Understanding Dynamic Scenes in Egocentric 4D Point Clouds

Junsheng Huang[1], Shengyu Hao[2], Bo-Cheng Hu[1], Hongwei Wang[1]*, Gaoang Wang[1]*

[1]Zhejiang University, China
[2]China Tower Corporation Limited, Hangzhou Science and Technology Innovation Center, China
junsheng.24@intl.zju.edu.cn, victorhsy7@gmail.com, thebrandonhu@gmail.com,
hongweiwang@intl.zju.edu.cn, gaoangwang@intl.zju.edu.cn

Abstract

Understanding dynamic 4D scenes from an egocentric perspective—modeling changes in 3D spatial structure over time—is crucial for human–machine interaction, autonomous navigation, and embodied intelligence. While existing egocentric datasets contain dynamic scenes, they lack unified 4D annotations and task-driven evaluation protocols for fine-grained spatio-temporal reasoning, especially on motion of objects and human, together with their interactions. To address this gap, we introduce **EgoDynamic4D**, a novel QA benchmark on highly dynamic scenes, comprising RGB-D video, camera poses, globally unique instance masks, and 4D bounding boxes. We construct **927K** QA pairs accompanied by explicit Chain-of-Thought (CoT), enabling verifiable, step-by-step spatio-temporal reasoning. We design 12 dynamic QA tasks covering agent motion, human–object interaction, trajectory prediction, relation understanding, and temporal–causal reasoning, with fine-grained, multidimensional metrics. To tackle these tasks, we propose an end-to-end spatio-temporal reasoning framework that unifies dynamic and static scene information, using instance-aware feature encoding, time and camera encoding, and spatially adaptive down-sampling to compress large 4D scenes into token sequences manageable by LLMs. Experiments on **EgoDynamic4D** show that our method consistently outperforms baselines, validating the effectiveness of multimodal temporal modeling for egocentric dynamic scene understanding.

1 Introduction

With the rapid advancement of embodied AI and human–machine interaction technologies, understanding dynamic 4D scenes from an egocentric viewpoint—integrating 3D spatial dimensions and the temporal dimension—has emerged as a pivotal research challenge. Unlike conventional third-person video analysis (Pang et al. 2021; Song et al. 2021; Pan et al. 2021; Li, Niu, and Zhang 2022), egocentric videos exhibit high dynamics, frequent scene changes, and rich interactive behaviors, requiring models not only to capture the wearer's movement but also to perceive and reason about surrounding people, objects, and their evolving relationships (Li et al. 2021; Yu et al. 2023; Fan et al. 2017; Lee, Lee, and Choi 2023; Lin et al. 2022; Pramanick et al. 2023). Applications such as robotic perception, AR, and autonomous driving demand efficient and accurate egocentric scene understanding to empower the next generation of embodied agents.

Although landmark egocentric datasets Ego4D (Grauman et al. 2022), EgoExo4D (Grauman et al. 2024), HD-Epic (Perrett et al. 2025), and 3D datasets (Dai et al. 2017; Wang et al. 2024a; Chen, Chang, and Nießner 2020) have driven progress in action recognition, object locating, and 3D scene analysis, they suffer from the following key limitations: (1) Incomplete 4D annotations and lack of dynamic content: Egocentric datasets often lack temporally aligned 3D bounding boxes and trajectories, while 3D datasets focus on static scenes without moving objects or agents—limiting the study of real-world dynamics. (2) Limited evaluation of temporal reasoning: Existing benchmarks emphasize short-term or moment-based tasks and lack protocols to assess reasoning over continuous object motion or interaction. (3) Incomplete multimodal evaluation: While some works explore temporal scene graphs (Yang et al. 2023), they focus on representation construction rather than end-to-end multimodal reasoning, and do not support QA-based evaluation of dynamic 4D scenes.

We propose **EgoDynamic4D**, the first egocentric QA benchmark for highly dynamic 4D scene understanding. By refining ADT (Pan et al. 2023) and THUD++ (Tang et al. 2024) annotations, we create a unified, multimodal dataset with RGB-D video, camera poses, globally unique instance masks, and 4D bounding boxes across indoor scenes. It includes 12 QA tasks covering scene descriptions, momentary and durative dynamics, with rich evaluation metrics. To improve reasoning and transparency, we equip QA pairs with explicit Chain-of-Thought (CoT), enabling step-by-step spatio-temporal explanations that aid training and allow interpretable intermediate results.

We also design an end-to-end spatio-temporal reasoning model that compresses long 4D sequences into LLM-compatible tokens via instance-aware encoding and adaptive down-sampling. Experiments show our method with CoT supervision outperforms prior baselines, setting a new standard for egocentric dynamic scene understanding.

Our main contributions are:

1. We introduce **EgoDynamic4D**, a novel highly dynamic 4D scene QA benchmark with 927K QA pairs and ex-

*Corresponding author.

