

LLAVA-4D: EMBEDDING SPATIOTEMPORAL PROMPT INTO LMMS FOR 4D SCENE UNDERSTANDING

—*Supplementary Material*—

Hanyu Zhou¹, Gim Hee Lee¹

¹ School of Computing, National University of Singapore
{hy.zhou, gimhee.lee}@nus.edu.sg

In this supplementary material, we provide additional technical details and experimental validations in terms of our dataset, method, and comparison. For our dataset, we further provide 2D&3D&4D data distributions of our Chat4D dataset in Sec. 1. For our method, we list the nomenclature in Sec. 2, and describe the potential applications in Sec. 3. We also conduct extensive experiments to demonstrate the effectiveness of different components of our framework in Sec. 4, including spatiotemporal features in Sec. 4.1-4.2, optical flow effect in Sec. 4.3, training strategy in Sec. 4.4, applicability for camera settings in Sec. 4.5, comparison of different input modalities in Sec. 4.6, frame rate discussion in Sec. 4.7, input view discussion in Sec. 4.8, model robustness in Sec. 4.9. We further provide the potential solutions to the limitation in Sec. 5. For our comparison, we conduct extensive quantitative comparison experiments on 2D, 3D and 4D benchmarks in Sec. 6.1-6.3, and provide more visual comparisons of 3D and 4D scene understanding in Sec. 6.4 and 6.5.

1 DETAILS OF OUR CHAT4D DATASET

In Table 1, we provide details of our Chat4D dataset, where 2D, 3D and 4D vision-language data settings cover dense caption (DC), visual question answering (VQA) and visual grounding (VG) tasks. For the 2D datasets, the data sources of the DC task include WebVid-10M (Bain et al., 2021), Panda-70M (Chen et al., 2024b), etc, totaling 121.7k of samples; the data sources of the VQA task include Grand (Rasheed et al., 2024), EgoQA (Grauman et al., 2022), etc, totaling 107.0k of samples; the data sources of the VG task include VTimeLMM-Stage2 (Huang et al., 2024a), DiDeMo (Anne Hendricks et al., 2017), etc, totaling 101.2k of samples. For the 3D datasets, the data sources of the DC task include MMScan (Lyu et al., 2024) and Scan2Cap (Chen et al., 2021), totaling 107.6k of samples; the data sources of the VQA task include ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2022), totaling 114.8k of samples; the data sources of the VG task include Multi3DRefer (Zhang et al., 2023) and ScanRef (Chen et al., 2020), totaling 102.2k of samples. For the 4D dataset, the visual data sources for these three tasks are the same 4D scene reconstruction datasets: iPhone (Gao et al., 2022), N3DV (Li et al., 2022), DAVIS (Perazzi et al., 2016), etc. The corresponding text instructions are generated by GPT, totaling 81.9k of samples for DC, 76.3k of samples for VQA, and 66.4k of samples for VG. We use 2D & 3D DC and VQA data for content alignment in training stage 1, 2D & 3D VG data for spatiotemporal coordinate alignment in training stage 2, and 4D data for instruction fine-tuning in training stage 3. We selected 2,000 samples for each task as test set to evaluate the 4D scene understanding performance of LMMS. We believe that our Chat4D dataset will become a foundation benchmark for 4D scene understanding training and evaluation, and promote the development of scene understanding in the LMM community.

2 LIST OF NOMENCLATURE

We provide the list of nomenclature in Table 2 to improve the readability of the whole paper.

3 POTENTIAL APPLICATIONS OF OUR LLAVA-4D

Unlike existing 3D LMMS that mainly focus on static scene understanding, our LLaVA-4D is a scalable and general 4D LMM with spatiotemporal awareness, capturing the dynamic patterns. In real-world settings, scenes typically contain both dynamic objects and static backgrounds. Existing 3D

Table 1: Data composition of our Chat4D dataset. The red/green/blue highlight 2D/3D/4D data.

