OccDirector: Language-Guided Behavior and Interaction Generation in 4D Occupancy Space

Zhuding Liang[1⋆], Tianyi Yan[1*], Dubing Chen[1], Jiasen Zheng[2], Huan Zheng[1], Cheng-zhong Xu[1], Yida Wang[3], Kun Zhan[3], and Jianbing Shen[1✉]

[1] SKL-IOTSC, CIS, University of Macau, Macau, China
[2] Northwestern University, Evanston, United States
[3] Li Auto Inc, Beijing, China

Abstract. Generative world models increasingly rely on 4D occupancy for realistic autonomous driving simulation. However, existing generation frameworks are limited in two respects: geometric-condition-based methods depend on explicit trajectories and cannot generalize to free-form instructions, while text-based methods operate at a coarse attribute level and fail to orchestrate complex, sequential multi-agent interactions. We propose OccDirector, a framework that generates 4D occupancy dynamics conditioned solely on natural language scripts, without requiring any geometric priors. OccDirector encodes free-form scripts via a frozen VLM and processes occupancy tokens through a Spatio-Temporal MMDiT backbone. A history-prefix anchoring strategy further ensures that generated sequences remain consistent with any provided context over long horizons. To support training and evaluation, we introduce OccInteract-85k, a dataset of 85k clips annotated with multi-level language instructions spanning static layouts to intricate multi-agent behaviors, alongside a novel VLM-based evaluation benchmark. Extensive experiments demonstrate state-of-the-art generation quality and strong instruction-following fidelity across diverse driving scenarios.

Keywords: 4D Occupancy Generation · Language-Guided Scene Generation · Multi-Agent Interaction Modeling · Vision-Language Model Conditioning · Spatio-Temporal Transformer

1 Introduction

The evolution of autonomous driving (AD) systems [6,9,23] is increasingly relying on Generative World Models [11,34,58] that can simulate diverse and realistic driving scenarios for closed-loop training and validation. Among various scene representations, 4D Occupancy (space-time occupancy) [17,28] has emerged as

⋆ Equal Contribution.
✉ Corresponding author: *Jianbing Shen*. This work was supported by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0134/2025/RIA2 and 0102/2023/RIA2, and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

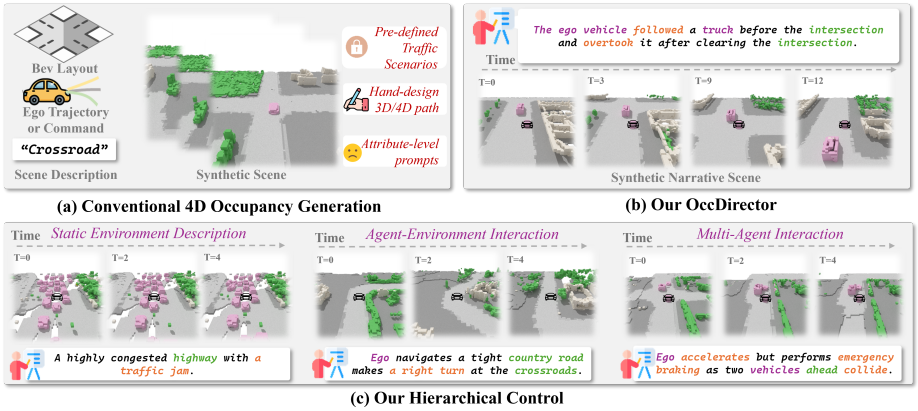

Fig. 1: Script-driven behavior and interaction generation with OccDirector. (a) vs. (b): Comparison between previous methods and our script-driven approach. (c) Hierarchical generation capabilities across Static, Agent-Environment, and Multi-Agent levels. The ego vehicle is marked with a *black car icon. Note: Due to raw data limitations Sec. 4.1, some results may exhibit missing ego voxels or sensor blind spots.*

the most unified and comprehensive representation. Unlike 3D bounding boxes which oversimplify object geometry [39,41], or LiDAR point clouds [1,46] which suffer from sparsity and occlusion, 4D occupancy grids offer a dense, geometry-aware, and temporally consistent description of dynamic environments. Consequently, synthesizing high-fidelity 4D occupancy rollouts has become a critical frontier in the computer vision community [8,55].