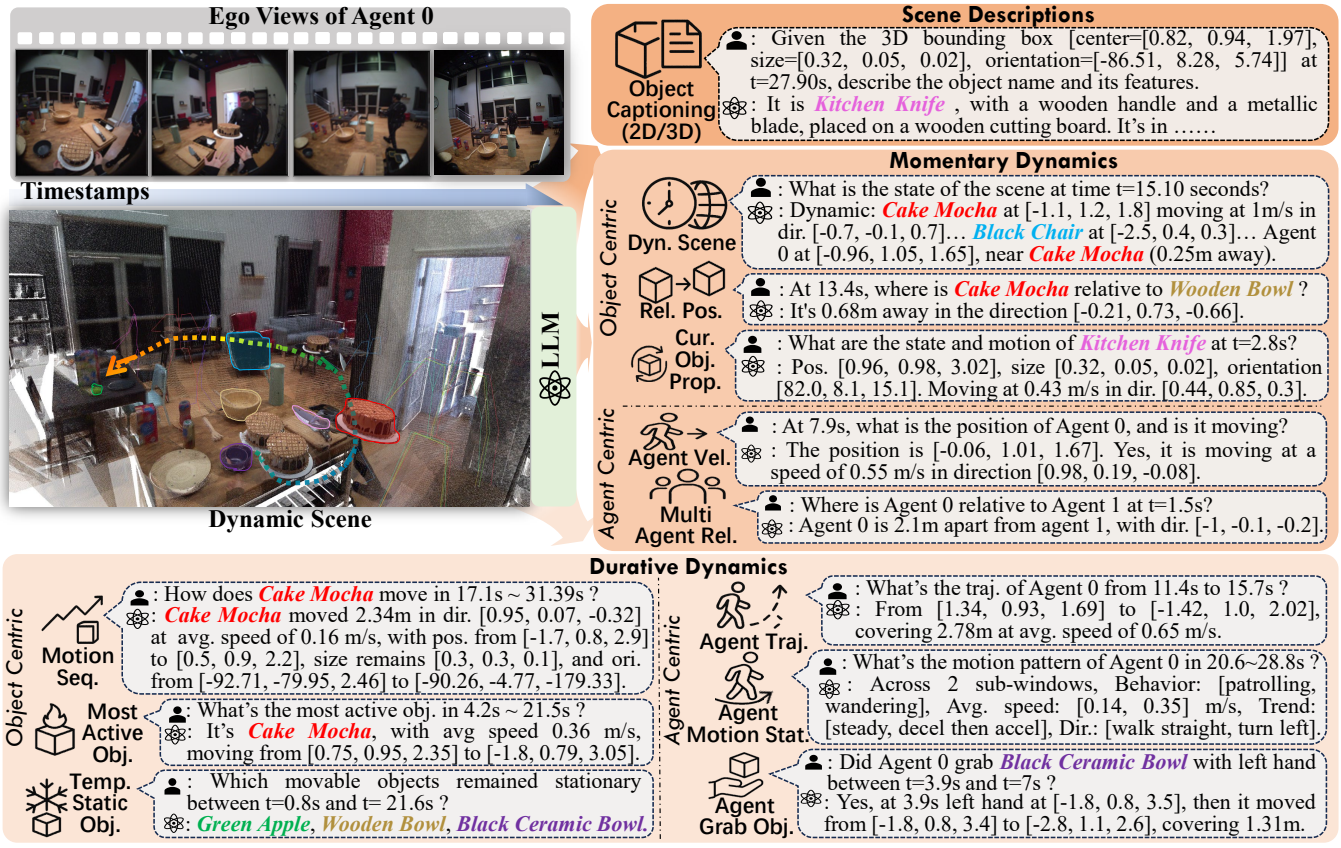


Figure 1: We introduce a novel QA benchmark **EgoDynamic4D** and an end-to-end spatio-temporal reasoning framework.

plicit CoT, covering 12 task types and multimodal data.

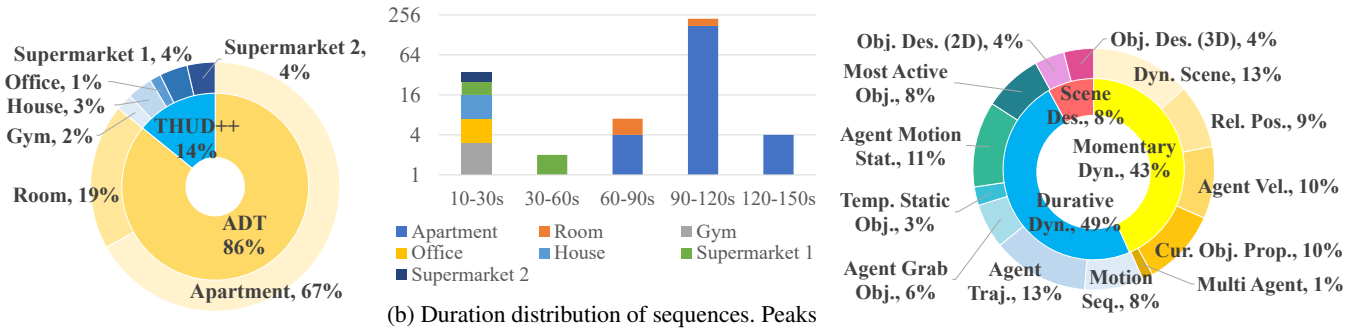
2. We propose an end-to-end spatiotemporal reasoning framework using instance-aware feature encoding, time and camera encoding, and adaptive down-sampling to efficiently handle large 4D data for LLMs.
3. We conduct extensive experiments on EgoDynamic4D using representative models. Our method significantly improves performance and provides interpretable reasoning, establishing a strong baseline for future research.

2 Related Work

Egocentric Video Datasets Egocentric datasets have catalyzed advances in action recognition and object tracking. EpicKitchens (Damen et al. 2018) provides first-person kitchen activity videos but lacks 3D spatial annotations. HD-Epic (Perrett et al. 2025) adds some object localization and short-term action data, but offers only sparse MPS point clouds without dense depth videos, limiting 4D reconstruction. Large-scale datasets like Ego4D (Grauman et al. 2022), EgoExo4D (Grauman et al. 2024) and EgoLife(Yang et al. 2025) omit 3D bounding boxes per frame, which constrains dynamic object reasoning. (Linghu et al. 2024) focuses on contextual reasoning in 3D scenes rather than dynamic objects. In contrast, ADT (Pan et al. 2023) and THUD++ (Tang et al. 2024) annotate each frame with RGB-D data, object poses, 2D/3D bounding boxes, and camera trajectories, supporting dynamic scene understanding. EgoDynamic4D builds reorganizing these 2 datasets to create a unified benchmark for dynamic 4D QA tasks.

3D LLMs 3DLLM (Hong et al. 2023) and 3UR-LLM (Xiong et al. 2025) take point clouds as inputs for tasks like description and localization. LLaVA3D (Zhu et al. 2024) and Video3DLLM (Zheng, Huang, and Wang 2025) introduce 3D patches by embedding 2D CLIP features (Radford et al. 2021) with 3D positional encoding. GPT4Scene (Qi et al. 2025) uses BEV images and consistent object IDs to reason about spatial relations. LSceneLLM (Zhi et al. 2025) employs LLM attention to identify relevant regions for fine details. Chat-3D (Wang et al. 2023) and Chat-Scene (Huang et al. 2024) use object identifiers to enable interactive understanding. While effective on static 3D scenes, these approaches lack mechanisms to capture temporal dynamics for 4D reasoning.

4D Scene Understanding Models 4D comprehension requires integrating RGB-D information and cross-time reasoning. PSG4D (Yang et al. 2023) offers a structured 4D scene representation, (Wu et al. 2025) leverages 2D annotations to enhance 4D learning via a 3D mask decoder–based LLM. Both focus on graph generation without direct support for QA tasks, whereas our work proposes a unified framework that introduces 4D dynamic scene QA for LLMs.



(a) Data sources that provide videos captured in different scenes.

(b) Duration distribution of sequences. Peaks at 10-30s (38 sequences) and 90-120s (224 sequences).

(c) QA pairs in different tasks. We define 12 tasks in 3 domains.

Figure 2: Distribution of data in EgoDynamic4D.

