What to Say and When to Say it: Live Fitness Coaching as a Testbed for Situated Interaction – Author Response –

1 Additional Results

Table 1: Evaluation of the effect of pre-training the baseline STREAM-VLM on FIT-COACH dataset (fitness feedbacks and fitness questions from the short-clips).

Method	METEOR↑	ROUGE-L↑	BERT↑	LLM-Acc.↑	T-F-Score↑
STREAM-VLM (w/o Pre-training)	0.095	0.087	0.858	2.16	0.52
STREAM-VLM	**0.125**	**0.116**	**0.863**	**2.56**	**0.59**

Table 2: Evaluation of models fine-tuned with the FIT-COACH dataset on the FIT-COACH benchmark. (†indicates results of non-interactive models evaluated at regular intervals, *indicates human evaluation is conducted on a smaller set of 200 feedbacks.)

Evaluation Model	Socratic-Llama-2-7B†	Video-ChatGPT [36] (fine-tuned)†	STREAM-VLM	STREAM-VLM (w/o 3D CNN)	STREAM-VLM (w/o Action-Tokens)†
Human*	2.63	2.59	**2.80**	2.51	2.71
Mixtral-Instruct-0.1 [25]	2.39	2.42	**2.56**	2.17	2.56
LLaMA-3-8B-Instruct [**?**]	1.74	1.82	**1.90**	1.62	1.89
LLaMA-3-70B-Instruct [**?**]	2.17	2.33	**2.45**	2.11	2.41

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.