Despite impressive progress in 3D/4D generation, existing occupancy generation frameworks [3,22,44] heavily rely on explicit geometric conditions—such as BEV semantic layouts, discrete navigation command, or predefined ego-vehicle trajectories—to control the generative process. As illustrated in Fig. 1(a), methods like [3, 14, 36] depend on rigid explicit trajectories or simplistic attribute prompts. While offering precise spatial constraints, these conditions are fundamentally inflexible and unscalable. Users must laboriously hand-design accurate 3D paths, a process that cannot intuitively express high-level intents. In contrast, our OccDirector (Fig. 1(b,c)) acts as a "scenario director", interpreting complex natural language scripts to orchestrate dynamic driving scenes, such as *"congested traffic jam"* or *"emergency braking"*. To date, generating dynamic 4D occupancy purely guided by natural language remains an unsolved challenge.

Pioneering text-driven 4D occupancy generation faces a significant challenge: the Semantic-Spatiotemporal Gap. Language instructions inherently rely on relational structures and sequential logic to describe abstract behaviors. For instance, consider the instruction: *"The ego vehicle followed a truck before the intersection and overtook it after clearing the intersection."* Translating such high-level temporal logic into occupancy generation is non-trivial, as it demands rigor-

ous spatio-temporal consistency and physically plausible evolution at the voxel level. However, most existing autonomous driving diffusion models [12, 45, 58] fail to bridge this gap. Relying on generic text encoders [31, 32] that function largely as "bag-of-words" extractors, these models lack the compositional reasoning required to parse complex interactions [19, 56]. Consequently, they often produce implausible collisions, discontinuous motions, or topological violations when processing intricate instructions [24].

To this end, we introduce OccDirector, taking a decisive step toward language-driven scenario orchestration. Unlike prior works constrained by geometric priors, OccDirector establishes a new paradigm of purely text-guided 4D occupancy generation, bridging the gap from appearance synthesis to complex agent behavior and interaction generation. By replacing rigid trajectory inputs with a director-like natural language interface, OccDirector enables comprehensive controllability along three complementary axes, as visualized in Fig. 1(c): (i) **Static Environment**, synthesizing diverse topologies like *"traffic jams"* via semantic composition; (ii) **Agent-Environment Interaction**, executing topology-constrained maneuvers such as yielding at intersections; and (iii) **Multi-Agent Interaction**, generating physically plausible reactive behaviors like *"emergency braking to avoid collisions"*. Technically, OccDirector leverages a powerful Vision-Language Model (VLM) as the text encoder to deeply comprehend procedural semantics. The joint text-occupancy sequence is modeled using a scalable spatio-temporal MMDiT equipped with Spatio-Temporal Separated Attention (STSA) and a history-prefix token-replace strategy for interaction-consistent rollouts. Furthermore, to overcome the complete absence of paired text-4D occupancy data, we develop an automated data engine that translates real-world logs and simulator data into abundant, logic-rich occupancy-language pairs.

Our contributions can be summarized as follows:

- We propose OccDirector, a pioneering framework for text-guided 4D occupancy generation, enabling director-like control over agent behaviors without rigid geometric priors.
- We design a VLM-driven Spatio-Temporal MMDiT with STSA and a history-prefix anchoring strategy, effectively bridging the semantic-spatiotemporal gap for consistent interactions.
- We introduce OccInteract-85k, a large-scale dataset with hierarchical language instructions ranging from static layouts to complex interactions, alongside a novel VLM-based benchmark.
- Extensive experiments demonstrate that OccDirector achieves state-of-the-art generation quality and unprecedented instruction-following capabilities.