Training Stage	Task	# of Samples	Data Source
Stage 1 Content Alignment	2D DC	121.7k	WebVid-10M [1], Panda-70M [2], InternVid-10M [3], Valley [4]
	2D VQA	107.0k	GranD [5], ANet-RTL [6], MUSE [7], EgoQA [8], NExT-QA [9]
	3D DC	107.6k	MMScan [10], Scan2Cap [11]
	3D VQA	114.8k	ScanQA [12], SQA3D [13]
Stage 2 Coordinate Alignment	2D VG	101.2k	VTimeLLM-Stage2 [14], GranD [15], DiDeMo [16], HiREST [17], Glamm [18]
	3D VG	102.2k	Multi3DRefer [19], ScanRef [20]
Stage 3 Instruction Fine-Tuning	4D DC	81.9k	iPhone [21], HyperNeRF [22], N3DV [23], PanopticSports [24], DAVIS [25], Immersive [26]
	4D VQA	76.3k	iPhone [21], HyperNeRF [22], N3DV [23], PanopticSports [24], DAVIS [25], Immersive [26]
	4D VG	66.4k	iPhone [21], HyperNeRF [22], N3DV [23], PanopticSports [24], DAVIS [25], Immersive [26]

[1] (Bain et al., 2021); [2] (Chen et al., 2024b); [3] (Wang et al., 2023); [4] (Luo et al., 2023); [5] (Rasheed et al., 2024); [6] (Huang et al., 2024b); [7] (Ren et al., 2024); [8] (Grauman et al., 2022); [9] (Xiao et al., 2021); [10] (Lyu et al., 2024); [11] (Chen et al., 2021); [12] (Azuma et al., 2022); [13] (Ma et al., 2022); [14] (Huang et al., 2024a); [15] (Rasheed et al., 2024); [16] (Anne Hendricks et al., 2017); [17] (Zala et al., 2023); [18] (Rasheed et al., 2024); [19] (Zhang et al., 2023); [20] (Chen et al., 2020); [21] (Gao et al., 2022); [22] (Park et al., 2021); [23] (Li et al., 2022); [24] (Luiten et al., 2024); [25] (Perazzi et al., 2016); [26] (Broxton et al., 2020)

Table 2: List of Nomenclature.

Variable Notations	Description	Dimension
$[x, y, z]$	3D world coordinate of a point	$\mathbb{R}^3$
$PE(x, y, z)$	Fourier-based 3D position encoding	$\mathbb{R}^D$
$TE(t)$	Fourier-based 1D time encoding	$\mathbb{R}^D$
w_p	Learnable projection matrix for 4D prompt	$\mathbb{R}^{D \times 2D}$
p_{4D}	4D coordinate encoding	$\mathbb{R}^D$
f	Visual features extracted from multi-view videos	$\mathbb{R}^{N \times D}$
f_s	Spatial component of visual features	$\mathbb{R}^{N \times D}$
f_t	Temporal component of visual features	$\mathbb{R}^{N \times D}$

LMMs struggle to perceive the motion states of objects, which limits their effectiveness in practical applications. Our method addresses this critical limitation by enabling true 4D spatiotemporal awareness, which is an essential capability for many real-world tasks. For example:

- In robotics, our model supports planning by understanding the dynamic surroundings.
- In Augmented Reality (AR), our model enables motion tracking for more immersive experiences.
- In autonomous driving, our model identifies the position and speed of moving objects and grounds language to the evolving road scene.

These examples highlight the significant practical value of our LLaVA-4D framework.

4 ABLATION STUDY AND DISCUSSION

4.1 IMPORTANCE OF SPATIOTEMPORAL FEATURES

In Table 3, we compare the impacts of disentangled spatial and temporal features with visual features directly encoded from multi-view videos on 4D scene understanding. In the 3D benchmark evaluation, the disentangled spatiotemporal features and visual features maintain similar contributions to 3D scene understanding, *i.e.*, spatial characteristics of static backgrounds. In the 4D benchmark evaluation, compared with visual features, the disentangled spatial and temporal features significantly improve the spatial and temporal understanding performance by a large margin, respectively. This verifies that the disentangled spatiotemporal features have discriminative representations for dynamic objects and static backgrounds, and are more suitable for 4D scene understanding.

Table 3: Discussion on the effects of various feature types for visual representation.

Feature type	3D ScanQA [†]		3D ScanRef [‡]	4D Chat4D			
	C [†]	B-4 [†]		C [†]	B-4 [†]	SAcc@0.5 [†]	TAcc [†]
w/ Visual features	92.4	16.5	51.9	88.3	15.3	52.7	49.4
w/ Disentangled spatial features	95.7	17.3	53.0	90.2	16.4	56.0	51.1
w/ Disentangled temporal features	95.5	17.1	52.1	89.8	16.1	53.2	53.5
w/ Disentangled spatiotemporal features	97.8	17.9	53.2	93.5	17.2	58.9	54.6

[†] (Azuma et al., 2022); [‡] (Chen et al., 2020)

Table 4: Discussion on the impacts of spatial encoding and temporal encoding.