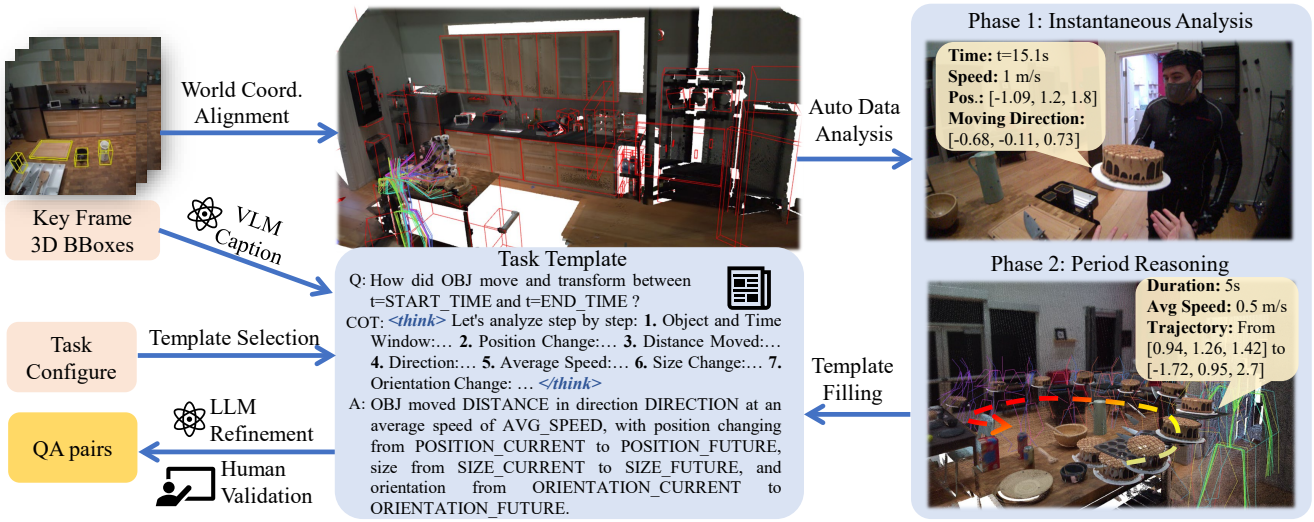


Figure 3: QA generation pipeline. Given RGB-D sequences with aligned 3D bounding boxes and poses, we extract spatial-temporal properties, apply template-based CoT reasoning, and refine questions via LLMs and human validation.

Spatio-Temporal CoT CoT has been extended to reasoning in vision, like (Yuan et al. 2023) and (Yuan, Cai, and Huang 2024). Video-CoT (Wang et al. 2024b) introduces QA pairs with CoT to benchmark reasoning over temporal video content. SpatialCoT (Liu et al. 2025) targets embodied spatial reasoning by aligning vision–language inputs with spatial coordinates. (Zeng et al. 2025) extends this idea to autonomous driving. Likewise, EgoDynamic4D provides rich CoT data for spatio-temporal reasoning.

3 EgoDynamic4D Benchmark

3.1 Overview

The **EgoDynamic4D** benchmark in Figure 1 interprets complex scene dynamics, track moving objects, and analyze interactions from a first-person perspective. Given an input 4D egocentric scene $S = \{F_t\}_{t=1}^{T}$, where $F_t = (I_t, D_t)$ represents a frame at timestamp t with image I_t, depth map D_t, and a natural language question Q about the scene, the model predicts answer A. The answer A can be a categorical label, numerical value, or descriptive text, depending on the specific type of question. We organize the QA tasks into three domains: **Scene Descriptions** capture semantics of the environment. This includes *object-captioning*, which requires understanding of objects and their surroundings given 2D/3D bounding boxes at specific time. **Momentary Dynamics** focus on real-time spatial relations and short-term motion cues. This domain includes object-centric tasks such as *dynamic-scene*, *relative-position*, and *current-object-property*, as well as agent-centric tasks such as *agent-velocity* and *multi-agent-relation*. **Durative Dynamics** address longer-term changes and interactions across time. Object-centric tasks include *temporary-static-objects*, *most-active-object*, and *motion-sequence*, while agent-centric tasks include *agent-trajectory*, *agent-grab-object*, and *agent-motion-status*. Detailed definitions are shown in the appendix.

3.2 QA Data Construction

The **EgoDynamic4D** dataset integrates and enriches two egocentric datasets: ADT (real-world)(Pan et al. 2023) and

Dataset	Annot. level	Scale	RGB-D	4D Boxes	Ego.	Dyn.	CoT
HOT3D (Banerjee et al. 2025)	Partial 3D	833 minutes video	Partial	Partial	✓	✓	✗
SQA3D (Ma et al. 2023)	Static 3D	33.4K questions	Partial	✗	Partial	✗	✗
ScanQA (Azuma et al. 2022)	Static 3D	41K QA pairs	✗	✗	✗	✗	✗
EQA (Yu et al. 2019)	Sparse	5K questions	✗	✗	✓	Partial	✗
EmbodiedScan (Wang et al. 2024a)	Static 3D	1M prompts	✓	✗	✓	Partial	✗
HD-EPIC (Perrett et al. 2025)	Partial 3D	26K questions	✗	Partial	✓	✓	✗
ADL4D (Zakour et al. 2024)	Partial 3D	1.1M frames	✓	Partial	✓	✓	✗
Ego4D (Grauman et al. 2022)	Sparse	3,670h, 5K MCQs	✗	✗	✓	✓	✗
EgoVQA (Fan 2019)	None	600 pairs	✗	✗	✓	✓	✗
EgoTextVQA (Zhou et al. 2025)	None	7K questions	✗	✗	✓	✓	✗
EgoExo4D (Grauman et al. 2024)	Sparse	1,286h	Partial	✗	✓	✓	✗
EgoDynamic4D (Ours)	Full 4D	927K QA pairs	✓	✓	✓	✓	✓

Table 1: Comparison of EgoDynamic4D with spatio-temporal datasets.

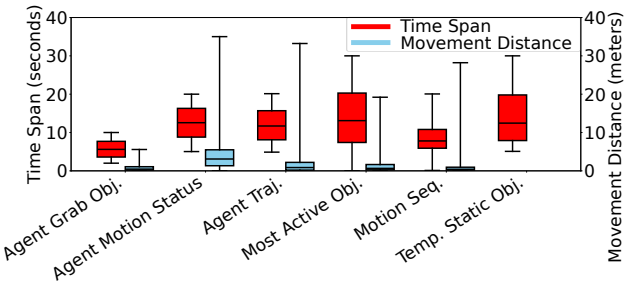


Figure 4: Time span and motion distance distribution.

THUD++ (synthetic scenes)(Tang et al. 2024), comprising a total of 275 carefully selected sequences, as shown in Figure 2a. While the overall number of sequences is modest compared to large-scale vision datasets, each sequence was manually selected to ensure rich dynamics and diversity, and is densely annotated with per-frame 3D bounding boxes and fine-grained QA supervision. The ADT subset contains 236 sequences recorded in real apartment and room environments, covering over 300 static and dynamic objects with high-quality per-frame 3D annotations. The THUD++ subset consists of 39 sequences from synthetic environments (e.g., houses, gyms, offices, supermarkets), with over 100 annotated objects. As shown in Figure 2b, the dataset spans a wide range of durations: shorter clips (10–30s) capture momentary dynamics, while longer ones (up to 150s) enable reasoning over extended temporal contexts.

