OccLLaMA: A Unified Occupancy-Language-Action World Model for Understanding and Generation Tasks in Autonomous Driving

Julong Wei[1], Shanshuai Yuan[1], Pengfei Li[2], Xinyi Quan[1], Lei Tai[4], Jieru Zhao[3], Zhongxue Gan[1], Wenchao Ding[1]

***Abstract*— Scene understanding via multi-modal large language models (MLLMs) and scene forecasting with world models have advanced the development of autonomous driving. The former maps visual inputs to driving-specific outputs, neglecting spatial reasoning and world dynamics. The latter captures world dynamics, lacking comprehensive scene understanding. In contrast, humans seamlessly integrate understanding, forecasting, and decision-making via multi-modal representations, avoiding misalignment and complexity. To this end, we propose OccLLaMA, a unified occupancy-language-action world model for multi-task learning. It uses semantic occupancy as a unified and modality-agnostic 3D visual representation, effectively integrating spatial scene understanding and scene forecasting. We further introduce a novel scene tokenizer tailored for occupancy, enabling a unified representation manner for multi-task across understanding and generation. Furthermore, we enhance LLM, specifically LLaMA, to enable end-to-end multi-task learning within a unified auto-regressive framework. Extensive experiments demonstrate that OccLLaMA not only achieves competitive performance on multi-task,including scene understanding, occupancy forecasting and motion planning, but also significantly enhances motion planning performance by the integration of multi-task learning, showcasing its effectiveness and potential as a foundation model for autonomous driving.**

I. INTRODUCTION

Multi-modal Large Language Models (MLLMs), as foundation models trained on massive internet-scale datasets, offer a new perspective for autonomous driving (AD). MLLMs combine extensive "world knowledge" with advanced reasoning capabilities, which is exactly what traditional AD models lack [1]. However, fully autonomous driving based on MLLMs has yet to become a reality. The reason is that existing MLLMs are good at scene understanding but are poor at planning specific actions, neglecting the dynamics of the world and the relations between action and world dynamics. In contrast, humans possess a world model that enables them to understand scenes, simulate the future and plan actions simultaneously. Therefore, exploring how to construct a human-like world model is essential for advancing autonomous driving.

There has seen extensive research on world models for autonomous driving. However, **the precise definition of a world model for autonomous driving remains an open question.** Current world models for autonomous driving primarily focus on sensor-based prediction tasks, such as video prediction [2], point cloud prediction [3] and occupancy prediction [4]. Yet these models fail to simultaneously achieve language reasoning, scene forecasting, and action-based interaction with the real world, which are precisely defining attributes of human intelligence. Therefore, we target at a model capable of unifying vision, language, and action (VLA) modalities and incorporating advanced spatial-temporal scene understanding capabilities.

However, two critical challenges must be solved for building such a VLA world model. The first is to build a general 3D visual representation that facilitates both understanding and world evolution, and the second is to design a multi-modal framework capable of accommodating VLA modal-

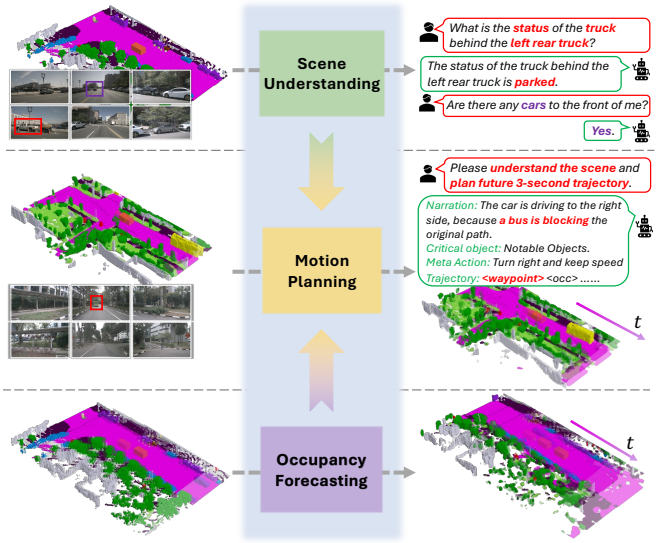

Fig. 1: Humans make decisions by simultaneously understanding the scene and forecasting future world dynamics through internal multi-modal representations. Inspired by this, we propose OccLLaMA, a unified world model integrating occupancy-language-action modalities for multiple tasks in autonomous driving. Moreover, OccLLaMA constructs a human-like motion planning process by integrating scene understanding as a prerequisite task and alternately forecasting scenes and planning waypoints. The significantly enhanced performance achieved by OccLLaMA highlights the effectiveness of multi-task learning and integration in improving driving capability.

[1]Academy for Engineering and Technology, Fudan University, China. [2]Institute for AI Industry Research, Tsinghua University, China. [3]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. [4]Horizon Robotics.

Emails: {jlwei23, ssyuan23, xyquan24}@m.fudan edu.cn, {ganzhongxue, dingwenchao}@fudan.edu.cn, li-pf22@mails.tsinghua.edu.cn, zhao-jieru@cs.sjtu.edu.cn, onlytailei@hotmail.com

ities, integrating world modeling, scene understanding and planning. In recent years, semantic occupancy (Occ) has gained significant attention as a general 3D visual representation. It can capture fine-grained 3D structures while incorporating high-level semantic information, making it highly suitable for spatial-semantic alignment. This motivates us to utilize Occ as the intermediate representation, which smoothly connects world semantics and 3D geometric representation. On the other hand, *MLLMs has been validated to be poor at spatial understanding*. It becomes an interesting question that whether LLMs can learn spatial-temporal reasoning by predicting 3D occupancy tokens. This can potentially open up new possibilities of conducting 4D scene understanding through LLMs. Moreover, LLMs also allow for incorporating planning capabilities into next-token prediction, which is highly promising for a unified and clean VLA framework.

