VidComposition: Can MLLMs Analyze Compositions in Compiled Video?

Yolo Y. Tang*, Junjia Guo*, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, JIng Bi, Zeliang Zhang, Pooyan Fazli, Chenliang Xu

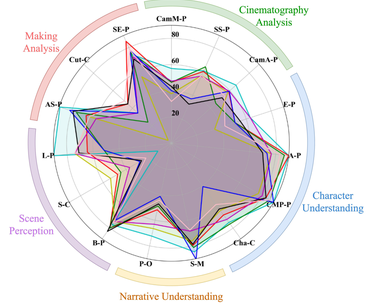
University of Rochester · Arizona State University

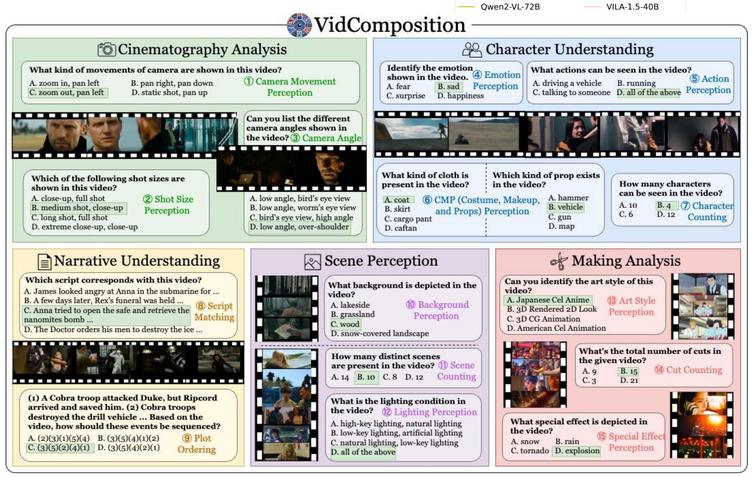
CVPR Nashville JUNE 11-15, 2025

Introduction

- We introduce VidComposition, a novel and high-quality benchmark for evaluating fine-grained video composition understanding in MLLMs.
- We comprehensively evaluate 33 MLLMs for video understanding with VidComposition. The results show the challenging nature of VidComposition and the substantial gap between MLLMs' and humans' capabilities in video composition understanding.
- We analyze the critical factors that influence the performance of MLLMs systematically, providing potential directions for model improvement and future advancements.

Benchmark	I/V	#Data	#Task	Compositional QA	Compiled Videos	Fine-Grained	Anno.
Winoground [49]	I	400	8	✓	-	✗	M
MME [11]	I	1.1k	14	✗	-	✓	M
MMBench [34]	I	1.7k	20	✗	-	✓	A+M
MMComposition [17]	I	4.3k	13	✓	-	✓	M
MSVD-QA [55]	V	504	5	✗	✗	✗	A
MSRVTT-QA [55]	V	2.9k	5	✗	✗	✗	A
TGIF-QA [19]	V	9.6k	4	✗	✗	✗	A
TVQA [22]	V	2.2k	8	✓	✓	✗	A+M
ActivityNet-QA [59]	V	5.8k	4	✗	✗	✗	M
NExT-QA [54]	V	1k	8	✗	✗	✗	A
AutoEval-Video [6]	V	327	9	✗	✗	✗	A+M
Video-Bench [41]	V	5.9k	10	✗	✗	✗	A+M
LVBench [51]	V	500	6	✗	✗	✗	M
MVBench [26]	V	3.6k	20	✗	✗	✓	A+M
MovieChat-1k [45]	V	100	8	✗	✓	✓	M
TempCompass [33]	V	410	4	✗	✗	✓	M
Video-MME [13]	V	900	12	✗	✗	✓	M
VIDCOMPOSITION	V	982	15	✓	✓	✓	M