Spatial encoding		Temporal encoding		C [†]	B-4 [†]	SAcc@0.5 [†]	TAcc [†]
Dis. spatial feature	3D pos encoding	Dis. temporal feature	1D time encoding				
✓	×	×	×	80.5	11.8	–	–
✓	✓	×	×	90.1	16.4	56.1	–
×	×	✓	×	79.3	11.3	–	–
×	×	✓	✓	84.2	14.1	–	52.2
✓	✓	✓	✓	93.5	17.2	58.9	54.6

Table 5: Ablation study on the effect of optical flow on coordinate encoding for scene understanding.

Coordinate encoding	C [†]	B-4 [†]	SAcc@0.5 [†]	TAcc [†]
w/o Optical flow	90.8	16.7	57.1	47.3
w/ Optiacl flow	93.5	17.2	58.9	54.6

4.2 COMPONENT-/ELEMENT-LEVEL ABLATION ON SPATIAL AND TEMPORAL ENCODING

As shown in Table 4, we conduct a component-level and element-level ablation study to analyze the impact of spatial and temporal encoding. In this study, the spatial encoding consists of the spatial features from spatiotemporal disentanglement and the 3D position encoding from the 4D coordinate prompt. The temporal encoding includes the temporal features from spatiotemporal disentanglement and the 1D time encoding from the 4D coordinate prompt.

From a component-level perspective, spatial encoding alone supports strong spatial understanding, and temporal encoding alone improves temporal understanding. Combining both significantly enhances overall model performance. From an element-level perspective, disentangled spatial and temporal features offer a coarse understanding of the scene. 3D position encoding and 1D time encoding are crucial for fine-grained spatiotemporal reasoning. Concurrently, both spatial and temporal encoding are essential for effective 4D scene understanding.

4.3 EFFECT OF OPTICAL FLOW ON COORDINATE ENCODING

We also discuss the impact of encoding optical flow into the 4D coordinate on LMMs for 4D scene understanding. As shown in Table 5, embedding optical flow into the 4D coordinate can significantly improve the 4D scene understanding performance of our model. The main reason is that optical flow reflects the motion information of objects in the scene along the temporal dimension, and encoding this motion feature into the 4D coordinate is conducive to enhancing the dynamic awareness of visual representation, improving the ability of the LMM to understand dynamic patterns in the scene.

4.4 COMPARISON BETWEEN MULTI-STAGE AND SINGLE-STAGE TRAINING.

We conduct an evaluation comparing different training pipelines in Table 6. The results indicate that our model with multi-stage training significantly outperforms our single-stage version. There are two reasons for this improvement: 1) The tasks modeled by LMMs are typically heterogeneous in nature. 2) The input modalities have distinct feature distributions. Single-stage training is more vulnerable to

Table 6: Discussion on the impacts of training pipeline for model performance.

Training pipeline	C $\uparrow$	B-4 $\uparrow$	SAcc@0.5 $\uparrow$	TAcc $\uparrow$
Single-stage training	86.7	14.5	49.3	43.8
Multi-stage training	93.5	17.2	58.9	54.6

Table 7: Discussion on the generalization of our model for various camera settings.

Camera setting	C $\uparrow$	B-4 $\uparrow$	SAcc@0.5 $\uparrow$	TAcc $\uparrow$
w/ Single static camera	94.1	17.5	58.5	55.3
w/ Single dynamic camera	94.1	17.5	58.3	55.8
w/ Dynamic multi-camera	94.3	17.4	59.7	55.8
w/ Asynchronous multi-camera	92.4	16.2	54.9	52.0
w/ Synchronized multi-camera	94.5	17.8	60.4	56.2

conflicting training signals and gradient interference from complex tasks. In contrast, multi-stage training leads to better overall performance by promoting stable convergence at each stage.

4.5 GENERALIZATION TO DIFFERENT CAMERA SETTINGS.

Although we introduce the entire model using synchronized multi-camera inputs as an example, we also discuss the applicability of our model to different camera settings including single static camera, single dynamic camera, dynamic multi-camera, and asynchronous multi-camera. When applied to these camera settings, the visual encoding processes in our framework, *i.e.* spatiotemporal disentanglement, 4D coordinate encoding and spatiotemporal fusion still work well.