Based on above observations, we propose OccLLaMA, an occupancy-language-action generative world model. As illustrated in Figure 1, OccLLaMA uses semantic occupancy as the 3D visual representation to integrate understanding and evolution through an auto-regressive model, unifying VLA-related tasks including scene understanding, occupancy forecasting and motion planning. Specifically, we introduce a general scene tokenizer to efficiently discretize and reconstruct occupancy scenes, addressing sparsity and class imbalance. Then, we align occupancy-language-action modalities into a unified space. Furthermore, we enhance LLM, specifically LLaMA [5], to perform spatial-temporal scene understanding and planning on the unified space as a human-like VLA world model.

We summarize our contributions as follows:

- A unified occupancy-language-action world model, OccLLaMA, which integrates scene understanding, occupancy forecasting and motion planning tasks within an auto-regressive framework.
- A general scene tokenizer that efficiently discretizes and reconstructs occupancy scenes, addressing sparsity and class imbalance while preserving fine-grained spatial and semantic details.
- Extensive experiments compared to SOTA methods from multiple tasks, showing competitive performance across all different tasks, including scene understanding, occupancy forecasting, and motion planning.
- A demonstration of the significant improvement in motion planning through multi-task learning, highlighting the effectiveness of integrating multiple tasks within a unified framework.

II. RELATED WORK

A. VLM Model for Autonomous Driving

The advancement of MLLMs has introduced new paradigms in autonomous driving, particularly in enabling more explainable driving behavior [6], [7] and improving generalizability via end-to-end learning frameworks. DriveGPT4 [8] and CarLLaVA [9] achieve promising results through the fusion of images and LLMs. However, these methods, which primarily rely on 2D image data, could benefit from stronger 3D scene understanding. OmniDrive [10] builds upon this by using sparse queries to generate 3D visual representations, resulting in improved performance in complex environments and dynamic driving scenarios. BeVLM [11] employs an adapter module to align BEV features with LLMs, enhancing spatial understanding and the accuracy of prediction. Other approaches [12], [7] utilize both image and LiDAR modalities as inputs but face challenges in effectively aligning these modalities. In contrast, Semantic Occupancy provides aligned semantic and spatial information, which offers a promising solution to the modality alignment problem.

B. World Model for Autonomous Driving

World models aim to predict future scenes, including spatial structures and object dynamics, based on agent action and observation [13]. In autonomous driving, world models are primarily utilized for generating synthetic training data and supporting decision-making tasks. Visual world models [14], [15] using image representations have been successful in predicting images or videos of driving scenes and simulating driving environments. Several methods [16], [17] utilize 3D point cloud representations, while enhancing the understanding of dynamic, real-world environments by capturing richer spatial information, but they still lack semantic information. Recent advancements focus on leveraging multi-modal sensor setups to build more comprehensive world models. Muvo [18] takes both camera and lidar data as inputs to learn a sensor-agnostic geometric representation of the world. BEVWorld [19] combines multi-modal sensor data into a unified BEV space, enabling future scene prediction using a latent diffusion model. Yet, challenges persist in aligning features from these different modalities. A promising direction for future research is the integration of 3D scene representations with semantic understanding, which could enable more accurate and context-aware predictions.

III. METHOD

As shown in Figure 2, we introduce **OccLLaMA**, a unified occupancy-language-action world model with the Scene Tokenizer (Section III-A) and the Unified World Model (Section III-B). Furthermore, we present our training strategy for tokenizer and multi-task learning (Section III-C).

A. Scene Tokenizer

Occupancy data $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$ represents the ego vehicle surrounding environment as a voxelized $H \times W \times D$ grid, where each cell is assigned a semantic label. The label set $\mathbf{S}$ includes $N-1$ non-air categories $\mathbf{S}_n$ and air category o_a.

$$\mathbf{S} = \{o_1, o_2, ...o_{N-1}, o_a\} = \{\mathbf{S}_n, o_a\} \tag{1}$$

Over 90% of the cells are assigned the low-value o_a, causing extreme sparsity and class imbalance. We thus propose an occupancy-specific VQVAE with a sparse encoder and decoupled decoder, extending existing methods [4].