2 Related Work

2.1 Generative World Models for Autonomous Driving

Generative world models have emerged as a paradigm shift in autonomous driving simulation, enabling the synthesis of diverse driving scenarios for closed-loop

training [16, 34, 45, 58]. Early approaches predominantly focused on 2D Video Generation [12,13,18], leveraging diffusion models to produce photorealistic camera views. While visually compelling, these perspective-view methods often lack explicit 3D geometric consistency and struggle to support downstream planning tasks that require precise spatial reasoning.

Recently, 4D Occupancy has been established as a unified representation bridging visual fidelity and geometric explicitness [41, 51]. By representing the world as a dense, voxelized spatiotemporal volume, it supports physically grounded simulation. However, existing occupancy world models [57, 59] primarily focus on *predicting* future states from past observations (forecasting). They lack the generative flexibility to create new scenarios from scratch based on user intent, limiting their utility for controllable simulation.

2.2 Conditioned 4D Occupancy Generation

Unlike forecasting, occupancy generation aims to synthesize plausible 4D scenes conditioned on specific inputs. Current state-of-the-art frameworks heavily rely on Explicit Geometric Conditions. For instance, methods like OccSora [44] and DynamicCity [3] condition the diffusion process on HD maps, 3D semantic layouts, or pre-defined agent trajectories. While these strong priors ensure structural validity, they impose a rigid generation paradigm: users must provide precise, low-level geometric inputs (e.g., drawing a spline for a lane change) to control the output. This "trajectory-engineering" process is labor-intensive and non-intuitive. In contrast, OccDirector proposes a paradigm shift from geometry-driven to language-driven generation. We eliminate the need for pre-defined trajectories, allowing users to orchestrate scene evolution through high-level semantic scripts (e.g., *"an aggressive cut-in"*), thereby offering a more scalable interface for scenario design.

2.3 Text-Conditioned Scene Generation and Fine-Grained Control

The integration of natural language into generative models has significantly advanced image and video synthesis [26, 35, 40, 43]. In the autonomous driving domain, however, language control remains largely limited to high-level attributes. Existing works [7, 11, 30, 58] typically utilize text prompts to control Global Attributes (e.g., weather conditions, time-of-day) or static scene composition. While some approaches like UniScene [22] incorporate text, they primarily use it for texture rendering or style transfer, leaving the core spatiotemporal dynamics (i.e., how agents move) determined by other modalities. Other works like X-Scene [54] adopt a hierarchical approach, converting text into intermediate layouts or bounding boxes before generation. Although this bridges language and static layout, it may struggle to capture the continuous nuances of dynamic interactions.

In this work, we introduce OccDirector, which pioneers end-to-end text-to-4D occupancy generation. Unlike methods that treat text as a style modifier or rely on intermediate layout proxies, we learn a direct mapping from procedural

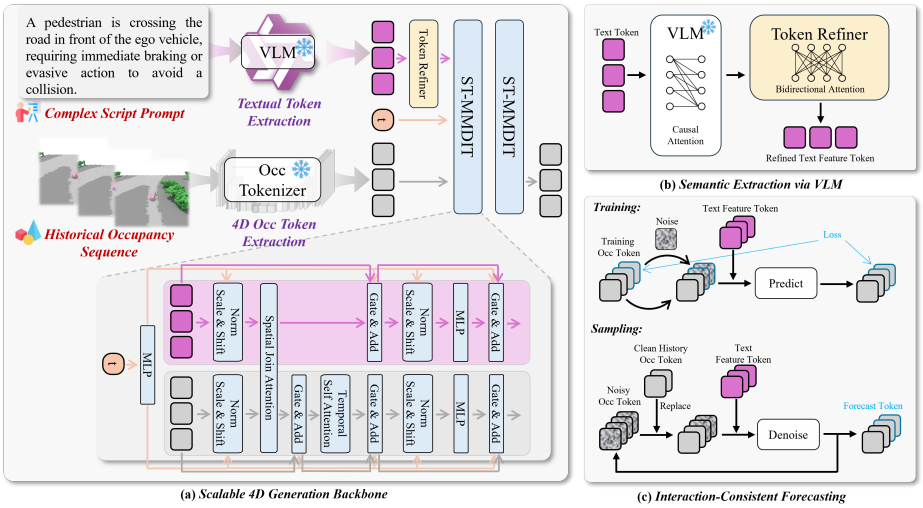

Fig. 2: Model Architecture of OccDirector. (a) Scalable 4D Generation Backbone: A Spatio-Temporal MMDiT processes text and occupancy tokens via dual streams, utilizing separated spatial and temporal attention. **(b) Semantic Extraction:** A frozen VLM and a Token Refiner bridge the semantic-spatial gap. **(c) Interaction-Consistent Forecasting:** The history-anchoring strategy ensures the generated future strictly adheres to the context during both training and sampling.

