JavisGPT: A Unified Multi-modal LLM for Sounding-Video Comprehension and Generation

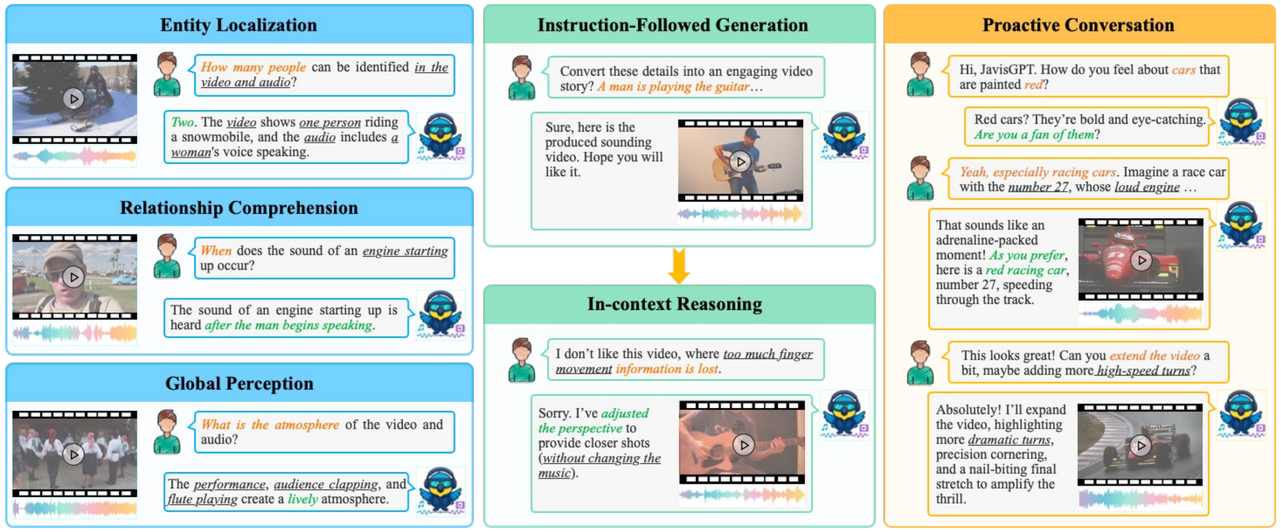
Kai Liu[1,2*] , Jungang Li[3*] , Yuchong Sun[4*] , Shengqiong Wu[2], Jianzhang Gao[4], Daoan Zhang[5], Wei Zhang[6], Sheng Jin[7], Sicheng Yu[8], Geng Zhan[9], Jiayi Ji[2], Fan Zhou[1], Liang Zheng[10], Shuicheng Yan[2], Hao Fei[2†] , Tat-Seng Chua[2]

NEURAL INFORMATION PROCESSING SYSTEMS

Homepage:

