	0.945	0.792	0.718	0.704	0.680
Object#2	person	human	someone	somebody	personnel
	0.930	0.835	0.792	0.732	0.721
Object#3	stroller	baby	pram	infant	cradle
	0.899	0.696	0.691	0.660	0.608
Object#4	stroller	pram	baby	infant	toddler
	0.899	0.701	0.666	0.630	0.589
Object#5	stroller	baby	infant	pram	toddler
	0.886	0.772	0.688	0.717	0.709
Object#6	child	kid	baby	youngster	infant
	0.821	0.773	0.737	0.717	0.709
Object#7	cradle	item	gear	stuff	outfit
	0.638	0.614	0.608	0.606	0.598

Ground truth: a woman is talking about a baby stroller
Ours: a woman is demonstrating a stroller

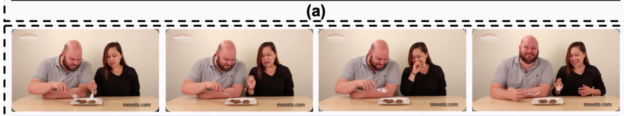
(a)



Object#0	woman	lady	female	feminie	wife
	0.995	0.901	0.864	0.710	0.686
Object#1	girl	female	woman	lady	girlfriend
	0.926	0.810	0.778	0.756	0.752
Object#2	person	human	someone	somebody	personnel
	0.878	0.775	0.751	0.718	0.683
Object#3	cooking	food	dish	dishware	cook
	0.786	0.762	0.760	0.731	0.729
Object#4	cooking	dish	food	dishware	cookware
	0.782	0.758	0.752	0.733	0.727
Object#5	cooking	dish	food	dishware	cookware
	0.782	0.758	0.752	0.733	0.727
Object#6	kitchen	pot	dish	bowl	cookware
	0.657	0.638	0.636	0.635	0.627
Object#7	container	pot	bowl	dish	item
	0.653	0.652	0.645	0.626	0.625

Ground truth: a woman is explaining how to prepare a dish
Ours: a woman in a kitchen is talking about how to prepare a dish

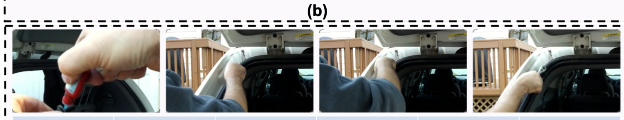
(b)



Object#0	woman	lady	female	feminie	wife
	0.993	0.892	0.868	0.710	0.681
Object#1	man	male	gentleman	men	guy
	0.992	0.776	0.758	0.715	0.663
Object#2	girl	daughter	female	schoolgirl	girlfriend
	0.945	0.758	0.744	0.728	0.707
Object#3	food	meal	cuisine	snack	dish
	0.925	0.835	0.783	0.699	0.692
Object#4	food	meal	cuisine	snack	lunch
	0.918	0.816	0.767	0.675	0.666
Object#5	person	human	someone	somebody	people
	0.880	0.818	0.754	0.723	0.707
Object#6	food	meal	cuisine	snack	lunch
	0.867	0.767	0.736	0.671	0.667
Object#7	item	food	stuff	meal	dish
	0.630	0.623	0.622	0.613	0.602

Ground truth: a man and women tasting food
Ours: a man and a woman are eating food

(c)



Object#0	man	male	gentleman	men	guy
	0.992	0.765	0.752	0.690	0.659
Object#1	person	human	someone	somebody	personnel
	0.965	0.844	0.818	0.760	0.739
Object#2	car	automobile	vehicle	sedan	suv
	0.929	0.854	0.835	0.767	0.682
Object#3	car	automobile	vehicle	sedan	coupe
	0.904	0.833	0.823	0.749	0.686
Object#4	car	vehicle	automobile	sedan	coupe
	0.879	0.813	0.811	0.729	0.695
Object#5	car	vehicle	automobile	coupe	sedan
	0.788	0.764	0.736	0.690	0.672
Object#6	person	woman	lady	human	female
	0.764	0.760	0.749	0.721	0.694
Object#7	gear	equipment	vehicle	apparatus	coupe
	0.690	0.676	0.675	0.666	0.652

Ground truth: a person is fixing a car
Ours: a man is fixing a car

(d)

Figure 1. Illustration of principal objects predicted by the entity module. See Section 1 for more details.

