Look Before you Speak: Visually Contextualized Utterances
Supplementary Material

Paul Hongsuck Seo Arsha Nagrani Cordelia Schmid
Google Research
{phseo,anagrani,cordelias}@google.com

In this document, we describe further experiments ablating our model (Section A), provide further insight into the experimental set up of the tasks used in the paper (Section B), display more qualitative results (Section C), and analyse the effect of ASR noise in HowToFUP with a brief manual study (Section D).

A. Further Ablations on HowToFUP

We ablate our model described in Section 4 of the main paper, varying (i) the number of CoTRM blocks S (described in Section 4.1.3 of the main paper), (ii) the MLM loss, and (iii) the visual input feature type (scene features only vs. combined features) in Table A. With scene features only, and without the MLM loss, performance degrades rapidly as S is increased (almost 4% drop). Using combined features (object and scene) however, prevents this performance drop, as does adding in the MLM loss. Adding both together, gives the best performance with $S = 4$, suggesting that the gains are complementary.

Table A: Ablations of our network on HowToFUP. We vary the value of S and show results with and without the masked language modeling (MLM) loss.

		w/o MLM		w/ MLM	
Methods	S	R@1	R@5	R@1	R@5
Scene features only	1	65.43	86.52	66.73	87.10
	2	65.64	86.67	67.13	87.42
	4	61.05	83.58	67.15	87.44
Combined features	1	66.74	87.30	67.70	87.95
	2	**66.82**	**87.38**	67.79	88.00
	4	66.53	87.18	**68.34**	**88.28**

B. Configurations in Different Tasks

In the main paper, we show results for 3 different tasks, Future Utterance Prediction, Next Step Prediction, and Video Question Answering. Here we describe the different setups for each one, in particular the inputs and outputs of our model (depicted visually in Figure A).

Future utterance prediction In the default configuration for FUP, our model ingests video frames and transcribed speech, and expects a set of next utterance candidates. Our model then returns a score for each candidate computed by a dot product of the input multimodal feature with each candidate.

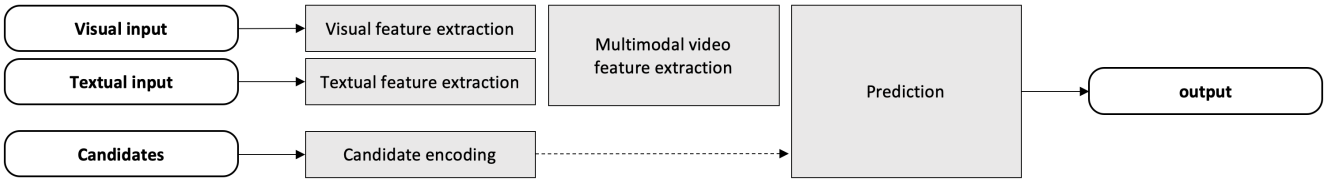
Next step prediction Since this task is formulated as a classification task, instead of encoding a set of FUP candidates, we use a two-layered classifier for prediction. Inputs are video frames and transcribed speech, and output is a softmax 735-way classifier prediction. Note that the new classifier module in this task cannot be initialized with the pretrained weights on HowToFUP and is trained from scratch.

Video question answering The goal here is to answer a question given an input video. Compared to the other tasks, there is an input question in addition to video frames and corresponding transcribed speech. We simply concatenate this additional input question to the transcript and feed the concatenated string as a single textual input to our model. Note that some videos do not contain any speech and, in such cases, the model simply takes the question only.

It is common to formulate VideoQA as a classification task using the most frequent answers as target classes. We instead adopt the formulation of answer ranking as in FUB where all possible answers in the training set are encoded using the candidate encoder and scored by the softmax normalized dot-product. In other words, we extract all possible answers from the training set, treat them as candidate answers, and select the best candidate using the same type of a candidate encoder as in FUP.

C. Further Qualitative Results

We present additional qualitative examples for HowToFUP in Figure B and C.



Task	Visual Input	Textual Input	Candidates	Prediction module	Output
FUP	video frames	transcript	100 pre-selected candidates	dot-product	scores for utterance candidates
NSP	video frames	transcript	None	two-layered classifier	scores for step classes
VideoQA	video frames	transcript + question	all answers in dataset	dot-product	scores for all answers in dataset

Figure A: Task-specific input output configurations for CoMVT. Note that for NSP, we do not use a candidate encoder since the prediction module is a classifier.