The EgoDynamic4D dataset is constructed through a multi-stage pipeline (Figure 3). We first extract synchronized RGB-D frames, 6-DoF camera poses, and aligned 3D bounding boxes for all objects, registering them into a shared world coordinate system for consistent spatio-temporal grounding. Then, we process different domains as follows:

Scene Description Given 2D projections of 3D bounding boxes on reference frames, we condition Qwen2.5-VL (Bai et al. 2025) on cropped RGB inputs and depth-aware spatial context to generate object-centric and scene-level captions.

Momentary and Durative Dynamics For dynamic reasoning tasks, we adopt a two-phase QA generation process:

- *Frame-level Analysis:* Key frames and adjacent timestamps are analyzed to compute instantaneous properties (e.g., position, velocity, distance, direction) for each object. Using 5 curated templates, we produce QA pairs that query spatial relations and short-term changes.
- *Temporal Reasoning:* Longer temporal windows are processed using sliding-window analysis to capture object trajectories, agent-object interactions, and causal dynamics. This stage uses 6 additional templates tailored for questions involving temporal comparisons, forecasting, and interaction reasoning.

Each generated QA pair is refined by LLM to improve fluency, consistency, and diversity. We further perform human-in-the-loop verification on the entire corpus to ensure factual correctness and linguistic quality.

3.3 Data Distribution

The EgoDynamic4D dataset is distinguished by its diverse and balanced coverage of different kinds of QA tasks and spatial-temporal spans, reflecting the dynamic complexity of egocentric 4D scenes.

Category Distribution As shown in Figure 2c, the dataset maintains a balanced distribution across 12 task categories, covering both short-term and long-term dynamics. This design enables comprehensive multimodal reasoning from a first-person perspective, with emphasis on object-centric, agent-centric, and agent-object interactions understanding.

Spatio-temporal Distribution Figure 4 illustrates the distribution of motion distances (blue) and temporal spans (red) across five representative tasks. Most object motions occur within 0–10 meters, while the *most-active-object* task reaches up to 35 meters. Temporally, most events span 10–20 seconds, with longer durations (up to 30 seconds) supporting complex temporal reasoning.

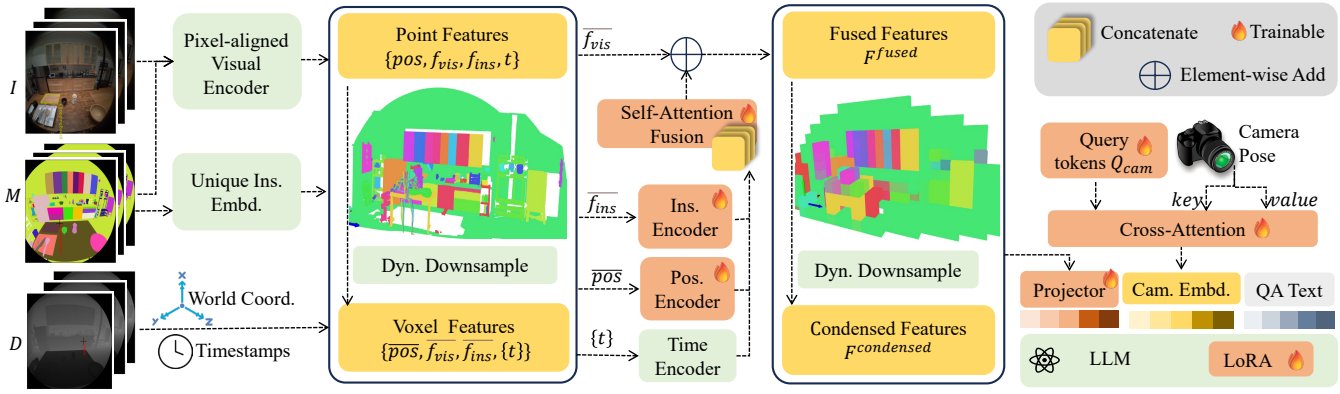


Figure 5: The end-to-end framework which encode dynamic 4D scenes based on the egocentric videos.

3.4 EgoDynamic4D vs Prior Datasets

EgoDynamic4D contains a novel benchmark of 927K QA pairs designed for 12 challenging tasks. Table 1 compares EgoDynamic4D with representative spatio-temporal datasets across key dimensions. Annotation Level (Annot. Level) reflects the granularity of object tracking: Full 4D denotes per-frame 3D bounding boxes over time; other levels (e.g., Partial 3D, Static 3D, Sparse, None) indicate decreasing spatial-temporal completeness. Across all frames in EgoDynamic4D, approximately 31.3% contain at least one dynamic object, with an average of 0.57 dynamic objects and agent interactions present per second. This high temporal density of dynamic events ensures rich supervision for modeling complex egocentric 4D scene dynamics.

4 Methodology

We design an end-to-end framework to encode egocentric 4D scenes by fusing spatial and temporal information. The pipeline (Figure 5) comprises three stages: instances and timestamps enhanced point-level feature extraction, feature fusion, and projection into tokens for LLM inference. During the whole process, we propose 3 novel component: global unique instance embeddings, voxel timestamps encoding and camera embedding.

4.1 Instance and Timestamps Enhanced Point-level Feature Extraction

To capture fine-grained semantics, we extract per-pixel features from all RGB frames and project them into the 4D dynamic point cloud P. Considering the i-th frame in a sequence with T frames in total, let $I^i \in \mathbb{R}^{H\times W\times 3}$, $M^i \in \mathbb{R}^{H\times W}$, $D^i \in \mathbb{R}^{H\times W}$ denote the RGB image, the corresponding instance segmentation map and depth map, respectively.

Pixel-aligned Visual Encoding Inspired by ConceptFusion (Li et al. 2022) and 3DLLM (Hong et al. 2023), we adapt a similair method on our 4D point cloud. We first use a pretrained vision encoder to compute the global feature representation for the frame:

$$F^i_{global} = \text{Pool}(\text{Enc}_{vis}(I^i)) \in \mathbb{R}^{d_{vis}}, \quad (1)$$

where d_{vis} is the dimension of the vision encoder. For segmentations of N instances in M^i, we clip corresponding regions $I^i_j \in \mathbb{R}^{h\times w}$ $(j \in \{1,2,...,N\}, h < H, w < W)$, then compute local features similar to Eq 1, getting $F^i_j \in \mathbb{R}^{d_{vis}}$. We get the point cloud of the i-th frame in world coordinate space P^i . Then, for every single point in the local region P^i_j corresponding to the j-th instance segment, we obtain the vision feature f_{vis} by weighted average:

$$f_{vis} = sim^i_j \cdot F^i_{global} + (1 - sim^i_j) \cdot F^i_j,$$
$$\text{where } sim^i_j = \frac{F^i_j \cdot F^i_{global}}{\|F^i_j\| \cdot \|F^i_{global}\|}. \quad (2)$$

Unique Instance Embedding We assign each instance a unique embedding vector sampled from $\mathcal{N}(0, I)$, and propagate it across all corresponding points in the segmentation mask. Let d_{ins} be the embedding dimension, then each point's instance encoding is $f_{ins} \in \mathbb{R}^{d_{ins}}$. In high-dimensional spaces, randomly sampled vectors are nearly orthogonal with high probability (Ghojogh et al. 2021), enabling a large number of identities to be distinguished. This enriches point features with instance identity information.