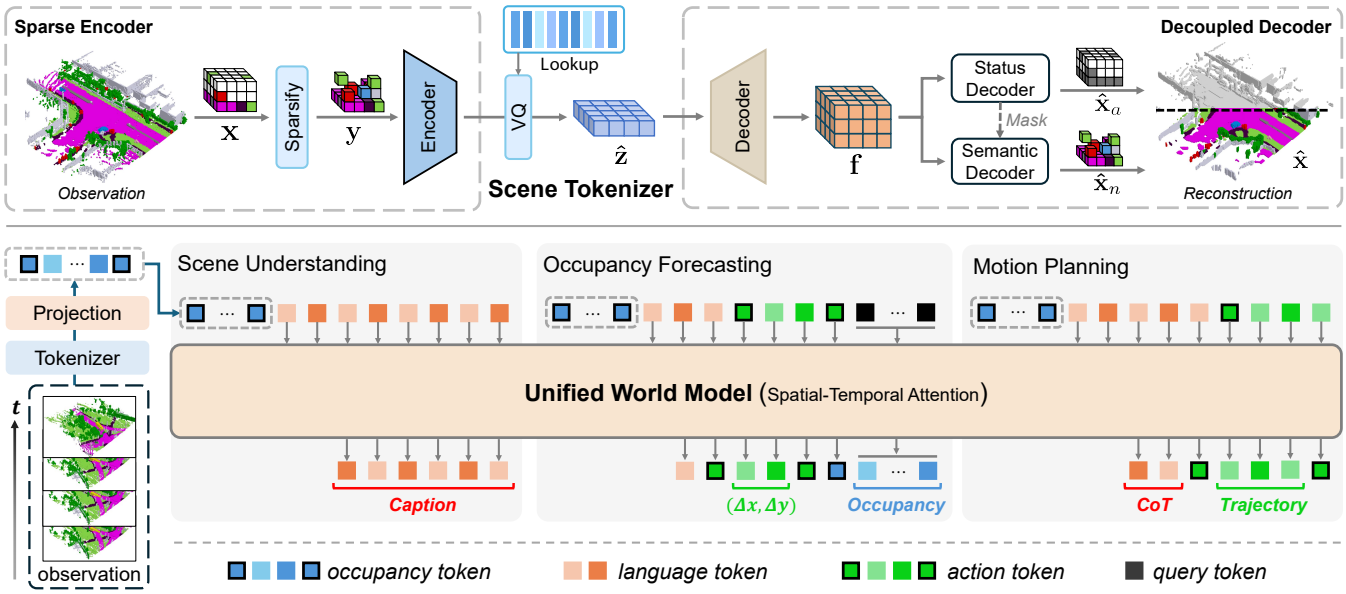

Fig. 2: **Overview of the OccLLaMA Architecture.** The Scene Tokenizer and Unified World Model are core components of OccLLaMA. The Scene Tokenizer employs a sparse encoder and decoupled decoder to efficiently tokenize the occupancy scene, addressing data sparsity and class imbalance. The Unified World Model integrates occupancy-language-action modalities within a unified discrete auto-regressive framework, supporting multi-task learning in autonomous driving.

Sparse Encoder Inspired by point cloud encoding methods, we design a sparse encoder for efficient occupancy compression. Specifically, we construct $H \times W$ pillars by vertically stacking the occupancy data $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$. Within each pillar, the M non-air cells are modeled as point clouds p, where each point is defined by its height $d \in \{0, ..., D\}$ and semantic label $l \in \mathbf{S}_n$. This design enables us to sparsify $\mathbf{x}$ into $\mathbf{y}$ by representing each pillar as a set $\mathbf{P}$ of points p along Bird-Eye-View (BEV) direction.

$$\mathbf{x} \xrightarrow{\text{sparsify}} \mathbf{y} \in \mathbf{P}^{H \times W} = \{(d_m, l_m)\}_M^{H \times W} \tag{2}$$

We then aggregate the pseudo point cloud features using pillar embedding [20], [21] and employ a swin-transformer block [22] to generate the BEV feature map:

$$\mathbf{z} = E(\mathbf{y}) \in \mathbb{R}^{\frac{H}{r} \times \frac{W}{r} \times c} \tag{3}$$

where r is down-sampling rate and c is the feature dimension.

Quantification To obtain discrete tokens, we map $\mathbf{z}$ into a learnable codebook $\mathbf{Z} = \{\hat{z}_i\}_{i=1}^K$ containing K vector entries. Specifically, the vector quantization (VQ) function $Q(\cdot)$ replaces each z_i to the nearest codebook entry $\hat{z}_k$:

$$\hat{\mathbf{z}} = Q(\mathbf{z}) := \arg\min_{\hat{z}_k \in \mathbf{Z}} \|z_i - \hat{z}_k\|_2, z_i \in \mathbf{z} \tag{4}$$

Decoupled Decoder The decoder backbone reconstructs dense features $\mathbf{f} \in \mathbb{R}^{H \times W \times D \times C}$ from $\hat{\mathbf{z}}$ via deconvolution blocks and up-sampling layers [4]. To address class imbalance, we decouple the decoding of the occupancy status and semantics, where the former indicates whether a voxel is occupied or empty and the latter specifies the category of occupied voxels. Specifically, we define a non-air category o_n and an occupancy status mask $\mathbf{M}$ to obtain decoupled ground truth of occupancy status $\mathbf{x}_a$ and semantics $\mathbf{x}_n$:

$$\mathbf{M}(o) = \begin{cases} \text{True}, & o \in \mathbf{S}_n \vee o = o_n \\ \text{False}, & o = o_a \end{cases} \tag{5}$$

$$\mathbf{x}_a = o_n \mathbf{M}(\mathbf{x}) + o_a \overline{\mathbf{M}}(\mathbf{x}) \tag{6}$$

$$\mathbf{x}_n = \mathbf{x}\mathbf{M}(\mathbf{x}) \tag{7}$$

Then we instantiate lightweight classification heads to decode occupancy status $\hat{\mathbf{x}}_a = h_a(\mathbf{f})$ and semantics $\hat{\mathbf{x}}_n = h_n(\mathbf{f})$, which are combined to form the final reconstruction result $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}_a \overline{\mathbf{M}}(\hat{\mathbf{x}}_a) + \hat{\mathbf{x}}_n \mathbf{M}(\hat{\mathbf{x}}_a) \tag{8}$$