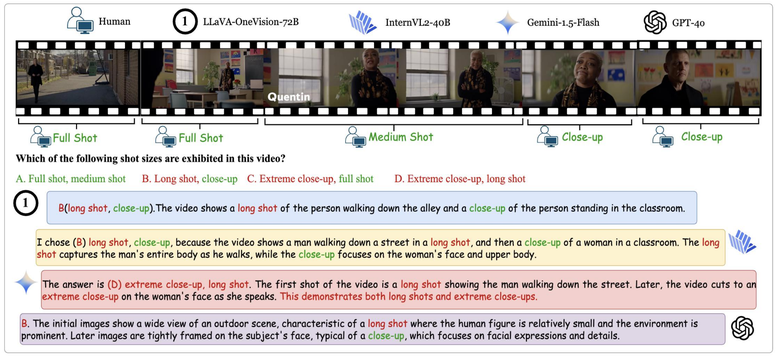
Main Results and Analysis

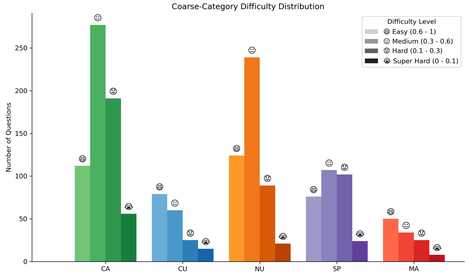
Method	Cinematography Analysis			Character Understanding				Narrative Underst.		Scene Perception			Making Analysis			Overall
	CamM-P	SS-P	CamA-P	E-P	A-P	CMP-P	Cha-C	S-M	P-O	B-P	S-C	L-P	AS-P	Cut-C	SE-P	
Human	84.1	85.4	80.0	82.6	92.3	92.9	94.1	97.0	97.5	94.4	80.2	81.8	85.7	87.5	94.7	86.26
LLaVA-OneVision-72B [24]	**57.1**	60.5	**66.3**	**63.3**	**90.0**	**90.0**	74.6	84.7	**72.4**	76.9	12.0	**90.3**	**89.5**	34.5	74.1	**63.31**
InternVL2-40B [7]	46.6	60.0	58.9	51.0	**90.0**	83.3	67.6	79.6	51.9	80.0	47.4	64.5	68.4	44.8	**85.2**	60.73
InternVL2-76B [7]	46.6	**63.9**	45.5	51.0	86.7	86.7	**76.1**	81.3	49.0	80.0	44.5	51.6	76.3	24.1	75.9	58.73
Qwen2-VL-72B [50]	37.9	56.6	57.9	34.7	76.7	76.7	63.4	76.2	66.5	73.8	53.6	74.2	65.8	3.4	55.6	58.68
Video-LLaMA2-72B [9]	47.5	56.1	59.4	**63.3**	76.7	80.0	71.8	68.5	63.2	73.8	36.8	74.2	60.5	34.5	72.2	58.62
InternVL2-8B [7]	55.3	56.6	59.4	44.9	80.0	83.3	59.2	67.2	40.6	78.5	32.5	64.5	52.6	31.0	72.2	54.63
GPT-4o [1]	40.2	37.1	59.4	51.0	73.3	**90.0**	40.8	**90.6**	41.4	72.3	27.8	51.6	81.6	34.5	77.8	52.93
Gemini-1.5-Flash [44]	43.4	32.7	52.0	55.1	70.0	48.4	62.0	78.7	47.3	**83.1**	26.3	71.0	78.9	44.8	70.4	52.40
VILA-1.5-40B [29]	32.0	56.6	54.5	42.9	83.3	86.7	57.7	67.7	44.4	75.4	22.5	74.2	65.8	48.3	77.8	51.23
GPT-4o mini [1]	33.8	49.8	50.5	49.0	80.0	**90.0**	31.0	79.6	41.4	66.2	26.8	61.3	76.3	20.7	79.6	50.23
Gemini-1.5-Pro [44]	33.8	51.7	51.5	51.0	73.3	80.0	70.4	47.7	36.4	73.8	37.3	71.0	84.2	**58.6**	75.9	49.36
Qwen2-VL-7B [50]	20.1	46.8	37.1	38.8	70.0	76.7	54.9	73.2	49.4	72.3	52.2	61.3	42.1	17.2	70.4	49.30
Oryx-7B [35]	34.7	54.1	57.4	57.1	80.0	73.3	66.2	48.5	34.7	73.8	41.6	61.3	39.5	20.7	66.7	48.77
Gemini-1.5-Flash-8B [44]	43.4	45.9	56.9	36.7	70.0	76.7	35.2	69.8	36.0	73.8	26.8	48.4	71.1	24.1	64.8	48.59
Video-LLaMA2.1 [9]	44.3	35.6	39.6	51.0	76.7	83.3	50.7	60.9	45.2	75.4	35.4	58.1	36.8	20.7	81.5	47.77
