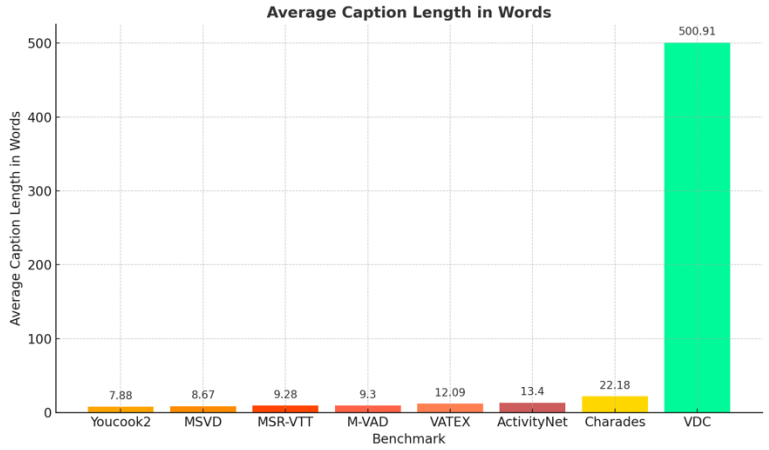
AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark

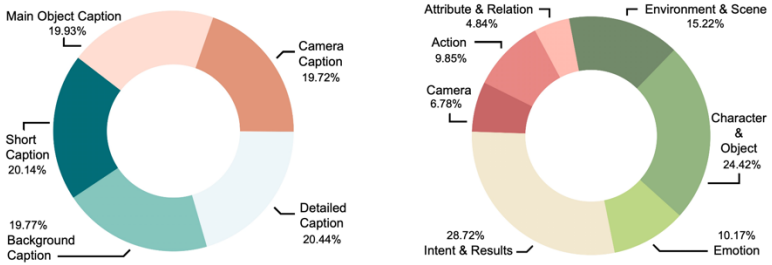
Wenhao Chai[1,2] Enxin Song[*] Yilun Du[4] Chenlin Meng[2,3] Vashisht Madhavan[2] Omer Bar-Tal[2] Jenq-Neng Hwang[1] Saining Xie[5] Christopher D. Manning[3]

[1]University of Washington [2]Pika Labs [3]Stanford University [4]Harvard University [5]New York University

New Task! Video Detailed Captioning



Video Source	# Sample	Proportion	Duration (sec.)	Ave. Length (sec.)	Ave. # Keyframe
Panda-70M (Chen et al., 2024e)	229	22.25%	5,714	24.95	7.18
Ego4D (Grauman et al., 2022)	196	19.05%	10,935	55.79	19.46
Mixkit (Mixkit, 2024)	197	19.14%	3,261	16.55	6.58
Pixabay (Pixabay, 2024)	199	19.34%	4,748	23.86	8.99
Pexels (Pexels, 2024)	208	20.21%	4,343	20.88	7.99
Total	1,027	–	29,001	28.18	10.43

