

# Supplementary Material: A Simple Baseline for Audio-Visual Scene-Aware Dialog

Idan Schwartz<sup>1</sup>, Alexander Schwing<sup>2</sup>, Tamir Hazan<sup>1</sup>  
<sup>1</sup>Technion      <sup>2</sup>UIUC

`idanschwartz@gmail.com, aschwing@illinois.edu, tamir.hazan@technion.ac.il`

In this supplementary we present additional qualitative results in Fig. 1. We observe the attention to focus on the unique aspects of the video.

## References

	<p>Q: what is on the tv ?            GT: i can 't see but . he picks up the remote for the tv at the end of the video            Ours: i can 't see, but it is hard to tell what is on the tv  <b>q+h+vgg-spatial+audio:</b> i can 't really tell what is on the tv but i can 't see what is on the tv  <b>i3d-rgb-temporal:</b> i can 't tell what is on the tv  <b>q+h+att:</b> i can 't see what is on the tv  <b>baseline:</b> i can 't tell what is in the video</p>		<p>Q: does she open the cabinet ?            GT: she does , but you don 't see her doing it . just leaning over .            Ours: no , she does not open the cabinet .  <b>q+h+vgg-spatial+audio:</b> no , she doesn 't open the door .  <b>i3d-rgb-temporal:</b> no , she doesn 't open the cabinet .  <b>q+h+att:</b> no , she does not open the door .  <b>baseline:</b> no , she does not open the door .</p>
	<p>Q: how does the story end ?            GT: the man sneezes several times and walk out of the doorway .            Ours: the video ends with him walking out of the room .  <b>q+h+vgg-spatial+audio:</b> he walks out of the room and the video ends .  <b>i3d-rgb-temporal:</b> it ends with him still holding the box .  <b>q+h+att:</b> the video ends with the man sitting on the bed .  <b>baseline:</b> the video ends with the man walking down the stairs .</p>		<p>Q: was he already in the room or did he walk into the room ?            GT: he was already in the room .            Ours: he was already in the room when the video started .  <b>q+h+vgg-spatial+audio:</b> he was already in the room .  <b>i3d-rgb-temporal:</b> he was already in the room at the beginning of the video .  <b>q+h+att:</b> he was already in the room at the beginning of the video  <b>baseline:</b> he is already in the room the whole time .</p>
	<p>Q: is she unhappy she has sandwich ?            GT: yes , she looks like she is unhappy about having that to eat .            Ours: no , she doesn 't have any expression  <b>q+h+vgg-spatial+audio:</b> yes , she has a sandwich on it .  <b>i3d-rgb-temporal:</b> no , she is not taking anything out of the video .  <b>q+h+att:</b> yes , that 's what she does .  <b>baseline:</b> no , she is not doing anything in the video</p>		<p>Q: how old would you say that she is ?            GT: i think that she 's in her mid 20s to early 30s            Ours: she appears to be in her mid to late 20 ' s early 30 ' s  <b>q+h+vgg-spatial+audio:</b> i would say in her twenties  <b>i3d-rgb-temporal:</b> she looks to be in her mid twenties .  <b>q+h+att:</b> i would say she is in her late 20 ' s or early 30 ' s .  <b>baseline:</b> she appears to be in her late 20 ' s or early 30 ' s or early 30 ' s .</p>
	<p>Q: is the camera stationary ?            GT: yes , the camera is stationary .            Ours: yes , the camera is stationary .  <b>q+h+vgg-spatial+audio:</b> yes , the camera is stationary .  <b>i3d-rgb-temporal:</b> no , the camera is still in the background .  <b>q+h+att:</b> yes , the camera is stationary at the end of the video .  <b>baseline:</b> no , the camera stay s on the whole time .</p>		<p>Q: did he ever move form that spot ?            GT: no he didnt move from there            Ours: no he did not do that  <b>q+h+vgg-spatial+audio:</b> no he didn't leave the room  <b>i3d-rgb-temporal:</b> no he didnt speak in the video  <b>q+h+att:</b> nothing else in the video  <b>baseline:</b> no he didnt leave the room</p>
	<p>Q: what room does this see to be ?            GT: it looks like a bedroom .            Ours: it appears to be a bedroom .  <b>q+h+vgg-spatial+audio:</b> it appears to be a bedroom .  <b>i3d-rgb-temporal:</b> it appears to be a living room .  <b>q+h+att:</b> it looks like a living room .  <b>baseline:</b> it looks like a living room .</p>		<p>Q: what does she do after rinsing out her mouth ?            GT: the video doesn 't seem to show what happens next .            Ours: she walks over to a mirror and then sneezes  <b>q+h+vgg-spatial+audio:</b> she walks out of the room and the video ends .  <b>i3d-rgb-temporal:</b> she walks out of the room and the video ends  <b>q+h+att:</b> she walks over to the window and turns off the light .  <b>baseline:</b> she takes a bite of the sandwich .</p>
	<p>Q: what is the first thing that happens ?            GT: the first thing is the man sitting down .            Ours: the man walks into the room and sits down .  <b>q+h+vgg-spatial+audio:</b> the man is sitting on the floor .  <b>i3d-rgb-temporal:</b> the man walks into the room and sits on the bed .  <b>q+h+att:</b> a man walks into the room .  <b>baseline:</b> the man walks into the room with a broom .</p>		<p>Q: okay ,so the man is in the living room eating a sandwich and drinking ,he runs into the bathroom and looks into a mirror ?            GT: no he was in kitchen i think and runs into living room            Ours: yes that 's all he does in the video  <b>q+h+vgg-spatial+audio:</b> yes that is what happens in the video  <b>i3d-rgb-temporal:</b> yes that is what happens in the video  <b>q+h+att:</b> yes that is what happens in the video  <b>baseline:</b> yes , that is what happens in the video</p>
	<p>Q: what is the man sweeping ?            GT: the man is sweeping the entry way floor .            Ours: he is cleaning the floor with a broom .  <b>q+h+vgg-spatial+audio:</b> the man is holding a broom .  <b>i3d-rgb-temporal:</b> he is sweeping the floor .  <b>q+h+att:</b> the man is sitting in a chair .  <b>baseline:</b> he is looking at his phone .</p>		<p>Q: so he is basically checking himself out in the mirror ?            GT: yes he is , he is posing and bending his knees to look at himself            Ours: yes he seems to be looking at himself in the mirror  <b>q+h+vgg-spatial+audio:</b> yes that 's all he does in the video  <b>i3d-rgb-temporal:</b> yes that is all that he does in the video  <b>q+h+att:</b> yes that is what happens in the video  <b>baseline:</b> yes that is what he does</p>
	<p>Q: what does she do with the bottle ?            GT: she sets it on the counter            Ours: she puts it down on the counter  <b>q+h+vgg-spatial+audio:</b> she puts it down on the table  <b>i3d-rgb-temporal:</b> she puts it on the table and then puts it on the table  <b>q+h+att:</b> she picks it up and puts it on the table  <b>baseline:</b> she puts it on the table</p>		<p>Q: does she look happy ?            GT: yes she looks like she is very entertained by what she is looking at            Ours: yes she does look happy  <b>q+h+vgg-spatial+audio:</b> yes she seems very happy  <b>i3d-rgb-temporal:</b> no she is not happy or sad  <b>q+h+att:</b> she does not look happy or sad  <b>baseline:</b> no she does not have any emotion</p>

Figure 1: Additional qualitative results.