language instructions to continuous spatiotemporal voxel flows, enabling fine-grained control over multi-agent interactions and maneuvers purely from natural language.

3 OccDirector: Language-Conditioned 4D Occupancy Interaction Generation

3.1 Method Overview

OccDirector is designed for *script-driven behavior and interaction generation* within 4D occupancy rollouts (Sec. 1). Formally, given a natural language script π and an optional history context $\mathbf{O}_{1:h}$, our goal is to generate a future 4D occupancy sequence $\mathbf{O}_{h+1:F}$ that strictly adheres to the spatial layout and dynamic interaction intents specified in π. Fig. 2 illustrates our architecture.

3.2 Joint Latent Representation

Directly modeling raw 4D occupancy data and complex natural language is computationally intractable and semantically challenging. To map both modalities into a unified, efficient latent space, we formulate our input encoding stage with three specialized components.

Geometry-Aware Occupancy Tokenization. To alleviate the extreme memory and computational footprint of 4D occupancy sequences $\mathbf{O}_{1:F} \in \{0, \dots, K-1\}^{F \times H \times W \times D}$, we require a geometry-aware compression mechanism. Building upon the discrete auto-encoding principles of DOME [14], we adopt a frozen Occ-VAE. This module maps the discrete 4D rollout into a compact, continuous latent grid, which is subsequently patchified and projected into a sequence of tokens $\mathbf{H} \in \mathbb{R}^{F \times S \times d_{\text{model}}}$, where F is the temporal length and S denotes the number of 3D spatial tokens per frame. Unlike DOME's implementation, we pretrain the Occ-VAE with variable-length rollouts, enabling flexible generation horizons.

Procedural Semantic Extraction via VLM. Effective script-driven generation requires controlling dynamic *interaction intents* (e.g., yielding, cut-in) over time. Standard CLIP-style embeddings [31,32] often lack the relational and procedural semantics needed for such fine-grained control. Recent advancements have demonstrated that leveraging LLMs or VLMs for text feature embedding yields superior performance compared to CLIP-style embeddings [40,49,50,53]. To capture these complex global relations, as shown in Fig. 2(b), we utilize a frozen Vision-Language Model (VLM) text encoder (Qwen3-VL-2B [52]). Given a prompt π, the encoder extracts dense, high-dimensional representations $\tilde{\mathbf{C}} \in \mathbb{R}^{L \times d_{\text{text}}}$, where L is the sequence length.

Token Refiner. While VLM features provide rich semantics, they are inherently suboptimal for direct generative control in 4D spatial-temporal layouts due to modality gaps. To bridge this semantic-spatial divide, we introduce a lightweight bidirectional token refiner (Fig. 2(b)) [20, 27]. This module, denoted as Refine$(\cdot)$, employs a shallow transformer alternating between full self-attention and MLP blocks. It projects the raw features to the model dimension, $\mathbf{C} = \text{Refine}(\tilde{\mathbf{C}}) \in \mathbb{R}^{L \times d_{\text{model}}}$, producing text tokens with enhanced compositional alignment for occupancy generation.

3.3 Scalable 4D Generation Backbone

Modeling environment constraints and multi-agent interactions in 4D space demands capturing long-range dependencies; failure to do so results in accumulated errors, such as geometry violations or implausible discontinuities. However, the cubic complexity of 3D spatial tokens (S) multiplied by the temporal dimension (F) makes standard full attention computationally prohibitive.

