

Supplementary Information

A multi-task video reasoning dataset across semantic, spatial, and temporal categories with four difficulty levels

Yiqing Shen, Chenjia Li, Chenxiao Fan, Mathias Unberath

Bracketed numbers in the tables below (e.g. [22]) refer to the reference list of the main article.

Supplementary Table 1. Performance comparison on reasoning segmentation queries from RVTBench, where our proposed RVTagent demonstrates substantial improvements over existing approaches. “ft” stands for fine-tuned model.

Model	$\mathcal{J}$ (%) ($\uparrow$)												$\mathcal{F}$ (%) ($\uparrow$)												$\mathcal{J\&F}$ (%) ($\uparrow$)															
	Semantic				Spatial				Temporal				Semantic				Spatial				Temporal				Semantic				Spatial				Temporal							
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4				
LISA-7B [4]	10.77	10.86	11.06	10.43	10.24	10.09	10.60	10.26	21.59	13.97	13.63	11.29	7.21	7.21	7.42	6.55	6.69	6.57	7.07	6.45	16.42	9.74	9.23	7.01	8.99	9.04	9.24	8.49	8.47	8.33	8.83	8.36	19.06	11.86	11.43	9.15				
LISA-13B [4]	11.49	11.79	11.82	10.89	11.49	10.69	11.35	10.61	22.80	16.08	14.14	11.81	7.88	8.18	8.13	7.30	7.71	7.25	7.80	7.13	16.53	11.42	9.65	7.58	9.69	9.98	9.97	9.09	9.60	8.97	9.57	8.87	19.66	13.75	11.89	9.70				
LISA++ [11]	10.34	10.49	10.79	9.87	9.93	9.71	10.44	9.54	24.79	15.07	13.07	11.10	6.72	6.78	7.08	6.23	5.79	6.18	6.95	6.03	18.14	10.53	8.67	6.90	8.53	8.63	8.94	8.05	7.86	7.94	8.69	7.79	21.46	12.80	10.87	9.00				
VISA [5]	3.94	3.55	3.34	2.97	4.54	3.89	3.77	3.13	7.41	2.93	2.95	2.57	8.89	9.80	9.56	9.02	10.13	9.79	9.06	8.74	13.08	10.34	9.13	9.54	6.42	6.53	6.45	6.00	7.34	6.84	6.42	5.94	10.25	6.67	6.04	6.06				
SegLLM [22]	9.52	9.48	9.62	9.37	9.57	9.16	9.44	9.14	19.95	12.90	11.72	10.64	12.99	13.17	13.22	12.92	12.93	12.87	13.03	12.65	24.40	16.82	15.69	14.49	11.25	11.33	11.42	11.14	11.25	11.01	11.24	10.90	22.18	14.86	13.71	12.57				
SAM4MLLM [23]	10.37	11.38	11.54	11.88	11.63	11.05	11.14	11.53	15.89	13.23	13.34	13.05	7.33	8.20	8.49	8.79	8.44	7.91	8.12	8.51	12.30	9.40	9.49	9.35	8.85	9.79	10.02	10.34	10.03	9.48	9.63	10.02	14.10	11.31	11.42	11.20				
LaSagnA [24]	7.77	7.88	7.97	7.02	9.25	7.69	7.86	6.89	14.78	10.49	9.73	7.48	10.8	10.95	11.21	10.23	12.26	10.71	11.01	10.03	20.75	14.26	13.38	11.22	9.29	9.42	9.59	8.63	10.75	9.20	9.43	8.46	17.76	12.37	11.56	9.35				
LISA-7B (ft) [4]	45.77	44.86	42.06	40.43	43.24	42.09	40.60	38.26	41.59	38.97	36.63	34.29	42.21	41.21	39.42	36.55	40.69	39.57	38.07	35.45	39.42	35.74	33.23	31.01	43.99	43.04	40.74	38.49	41.97	40.83	39.34	36.86	40.51	37.36	34.93	32.65				
LISA-13B (ft) [4]	48.49	47.79	45.82	43.89	46.49	45.69	44.35	42.61	44.80	42.08	39.14	36.81	45.88	45.18	43.13	41.30	44.71	43.25	41.80	40.13	42.53	39.42	36.65	34.58	47.19	46.49	44.48	42.60	45.60	44.47	43.08	41.37	43.67	40.75	37.90	35.70				
VISA (ft) [5]	51.20	49.45	47.70	45.15	48.75	47.20	45.55	43.10	44.40	41.85	39.55	37.20	49.10	47.70	46.05	43.55	47.25	45.80	44.15	41.70	42.60	40.05	37.75	35.40	50.15	48.58	46.88	44.35	48.00	46.50	44.85	42.40	43.50	40.95	38.65	36.30				
RVTagent	74.80	72.65	70.30	67.95	72.40	70.15	67.85	65.40	68.75	66.20	63.90	61.55	72.60	70.45	68.10	65.75	70.20	67.95	65.65	63.20	66.55	64.00	61.70	59.35	73.70	71.55	69.20	66.85	71.30	69.05	66.75	64.30	67.65	65.10	62.80	60.45				