- For *a single static camera* setting, the number of views is simply reduced to one. During spatiotemporal disentanglement, we perform correlation only on intra-frame visual features by treating them as spatial features. All other encoding processes remain unchanged.
- For *a single dynamic camera* setting, the main difference is the additional background motion caused by the ego movement. This motion is different from the independent motion of the object. To alleviate this issue, we introduce a motion threshold to filter out background motion during 4D coordinate encoding. During spatiotemporal disentanglement, we just construct temporal features. All other encoding processes remain unchanged.
- For *a dynamic multi-camera* setting, we also introduce a motion threshold to perform motion segmentation for the independent moving object during 4D coordinate encoding. All other parts of the encoding process remain unchanged.
- For *an asynchronous multi-camera* setting, spatiotemporal disentanglement and 4D coordinate encoding may introduce a shift in the feature distribution. However, the spatiotemporal fusion module mitigates this risk by using a cross-attention mechanism to model temporal relationships between frames that enable the model to learn inter-frame dependencies.

Table 7 shows the experimental results to verify the generalization of our model on the above camera settings. First, the performance metrics of our model with the single camera setting and dynamic multi-camera settings are only slightly lower than our model with the synchronized multi-camera setting. Second, although the asynchronous multi-camera setting leads to a drop in the performance of our model to some extent, it is still within an acceptable error range. The results show that our model has strong generalization to different camera settings.

4.6 POINT-BASED LMMS *v.s.* IMAGE-BASED LMMS

In Table 8, we introduce the 4DGS method (Wu et al., 2024) to reconstruct dynamic scene point clouds from multi-view videos, and compare the 4D scene understanding performance of our LLaVA-4D model with point clouds and images as input. The results show that the image-based LMM outperforms the point-based LMM in 4D scene understanding. This is because point-based LMM

Table 8: Discussion on the effect of various vision modalities on scene understanding.

Vision modality	C $\uparrow$	B-4 $\uparrow$	SAcc@0.5 $\uparrow$	TAcc $\uparrow$
Ours w/ point clouds	86.2	14.8	49.3	48.5
Ours w/ images	93.5	17.2	58.9	54.6

Table 9: Discussion on the effect of frame rates on model performance and efficiency.

Frame rate (FPS)	C $\uparrow$	B-4 $\uparrow$	SAcc@0.5 $\uparrow$	TAcc $\uparrow$	Storage (GB)
3	93.6	16.8	58.4	51.8	1.1
8	93.9	17.1	59.1	55.0	3.0
15	94.3	17.4	59.8	55.7	5.6
30	94.5	17.8	60.4	56.2	11.3
50	94.6	17.9	60.6	56.4	18.8
70	94.7	17.9	60.6	56.5	26.3

Table 10: Discussion on the impact of input views on model performance and efficiency.

Views (n)	SAcc@0.5 $\uparrow$	S_Gain	TAcc $\uparrow$	T_Gain	Storage (GB)
2	58.7	0.0	55.5	0.0	2.3
6	59.8	1.1	56.0	0.5	6.8
10	60.4	0.6	56.2	0.2	11.3
14	60.6	0.2	56.3	0.1	15.8
18	60.7	0.1	56.3	0.0	20.3

relies heavily on the precision of the reconstruction technology, while existing reconstruction methods still face huge challenges for 4D dynamic scene reconstruction.

4.7 DISCUSSION ON DIFFERENT FRAME RATES

In Table 9, we provide an ablation study on how video frame rate affects model performance. First, our model does not require high frame rates, *e.g.* 70 FPS to function effectively. It operates robustly within typical range of frame rates, *e.g.* 15–30 FPS without incurring significant computational overhead. Second, our model shows only a slight performance drop in temporal grounding at low frame rates, *e.g.* 3 FPS, while maintaining strong performance across other scene understanding metrics. The underlying reason lies in the design of our 4D coordinate encoding of position + time + motion, which is inherently redundant and error-tolerant. The temporal prompt is not fully compromised even when optical flow becomes less accurate at low frame rates. This is because the spatial position and time remain well-defined and continue to provide reliable information.