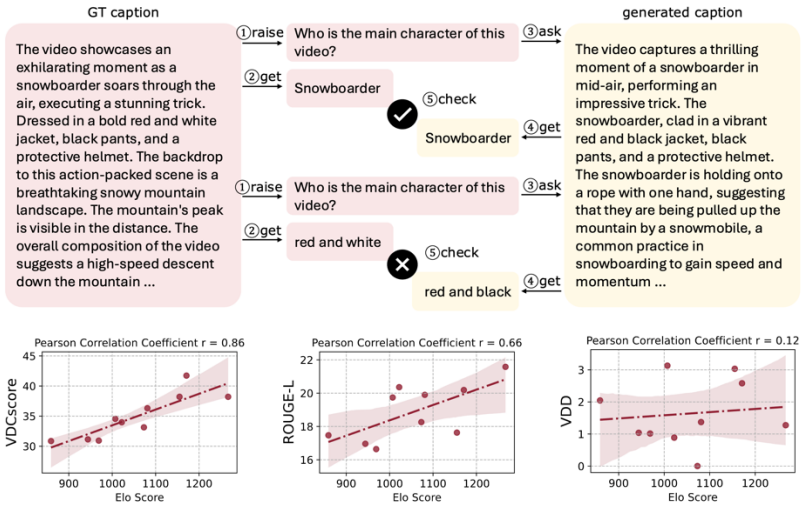


VDC Benchmark

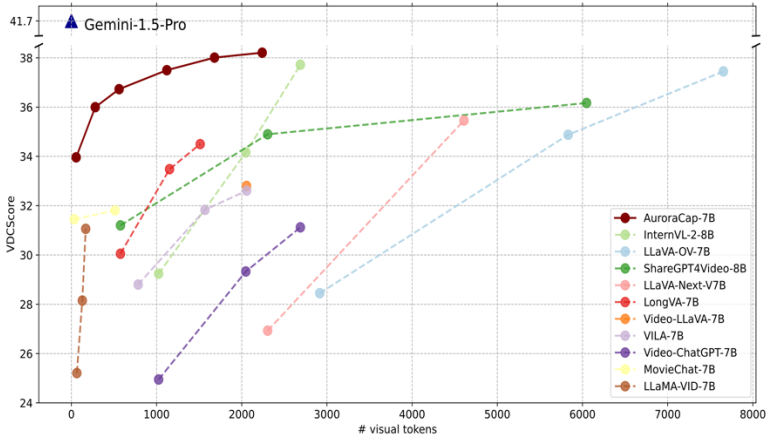
We care about both performance and efficiency.

Model Family	# F	TPF	Avg. VDCscore		Detailed		Camera		Short		Background		Object	
			Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
Aria-3.5B×8	64	128	42.5	2.2	47.3	2.4	42.4	2.2	33.2	1.8	40.9	2.1	48.9	2.5
Gemini-1.5 Pro	unk.	unk.	41.7	2.1	43.1	2.2	38.7	2.0	35.7	1.9	43.8	2.2	47.3	2.4
LLaVA-Video-7B	64	676	39.0	2.0	35.0	1.8	46.1	2.3	32.8	1.7	37.6	1.9	46.2	2.4
AuroraCap-7B	10	224	38.2	2.0	41.3	2.1	43.5	2.3	32.1	1.7	35.9	1.8	39.0	2.0
InternVL-2-8B	10	256	37.7	2.0	34.9	1.8	39.1	2.1	33.0	1.7	37.5	1.9	44.2	2.2
LLaVA-OV-7B	10	729	37.5	1.9	41.2	2.1	37.8	2.0	32.6	1.7	37.4	1.9	38.2	2.0
LLaVA-1.6-7B	10	576	36.5	1.9	36.5	1.9	36.5	1.9	31.9	1.6	37.6	1.9	36.0	1.9
ShareGPT4Vide...	10	576	36.2	1.9	35.6	1.8	33.3	1.8	39.1	1.9	35.8	1.8	37.1	1.9
LLaVA-1.6-13B	10	576	35.9	1.9	36.2	1.9	35.6	1.9	31.9	1.7	38.9	2.0	36.6	1.9
LLaVA-NeXT-V7B	8	576	35.5	1.9	33.8	1.8	39.7	2.1	30.6	1.6	36.5	1.9	36.5	1.9
LLaVA-1.5-13B	10	576	34.8	1.8	33.0	1.7	39.0	2.1	30.9	1.6	34.8	1.8	36.3	1.8
LongVA-7B	10	144	34.5	1.8	27.9	1.5	35.3	1.9	31.9	1.6	36.4	1.9	41.0	2.1
LLaVA-1.5-7B	10	576	34.0	1.8	33.4	1.7	38.4	2.0	28.6	1.5	34.9	1.8	34.6	1.7

VDCscore



AuroraCap Model



Powered by Token Merging

More than a thousand tokens per second — do we really need that many visual tokens to understand videos? NO! We need only few.