Loss The tokenizer loss comprises two components: reconstruction loss $\mathcal{L}_r$ and vector-quantized loss $\mathcal{L}_{vq}$ following Occworld [4]. The reconstruction loss $\mathcal{L}_r$ combines Cross-Entropy $\mathcal{L}_{ce}$ and Lovasz-Softmax$\mathcal{L}_{ls}$, with decoupled supervision for occupancy status and semantics separately.

$$\mathcal{L} = \lambda_1 \mathcal{L}_r(\mathbf{x}_n, \hat{\mathbf{x}}_n) + \lambda_2 \mathcal{L}_r(\mathbf{x}_a, \hat{\mathbf{x}}_a) + \mathcal{L}_{vq} \tag{9}$$

$$\mathcal{L}_r = \mathcal{L}_{ce} + \lambda_3 \mathcal{L}_{ls} \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ serve as balancing factors.

B. Unified World Model

OccLLaMA integrates the understanding and generation of occupancy-language-action modalities within a unified discrete auto-regressive world model, enabling multi-task joint learning in autonomous driving. The proposed architecture, as illustrated in Figure 2, is detailed as follows:

Joint Vocabulary We establish a joint vocabulary $\mathbf{V}$ that integrates multimodal tokens to serve as the foundation

for multi-task learning. Specifically, $\mathbf{V}$ includes the following: **1)Language Tokens:** We initialize $\mathbf{V}$ with the vocabulary from the pretrained LLM. **2)Occupancy Tokens:** We define {<occ1>,...,<occK>} that correspond to the vector indices of the codebook $\mathbf{Z}$ in Section III-A, enabling scene reconstruction via tokenizer based on specific tokens combinations. **3)Action Tokens:** Waypoints are discreted into 256 bins and {<bin1>,...,<bin256>} are defined to represent fine-grained numerical actions following DriveLM [23]. **4)Special Tokens:** We further add {<occ>,</occ>,<act>,</act>} to delineate modality boundaries and <que_i> as learnable scene queries.

Model Architecture OccLLaMA is constructed based on a pretrained LLM, comprising an embedding layer, backbone, and lm-head, with the dimensions expanded to accommodate $\mathbf{V}$. For occupancy tokens input, we employ a lightweight projection layer that aligns scene features with language embedding space as demonstrated in prior work [24]. For output, all three modalities are converted to the discrete probability distribution over $\mathbf{V}$. Therefore, Cross-Entropy can be used as a uniform loss function for multi-task learning.

Spatial-Temporal Attention Language and action modalities are inherently sequential, making the internal tokens naturally suitable for temporal attention implemented via causal masks in the LLM. However, the occupancy modality lacks intrinsic contextual dependencies and is typically modeled with bidirectional attention to capture spatial relationships. Therefore, we introduce the spatial-temporal attention that applies temporal attention both within the language and action modalities, and across different modalities, while employing spatial attention within the occupancy modality.

Next Token or Scene Prediction Due to the inherent differences in attention mechanism and the quantity gap between the occupancy modality and others, the prediction mechanism operates in two modes: next token or scene prediction. For language and action modalities, we follow the standard practice of sampling next token from the predicted distribution. For occupancy modality, we initialize learnable queries equal to the number of an occupancy scene tokens, enabling the prediction of the entire scene distribution in a single forward pass. The implementation details are provided in Algorithm 1. This design not only aligns with the bidirectional dependencies of the occupancy modality but also significantly reduces inference costs.

Algorithm 1 Next Token or Scene Prediction

Input: Prompt $\mathbf{x} = \{x_i\}_1^n,\ x_i \in \mathbf{V}$
Params: Max Length L, Scene Size S
Output: Completed Reply $\mathbf{x}$

Initialize OccLLaMA as $\mathbf{M}$
Initialize Queries as $\mathbf{q} = [\texttt{<que\_i>}\ \forall\ i \in [1, S]]$
while $\mathbf{x}_{-1} \neq$ <end> **and** $|\mathbf{x}| < L$ **do**
  if $\mathbf{x}_{-1} =$ <occ> **then**
    $\mathbf{x} \leftarrow \mathbf{x} \cup \mathbf{q}$
    $\mathbf{x}_{[-S:]} \leftarrow \mathbf{M}(\mathbf{x})_{[-S:]}$
    $\mathbf{x} \leftarrow \mathbf{x} \cup$ </occ>
  else
    $\mathbf{x} \leftarrow \mathbf{x} \cup \mathbf{M}(\mathbf{x})_{-1}$
  end if
end while
return $\mathbf{x}$

C. Training strategy

Scene Tokenizer Training We initally train scene tokenizer with the objective function defined in Equation (9). After training, the parameters are frozen and seamlessly integrated into the downstream training pipeline.

Multi-Task Pretraining During the pretraining stage, we generate a set of prompt-answer pairs, leveraging historical 4-frame occupancy scenes (occupancy modality), driving command (language modality) and ego status (action modality) [25]. These pairs encompass five distinct tasks: **1)Detection Task:** Identifying object categories and 2D positions, with categories in language modality and positions (nearest action bin) in action modality. **2)Counting Task:** Providing counts of specific categories, with counts in language modality. **3)Prediction Task:** Predicting the 3-second future trajectory of specific objects, with positions and trajectory in action modality. **4)Planning Task:** Predicting the 3-second future trajectory of ego vehicle, with the trajectory in action modality. **5)Forecasting task:** Forecasting the next waypoint and occupancy scene alternatively, with waypoint in action modality and scene in occupancy modality.