4.8 IMPACT OF INPUT VIEW NUMBER ON MODEL PERFORMANCE.

In Table 10, we construct an analysis experiment on the impact of input views, where we choose our 4D dataset as evaluation benchmark, SAcc and TAcc as performance metrics, corresponding gains as uncertainties around marginal gains, and storage cost as data collection cost. First, increasing the number of input views steadily improves model performance, though the improvement is nonlinear. Second, the performance gain becomes marginal when the number of input views exceeds 14. Concurrently, the storage cost continues to grow linearly. Third, the optimal configuration is to use 10 input views, where a favorable trade-off between performance, gain, and data collection cost is achieved. Importantly, our method is a general framework that supports an arbitrary number of input views. It remains effective even under sparse-view conditions, and therefore makes it practical for real-world deployment.

Table 11: Discussion on the robustness of our model to 4D coordinate errors.

Scene setting	Coordinate errors from SfM & MVS (%)	C↑	B-4↑	SAcc@0.5↑	TAcc↑
Clean scene	0	93.5	17.2	58.9	54.6
Low light	-6.7	92.1	16.3	54.8	53.5
Fast motion	-5.3	92.6	16.5	55.0	53.9

4.9 ROBUSTNESS TO 4D COORDINATE ERRORS FROM SFM AND MVS

In our framework, the 4D coordinate prompt relies on some external modules such as SfM and MVS. Our overall model remains robust despite these modules can introduce errors under challenging conditions such as low lighting or fast motion. This robustness comes from the role of the 4D spatiotemporal coordinate, which serves as a coarse prompt. Similarly to NeRF models, it emphasizes relative position and temporal ordering instead of requiring precise coordinates. Even when the 4D coordinates contain moderate errors, our model still learns meaningful structural and temporal cues.

We also conduct an analysis to evaluate the impact of inaccuracies from these external modules on the 4D scene understanding performance of our LLaVA-4D. The results are shown in Table 11, where low-light and fast-motion conditions are simulated from clean Chat4D data using contrast reduction and motion blur, respectively. Compared to clean scenes, the 4D coordinates estimated by SfM and MVS do exhibit noticeable errors under low-light and fast-motion conditions. However, the overall 4D scene understanding performance of our model remains largely stable. We observe only a slight drop in the spatial grounding metric SAcc, and this decrease stays within an acceptable error range. This result highlights the robustness of our model to moderate coordinate inaccuracies.

5 LIMITATION AND POTENTIAL SOLUTION

Our LLaVA-4D has limitations in terms of methodological details and scene settings. On methodological details, our model has achieved success in end-to-end inference for scene understanding, but the preprocessing stage related to 4D coordinate construction relies on external modules such as SfM. This slightly reduces the overall efficiency. However, these external modules are design choices that are specific to only the current implementation. Given the absence of existing 4D LMMs, our work pioneers the exploration of applying large multimodal models to 4D scene understanding. We adopt reliable external modules in the construction of spatiotemporal prompts to allow rapid validation on the offline datasets. These external modules are modular, replaceable, and do not define or constrain the overall framework. In the future, we adopt end-to-end geometric models, *e.g.* MonST3R (Zhang et al., 2024) to replace these external components. This will improve the practicality of our model.

On scene setting, our LLaVA-4D can understand most regular 3D & 4D scenes, but there are still some rare challenging scenes where it performs poorly, *e.g.* dynamic objects with fast motion. Such high-dynamic patterns typically lead to motion blur and discontinuous trajectories in frame imaging due to long exposure and low temporal resolution of the frame camera, which weakens the discriminative spatiotemporal features of the scene and limits the scene understanding capability of our model. In contrast, an event camera (Gallego et al., 2020) is a neuromorphic sensor with high temporal resolution that asynchronously triggers signals based on brightness changes. Event-based imaging compensates for the limitations of frame-based cameras in high-dynamic scenes. To leverage this advantage, we propose embedding event streams into vision input and 4D coordinate encoding. First, event streams can be integrated with frame videos to enrich temporal feature representations. Second, they can replace optical flow to improve the dynamic awareness of 4D coordinate encoding.

6 COMPARISON WITH STATE-OF-THE-ART MODELS

6.1 QUANTITATIVE COMPARISON ON 2D BENCHMARKS

In Table 12, we compare the basic visual understanding capabilities of 2D LMMs and 3D LMMs on several typical 2D benchmarks. The results show that our method can maintain the same level of image understanding as other competing LMMs. This also indirectly proves that our model can obtain the initial multimodal spatiotemporal understanding capability from the 2D dataset.