To resolve this scalability bottleneck while maintaining strict text-to-layout alignment, we formulate a double-stream transformer backbone (Fig. 2(a)) that builds upon the MMDiT architecture [10]. This design preserves a dedicated text pathway, allowing for repeated, deep interactions between text tokens $\mathbf{C}$ and occupancy tokens $\mathbf{H}$ to strengthen fine-grained conditioning.

To efficiently process the massive number of occupancy tokens, we factorize the attention mechanism via Spatio-Temporal Separated Attention (STSA) [29]. Specifically, each transformer block decomposes the operation into two sequential steps: (i) *Frame-wise 3D spatial joint attention*, which models interactions between text tokens and spatial occupancy tokens within each independent frame

to ensure script-to-layout grounding; followed by (ii) *Patch-wise temporal self-attention*, applied solely across the temporal axis of occupancy tokens to propagate dynamics smoothly over time. To align with this factorization, we apply 2D spatial RoPE [37] on the $H \times W$ grid within each frame (with the Z-axis folded into the channel dimension), and 1D temporal RoPE across frames.

3.4 Interaction-Consistent Forecasting via History Anchoring

To train OccDirector for stable and high-quality continuous latent generation, we adopt the Continuous Normalizing Flow (CNF) framework via Flow Matching [25]. We define $t = 0$ as the ground-truth clean data and $t = 1$ as the pure noise. Let $\mathbf{x}_0$ denote the ground-truth latent occupancy tokens and $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the standard Gaussian noise. We construct a linear probability path $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, where $t \sim \mathcal{U}(0,1)$. The corresponding vector field is given by $v_t(\mathbf{x}) = \mathbf{x}_1 - \mathbf{x}_0$. The base objective trains the model v_θ to regress this vector field conditioned on the refined text tokens $\mathbf{C}$.

Training with History-Prefix Anchoring. For behavior generation, users often specify partial layouts or history contexts $\mathbf{O}_{1:h}$. Maintaining strict consistency with these constraints is critical, as small deviations in early frames can compound into collisions or interaction drift. Rather than relying on training-free heuristics during inference (which often yield temporal inconsistencies), we explicitly equip the model with forecasting capabilities during the training phase.

To achieve this, we introduce a stochastic *token-replace anchoring* strategy, visualized in Fig. 2(c) [20]. During training, we randomly sample a history horizon h. With a certain probability (similar to classifier-free guidance dropping), we replace the noisy history prefix $\mathbf{x}_t^{(1:h)}$ with the perfectly clean ground-truth tokens $\mathbf{x}_0^{(1:h)}$. Let $\tilde{\mathbf{x}}_t$ denote this partially clean, partially noisy input sequence. To ensure the model focuses entirely on forecasting the unknown dynamics, we apply a masked loss that only penalizes the predictions on the future frames $(h+1:F)$:

$$\mathcal{L}_{\text{Anchor}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\left\| v_\theta(\tilde{\mathbf{x}}_t, \mathbf{C}, t)^{(h+1:F)} - (\mathbf{x}_1 - \mathbf{x}_0)^{(h+1:F)} \right\|_2^2\right]. \tag{1}$$

This formulation forces the network to learn the conditional transition dynamics from a clean history to a noisy future, strictly guided by the script $\mathbf{C}$.

Inference via Step-wise Anchoring. During the inference (ODE sampling) phase, we start from pure noise $\mathbf{x}_1$ and integrate backwards to $t = 0$. To enforce robust controllability and ensure the generated sequence perfectly matches the user-provided history $\mathbf{H}_{1:h}$, we apply step-wise anchoring. At every integration step t, after the solver updates the sequence, we forcefully overwrite the history prefix with the clean condition $\mathbf{x}_{t-\Delta t}^{(1:h)} \leftarrow \mathbf{H}_{1:h}$. Because the model has been explicitly trained to condition on clean history prefixes via Eq. (1), this step-wise replacement seamlessly anchors the generation process without introducing out-of-distribution artifacts, ensuring that the future dynamics $\mathbf{x}_0^{(h+1:F)}$ remain strictly coherent with the provided context.