Instruction Finetuning During the finetuning stage, we utilize NuScenesQA [26] as the primary language-related dataset to finetune the pretrained model, aiming to quantitatively evaluate its spatial scene understanding capability. Furthermore, we extend the Chain-of-Thought (CoT) of GPT-Driver [27] by incorporating narration, reasoning, notable objects, potential effects, driving action, and trajectory, using the Nu-X and Driving Command datasets [25]. The extension integrates scene understanding as a prerequisite task for motion planning, enabling us to explore the interactions among multiple tasks in Section IV-C.

IV. Experiments

A. Experimental Settings

Datasets and Metrics We conduct main experiments on the NuScenes, utilizing occupancy from Occ3D [28]. And we involve three language datasets: Nu-X, Command, and NuScenesQA. Nu-X and Command in Hint-AD [25] provide diverse linguistic expressions for narration and reasoning on NuScenes. NuScenesQA [26] is a visual question-answering dataset covering five categories: existence, counting, query-object, query-status, and comparison, which are further classified into H0 and H1 by complexity. For language tasks, we adopt standard caption metrics, CIDEr, BLEU, METEOR and Rouge for Nu-X, Accuracy for NuScenesQA and Command. Additionally, we focus on mIoU and IoU for the occupancy forecasting task, as well as L2 precision and collision rate for the motion planning task.

NuSceneQA

Q: There is a trailer; what status is it? (status)
GT: Parked **Pred:** Parked

Q: Are there any cars? (exist)
GT: Yes **Pred:** Yes

Q: What number of other cars in the same status as the pedestrian that is to the front of the trailer? (count)
GT: 3 **Pred:** 4

Q: The parked thing to the front of the moving truck is what? (object)
GT: Truck **Pred:** Truck

CoT

Narration: the car accelerates rapidly down the open freeway
Reasoning: to take advantage of the unoccupied space in front of it, allowing for a smooth and uninterrupted journey
Notable objects: None
Potential effects: None
Driving action: turn right and accelerate
Trajectory: [(0.04, 4.67), (0.03, 9.38), (0.04, 13.15), (0.01, 17.97), (-0.06, 22.79), (-0.15, 27.65)] **GT**

Narration: the car accelerates on the road
Reasoning: No obstacles in front of the ego vehicle
Notable objects: None
Potential effects: None
Driving action: forward and accelerate
Trajectory: [(0.02, 3.22), (0.04, 5.76), (0.04, 10.28), (0.04, 14.30), (0.06, 16.88), (0.06, 20.92)]
Pred

Fig. 3: **Qualitative Results of Scene Understanding.** OccLLaMA enables scene understanding with spatial reasoning based on occupancy observation and enhances motion planning as a prerequisite chain-of-thought, as analyzed in Section IV-C.

Method	Input	Nu-X				NuScenesQA			Command
		C	B	M	R	H0	H1	All	Acc.
GPT-4o	Image + 6-shot examples	19.0	3.95	10.3	24.9	42.0	34.7	37.1	75.4
Gemini 1.5		17.6	3.43	9.3	23.4	40.5	32.9	35.4	80.9
Lidar-LLM	Lidar	-	-	-	-	53.9	45.7	48.6	-
ADAPT	BEV	17.7	2.06	12.8	**27.9**	51.0	44.2	46.4	79.3
BEV+Adapter		18.6	3.47	11.3	24.5	51.8	45.6	47.7	81.1
BEVDet+MCAN	BEV + bounding boxes	13.2	2.91	10.3	24.5	56.2	46.7	49.9	80.7
TOD^{3}Cap		14.5	2.45	10.5	23.0	53.0	45.1	49.0	78.2
Hint-VAD	BEV + inter.	22.4	**4.18**	**13.2**	27.6	**55.4**	48.0	50.5	**82.3**
Ours	Occ	**23.8**	3.96	12.4	25.7	55.3	**51.9**	**53.0**	81.3

TABLE I: **Quantitative Comparison of Scene Understanding.** OccLLaMA achieves state-of-the-art performance on NuScenesQA and competitive performance on Nu-X and Command, relying on less occupancy-based observation. The highest and second-highest performances are indicated by **bold** and underline, respectively.

Implementation Details We set the language model backbone as LLaMA-3.1-8b and the scene tokenizer parameters as $50 \times 256 \times 2048$. Scene tokenizer is trained with learning rate of 10^{-4}, batch size of 4, $\lambda_1 = 2$, $\lambda_2 = 2$, and $\lambda_3 = 0.5$, while Generative World Model is trained with learning rate of 10^{-4} and batch size of 1 in pre-training stage, 5×10^{-5} and 4 in each instruction tuning stage. The Scene tokenizer undergoes 100 epochs on 8 RTX 4090 GPUs, while the Generative World Model undergoes 10 epochs in the pre-training stage and 5 epoch in each instruction tuning stage on 8 V100 GPUs.