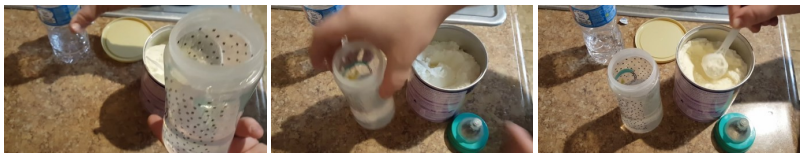
Inputs (video frames and utterances)	Prediction (future utterance)	
 **Transcript**: This should get him through the night if we're lucky. Usually he sleeps of the night.	*We'll see.* *So I pushed it forward and I kind of I had to five freehand.* *So we need three scoops of the **formula**.* *I just thought it was interesting that we have some volunteer corn coming up.* . . .	✓
 **Transcript**: Everyone was asking me why have you not been to him this stead of stuff? I don't know what else came in to me yesterday.	*This is a jewelry making tool.* *I keep the seam gauge in my toolbox.* *I saw the **fish** in my fridge, and I thought what should I do with this **fish** now, and I remember Wow in Nigeria.* *I said hello just Curtis or as well be should be sue and I put in another ticket as it stated.* . . .	✓
 **Transcript**: So I decided to make a small scar for my cow and you know, one of those things that you put around your neck not a move on just the one that goes around your neck.	*I am super full right now, but this is also good.* *So you're going to **knit the first stitch** bring your yarn to the front as to make a yarn over and you **knit** the next two together.* *Now the reason why I chose snakeskin, it's because I've read I've been doing a lot of research on it, but I read that snakes can't have a lot of properties for the skin.* *He got to this stick to that bag to this stick.* . . .	✓
 **Transcript**: It's the best sugar then you just that'll put some sugar in there to sweeten it up. That's a beautiful way to start your day.	*You want to make it diet use sweet and low beautiful **noodle kugel**.* *They were like, oh it was a lot of fun.* *Um, the only reason why I want to do this weight loss drink today, I get the request for it.* *You can finish eating buddy.* . . .	✓

Figure B: **Qualitative results on HowToFUP.** On the right, we show the results of the baseline model that uses text inputs only (highlighted in red) and our multimodal model (highlighted in green). The GT utterance has a ✓ next to it. In many of these cases, the correct future utterance refers to an object which can only be known from the visual context (highlighted in bold). Note how both the speech and the visual frames provide complementary information, that we cannot learn from a single modality alone, helping to paint a complete picture. The ASR has mistakes ('scar' in row 3 should refer to 'scarf').

Inputs (video frames and utterances)	Prediction (future utterance)	
 **Transcript**: The fact is, you know, got who capacity on the turkey map and the isolated scriptures on it just means it's quite an ideal community.	*They have a tipple chassé ten minutes will be I said, it's three people three hungry people to hungry people and poor little people.* *So I thought we would give her a try this week.* *So that's it anyway an overview of the **extension lead with six amperes**.* *Now, what I'm doing is I've got this section of four inch pipe right here.* . . .	✓
 **Transcript**: You can cut into a queue about half inch thickness and say about one and half inch long but doesn't matter was the shape. Sometimes they're too big for your leave or too small for your lip.	*It comes with caps.* *So it was that kind of a medium-high.* *You can add more **meat**.* *You're going to continue to pick these and for gonna further a further like decorating.* . . .	✓
 **Transcript**: It is 5 1 0 to 9 3 nice and light and fresh in terms of the fragrance, which is so lovely..	*I'll show you how to change it by an industrial scene machine.* *You're actually going to love **washing those delicates** again.* *I mean if I was, you know, not in a hurry or whatever and I was the scent spray it on let it set for I actually tend to one time.* *It's a violent death.* . . .	✓
 **Transcript**: So instead of that I decided to try this sounds too good to be true method to hopefully achieve the same results.	*So I'm gonna go ahead and add 2/3 of our **chocolate** to the bowl and we're just gonna do this by eye and like I said, we will reserve 1/3 of the **chocolate** to add later and to melt the **chocolate** what we'...* *Nine ten.* *So I just put it all together all it's going to be be 2.2 percent.* *It's loose and falls.* . . .	✓
 **Transcript**: So the cut ends are secured because it's tackled inside we're good.	*So let me grab my real dressed hands to study the **stitches** are around like one inch apart.* *We got strawberry pina colada.* *So you reposition your hand on the other side to do the same thing twist up and back now we flip it on its back.* *So that's the theory anyway, would you just put it straight in the soil?* . . .	✓

Figure C: **Further qualitative results on HowToFUP:** On the right, we show the results of the baseline model that uses text inputs only (highlighted in red) and our multimodal model (highlighted in green). The GT utterance has a ✓ next to it.

D. ASR error analysis in HowToFUP

It is a well known fact that the HowTo100M dataset is noisy, and because ASR is obtained via an automatic method, there is the potential for error. We briefly investigate the quality of transcripts by *manually correcting* ASR mistakes in the next utterances of 100 random samples from HowToFUP. We observe a word error rate of 3.3%, and note that at a sentence level - 80.0% of the automatic transcripts are correct while the others contain only small (nominal) mistakes (*e.g.*, *want it* $\rightarrow$ *wanted*). To quantify the impact of these ASR errors on our model performance, we evaluate our model on the 100 samples with and without these corrected transcripts for the task of future utterance prediction. The model selects the identical candidates in both settings except for 2 samples (98 correct). This indicates that the few errors introduced by ASR do not make a significant difference, particularly at scale.