4 OccInteract-85k: A Language-Conditioned Occupancy Dataset

To train OccDirector with language-conditioned occupancy prediction, we require paired supervision between (i) 4D occupancy representations and (ii) natural-language descriptions that emphasize *geometry, topology, and interaction intent* rather than appearance. To this end, we introduce **OccInteract-85k**, a dataset constructed via an automated pipeline that synthesizes three complementary corpora: *static layout captions* (OccInteract-Static), *environment-interaction clips* (OccInteract-Env), and *agent-interaction simulations* (OccInteract-Agent). All annotations are generated with structured prompting and enforced output schemas, enabling scalable production and automatic filtering.

4.1 Data Curation Pipeline

Our pipeline integrates three data sources to balance realism, scale, and diversity: (i) *Real-world Logs.* We leverage UniOcc [47] to aggregate real-world autonomous driving logs, providing unified occupancy representations with high photorealism. (ii) *Large-scale Simulation.* To overcome the scalability limits of real logs, we collect extensive simulator-based data following the CarlaSC protocol [48], enabling diverse environmental configurations. (iii) *Safety-critical Scenarios.* To target long-tail and safety-critical interactions, which are often absent in real logs. Therefore we generate text-defined scenarios using TTSG [33]. These are rendered via CarlaSC, ensuring coverage of high-risk edge cases.

Standardization. We standardize these raw sources into paired *occupancy–language* training instances. Following UniOcc [47], we standardize all data sources into a unified occupancy grid format. Then, we render semantic occupancy grids into BEV sequences and prompt a frozen VLM (`qwen3-vl-235b-a22b-instruct`) to generate schema-constrained, machine-parseable JSON captions. This process yields annotations at three temporal granularities: (i) single-frame layout captions for learning geometric priors, (ii) clips around key maneuvers for learning behavior-conditioned structure, and (iii) specified multi-agent plans for generating diverse interactions, forming the sub-datasets detailed below. Table 1 summarizes the statistics. Concretely, all sources are unified to a $200\times200\times16$ voxel grid (0.4 m/voxel, X/Y: ±40 m, Z: -3.2–3.2 m) sampled at 2 Hz, with semantic labels remapped to a shared 11-category ontology following UniOcc [47].

4.2 Dataset Composition

OccInteract-Static: Static Layout Priors. This subset pairs single-frame BEV semantic occupancy snapshots with ego-centric layout captions. The annotations focus on road topology (e.g., lane structures, intersections) and drivable regions, providing fundamental geometric priors independent of temporal dynamics.

Table 1: Dataset Statistics. Overview of the processed corpora. Instances denote single frames for OccInteract-Static and video clips for OccInteract-Env / OccInteract-Agent.

Dataset	# Instances	Total Duration
OccInteract-Static	73.5k	N/A (Single-frame)
OccInteract-Env	7.6k	128 h
OccInteract-Agent	4.6k	12 h

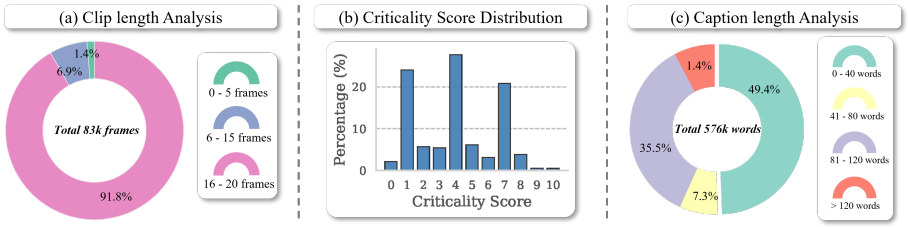

Fig. 3: Data Statistics of OccInteract-Agent. (a) Distribution of clip lengths. (b) Histogram of VLM-predicted criticality scores. (c) Distribution of caption lengths.

OccInteract-Env: Behavior-Conditioned Dynamics. Targeting key maneuvers at junctions, this subset provides clip-level supervision. Each instance includes a discrete ego-action command (straight/left/right) and a description of the local topology. This facilitates learning how occupancy structures evolve conditioned on ego-behavior.