Supplementary Table 2. Performance comparison on reasoning grounding task with RVTBench.

Model	cIoU (%) (↑)												gIoU (%) (↑)												AP@50 (%) (↑)											
	Semantic				Spatial				Temporal				Semantic				Spatial				Temporal				Semantic				Spatial				Temporal			
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4
LISA-7B [4]	8.58	8.71	9.09	8.89	8.68	8.52	8.73	8.85	10.02	10.29	9.64	8.97	9.28	9.37	9.06	9.11	9.62	9.46	9.39	9.07	11.67	10.51	10.30	9.02	6.03	6.20	6.41	5.70	6.42	6.06	6.04	5.63	8.35	7.67	7.12	5.78
LISA-13B [4]	9.22	9.24	9.66	8.97	9.55	9.19	9.32	8.86	11.04	11.40	10.21	9.01	9.81	10.01	10.10	9.37	10.13	10.01	9.94	9.34	11.79	11.36	10.36	9.45	6.67	6.96	6.93	5.99	7.23	7.05	6.89	5.91	9.60	8.61	7.02	6.16
LISA++ [11]	8.68	9.18	9.56	9.62	8.77	8.84	9.07	9.74	13.69	11.09	9.76	10.40	9.15	9.14	9.53	8.70	9.21	9.18	9.44	8.72	12.40	10.63	10.02	9.21	6.13	6.53	6.88	5.92	6.21	6.62	6.79	5.96	9.60	8.07	7.69	6.39
VISA [5]	13.30	14.41	14.79	14.47	12.37	14.48	15.98	14.41	4.67	16.20	14.30	15.48	14.97	16.05	15.57	16.25	15.28	15.48	16.05	16.40	13.25	17.08	16.09	15.97	7.50	9.10	8.47	8.91	9.47	8.21	8.84	9.05	7.28	7.65	9.12	8.58
SegLLM [22]	8.94	9.77	10.01	10.60	10.07	9.52	9.64	10.69	13.51	11.81	10.85	11.90	11.07	11.39	11.47	11.44	11.70	11.53	11.51	11.40	15.19	12.67	12.47	11.84	9.60	9.81	9.98	9.68	10.30	9.88	10.05	9.60	14.78	11.45	11.14	10.14
SAM4MLLM [23]	8.67	9.39	9.43	9.77	9.04	9.35	8.90	9.74	6.75	7.88	9.53	10.68	10.85	11.26	11.52	11.71	10.92	11.29	11.33	11.57	12.52	12.18	11.98	12.47	8.27	8.90	9.19	9.26	8.38	8.78	8.99	9.09	11.16	10.17	9.74	10.18
LaSagnA [24]	10.18	9.92	10.01	9.25	10.06	9.92	9.75	9.22	10.52	10.47	10.39	9.71	10.55	10.26	10.14	9.38	10.65	10.36	10.21	9.43	12.45	11.55	11.06	9.68	8.03	7.85	7.83	6.44	8.16	8.07	7.79	6.49	9.50	9.30	8.46	6.72
LISA-7B (ft) [4]	42.85	41.20	39.45	37.60	40.30	38.85	37.20	35.45	36.65	35.10	33.25	31.70	40.95	39.30	37.55	35.70	38.40	36.95	35.30	33.55	34.75	33.20	31.35	29.80	38.45	36.80	35.05	33.20	35.90	34.45	32.80	31.05	32.25	30.70	28.85	27.30
LISA-13B (ft) [4]	45.95	44.30	42.55	40.70	43.40	41.95	40.30	38.55	39.75	38.20	36.35	34.80	44.05	42.40	40.65	38.80	41.50	40.05	38.40	36.65	37.85	36.30	34.45	32.90	41.55	39.90	38.15	36.30	39.00	37.55	35.90	34.15	35.35	33.80	31.95	30.40
VISA (ft) [5]	49.05	47.40	45.65	43.80	46.50	45.05	43.40	41.65	42.85	41.30	39.45	37.90	47.15	45.50	43.75	41.90	44.60	43.15	41.50	39.75	40.95	39.40	37.55	36.00	44.65	43.00	41.25	39.40	42.10	40.65	39.00	37.25	38.45	36.90	35.05	33.50
RVTagent	69.85	67.40	65.25	62.90	67.30	65.05	62.80	60.55	63.75	61.50	59.35	57.10	67.95	65.50	63.35	61.00	65.40	63.15	60.90	58.65	61.85	59.60	57.45	55.20	65.45	63.00	60.85	58.50	62.90	60.65	58.40	56.15	59.35	57.10	54.95	52.70