B. Single-Task Performance Evaluation

Scene Understanding We first compare the scene understanding performance of OccLLaMA with the methods based on different input formats. LidarLLM [29] integrates point cloud into the language model. ADAPT [6] and BEV+Adapter [30] rely on BEV features built from images. BEVDet+MCAN [31] and TOD^{3}Cap [7] further require detection bounding boxes. Hint-AD [25] integrates intermediate features into the language model. As illustrated in Table I, OccLLaMA achieves state-of-the-art on the NuScenesQA benchmark benefiting from the 3D occupancy modality as input, which enhances its ability to handle spatial reasoning tasks. Notably, Nu-X contains tasks that depend on raw categories and attributes absent in occupancy input (e.g., traffic signs and object colors), presenting an unfair challenge to OccLLaMA on Nu-X caption metrics. However, OccLLaMA still achieves competitive performance on the Nu-X

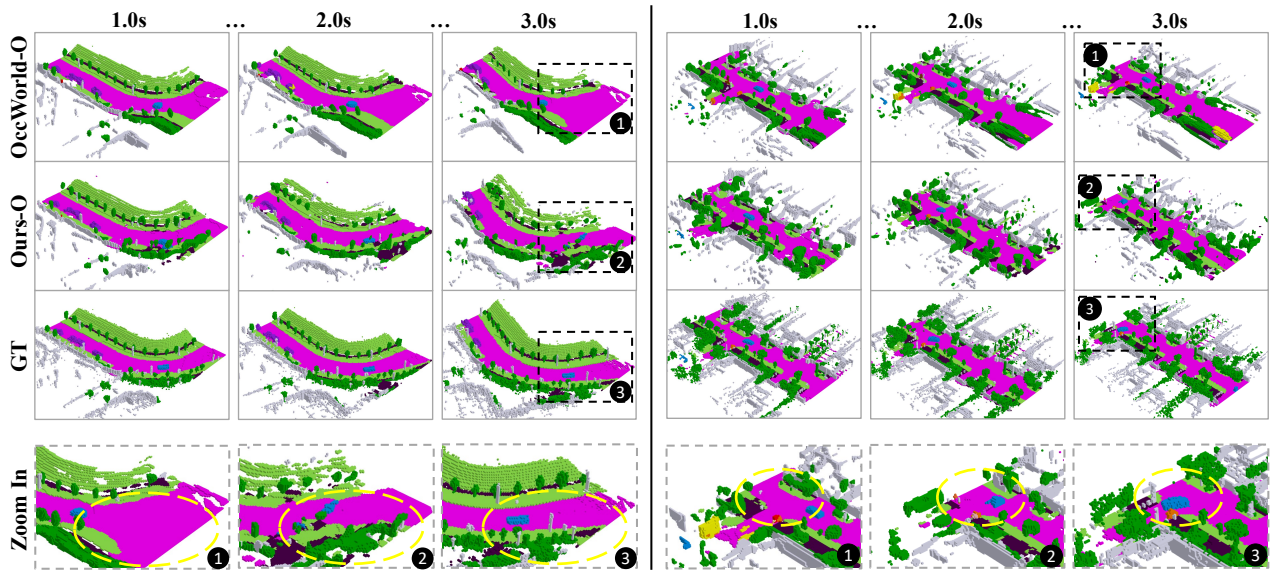

Fig. 4: **Qualitative Comparison of Occupancy Forecasting.** OccLLaMA demonstrates superior long-term forecasting performance, both static scenes evolution (Zoom In, Left) and dynamic objects motion (Zoom In, Right).

Method	Input	mIoU(%)↑					IoU(%)↑				
		Recon.	1s	2s	3s	Avg.	Recon.	1s	2s	3s	Avg.
OccWorld-D	Camera	18.63	11.55	8.10	6.22	8.62	22.88	18.90	16.26	14.43	16.53
OccWorld-F	Camera	20.09	8.03	6.91	3.54	6.16	35.61	23.62	18.13	15.22	18.99
PreWorld	Camera	-	12.27	9.24	7.15	9.55	-	23.62	21.62	19.63	21.62
Ours-F	Camera	37.38	10.34	8.66	6.98	8.66	38.92	25.81	23.19	19.97	22.99
Copy&Paste	Occ	66.38	14.91	10.54	8.52	11.33	62.29	24.47	19.77	17.31	20.52
OccWorld-O	Occ	66.38	**25.78**	15.14	10.51	17.14	62.29	**34.63**	25.07	20.18	26.63
Ours-O	Occ	**75.20**	25.05	**19.49**	**15.26**	**19.93**	**63.76**	34.56	**28.53**	**24.41**	**29.17**

TABLE II: **Quantitative Comparison of Occupancy Forecasting.** Recon. refers to the performance of the scene tokenizer. OccLLaMA achieves a competitive forecasting performance within 1-second interval and outperforms OccWorld within 3-second interval, highlighting its enhanced long-term forecasting capabilities.

benchmark. As shown in Figure 3, these results confirm the effectiveness of pretraining in aligning the occupancy and language modalities.

Occupancy Forecasting Accurate tokenization is crucial for occupancy forecasting performance. In Table II, we first compare our tokenizer with OccWorld [4], achieving SOTA tokenization performance, with 14.5% improvement in mIoU and 2.6% in IoU. This demonstrates that our tokenizer design better captures scene details by addressing the class imbalance. Table II further presents the occupancy forecasting performance of OccLLaMA compared to baseline models, OccWorld [4] and PreWorld [32], setting with ground-truth occupancy (-O) input and predicted results from FB-OCC predictor (-F) input. Our method achieves a competitive performance within 1-second interval and outperforms OccWorld within 3-second interval, highlighting its enhanced long-term forecasting capabilities. Ours-F achieves better IoU than the SOTA baseline PreWorld within 3-second interval, demonstrating its potential as an end-to-end world model. Notably, our method demonstrates scalability for improvement through zero-cost replacement with better predictors. In addition, Figure 4 further visualizes that our method forecasts more accurate details for both static scenes and dynamic objects.