VideoChat2 [26]	24.2	58.0	42.1	44.9	66.7	83.3	60.6	62.6	27.6	73.8	**55.0**	35.5	50.0	10.3	59.3	47.36
InternVL2-26B [7]	33.3	47.8	39.6	55.1	76.7	83.3	56.3	68.9	33.9	76.9	25.4	45.2	34.2	41.4	75.9	46.42
LongVA [62]	24.7	41.0	48.0	40.8	70.0	73.3	42.3	51.9	32.2	72.3	42.6	48.4	52.6	34.5	70.4	43.73
MiniCPM-V2.6 [56]	28.3	43.4	43.6	53.1	73.3	80.0	50.7	59.1	23.0	75.4	22.0	71.0	57.9	20.7	72.2	42.49
InternVL2-4B [7]	27.4	42.9	26.2	32.7	66.7	73.3	49.3	60.4	28.0	78.5	41.6	35.5	44.7	10.3	72.2	41.68
Video-LLaMA2.1-AV [9]	27.4	45.9	38.6	55.1	73.3	76.7	47.9	46.4	30.1	81.5	25.8	45.2	34.2	34.5	83.3	41.50
VILA-1.5-8B [29]	31.5	40.0	37.6	51.0	63.3	66.7	40.8	40.9	26.8	70.8	37.8	41.9	60.5	44.8	59.3	40.21
GPT-4-turbo [1]	23.7	37.1	35.1	46.9	63.3	80.0	25.4	54.9	36.4	50.8	29.7	64.5	39.5	44.8	70.4	39.85
LongLLaVA [53]	28.3	37.1	27.2	24.5	60.0	56.7	54.9	48.1	32.6	61.5	38.3	41.9	26.3	24.1	66.7	38.45
Kangaroo [31]	29.2	42.0	24.3	30.6	56.7	66.7	57.7	31.5	26.8	67.7	47.8	61.3	21.1	6.9	55.6	37.10
InternVL2-2B [7]	23.7	24.4	24.8	36.7	76.7	63.3	53.5	48.9	21.8	80.0	40.2	29.0	47.4	6.9	83.3	36.75
LongVILA [29]	25.1	35.6	40.6	40.8	80.0	60.0	38.0	32.8	25.1	76.9	20.6	64.5	50.0	37.9	79.6	36.46
AuroraCap [5]	34.7	42.0	40.6	46.9	53.3	56.7	35.2	35.3	22.2	63.1	22.5	32.3	50.0	31.0	59.3	36.28
Qwen2-VL-2B [50]	21.0	29.3	25.2	30.6	63.3	70.0	42.3	50.6	23.8	67.7	37.3	74.2	34.2	24.1	63.0	36.16
Video-LLaMA2-7B [9]	25.1	29.3	23.3	30.6	70.0	66.7	40.8	31.5	26.4	72.3	40.2	29.0	44.7	27.6	66.7	34.35
VILA-1.5-3B [29]	20.1	32.7	38.1	51.0	53.3	46.7	31.0	26.4	10.5	72.3	35.4	32.3	36.8	10.3	83.3	31.95
Video-LLaVA [28]	26.5	25.9	38.1	32.7	53.3	40.0	25.4	26.8	23.0	55.4	30.1	38.7	21.1	51.7	51.9	31.07
Chat-UniVi [20]	25.1	31.7	30.2	30.6	53.3	30.0	22.5	26.4	24.3	29.2	21.1	32.3	34.2	44.8	40.7	28.02
InternVL2-1B [7]	24.7	24.9	22.8	22.4	46.7	33.3	22.5	26.4	26.8	30.8	30.6	22.6	23.7	34.5	29.6	26.61
RANDOM	26.0	25.8	25.3	24.3	23.7	23.7	24.8	25.0	25.3	27.4	24.4	20.3	24.7	25.2	28.7	25.33