Supplementary Table 3. Performance comparison on reasoning VQA and summary tasks with RVTBench.

Task	Model	BLEU-4 (%) (†)												ROUGE-L (%) (†)												BertScore (%) (†)												CIDEr (%) (†)												
		Semantic				Spatial				Temporal				Semantic				Spatial				Temporal				Semantic				Spatial				Temporal																
		L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4													
VQA	GPT-o4-mini	2.25	1.15	0.80	0.42	1.63	1.13	0.90	0.45	0.00	1.21	0.58	0.39	29.99	22.07	20.23	19.02	29.38	21.89	20.41	19.25	26.82	21.30	19.61	18.74	74.25	71.57	68.98	69.04	72.87	71.40	69.16	69.28	72.20	71.34	68.80	69.34	70.76	49.97	25.50	9.74	46.92	43.33	26.44	10.36	56.07	31.16	20.39	10.87	
VQA	Gemini-2.0-flash-lite	1.13	0.76	0.37	0.20	1.54	0.63	0.45	0.20	0.00	0.33	0.30	0.13	21.27	17.24	14.61	13.99	23.04	17.38	14.88	14.35	22.21	15.96	14.63	13.50	65.28	63.20	57.90	54.17	63.96	62.67	57.97	54.38	65.03	60.35	57.29	53.77	34.05	24.72	5.55	1.07	33.84	21.18	6.34	1.34	46.97	14.98	5.11	0.88	
VQA	Claude3-Haiku	0.94	0.66	0.35	0.33	0.83	0.60	0.36	0.35	0.00	0.27	0.33	0.21	18.48	16.44	13.72	13.80	19.36	16.31	13.97	14.03	12.63	14.42	13.45	12.92	65.41	65.09	62.44	62.31	65.42	64.86	62.71	62.53	62.34	65.31	62.31	62.04	9.62	8.84	0.83	0.13	0.80	5.44	1.03	0.17	0.00	1.29	0.46	0.03	
VQA	Qwen2.5-omni [25]	0.89	0.48	0.40	0.41	1.28	0.59	0.44	0.42	0.00	0.53	0.45	0.35	17.72	13.44	12.71	13.64	20.73	13.48	12.89	13.88	10.48	12.99	12.42	13.08	61.74	59.57	57.60	58.39	62.43	59.48	57.86	58.52	55.65	60.16	57.64	58.40	6.09	3.35	1.25	0.35	8.09	3.84	1.21	0.43	0.00	1.90	0.93	0.22	
VQA	Janus-Pro-7B [26]	2.92	1.72	1.30	1.18	3.52	1.87	1.47	1.30	0.00	1.15	1.23	1.06	28.40	21.58	19.23	19.36	30.04	21.59	19.70	19.92	20.60	19.81	19.14	18.92	72.32	69.99	67.12	66.50	71.29	69.93	67.48	66.72	70.55	69.04	66.71	66.29	74.46	54.38	22.17	10.25	62.72	53.62	23.69	11.02	22.47	35.36	17.67	9.28	
