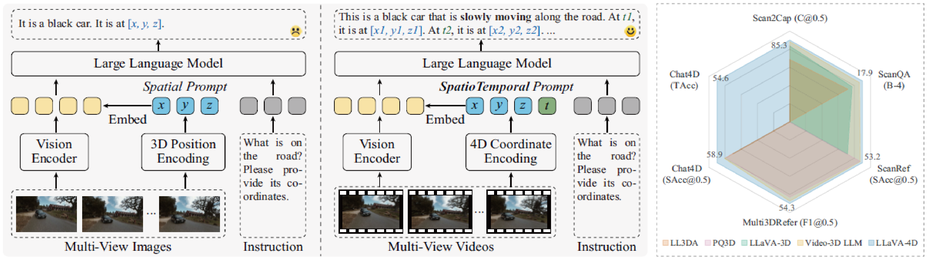
LLaVA-4D: Embedding SpatioTemporal Prompt into LMMs for 4D Scene Understanding

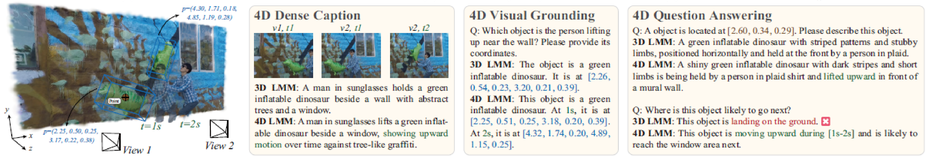
Hanyu Zhou, Gim Hee Lee*
National University of Singapore

Problem & Idea

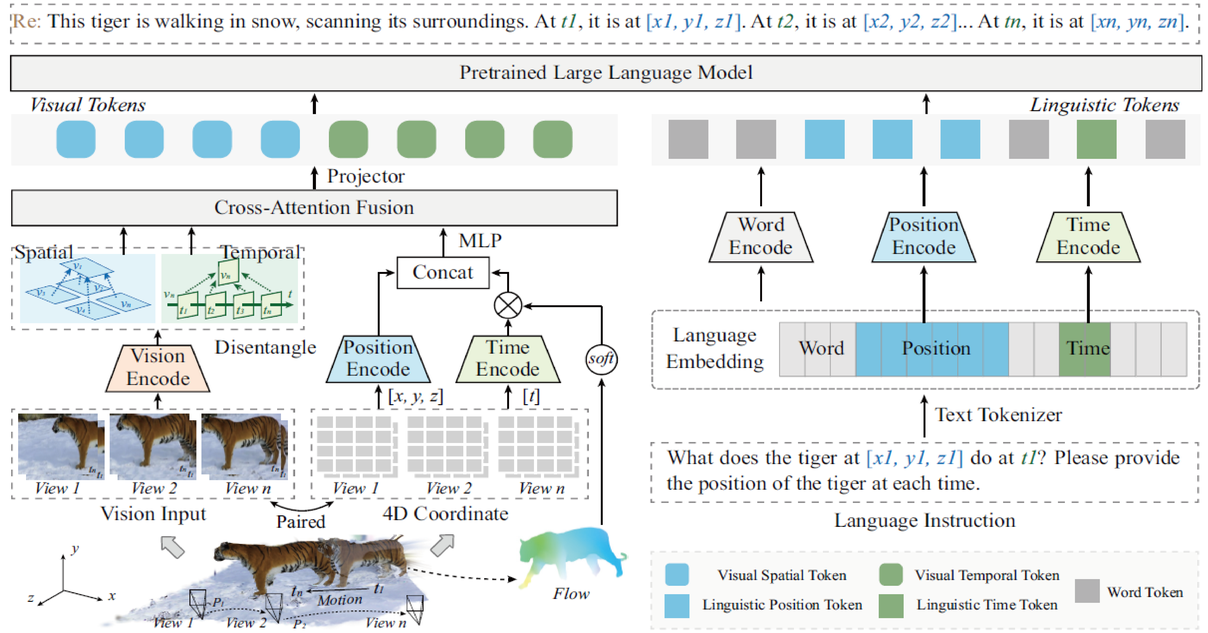
3D LMMs encode 3D positions as spatial prompts but overlook dynamic objects. Our 4D LMM framework embeds 4D coordinates: positions and timestamps as spatiotemporal prompts to capture both background and dynamic objects.



(a) 3D LMM (b) 4D LMM (Ours) (c) Performance on Multiple Benchmarks

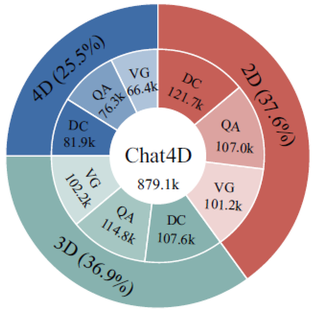


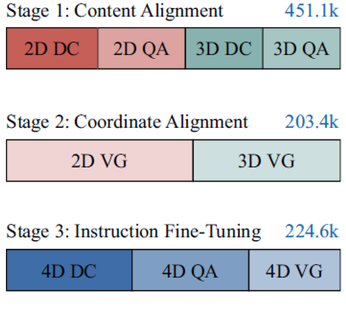
- **4D coordinate encoding:** Encode 3D position and 1D time with flow.
- **Vision embed:** Disentangle visual features into spatiotemporal features and embed the encoded 4D coordinates.
- **Language embed:** Align textual position and time with the fused vision embedding.