Timestamps Each point is marked with its timestamp t. Combining position $pos \in \mathbb{R}^3$ with visual features, instance embeddings, and timestamps, the final representation of each point p is $f_p = \{pos_p, f_{vis,p}, f_{ins,p}, t_p\}$. Since points within a segmented region share identical features per frame, the enhanced local point cloud representation is $S^i_j = \{f_p \mid p \in P^i_j\} \in \mathbb{R}^{|P|\times(3+d_{vis}+d_{ins}+1)}$, where $|P|$ denotes the number of points. Thus, the full point cloud representation S integrates 4D information.

4.2 Feature Fusion

To reduce the computational complexity of the sparse feature S, we apply an octree-based dynamic downsampling on spatial positions, and voxel features are aggregated per node. For each node, point positions, visual features, and instance embeddings are averaged, while all timestamps are collected into a set $\{t\}$. This yields voxels V, each represented by $f_v = \{\overline{pos_v}, \overline{f_{vis,v}}, \overline{f_{ins,v}}, \{t\}_v\}, v \in V$. The process reduces voxel count from 50M–300M to 100K–250K.

Subset	Method	Scene Des.	Momentary Dynamics					Durative Dynamics						**Overall BLEU-4**
		obj. caption (bleu4)↑	dyn. scene (f1)↑	rel. pos. (acc%)↑	curr. obj. prop. (acc%)↑	agent vel. (acc%)↑	multi agent rel. (acc%)↑	motion seq. (acc%)↑	most active obj. (acc%)↑	temp. static obj. (f1)↑	agent traj. (acc%)↑	agent motion status (acc%)↑	agent grab obj. (acc%)↑	
ADT	LLaVA-3D	0.072	0.290	42.56	30.12	23.07	29.94	25.78	28.62	0.718	24.21	24.30	12.30	0.388
	Video3DLLM	0.034	0.307	35.65	27.97	24.55	22.31	23.80	27.47	0.737	24.07	23.22	8.34	0.392
	VG-LLM	0.211	0.297	43.54	43.72	25.95	29.52	26.48	29.58	0.713	26.51	26.02	16.38	0.406
	3DLLM	0.033	0.003	30.48	14.52	20.49	14.78	17.69	2.35	0.502	6.96	18.18	5.96	0.345
	LL3DA	0.014	0.006	14.68	0.90	23.21	17.04	18.78	8.38	0.166	18.07	21.29	1.49	0.287
	Chat-Scene	0.000	0.104	39.60	0.00	8.25	29.39	0.00	23.28	0.628	8.13	0.00	8.98	0.187
	Ours	**0.244**	**0.455**	49.79	58.39	**31.32**	55.47	40.56	41.98	**0.763**	46.11	**29.94**	28.25	0.435
	Ours+CoT	0.238	0.454	**84.11**	**67.35**	19.33	**65.72**	**57.04**	**56.82**	0.726	**47.35**	27.69	**29.64**	**0.436**
THUD++	LLaVA-3D	0.034	0.152	10.99	9.46	45.91	–	11.01	10.03	0.436	0.80	37.60	–	0.370
	Video3DLLM	0.040	0.087	9.66	3.11	44.46	–	9.47	3.37	0.410	1.63	36.81	–	0.349
	VG-LLM	0.083	0.002	9.41	1.55	44.86	–	10.26	4.07	0.014	1.41	39.85	–	0.354
	3DLLM	0.009	0.000	9.82	0.00	36.69	–	10.88	0.69	0.000	3.57	31.87	–	0.312
	LL3DA	0.002	0.000	6.66	0.00	32.91	–	7.08	0.60	0.000	1.90	24.89	–	0.265
	Chat-Scene	0.000	0.024	13.70	0.00	20.55	–	0.00	19.12	0.386	3.42	0.00	–	0.185
	Ours	0.097	0.221	13.14	27.68	**59.24**	–	26.10	21.89	0.472	7.36	50.42	–	0.403
	Ours+CoT	**0.105**	**0.333**	**61.66**	**65.49**	48.59	–	**43.67**	**32.66**	**0.596**	**18.33**	**55.58**	–	**0.431**

Table 2: Baseline models, best results in **bold** and the second with underline.

Time Encoding We encode timestamps to capture dynamic evolution. Let $\{t_{v,1}, ..., t_{v,k}, ..., t_{v,K}\}$ be the K times when points appears in the voxel v. Then compute sinusoidal features $s_{v,k}$ of $t_{v,k}$ as Eq 3:

$$m \in \{0, 1, ..., \lfloor \tfrac{d_{vis}}{2} \rfloor - 1\},\ d_m = \exp\big(-\tfrac{\ln 10^4}{d_{vis}}\, m\big),$$
$$s_{v,k}^{2m} = \sin\big(t_{v,k} \cdot d_m\big),\ s_{v,k}^{2m+1} = \cos\big(t_{v,k} \cdot d_m\big). \tag{3}$$

Then we aggregate them via max and mean pooling as Eq 4, where α is the factor. This Fourier basis supports distinguishing intervals up to 10^4 s.

$$t_v^{emb} = \alpha \cdot \max_k s_{v,k} + (1-\alpha) \cdot \operatorname*{avg}_k s_{v,k} \in \mathbb{R}^{d_{vis}} \tag{4}$$

Instance Encoding and Feature Integration A learnable matrix $W_{ins} \in \mathbb{R}^{d_{vis} \times d_{ins}}$ projects instance embedding $\overline{f_{vis,v}}$ into hidden dimension, then it is fused with time embeddings and original position embeddings via self-attention (SA), adding to visual feature together.

$$f_v^{fused} = \overline{f_{vis,v}} + \mathrm{SA}\big([\, W_{ins} \cdot \overline{f_{ins,v}} \,\|\, t_v^{emb} \| \mathrm{Enc}_{pos}(\overline{pos_v})\,]\big) \tag{5}$$

After obtaining the fused voxel features $F^{fused} \in \mathbb{R}^{|V| \times d_{vis}}$, where $|V|$ is the number of voxels, we apply another downsampling to yield condensed representation $F^{condensed}$ ($\sim$1K tokens) for LLM processing.

Camera Embedding We compress the sequence of camera poses $(x, y, z, q_x, q_y, q_z, q_w) \in \mathbb{R}^7$ into a compact embedding using a learnable attention mechanism. A sequence of T poses are first project into hidden space $f_{cam} \in \mathbb{R}^{T \times d_{vis}}$, and then M learnable query tokens $Q_{cam} \in \mathbb{R}^{M \times d_{vis}}$ attends to f_{cam} through cross-attention (CA):

$$F_{cam} = \mathrm{CA}(Q_{cam}, f_{cam}, f_{cam}) \in \mathbb{R}^{M \times d_{vis}} \tag{6}$$

Finally, $F^{condensed}$ and F_{cam} are projected into LLM embedding space, then we use LoRA for efficient fine-tuning.