VQA	LISA-7B [4]	2.66	1.26	0.70	0.54	2.32	1.52	0.82	0.58	0.00	0.58	0.77	0.43	24.97	19.15	14.24	14.28	24.26	19.26	14.49	14.56	15.05	15.92	14.59	13.97	70.01	67.22	61.69	62.89	68.08	67.15	61.86	63.01	66.79	66.14	62.58	62.83	74.25	46.61	11.56	0.69	51.09	45.80	12.98	0.68	19.73	28.93	10.28	0.67	
VQA	LISA-13B [4]	2.72	1.64	0.74	0.39	2.82	1.90	0.88	0.39	0.00	0.80	0.69	0.28	25.12	19.79	15.01	13.52	23.59	19.75	15.46	13.67	18.63	17.71	14.91	13.08	70.80	68.60	63.34	62.21	68.09	68.36	63.73	62.21	68.96	68.20	62.88	62.07	73.90	53.43	14.83	0.41	49.55	51.78	15.97	0.47	23.24	33.67	9.08	0.45	
VQA	LISA++ [11]	2.50	1.24	0.75	0.46	3.03	1.19	0.91	0.49	0.00	0.59	0.60	0.33	25.88	19.20	15.63	14.47	28.39	19.01	16.20	14.67	16.36	16.43	14.70	13.87	70.55	67.67	63.35	62.00	70.05	67.69	63.82	62.05	67.68	66.35	62.59	61.93	72.89	45.88	15.75	3.78	55.86	42.75	18.14	4.12	28.55	27.28	11.91	4.12	
VQA	RVTagent	7.23	6.54	5.87	5.10	6.84	6.18	5.56	5.23	4.82	5.65	5.08	4.76	42.86	39.75	37.40	35.64	41.25	38.70	36.85	35.72	39.58	37.46	35.28	34.53	84.37	82.60	80.42	79.35	83.18	81.54	79.87	78.95	82.40	80.82	79.45	78.65	92.48	85.63	77.41	71.28	88.35	81.94	75.78	70.45	83.26	79.52	73.40	68.75	
Summary	GPT-o4-mini	0.47	0.61	0.74	0.43	0.60	0.76	0.76	0.42	0.25	0.88	0.71	0.40	20.67	20.66	21.41	21.22	21.11	20.90	21.32	21.12	18.43	20.45	20.74	20.61	68.73	69.14	69.33	69.63	68.96	69.21	69.24	69.58	67.18	68.59	68.36	68.88	7.96	6.61	6.45	4.71	7.48	6.01	6.13	4.41	1.86	5.38	2.94	3.24	
Summary	Gemini-2.0-flash-lite	0.62	0.69	0.79	0.42	0.59	0.71	0.79	0.42	0.32	0.70	0.76	0.40	18.72	18.61	19.90	17.53	18.36	18.51	19.85	17.45	17.77	17.75	19.34	17.41	61.16	61.45	62.52	60.26	61.28	61.28	62.32	60.22	60.34	60.90	61.78	59.71	3.55	2.73	1.73	0.29	2.86	1.85	1.55	0.30	0.61	1.72	0.68	0.30	