Table 12: Quantitative results of LMMs for scene understanding on 2D zero-shot benchmarks.

Methods		VQAv2 [6]	MMBench [7]	MME [8]	MM-Vet [9]
2D	MobileVLM [1]	47.5	59.6	1289	–
	Qwen-VL [2]	63.8	38.2	–	–
	Qwen-VL-Chat [2]	61.5	60.6	1488	–
	LLaMA-VID [3]	–	65.1	1521	–
	LLaVA-1.5 [4]	58.2	65.2	1511	31.1
3D	LLaVA-3D [5]	57.8	65.0	1502	30.9
4D	LLaVA-4D (Ours)	58.0	65.1	1510	31.1

[1] (Chu et al., 2023); [2] (Bai et al., 2023); [3] (Li et al., 2024); [4] (Liu et al., 2024a); [5] (Zhu et al., 2024a); [6] (Goyal et al., 2017); [7] (Liu et al., 2024b); [8] (Fu et al., 2024); [9] (Yu et al., 2023)

Table 13: Quantitative results of LMMs for scene understanding on different 3D benchmarks.

Methods		Scan2Cap [9]				ScanQA [10]			Multi3DRefer [11]		ScanRef [12]		SQA3D [13]	
		C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C↑	B-4↑	M↑	R↑	F1@0.25↑	F1@0.5↑	SAcc@0.25↑	SAcc@0.5↑	EM@1↑
3D	3D-LLM [1]	—	—	—	—	69.4	12.0	14.5	35.7	—	—	30.3	—	—
	Chat-3D v2 [2]	63.9	31.8	—	53.6	87.6	14.0	—	—	—	41.6	42.5	38.4	54.7
	LL3DA [3]	65.2	36.8	26.0	55.1	76.8	13.5	15.9	37.3	—	—	—	—	—
	3D-LLaVA [4]	78.8	36.9	27.1	57.7	92.6	17.1	18.4	43.1	—	—	—	—	54.5
	Grou. 3D-LLM [5]	70.6	35.5	—	—	72.7	13.4	—	—	45.2	40.6	47.9	44.1	—
	PQ3D [6]	80.3	36.0	29.1	57.9	87.8	—	17.8	42.9	—	50.1	57.0	51.2	54.6
	LLaVA-3D [7]	79.2	41.1	30.2	63.4	91.7	14.5	20.7	50.1	54.3	47.2	54.1	42.2	55.6
	Video-3D LLM [8]	83.8	42.4	28.9	62.3	102.1	16.2	19.8	49.0	58.0	52.7	58.1	51.7	58.6
4D	LLaVA-4D (Ours)	85.3	45.7	31.3	65.7	97.8	17.9	21.2	51.8	60.5	54.3	59.4	53.2	59.2

[1] (Hong et al., 2023); [2] (Huang et al., 2023); [3] (Chen et al., 2024a); [4] (Deng et al., 2025); [5] (Chen et al., 2024c); [6] (Zhu et al., 2024b); [7] (Zhu et al., 2024a); [8] (Zheng et al., 2024); [9] (Chen et al., 2021); [10] (Azuma et al., 2022); [11] (Zhang et al., 2023); [12] (Chen et al., 2020); [13] (Ma et al., 2022)

Table 14: Quantitative results of LMMs for scene understanding on different 4D benchmarks.

Methods		C↑	B-4↑	M↑	R↑	SAcc@0.5↑	TAcc↑
3D	3D-LLM [1]	61.6	11.5	12.3	33.5	31.4	–
	Chat-3D v2 [2]	81.8	13.7	–	–	39.5	–
	LL3DA [3]	72.3	11.9	14.9	36.8	46.2	–
	3D-LLaVA [4]	85.1	16.0	18.2	44.2	52.0	–
	Grounded 3D-LLM [5]	66.3	12.2	–	–	43.7	–
	PQ3D [6]	84.7	14.3	17.5	43.1	51.5	–
	LLaVA-3D [7]	87.4	14.8	19.4	47.5	45.6	–
	Video-3D LLM [7]	89.4	16.1	19.2	48.3	52.8	–
4D	LLaVA-4D (Ours)	93.5	17.2	21.0	50.2	58.9	54.6