OccInteract-Agent: Multi-Agent Interactions. This corpus covers a spectrum from routine driving to safety-critical near-misses. Each instance is paired with an interaction-focused caption and a *criticality score* ($s \in [0, 10]$), predicted by a frozen VLM judge (`qwen3-vl-235b-a22b-instruct`). We interpret this score as a proxy for risk: a low score corresponds to routine, non-conflicting interactions; medium indicates potential conflicts requiring caution (e.g., nearby vulnerable road users or occlusions); high reflects imminent hazards or near-collision situations. As shown in Fig. 3 (b), the scores exhibit a trimodal distribution (peaks at 1/4/7), effectively stratifying samples into low, medium, and high-risk regimes. This metadata enables curriculum learning and targeted sampling of rare safety-critical scenarios. Furthermore, Fig. 3 (a) and (c) present the statistics for clip durations and caption lengths.

4.3 Automated Quality Assessment

To close the loop between data generation and model learning, we implement a rubric-based VLM evaluator. This mechanism serves a dual purpose: filtering low-quality synthetic data and benchmarking model performance (see Sec. 5).

Table 2: Automated Data Validation. Mean rubric scores (1-5) assessing the quality of the generated dataset.

Dataset	Judge VLM	Completeness ↑	Structural ↑	Sem. Align. ↑	Mean ↑
OccInteract-Static	`gemini-3.0-pro`	4.16	4.20	3.84	4.08
	`gpt-5.2`	2.97	3.71	3.03	3.24
	`qwen3-vl-235b-a22b-instruct`	3.22	3.84	3.97	3.68
OccInteract-Env	`gemini-3.0-pro`	4.62	4.60	3.55	4.26
	`gpt-5.2`	3.21	3.78	3.01	3.33
	`qwen3-vl-235b-a22b-instruct`	3.32	3.81	3.66	3.60
OccInteract-Agent	`gemini-3.0-pro`	4.51	4.53	3.49	4.18
	`gpt-5.2`	3.21	3.72	2.58	3.17
	`qwen3-vl-235b-a22b-instruct`	3.30	3.64	3.32	3.42

Evaluation Protocol. We adopt a reference-free evaluation paradigm. Specifically, we render the generated semantic occupancy grids into BEV video clips and prompt a frozen VLM judge to score them against the source text descriptions. We define three evaluation axes, scored on an integer scale of $[1, 5]$: (i) *scene completeness* (coverage of relevant static/dynamic elements), (ii) *structural correctness* (physical plausibility and topological consistency), and (iii) *Semantic Alignment* (faithfulness to the prompt).

To ensure robustness, we employ three state-of-the-art VLMs as judges: `gemini-3.0-pro`, `gpt-5.2`, and `qwen3-vl-235b-a22b-instruct`. We sample 100 instances from each sub-dataset, with each instance evaluated five times to reduce variance.

Analysis. Table 2 presents a comprehensive evaluation across three axes. First, regarding *absolute quality*, `gemini-3.0-pro` consistently awards high scores (mean > 4.0), confirming that our synthesized occupancy scenes possess high completeness and physical plausibility. Second, regarding *scorer calibration*, we observe a systematic offset between judges; for instance, `gpt-5.2` is notably stricter (mean ≈ 3.2). We attribute this to the inherent visual abstraction of voxelized representations, which stricter VLMs may penalize compared to photorealistic imagery. Third, and most importantly, we demonstrate *pipeline robustness* through relative consistency. Despite the increasing complexity from static layouts (OccInteract-Static) to multi-agent interactions (OccInteract-Agent), the score variance under the same judge remains minimal (e.g., `gpt-5.2` means range tightly: 3.17–3.33). This stability indicates that our automated pipeline maintains uniform fidelity regardless of scenario complexity, validating the reliability of OccInteract-85k for large-scale training.

5 Experiments

5.1 Experimental Setup

Data. We evaluate on the same processed dataset described in Sec. 4, which contains about 85k training instances.