5 Experiments

5.1 Implementation Details

We conduct experiments on EgoDynamic4D (80% train, 20% test). Our model builds upon LLaVA3D (CLIP + LLaMA) (Zhu et al. 2024) with frozen backbones, unfreezing only the proposed modules (d_{ins}=8, M=8) and LoRA parameters (rank=8, alpha=16). We set sample fps to 5, and train using AdamW (learning rate 5e-5) for 2 epochs on 8 × RTX 4090 (24GB) with batch size 1 per GPU.

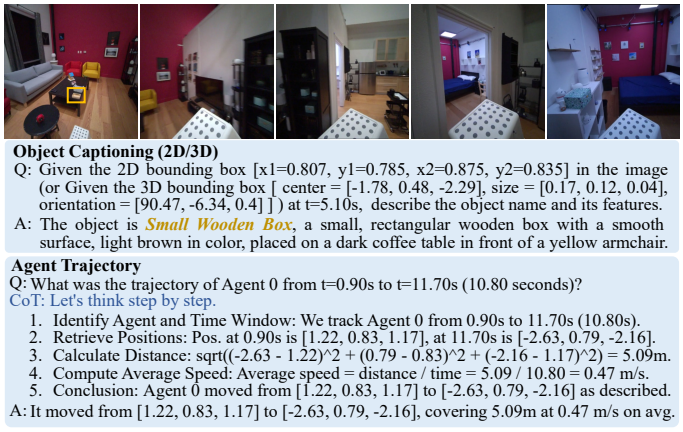
5.2 Results on EgoDynamic4D

We evaluate our method on the EgoDynamic4D dataset (ADT and THUD++ subsets), and benchmark it against representative baselines. Since 4D LLMs like LLaVA-4D (Zhou and Lee 2025) are not publicly available yet, we evaluated 3D LLMs including LLaVA-3D (Zhu et al. 2024), Video3DLLM (Zheng, Huang, and Wang 2025), LL3DA (Chen et al. 2024), Chat-Scene (Huang et al. 2024), 3DLLM (Hong et al. 2023) and VG-LLM (Zheng et al. 2025). To ensure a fair comparison under limited resources, we fine-tune all LLM models with the same LoRA configuration, while keeping other components in original implementation settings. Evaluation follows the benchmark's protocol (details in the appendix), with thresholds settings: speed error within $0.05\ m/s$, direction error within $0.5\ rad$, IOU above 0.1, distance and position errors within $0.1\ m$.

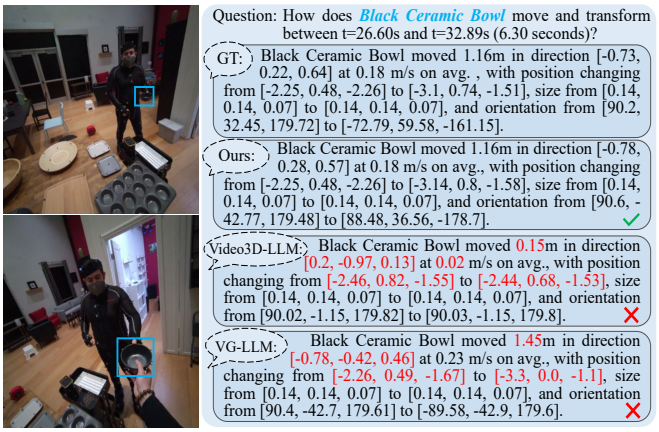
As shown in Table 2, our models consistently outperform all baseline methods on EgoDynamic4D across both subsets, demonstrating robust spatial-temporal reasoning capabilities. Our CoT-enhanced model achieves top performance in overall BLEU-4 and most tasks, showcasing its strength in capturing complex 4D scene dynamics and relationships. Our non-CoT model also surpasses baselines, delivering competitive results across all metrics. This consis-

Subset	Settings	Scene Des.	Momentary Dynamics					Durative Dynamics						**Overall BLEU-4**
		obj. caption (bleu4)↑	dyn. scene (f1)↑	rel. pos. (acc%)↑	curr. obj. prop. (acc%)↑	agent vel. (acc%)↑	multi agent rel. (acc%)↑	motion seq. (acc%)↑	most active obj. (acc%)↑	temp. static obj. (f1)↑	agent traj. (acc%)↑	agent motion status (acc%)↑	agent grab obj. (acc%)↑	
ADT	whole	**0.244**	**0.454**	**49.79**	**58.39**	**31.32**	**55.47**	**40.56**	**41.98**	**0.763**	**46.11**	**29.94**	**28.25**	**0.435**
	w/o c	0.239	0.450	49.16	48.72	30.34	54.11	39.52	40.74	0.757	43.82	28.58	24.70	0.432
	w/o c&i	0.237	0.433	48.50	48.39	30.12	52.49	37.47	39.60	0.757	42.75	28.03	24.59	0.429
	w/o c&i&t	0.234	0.346	48.37	37.30	26.72	42.41	31.95	33.37	0.739	31.22	25.55	19.34	0.411
	mlp	0.236	0.438	48.94	45.92	30.55	53.11	36.18	41.59	0.754	43.72	28.73	23.75	0.429
THUD++	whole	**0.097**	**0.221**	13.14	27.68	**59.24**	–	**26.10**	**21.89**	**0.472**	**7.36**	50.42	–	**0.403**
	w/o c	0.095	0.201	13.14	26.55	57.67	–	21.18	17.30	0.424	6.55	48.67	–	0.400
	w/o c&i	0.087	0.189	12.22	24.58	57.63	–	18.20	16.96	0.411	5.31	48.45	–	0.395
	w/o c&i&t	0.071	0.176	11.91	22.46	55.77	–	17.11	16.87	0.402	3.90	48.36	–	0.394
	mlp	0.093	0.195	12.99	**31.07**	59.10	–	20.04	18.86	0.448	6.28	**52.09**	–	0.399

Table 3: Ablation on camera (c), instance (i), time (t) embedding, and MLP feature fusion (mlp, with c&i&t).



(a) Examples of *object-captioning* and *agent-trajectory* tasks.



(b) Comparison of baselines on *motion-sequence* task.

Figure 6: Qualitative examples illustrating the complexity of spatial-temporal reasoning in EgoDynamic4D.

tent superiority highlights the effectiveness of our approach in handling diverse 4D environments, with CoT further enhancing precision in complex spatial-temporal tasks.