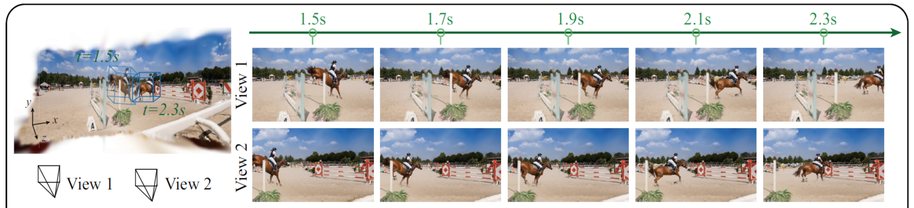
Dataset: Chat4D

Chat4D includes 2D/3D/4D vision-language training sets. Three-stage training: stages 1-2 use 2D/3D data for initialization; stage 3 uses 4D data for fine-tuning.





Comparison on Scene Understanding



User	What does the person do? Please provide its coordinates.
LLaVA-3D	A person is riding a horse, jumping over an obstacle in an outdoor equestrian arena. It is at [11.26, 2.10, -7.35, 12.48, 0.36, -7.10].
Video-3D LLM	The person is horseback riding in a sandy arena, jumping over a high obstacle during a show jumping competition, surrounded by fences and other jump setups. It is at [11.04, 2.05, -7.30, 12.39, 0.31, -7.05].
LLaVA-4D	This person is riding a brown horse in a sandy arena, approaching the hurdle, jumping high, clearing it mid-air, landing smoothly, and continuing toward the next obstacle. At 1.5s, it is at [11.15, 2.08, -7.32, 12.45, 0.30, -7.15]. At ..., it is at [...]. At 2.3s, it is at [12.80, 1.84, -6.80, 14.26, 0.12, -6.51].

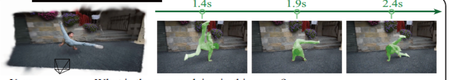
Ablation Study and Discussion

	Methods	3D Benchmark							4D Benchmark			
		Scan2Cap			ScanQA			Multi3DRefer	ScanRef	Chat4D (Ours)			
		C@0.5↑	B-4@0.5↑	M@0.5↑	C↑	B-4↑	M↑	F1@0.5↑	SAcc@0.5↑	C↑	B-4↑	SAcc@0.5↑	TAcc↑
3D	3D-LLM	–	–	–	69.4	12.0	14.5	–	–	61.6	11.5	31.4	–
	Chat-3D v2	63.9	31.8	–	87.6	14.0	–	41.6	38.4	81.8	13.7	39.5	–
	LL3DA	65.2	36.8	26.0	76.8	13.5	15.9	–	–	72.3	11.9	46.2	–
	3D-LLaVA	78.8	36.9	27.1	92.6	17.1	18.4	–	–	85.1	16.0	52.0	–
	Grounded 3D-LLM	70.6	35.5	–	72.7	13.4	–	40.6	44.1	66.3	12.2	43.7	–
	PQ3D	80.3	36.0	29.1	87.8	–	17.8	50.1	51.2	84.7	14.3	51.5	–
	LLaVA-3D	79.2	41.1	30.2	91.7	14.5	20.7	–	42.2	87.4	14.8	45.6	–
	Video-3D LLM	83.8	42.4	28.9	**102.1**	16.2	19.8	52.7	51.7	89.4	16.1	52.8	–
	Spatial-MLLM	–	–	–	91.8	14.8	18.4	–	–	–	–	–	–
	3UR-LLM	–	–	–	87.7	15.5	18.4	–	–	–	–	–	–
	GPT-4o w/ Co. Corr.	–	–	–	87.0	–	18.0	–	–	–	–	–	–
4D	LLaVA-4D (Ours)	**85.3**	**45.7**	**31.3**	97.8	**17.9**	**21.2**	**54.3**	**53.2**	**93.5**	**17.2**	**58.9**	**54.6**

Spatial Semantic

User	How many persons in the scene?
LLaVA-4D	Four.
User	Please segment these persons.
LLaVA-4D	Sure, [SEG].

Temporal Action



User	What is the person doing in this scene?
LLaVA-4D	This person is performing a dance move on a street, spinning on one hand, and switching leg positions mid-air.
User	Please segment this dance move.
LLaVA-4D	At 1.4s, [SEG]. At 1.9s, [SEG]. At 2.4s, [SEG].