Motion Planning In Table III, the motion planning performance of our model is extensively compared with several strong baselines, including end-to-end methods, LLM-based methods, and world models. Without manual label supervision or postprocessing, our method achieves SOTA performance on the L2 metric and competitive performance on the collision rate metric. In particular, OccLLaMA significantly outperforms the SOTA occupancy world models, OccWorld [4] and RenderWorld [33], under the settings of occupancy ground truth input (-O) and raw images input (-F or †) same as previous subsection. This highlights the effectiveness of multi-task learning in enhancing the performance of motion planning, which will be thoroughly discussed in Section IV-C.

C. Effectiveness of Multi-Task Learning

To validate the effectiveness of multi-task learning in enhancing the performance of motion planning, we conduct

Method	Input	Supervision	L2(m)↓				Collision(%)↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD	Camera	Map&Box&Motion&Track&Occ	0.48	0.96	1.65	1.03	0.05	0.17	**0.71**	**0.31**
VAD	Camera	Map&Box&Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
OccNet	Occ	Map&Box	1.29	2.31	2.98	2.25	0.20	0.56	1.30	0.69
GPT-Driver	Camera	Map&Box&Motion&Track&Occ	0.27	0.74	1.52	0.84	0.07	**0.15**	1.10	0.44
OmniDrive	Camera	Box&Centerline	0.40	0.80	**1.32**	0.84	0.04	0.46	2.32	0.94
OccWorld-F	Camera	Occ	0.45	1.33	2.25	1.34	0.08	0.42	1.71	0.73
RenderWorld†	Camera	None	0.48	1.30	2.67	1.48	0.14	0.55	2.23	0.97
Ours-F	Camera	Occ	0.38	1.07	2.15	1.20	0.06	0.39	1.65	0.70
OccWorld-O	Occ	None	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
RenderWorld-O	Occ	None	0.35	0.91	1.84	1.03	0.05	0.40	1.39	0.61
Ours-O	Occ	None	**0.25**	**0.64**	1.50	**0.80**	**0.03**	0.37	0.96	0.45

TABLE III: **Quantitative Comparison of Motion Planning.** OccLLaMA achieves SOTA performance on L2 metric and competitive results in collision rate metric without manual label supervision or postprocessing, compared to strong baselines including end-to-end methods, LLM-based methods, and world models.

Model	Pretraining	Forecasting	Understanding	L2(m)↓				Collision(%)↓			
				1s	2s	3s	Avg.	1s	2s	3s	Avg.
$\mathcal{M}_0$	✓			0.61	1.39	2.27	1.42	0.14	0.68	1.96	0.93
$\mathcal{M}_1$	✓	✓		0.37	1.02	2.03	1.14(↓0.28)	0.04	**0.24**	1.20	0.49(↓0.44)
$\mathcal{M}_2$	✓		✓	0.30	0.71	**1.47**	0.82(↓0.60)	0.04	0.39	**0.89**	**0.44**(↓0.49)
$\mathcal{M}_3$	✓	✓	✓	**0.25**	**0.64**	1.50	**0.80(↓0.62)**	**0.03**	0.37	0.96	0.45(↓0.48)

TABLE IV: **Effectiveness of Multi-Task Learning** Comparative evaluation of $\mathcal{M}_0$ to $\mathcal{M}_3$ highlights the performance improvements of motion planning achieved through occupancy forecasting and scene understanding tasks.

an extensive evaluation of OccLLaMA across various configurations from $\mathcal{M}_0$ to $\mathcal{M}_3$, as detailed in Table IV. $\mathcal{M}_0$ serves as the baseline, directly predicting future waypoints based on the pretrained model. $\mathcal{M}_1$ extends $\mathcal{M}_0$ by incorporating the occupancy forecasting task, enabling alternate prediction of future occupancy scenes and waypoints. In $\mathcal{M}_2$, the pretrained model is fine-tuned as outlined in Section III-C to integrate the scene understanding task, thereby establishing a chain-of-thought reasoning mechanism. Finally, $\mathcal{M}_3$ combines the alternate prediction strategy with the chain-of-thought reasoning, providing a comprehensive evaluation.

Compared to the baseline, $\mathcal{M}_1$ achieves a modest reduction in L2 precision and substantial reduction in collision rate, demonstrating its ability to enhance driving safety by explicitly predicting the evolution of future scenes. $\mathcal{M}_2$ achieves a significant reduction in both L2 precision and collision rate, demonstrating that the CoT reasoning driven by scene understanding, enhances driving accuracy and human alignment. $\mathcal{M}_3$ achieves a further reduction in L2 precision, demonstrating that our model can effectively accommodate various tasks without catastrophic forgetting, validating the generality and robustness of the proposed multi-task learning framework.

D. Ablation Study

Scene Tokenizer Parameters Table V evaluates the impact of different hyper-parameters on tokenization performance of the Scene Tokenizer, focusing on latent space resolution and codebook size. The result shows that a smaller codebook fails to capture scene distributions effectively, while a larger one leads to overfitting due to inefficient codebook utilization. Higher resolutions improve reconstruction accuracy but also elevate the forecasting burden by increasing the number of tokens required per scene. Based on these findings, we select a resolution of 50 and codebook size of 2048 to balance performance and computational efficiency.

Setting		Reconstruction	
Res.	Size	mIoU(%)↑	IoU(%)↑
25	2048	59.04	49.25
100	2048	79.18	66.36
50	1024	68.26	58.81
50	4096	70.94	61.03
50	2048	75.20	63.76

TABLE V: **Ablation of Tokenizer Parameters.** Res. refers to latent space resolution.