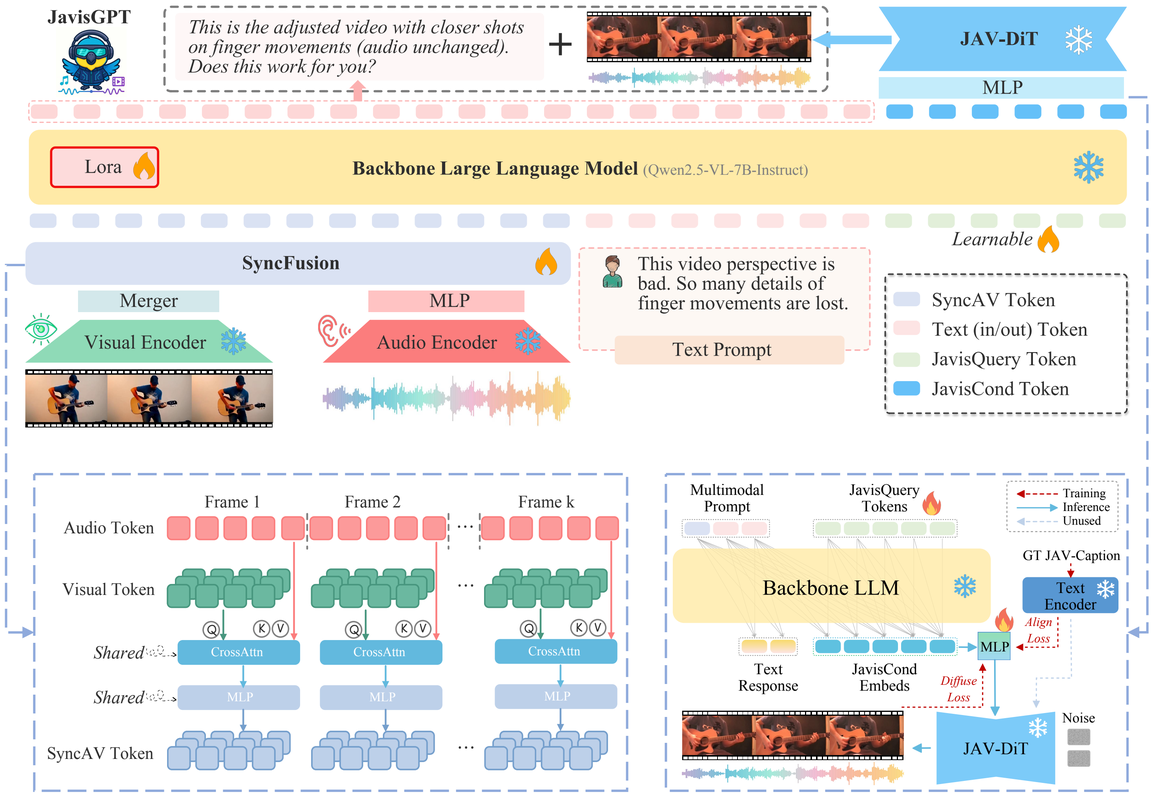
Motivation

Build the first unified MLLM for: (1) hierarchical spatiotemporal synchrony comprehension, (2) instruction-following audiovisual generation, and (3) multi-turn proactive conversation

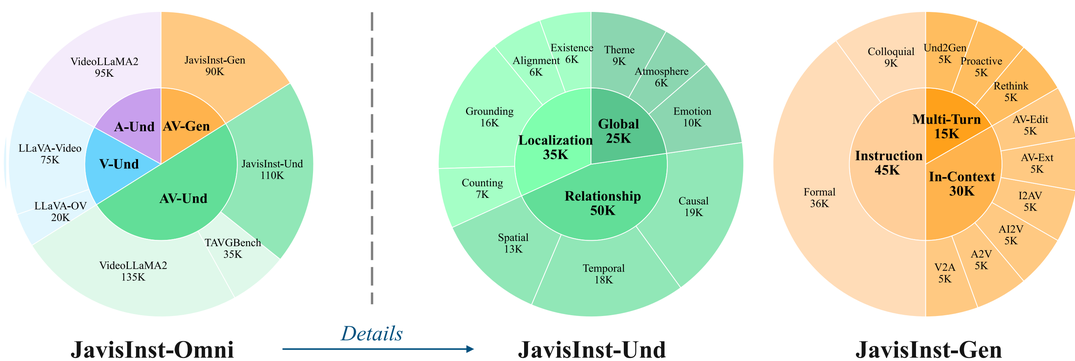


Methodology

Equip LLM with a DiT decoder, plus capture spatiotemporal synchrony



Develop a multimodal instruction tuning dataset for audiovisual U&G

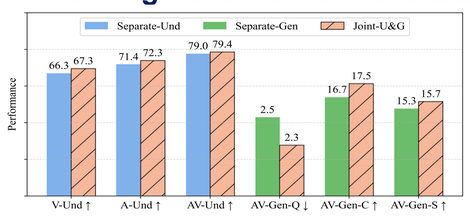


Evaluation

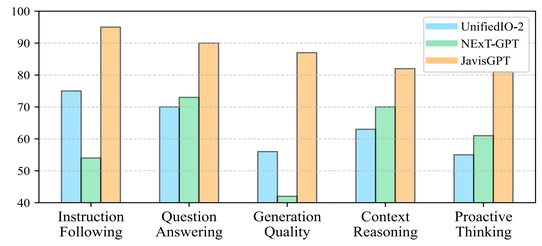
Comprehensive AudioVisual Understanding

Model (7-8B)	#Samples	Video			Audio		Audio-Video		
		ActivityNet	Perception	MVbench	ClothoAQA	TUT2017	AVQA	MU-AVQA	AVSD
NExT-GPT [68]	1.9M	21.5	33.7	27.9	26.9	6.4	25.3	19.3	30.8
UnifiedIO-2 [35]	9.2B	23.2	34.7	30.4	26.0	9.1	61.2	34.1	16.5
Macaw-LLM [37]	-	-	-	-	-	-	78.7	31.8	34.3
AV-LLM [55]	-	47.2	-	-	-	-	78.7	45.2	52.6
VideoLLaMA [77]	-	12.4	-	-	-	-	80.9	36.6	36.7
VideoLLaMA2 [13]	1.9M	50.2	51.4	54.6	70.1	78.4	-	79.2	57.2
VideoLLaMA2.1 [13]	1.9M	53.0	54.9	57.3	71.0	77.3	-	80.9	57.2
Qwen2.5-Omni [72]	-	57.2	70.4	66.9	**72.2**	78.3	91.5	79.9	**62.8**
JavisGPT (Ours)	1.5M	58.1	70.2	68.4	71.4	**82.1**	**93.8**	**82.1**	62.2

Joint Training Benefits Both Und.&Gen.



SOTA Joint Comprehension and Generation



Und. Brings Better Gen. Consistency