5.3 Ablation Study

Table 3 evaluates camera, instance, time embeddings, and also compares attention-based feature fusion (whole) with MLP-based fusion (mlp) on EgoDynamic4D. The full model achieves the highest overall BLEU-4, excelling in dynamic tasks. On THUD++, where motion patterns in *current-object-property* and *agent-motion-status* are less diverse than in ADT, the MLP-based fusion outperforms the full model, likely due to MLP's localized feature fusion preserving fine-grained details in less dynamic tasks, whereas attention's global integration may introduce noise. Ablations confirm that camera, instance, and time embeddings enhance spatial, object-specific, and temporal reasoning.

5.4 Qualitative Analysis

Figure 6a showcases examples from the dataset, including *object-captioning* given 2D or 3D bounding boxes, and *agent-trajectory* that is solved using CoT. Figure 6b compares our model against other baselines on *motion-sequence* task. The results highlight the challenging nature of EgoDynamic4D benchmark and the effectiveness of our method in capturing fine-grained dynamics.

6 Discussion and Conclusion

We introduced **EgoDynamic4D**, a novel benchmark for highly dynamic egocentric 4D scene understanding, featuring 927K QA pairs across 12 diverse task types that capture dense, fine-grained temporal dynamics. To support spatio-temporal reasoning, we proposed a unified framework incorporating egocentric pose embedding, global instance encoding, and Fourier-based timestamp embedding. An octree-based downsampling strategy ensures efficiency while preserving structural integrity. Experiments show strong performance, though challenges remain in generalization and robustness to degraded point clouds. We leave these directions for future work. EgoDynamic4D lays a new foundation for egocentric 4D understanding, with broad relevance to embodied AI and robotics.

Acknowledgments

This work was sponsored by the National Key R&D Program of China (No. 2024YFF0907803), the National Natural Science Foundation of China (No. 62576308), and the Fundamental Research Funds for the Central Universities (No. 226-2025-00167).

References

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Banerjee, P.; Shkodrani, S.; Moulon, P.; Hampali, S.; Han, S.; Zhang, F.; Zhang, L.; Fountain, J.; Miller, E.; Basol, S.; Newcombe, R.; Wang, R.; Engel, J. J.; and Hodan, T. 2025. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In *CVPR*.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*.

Chen, S.; Chen, X.; Zhang, C.; Li, M.; Yu, G.; Fei, H.; Zhu, H.; Fan, J.; and Chen, T. 2024. 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 (CVPR)*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*.

Fan, C. 2019. EgoVQA - An Egocentric Video Question Answering Benchmark Dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.

Fan, C.; Lee, J.; Xu, M.; Singh, K. K.; Lee, Y. J.; Crandall, D. J.; and Ryoo, M. S. 2017. Identifying First-Person Camera Wearers in Third-Person Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ghojogh, B.; Ghodsi, A.; Karray, F.; and Crowley, M. 2021. Johnson-Lindenstrauss lemma, linear and nonlinear random projections, random Fourier features, and random kitchen sinks: Tutorial and survey. *arXiv preprint arXiv:2108.04172*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; Martin, M.; Nagarajan, T.; Radosavovic, I.; Ramakrishnan, S. K.; Ryan, F.; Sharma, J.; Wray, M.; Xu, M.; Xu, E. Z.; Zhao, C.; Bansal, S.; Batra, D.; Cartillier, V.; Crane, S.; Do, T.; Doulaty, M.; Erapalli, A.; Feichtenhofer, C.; Fragomeni, A.; Fu, Q.; Gebreselasie, A.; González, C.; Hillis, J.; Huang, X.; Huang, Y.; Jia, W.; Khoo, W.; Kolár, J.; Kottur, S.; Kumar, A.; Landini, F.; Li, C.; Li, Y.; Li, Z.; Mangalam, K.; Modhugu, R.; Munro, J.; Murrell, T.; Nishiyasu, T.; Price, W.; Puentes, P. R.; Ramazanova, M.; Sari, L.; Somasundaram, K.; Southerland, A.; Sugano, Y.; Tao, R.; Vo, M.; Wang, Y.; Wu, X.; Yagi, T.; Zhao, Z.; Zhu, Y.; Arbeláez, P.; Crandall, D.; Damen, D.; Farinella, G. M.; Fuegen, C.; Ghanem, B.; Ithapu, V. K.; Jawahar, C. V.; Joo, H.; Kitani, K.; Li, H.; Newcombe, R. A.; Oliva, A.; Park, H. S.; Rehg, J. M.; Sato, Y.; Shi, J.; Shou, M. Z.; Torralba, A.; Torresani, L.; Yan, M.; and Malik, J. 2022. Ego4D: Around the World in 3, 000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; Byrne, E.; Chavis, Z.; Chen, J.; Cheng, F.; Chu, F.-J.; Crane, S.; Dasgupta, A.; Dong, J.; Escobar, M.; Forigua, C.; Gebreselasie, A.; Haresh, S.; Huang, J.; Islam, M. M.; Jain, S.; Khirodkar, R.; Kukreja, D.; Liang, K. J.; Liu, J.-W.; Majumder, S.; Mao, Y.; Martin, M.; Mavroudi, E.; Nagarajan, T.; Ragusa, F.; Ramakrishnan, S. K.; Seminara, L.; Somayazulu, A.; Song, Y.; Su, S.; Xue, Z.; Zhang, E.; Zhang, J.; Castillo, A.; Chen, C.; Fu, X.; Furuta, R.; González, C.; Gupta, P.; Hu, J.; Huang, Y.; Huang, Y.; Khoo, W.; Kumar, A.; Kuo, R.; Lakhavani, S.; Liu, M.; Luo, M.; Luo, Z.; Meredith, B.; Miller, A.; Oguntola, O.; Pan, X.; Peng, P.; Pramanick, S.; Ramazanova, M.; Ryan, F.; Shan, W.; Somasundaram, K.; Song, C.; Southerland, A.; Tateno, M.; Wang, H.; Wang, Y.; Yagi, T.; Yan, M.; Yang, X.; Yu, Z.; Zha, S. C.; Zhao, C.; Zhao, Z.; Zhu, Z.; Zhuo, J.; Arbeláez, P.; Bertasius, G.; Damen, D.; Engel, J.; Maria Farinella, G.; Furnari, A.; Ghanem, B.; Hoffman, J.; Jawahar, C. V.; Newcombe, R.; Park, H. S.; Rehg, J. M.; Sato, Y.; Savva, M.; Shi, J.; Shout, M. Z.; and Wray, M. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. In *Advances in Neural Information Processing Systems*.

Huang, H.; Chen, Y.; Wang, Z.; Huang, R.; Xu, R.; Wang, T.; Liu, L.; Cheng, X.; Zhao, Y.; Pang, J.; et al. 2024. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*.

Lee, D.; Lee, J.; and Choi, J. 2023. CAST: Cross-Attention in Space and Time for Video Action Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Li, J.; Niu, L.; and Zhang, L. 2022. From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Y.; Nagarajan, T.; Xiong, B.; and Grauman, K. 2021. Ego-Exo: Transferring Visual Representations from Third-person to First-person Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, K. Q.; Wang, J.; Soldan, M.; Wray, M.; Yan, R.; Xu, E. Z.; Gao, D.; Tu, R.-C.; Zhao, W.; Kong, W.; et al. 2022. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*.