[1] (Hong et al., 2023); [2] (Huang et al., 2023); [3] (Chen et al., 2024a); [4] (Deng et al., 2025); [5] (Chen et al., 2024c); [6] (Zhu et al., 2024b); [7] (Zhu et al., 2024a); [8] (Zheng et al., 2024)

6.2 QUANTITATIVE COMPARISON ON 3D BENCHMARKS

In Table 13, we also provide more evaluation metrics of competing methods on 3D benchmarks, including Scan2Cap (Chen et al., 2021), ScanQA (Azuma et al., 2022), SQA3D (Ma et al., 2022), Multi3DRefer (Zhang et al., 2023), and ScanRef (Chen et al., 2020). Combining the results in these tables, we can conclude that our LLaVA-4D has a better overall performance than other 3D LMMs. The main reason is that spatial features disentangled from multi-view images have stronger representation capabilities for spatial attributes and relationships, while ordinary visual features lack matching association across different views.

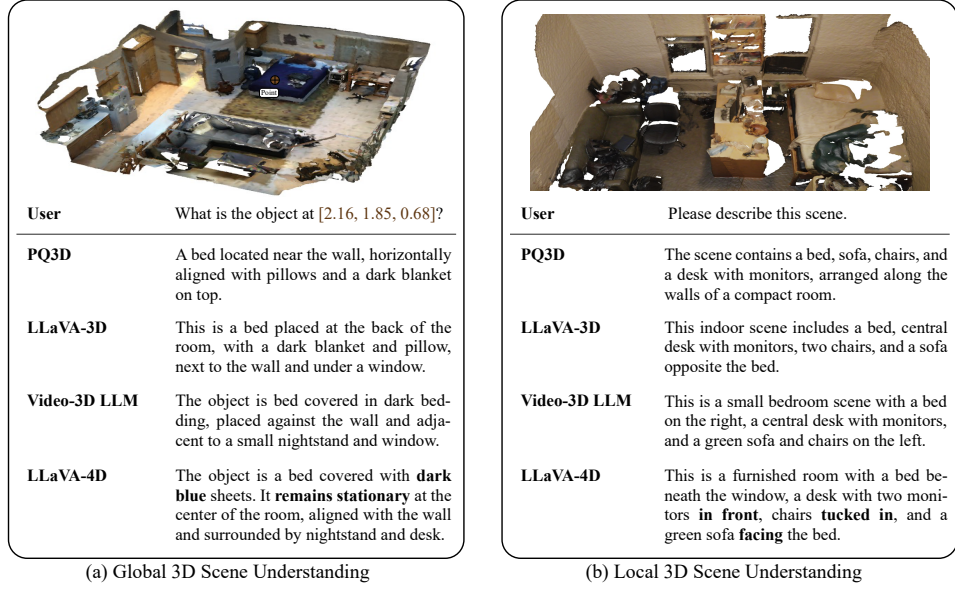


Figure 1: Visual comparison of LMMs on 3D scene understanding.

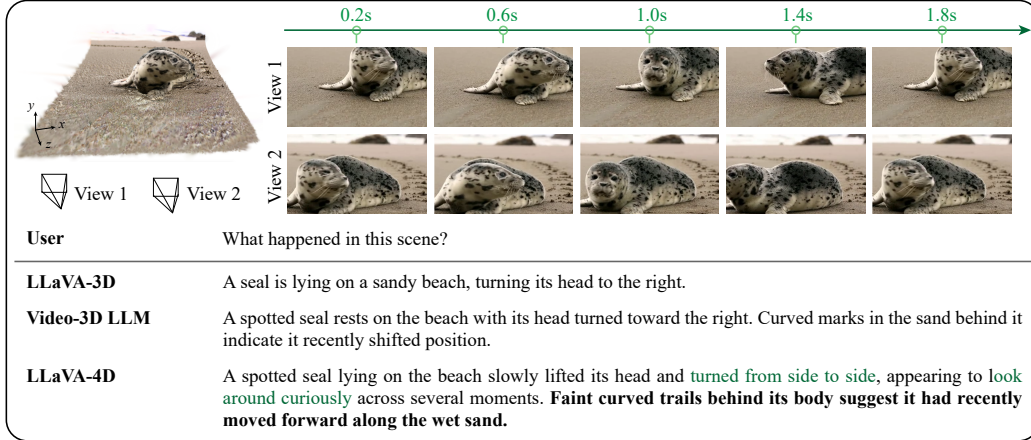


Figure 2: Visual comparison of LMMs on 4D scene understanding.