Summary	Claude3-Haiku	0.38	0.35	0.32	0.21	0.38	0.38	0.35	0.21	0.24	0.38	0.30	0.20	14.75	13.79	13.99	14.27	14.70	14.00	14.05	14.25	13.78	13.39	13.51	14.26	63.21	61.53	61.22	61.56	63.42	61.65	61.21	61.52	63.31	61.34	60.59	61.20	0.01	0.02	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Summary	Qwen2.5-omni [25]	0.50	0.86	1.07	0.98	0.52	0.87	1.08	0.98	0.43	0.86	1.03	0.98	17.00	17.41	19.13	20.17	16.88	17.46	19.14	20.10	16.33	16.72	18.63	20.14	60.75	61.34	61.83	63.28	60.85	61.23	61.83	63.26	60.78	60.83	61.38	63.45	1.91	1.61	2.11	0.53	1.38	1.64	1.89	0.54	0.81	0.57	1.76	0.49	
Summary	Janus-Pro-7B [26]	1.37	1.34	1.62	1.30	1.37	1.44	1.65	1.30	0.48	1.06	1.42	1.32	22.54	21.30	23.13	22.32	23.00	21.69	23.24	22.21	20.61	19.63	22.33	22.30	68.84	67.70	68.48	67.51	69.49	67.78	68.57	67.45	67.93	66.67	67.47	67.09	23.20	16.92	15.99	6.42	26.46	17.05	16.11	6.36	11.23	12.05	12.85	5.69	
Summary	LISA-7B [4]	0.33	0.35	0.45	0.29	0.52	0.40	0.47	0.28	0.41	0.49	0.53	0.33	11.94	11.01	11.19	8.58	14.56	11.76	11.65	8.59	14.82	13.28	12.71	9.54	58.06	57.01	56.87	53.71	61.25	57.75	57.23	53.72	62.23	59.43	58.12	54.40	4.12	2.17	0.87	0.01	7.18	2.34	0.78	0.01	4.22	2.50	1.10	0.01	
Summary	LISA-13B [4]	0.30	0.34	0.36	0.20	0.31	0.36	0.36	0.20	0.35	0.39	0.39	0.20	11.51	12.46	12.19	8.49	12.97	13.04	12.41	8.48	12.65	13.52	12.74	8.61	57.87	58.51	57.53	53.75	59.62	59.02	57.71	53.73	60.70	59.60	58.83	53.59	0.72	0.65	0.25	0.01	0.91	0.69	0.26	0.01	0.59	0.17	0.33	0.01	
Summary	LISA++ [11]	0.92	0.80	0.80	0.36	0.95	0.81	0.84	0.36	0.58	0.57	0.71	0.38	19.49	18.99	19.84	13.98	20.26	19.16	19.86	13.97	17.11	16.80	18.78	13.94	66.62	66.00	66.35	60.47	67.33	65.94	66.31	60.46	64.84	64.33	64.97	60.11	14.83	9.17	8.88	0.23	16.53	9.20	9.01	0.26	6.50	6.46	5.91	0.14	
Summary	RVTagent	5.68	5.23	4.95	4.56	5.42	5.08	4.83	4.50	4.95	5.12	4.78	4.42	38.76	36.94	35.87	34.25	37.85	36.18	35.40	34.10	36.73	35.42	34.86	33.94	82.45	80.87	79.35	78.10	81.78	80.24	78.86	77.95	80.63	79.14	78.23	77.48	85.76	79.45	72.83	67.40	82.54	76.32	71.45	66.84	77.38	73.52	70.18	65.72	