Linghu, X.; Huang, J.; Niu, X.; Ma, X.; Jia, B.; and Huang, S. 2024. Multi-modal Situated Reasoning in 3D Scenes. *Advances in Neural Information Processing Systems*.

Liu, Y.; Chi, D.; Wu, S.; Zhang, Z.; Hu, Y.; Zhang, L.; Zhang, Y.; Wu, S.; Cao, T.; Huang, G.; et al. 2025. SpatialCoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning. *arXiv preprint arXiv:2501.10074*.

Ma, X.; Yong, S.; Zheng, Z.; Li, Q.; Liang, Y.; Zhu, S.-C.; and Huang, S. 2023. SQA3D: Situated Question Answering in 3D Scenes. In *International Conference on Learning Representations*.

Pan, T.; Song, Y.; Yang, T.; Jiang, W.; and Liu, W. 2021. VideoMoCo: Contrastive Video Representation Learning with Temporally Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pan, X.; Charron, N.; Yang, Y.; Peters, S.; Whelan, T.; Kong, C.; Parkhi, O.; Newcombe, R.; and Ren, Y. C. 2023. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20133–20143.

Pang, B.; Peng, G.; Li, Y.; and Lu, C. 2021. PGT: A Progressive Method for Training Models on Long Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Perrett, T.; Darkhalil, A.; Sinha, S.; Emara, O.; Pollard, S.; Parida, K.; Liu, K.; Gatti, P.; Bansal, S.; Flanagan, K.; Chalk, J.; Zhu, Z.; Guerrier, R.; Abdelazim, F.; Zhu, B.; Moltisanti, D.; Wray, M.; Doughty, H.; and Damen, D. 2025. HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pramanick, S.; Song, Y.; Nag, S.; Lin, K. Q.; Shah, H.; Shou, M. Z.; Chellappa, R.; and Zhang, P. 2023. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Qi, Z.; Zhang, Z.; Fang, Y.; Wang, J.; and Zhao, H. 2025. GPT4Scene: Understand 3D Scenes from Videos with Vision-Language Models. *arXiv:2501.01428*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

Song, S.; Lin, X.; Liu, J.; Guo, Z.; and Chang, S.-F. 2021. Co-Grounding Networks with Semantic Attention for Referring Expression Comprehension in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tang, Y.-F.; andFang Xin Chen, C. T.; Zhang, W.-T.; Zhang, T.; Liu, Y.-J.; and Zeng*, L. 2024. Mobile Oriented Large-Scale Indoor Dataset for Dynamic Scene Understanding. In *Mobile Oriented Large-Scale Indoor Dataset for Dynamic Scene Understanding, submitted to IEEE International Conference Robotic and Automation, 2024*.

Wang, T.; Mao, X.; Zhu, C.; Xu, R.; Lyu, R.; Li, P.; Chen, X.; Zhang, W.; Chen, K.; Xue, T.; Liu, X.; Lu, C.; Lin, D.; and Pang, J. 2024a. EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y.; Zeng, Y.; Zheng, J.; Xing, X.; Xu, J.; and Xu, X. 2024b. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*.

Wang, Z.; Huang, H.; Zhao, Y.; Zhang, Z.; and Zhao, Z. 2023. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*.

Wu, S.; Fei, H.; Yang, J.; Li, X.; Li, J.; Zhang, H.; and Chua, T.-s. 2025. Learning 4D Panoptic Scene Graph Generation from Rich 2D Visual Scene. *arXiv preprint arXiv:2503.15019*.

Xiong, H.; Zhuge, Y.; Zhu, J.; Zhang, L.; and Lu, H. 2025. 3UR-LLM: An End-to-End Multimodal Large Language Model for 3D Scene Understanding. *arXiv preprint arXiv:2501.07819*.

Yang, J.; Cen, J.; Peng, W.; Liu, F., Shuai amd Hong; Li, X.; Zhou, K.; Chen, Q.; and Liu, Z. 2023. 4D Panoptic Scene Graph Generation. In *Advances in Neural Information Processing Systems*.

Yang, J.; Liu, S.; Guo, H.; Dong, Y.; Zhang, X.; Zhang, S.; Wang, P.; Zhou, Z.; Xie, B.; Wang, Z.; Ouyang, B.; Lin, Z.; Cominelli, M.; Cai, Z.; Zhang, Y.; Zhang, P.; Hong, F.; Widmer, J.; Gringoli, F.; Yang, L.; Li, B.; and Liu, Z. 2025. EgoLife: Towards Egocentric Life Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, H.; Cai, M.; Liu, Y.; and Lu, F. 2023. First- And Third-Person Video Co-Analysis By Learning Spatial-Temporal Joint Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 6631–6646.

Yu, L.; Chen, X.; Gkioxari, G.; Bansal, M.; Berg, T. L.; and Batra, D. 2019. Multi-Target Embodied Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuan, L.; Cai, Y.; and Huang, J. 2024. Few-shot joint multimodal entity-relation extraction via knowledge-enhanced cross-modal prompt model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8701–8710.

Yuan, L.; Cai, Y.; Wang, J.; and Li, Q. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, 11051–11059.

Zakour, M.; Nath, P. P.; Lohmer, L.; Gökçe, E. F.; Piccolrovazzi, M.; Patsch, C.; Wu, Y.; Chaudhari, R.; and Steinbach, E. 2024. Adl4d: Towards a contextually rich dataset for 4d activities of daily living. *arXiv preprint arXiv:2402.17758*.

Zeng, S.; Chang, X.; Xie, M.; Liu, X.; Bai, Y.; Pan, Z.; Xu, M.; and Wei, X. 2025. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685*.

Zheng, D.; Huang, S.; Li, Y.; and Wang, L. 2025. Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors. *arXiv e-prints*, arXiv:2505.24625.

Zheng, D.; Huang, S.; and Wang, L. 2025. Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhi, H.; Chen, P.; Li, J.; Ma, S.; Sun, X.; Xiang, T.; Lei, Y.; Tan, M.; and Gan, C. 2025. LSceneLLM: Enhancing Large 3D Scene Understanding Using Adaptive Visual Preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, H.; and Lee, G. H. 2025. LLaVA-4D: Embedding SpatioTemporal Prompt into LMMs for 4D Scene Understanding. *arXiv preprint arXiv:2505.12253*.

Zhou, S.; Xiao, J.; Li, Q.; Li, Y.; Yang, X.; Guo, D.; Wang, M.; Chua, T.-S.; and Yao, A. 2025. EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, C.; Wang, T.; Zhang, W.; Pang, J.; and Liu, X. 2024. LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness. *arXiv preprint arXiv:2409.18125*.