6.3 QUANTITATIVE COMPARISON ON 4D CHAT4D BENCHMARK

In Table 14, we also provide more detailed evaluations in our Chat4D benchmark. The results show that our model is far superior to other 3D LMMs in 4D scene understanding. Compared to competing 3D LMMs, our model introduces temporal encoding to enhance dynamic awareness. It further disentangles visual features into temporal components to improve the representation of temporal attributes and relationships across the scene. These designs significantly strengthen the 4D scene understanding ability of our model.

6.4 QUALITATIVE COMPARISON ON 3D BENCHMARKS

In Fig. 1, we provide more visual comparisons on 3D scene understanding, where we select some typical scenes as examples. Compared to the competing 3D LMMs, our LLaVA-4D is more fine-grained in representing spatial characteristics of the scene. This also verifies the effectiveness of the disentangled spatial features in our model for the spatial characteristic representation of the scene.

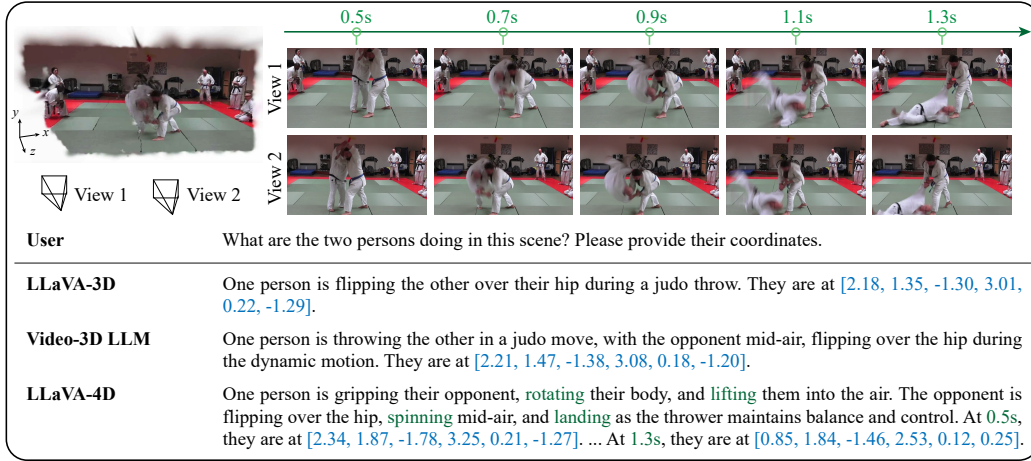


Figure 3: Visual comparison of LMMs on 4D scene understanding.

6.5 QUALITATIVE COMPARISON ON 4D CHAT4D BENCHMARK

In Fig. 2 and 3, we also provide more visual comparisons on 4D scene understanding. The results show that 3D LMMs cannot respond to timestamp and corresponding temporal information, while our method can understand the temporal content. This is because these 3D LMMs lack the representation of temporal characteristic, while our method introduces the spatiotemporal prompt to enhance the dynamic representation of 4D scenes.

REFERENCES

- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.
- Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 19129–19139, 2022.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.
- Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Trans. Graph.*, 39(4):86–1, 2020.
- Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.
- Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024a.
- Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m:

- Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.
- Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024c.
- Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3193–3203, 2021.
- Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 1(2):3, 2023.
- Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d llms with omni superpoint transformer. *arXiv preprint arXiv:2501.01163*, 2025.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. doi: 10.1109/TPAMI.2020.3008413.
- Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35: 33768–33780, 2022.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.
- Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.
- Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024a.
- De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024b.
- Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.
- Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5521–5531, 2022.
- Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.
- Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pp. 800–809. IEEE, 2024.
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.
- Ruiyuan Lyu, Jingli Lin, Tai Wang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, and Jiangmiao Pang. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations. *Advances in Neural Information Processing Systems*, 37:50898–50924, 2024.
- Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.
- Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.
- Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 724–732, 2016.
- Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13009–13018, 2024.
- Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26374–26383, 2024.
- Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.
- Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 20310–20320, 2024.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065, 2023.
- Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.
- Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15225–15236, 2023.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. *arXiv preprint arXiv:2412.00493*, 2024.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering llms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024a.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pp. 188–206. Springer, 2024b.