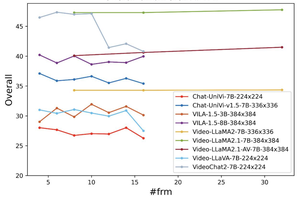
Models	#frm	LLM size	Res.	CA	CU	NU	SP	MA	Overall
Chat-UniVi [20]; Video-LLaVA [28]; VideoChat2 [26]	8	7B	224	31.68	42.22	31.02	40.11	42.98	34.94
Chat-UniVi-v1.5 [20]; LongVA [62]; Video-LLaMA2 [9]			336	33.6 +1.92	48.89 +6.67	32.42 +1.4	**44.26** +4.15	50.96 +7.98	38.04 +3.1
Video-LLaMA2.1 [9]; Video-LLaMA2.1-AV [9]			384	**36.9** +3.3	**55.28** +6.39	**46.0** +13.58	43.11 −1.15	**53.72** +2.76	**43.7** +5.66
Chat-UniVi [20]; Video-LLaVA [28]; VideoChat2 [26]	16	7B	224	29.66	39.81	31.86	28.52	34.71	31.52
Chat-UniVi-v1.5 [20]; LongVA [62]; Video-LLaMA2 [9]			336	32.75 +3.09	49.07 +9.26	32.21 +0.35	**44.92** +16.4	49.04 +14.33	37.67 +6.15
Video-LLaMA2.1 [9]; Video-LLaMA2.1-AV [9]			384	**37.06** +4.31	**57.22** +8.15	**45.68** +13.47	43.44 −1.48	**54.13** +5.09	**43.96** +6.29
LongVA [62]; Video-LLaMA2 [9]	32	7B	336	31.39	46.67	35.26	**44.26**	53.72	37.98
Video-LLaMA2.1 [9]; Video-LLaMA2.1-AV [9]			384	**38.5** +7.11	**59.72** +13.05	**45.47** +10.21	42.95 −1.31	**54.55** +0.83	**44.64** +6.66

Model	#frm	Res.	LLM size	Data volume	CA	CU	NU	SP	MA	Overall
Chat-UniVi [20]	8	224	7B	0.65M	25.56	34.44	23.37	25.9	36.36	26.73
VideoChat2 [26]				2M	**39.46** +13.9	**57.78** +23.34	**44.84** +21.47	**58.03** +32.13	**50.41** +14.05	**47.01** +20.28
Chat-UniVi-v1.5 [20]	8	336	7B	1.27M	37.38	47.78	26.53	37.38	46.28	36.11
LongVA [62]				1.32M	**37.54** +0.16	**51.67** +3.89	**41.89** +15.36	**49.51** +12.13	**56.2** +9.92	**43.73** +7.62
VILA-1.5 [29]	8	384	8B	1.21M	35.94	52.78	34.74	42.62	56.2	40.04
Video-LLaMA2.1-AV [9]			7B	3.35M	35.46 −0.48	52.78 +0	37.05 +3.31	39.34 −3.28	**58.68** +2.48	40.09 +0.05
Video-LLaMA2.1 [9]			7B	3.35M	**38.34** +2.88	**57.78** +5.0	**54.95** +17.9	**46.89** +7.75	48.76 −9.92	**47.3** +7.22
Chat-UniVi [20]	16	224	7B	0.65M	24.92	32.22	23.37	24.92	**38.84**	26.26
VideoChat2 [26]				2M	**39.3** +14.38	**60.0** +27.78	**44.84** +21.47	**31.8** +6.88	26.45 −12.39	**40.8** +14.54
Chat-UniVi-v1.5 [20]	16	336	7B	1.27M	34.5	48.89	25.89	40.98	42.98	35.4
LongVA [62]				1.32M	**37.86** +3.36	**51.11** +2.22	**41.89** +16.0	**47.87** +6.89	**53.72** +10.74	**43.32** +7.92
VILA-1.5 [29]	16	384	8B	1.21M	35.78	51.11	36.42	39.34	**60.33**	39.98
Video-LLaMA2.1-AV [9]			7B	3.35M	35.78 +0	56.67 +5.56	36.42 +0	40.0 +0.66	59.5 −0.83	40.62 +0.64
Video-LLaMA2.1 [9]			7B	3.35M	**38.34** +2.56	**57.78** +1.11	**54.95** +18.53	**46.89** +6.89	48.76 −10.74	**47.3** +6.68
Kangaroo [31]	64	448	8B	2.94M	31.79	51.67	29.05	**53.44**	33.06	37.1
MiniCPM-V [56]				8.32M	**38.18** +6.39	**60.0** +8.33	**40.84** +11.79	38.36 −15.08	**55.37** +22.31	**42.5** +5.4