Setting		Forecasting		Planning	
s.a.	a.t.	mIoU(%)↑	IoU(%)↑	L2(m)↓	Coll.(%)↓
✓	✓	**19.93**	**29.17**	1.14	0.49
✓		18.05	28.55	1.19	0.54
	✓	15.78	27.84	**1.12**	**0.48**

TABLE VI: **Ablation of Model Components.** s.a. refers to spatial attention. a.t. refers to action tokenization.

Model Components We evaluate the impact of key components of the unified world model on occupancy forecasting and motion planning based on the pretrained model. As illustrated in Table VI, the absence of spatial attention (w/o s.a.) results in all tokens performing temporal attention akin to the original LLM, while the absence of action tokeniza-

tion (w/o a.t.) replaces action-specific tokens with language modality for waypoint representation. The results demonstrate that action-specific tokens significantly enhance motion planning performance by enabling explicit representation. Furthermore, spatial attention proves essential for modeling spatial dependencies within the occupancy scene, yielding notable improvements in occupancy forecasting task.

V. Conclusion

We propose OccLLaMA, a unified occupancy-language-action world model for understanding and generation tasks in autonomous driving. By leveraging semantic occupancy as a modality-agnostic 3D representation, OccLLaMA integrates vision, language and action modalities into a single auto-regressive framework, enabling end-to-end multi-task learning. Furthermore, OccLLaMA constructs a human-like motion planning process by integrating scene understanding and occupancy forecasting. Extensive experiments demonstrate the superior performance of OccLLaMA across scene understanding, occupancy forecasting, and motion planning, highlighting the effectiveness of multi-task learning and integration in improving driving performance. In the future, we will explore quantization methods to address the inference delay caused by a large number of parameters.

References

[1] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[2] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 749–14 759.

[3] L. Zhang, Y. Xiong, Z. Yang, S. Casas, R. Hu, and R. Urtasun, "Learning unsupervised world models for autonomous driving via discrete diffusion," *arXiv preprint arXiv:2311.01017*, 2023.

[4] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," *arXiv preprint arXiv:2311.16038*, 2023.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[6] B. Jin, X. Liu, Y. Zheng, P. Li, H. Zhao, T. Zhang, Y. Zheng, G. Zhou, and J. Liu, "Adapt: Action-aware driving caption transformer," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[7] B. Jin, Y. Zheng, P. Li, W. Li, Y. Zheng, S. Hu, X. Liu, J. Zhu, Z. Yan, H. Sun, *et al.*, "Tod3cap: Towards 3d dense captioning in outdoor scenes," *arXiv preprint arXiv:2403.19589*, 2024.

[8] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, 2024.

[9] K. Renz, L. Chen, A.-M. Marcu, J. Hünermann, B. Hanotte, A. Karnsund, J. Shotton, E. Arani, and O. Sinavski, "Carllava: Vision language models for camera-only closed-loop driving," *arXiv preprint arXiv:2406.10165*, 2024.

[10] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.

[11] Y. Dong, H. Liang, M. Zhai, C. Li, M. Xia, X. Liu, M. Mo, J. Leng, J. Tao, and X. Gao, "Bevlm: Got-based integration of bev and llm for driving with language."

[12] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 120–15 130.

[13] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[14] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[15] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model," *arXiv preprint arXiv:2310.07771*, 2023.

[16] T. Khurana, P. Hu, D. Held, and D. Ramanan, "Point cloud forecasting as a proxy for 4d occupancy forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1116–1124.

[17] L. Zhang, Y. Xiong, Z. Yang, S. Casas, R. Hu, and R. Urtasun, "Learning unsupervised world models for autonomous driving via discrete diffusion," *arXiv preprint arXiv:2311.01017*, 2023.

[18] D. Bogdoll, Y. Yang, and J. Marius Zöllner, "Muvo: A multimodal generative world model for autonomous driving with geometric representations," *arXiv e-prints*, pp. arXiv–2311, 2023.

[19] Y. Zhang, S. Gong, K. Xiong, X. Ye, X. Tan, F. Wang, J. Huang, H. Wu, and H. Wang, "Bevworld: A multimodal world model for autonomous driving via unified bev latent space," *arXiv preprint arXiv:2407.05679*, 2024.

[20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[21] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[23] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," *arXiv preprint arXiv:2312.14150*, 2023.

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[25] K. Ding, B. Chen, Y. Su, H.-a. Gao, B. Jin, C. Sima, W. Zhang, X. Li, P. Barsch, H. Li, *et al.*, "Hint-ad: Holistically aligned interpretability in end-to-end autonomous driving," *arXiv preprint arXiv:2409.06702*, 2024.

[26] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4542–4550.

[27] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," 2023. [Online]. Available: https://arxiv.org/abs/2310.01415

[28] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, Y. Guo, and S. Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv preprint arXiv:2312.14074*, 2023.

[30] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[31] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[32] X. Li, P. Li, Y. Zheng, W. Sun, Y. Wang, and Y. Chen, "Semi-supervised vision-centric 3d occupancy world model for autonomous driving," 2025. [Online]. Available: https://arxiv.org/abs/2502.07309

[33] Z. Yan, W. Dong, Y. Shao, Y. Lu, L. Haiyang, J. Liu, H. Wang, Z. Wang, Y. Wang, F. Remondino, *et al.*, "Renderworld: World model with self-supervised 3d label," *arXiv preprint arXiv:2409.11356*, 2024.