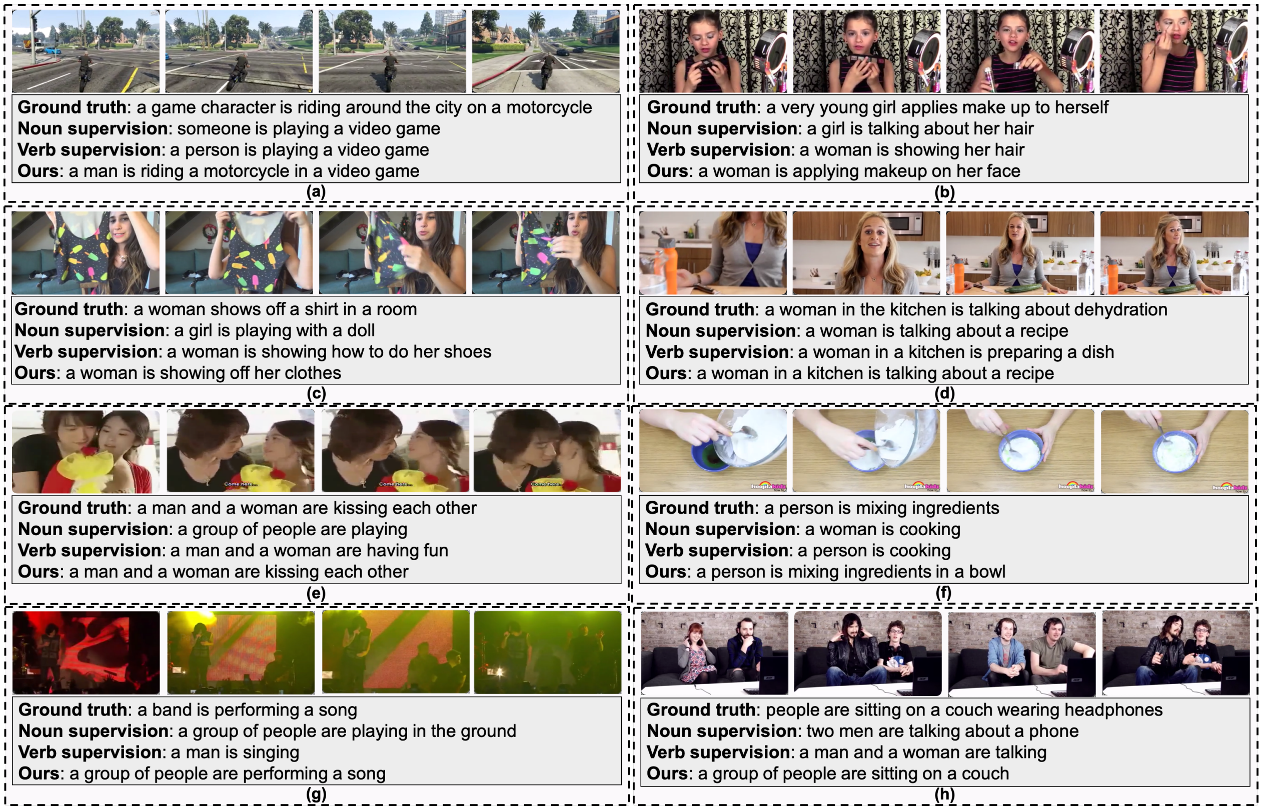


Figure 2. Examples of generated captions using two variants of our model. For comparison convenience, we also present the ground truth and the result of our model.

1. Illustration of Principal Objects

We show four MSR-VTT examples of what our entity module can learn in Figure 1. Since $\bar{\mathcal{E}} = \{\bar{e}_i\}_{i=1}^N$ are predicted linguistic embeddings of N principal objects, we compute the cosine similarity between $\bar{e}_i$ and each *entity* words in vocabulary. The top-5 results are presented.

In Figure 1 (a), $\bar{e}_0$ captures the entity "woman" in the video, which serves as the *subject* in the generated caption; $\bar{e}_1$ and $\bar{e}_6$ capture the entity "girl" and "child", which are discarded by other modules when generating the final caption; $\bar{e}_3$, $\bar{e}_4$ and $\bar{e}_5$ capture the "stroller", which serves as the *object* in the generated caption. In Figure 1 (b), $\bar{e}_0$ captures the *subject* "woman"; $\bar{e}_3$, $\bar{e}_4$ and $\bar{e}_5$ capture the *object* "dish" on the cutting broad; $\bar{e}_6$ captures the adverbial modifier "kitchen". Although many other video objects, such as knifes, paintings, cutting broad, and gas stove, also appear in the video, our entity module does not adopt them as principal objects. This demonstrates that our proposed entity module has the ability to select those video objects which are likely to be mentioned in captions.

2. Examples of Using Different Supervisions

We show the captions generated by two model variants, i.e., "noun supervision" and "verb supervision", on eight MSR-VTT videos in Figure 2, where the "noun supervision" replaces the *entity* with broader *nouns* to supervise our entity module, and the "verb supervision" replaces the *predicate* with *verb* to supervise our predicate module. We also present the generated captions of our model for comparison.

We observe that the captions generated by our model contain richer and more accurate content than the two variants. For instance, in Figure 2 (a), "noun supervision" misses the "motorcycle", and "verb supervision" generates *predicate* "playing video game" rather than more accurate "riding motorcycle". Similarly, in Figure 2 (b), "noun supervision" focuses on the girl's hair rather than her face. "verb supervision" incorrectly generates "showing her hair", without realizing that the action is "applying makeup". Our model outperforms "noun supervision" by generating more accurate *entities*. This is because *abstract nouns* in the ground truth, such as "hunger" and "satisfaction", have no corresponding video objects, and thus introduce noise for generating captions. Meanwhile, our model can predict more accurate video actions than "verb supervision" because using the *predicate* as supervision can keep the agreement between *verbs* and *verbs'* recipients.