Visualization







Model	Res.	LLM size	#frm					
			4	8	16	32	64	128
LongVA [62]	336	7B	40.74	**43.73**	43.32	41.62	39.98	39.39
LongVILA [29]	Dynamic	8B	33.00	35.87	35.11	36.46	**36.17**	36.11

Model	Res.	#frm	LLM size	CA	CU	NU	SP	MA	Overall
Qwen2-VL [50]	Dynamic	2 fps	2B	25.08	47.22	37.05	47.54	44.63	36.17
			7B	34.35 +9.27	56.67 +9.45	61.05 +24.0	57.38 +9.84	**48.76** +4.13	49.3 +13.13
			72B	**50.48** +16.13	**60.0** +3.33	**71.16** +10.11	**60.0** +2.62	46.28 −2.48	**58.68** +9.38
VILA-1.5 [29]	384	8	3B	26.84	43.89	19.16	37.38	47.11	29.84
			8B	**35.94** +9.1	**52.78** +8.89	**34.74** +15.58	**42.62** +5.24	**56.2** +9.09	**40.04** +10.2
		16	3B	26.04	41.67	23.37	34.75	48.76	30.13
			8B	**35.78** +9.74	**51.11** +9.44	**36.42** +13.05	**39.34** +4.59	**60.33** +11.57	**39.98** +9.85
Video-LLaMA2 [9]	336	32	7B	25.88	47.78	28.84	45.9	50.41	34.35
			72B	**54.15** +28.27	**71.67** +23.89	**65.68** +36.84	**48.52** +2.62	**59.5** +9.09	**58.62** +24.27
InternVL2 [8]	448	16	0.5B	24.12	28.33	26.53	29.84	28.93	26.61
			1.8B	24.28 +0.16	54.44 +26.11	35.16 +8.63	47.54 +17.7	53.72 +24.79	36.75 +10.14
			3.8B	32.11 +7.83	51.67 −2.77	44.0 +8.84	48.85 +1.31	48.76 −4.96	41.68 +4.93
			8B	**57.03** +24.92	62.78 +11.11	53.68 +9.68	45.57 −3.28	56.2 +7.44	54.63 +12.95
			20B	40.1 −16.93	63.89 +1.11	51.16 −2.52	38.36 −7.21	54.55 −1.65	46.42 −8.21
			34B	54.95 +14.85	69.44 +5.55	**65.47** +14.31	**56.07** +17.71	**70.25** +15.7	**60.73** +14.31
			70B	51.92 −3.03	**72.78** +3.34	64.84 −0.63	52.79 −3.28	63.64 −6.61	58.73 −2.0

https://yunlong10.github.io/VidComposition/
https://github.com/yunlong10/VidComposition/