Compared methods. We compare OccDirector with recent state-of-the-art 4D occupancy generation methods, including DOME [14], COME [36], and DynamicCity [3]. As discussed in Sec. 1, these traditional approaches rely on explicit geometric conditions, such as predefined ego-vehicle trajectories or semantic layouts, to control the generation process. To the best of our knowledge, OccDirector is the pioneering framework for purely text-guided 4D occupancy generation. Lacking direct off-the-shelf baselines, we adapt DynamicCity by replacing its geometric control tokens with text embeddings from pre-trained encoders, establishing two baselines: *DynamicCity + CLIP [31]* and *DynamicCity + T5 [32]*. Finally, we report results for two variants of our proposed model: (i) w/o history-prefix control, and (ii) w/ history-prefix control.

Implementation Details. We implement OccDirector using PyTorch, trained for 100k iterations with batch size 32, AdamW optimizer (lr=1×10^{-4}, β_1=0.9, β_2=0.95), and CFG dropout probability 0.15. The MMDiT backbone has 14 transformer layers (d_{model}=896) with a 2-layer Token Refiner. At inference we use 20 Euler steps with CFG scale 7.5.

Evaluation Metrics. Following previous work [3, 21, 54], we render the 3D/4D occupancy grids into BEV-view RGB images and clips. These rendered RGB representations are subsequently utilized to evaluate all single-frame and video-level visual quality metrics. Specifically, we report Inception Score (IS) [2], Fréchet Inception Distance (FID) [15], Kernel Inception Distance (KID) [4] and Fréchet Video Distance (FVD) [42] to assess fidelity and temporal coherence, alongside Precision and Recall for diversity. For instruction alignment, we employ the VLM-based instruction alignment score (Sec. 4.3). To ensure consistent and rigorous comparison, we fix `gemini-3.0-pro` as our sole judge VLM for all experiments. For evaluation, we uniformly sample 10K ground-truth 16-frame clips and generate a matching set via the CFG unconditional branch. Single-frame metrics (IS, FID, KID, Precision, Recall) are computed on 5 frames uniformly sampled per clip using Inception-v3 [38] features; video-level FVD and video Precision/Recall use the full 16-frame sequences with I3D [5] features.

5.2 Generation Quality Results

Quantitative results. We present the single-frame visual quality results in Tab. 3 and the video generation quality results in Tab. 4.

As shown in Tab. 3, OccDirector demonstrates highly competitive per-frame generation quality. It is important to note that traditional methods (e.g., DOME, COME, DynamicCity) rely on explicit geometric conditions (e.g., trajectories or layouts) which naturally constrain the generation space and aid in structural consistency. Despite OccDirector being designed for the more abstract text-to-4D task, our unconditional samples achieve an Inception Score (IS) of 3.40 and FID of 62, performance that is comparable to these geometric-guided upper bounds. Crucially, when compared to the direct text-adapted baselines (DynamicCity + CLIP/T5) which share the same functional goal, our method consistently

Table 3: Single-frame visual quality. We evaluate per-frame image fidelity and diversity using Inception Score (IS), Fréchet Inception Distance (FID), Kernel Inception Distance (KID), Precision and Recall.

Method	IS ↑	FID ↓	KID ↓	Precision ↑	Recall ↑
DOME [14]	3.21	67	0.06	0.38	**0.23**
COME [36]	3.26	77	0.06	**0.40**	**0.23**
DynamicCity [3]	**3.39**	**61**	**0.05**	0.31	**0.23**
DynamicCity + CLIP	3.39	65	**0.05**	0.32	**0.24**
DynamicCity + T5	3.36	66	**0.05**	0.31	0.22
OccDirector w/o history-prefix control	3.39	64	**0.05**	**0.34**	**0.24**
OccDirector w/ history-prefix control	**3.40**	**62**	**0.05**	0.33	0.23

Table 4: Video generation quality and instruction alignment. We evaluate video-level fidelity and diversity using Fréchet Video Distance (FVD), Precision and Recall. Instruction alignment is evaluated using the VLM-based rubric score (Sec. 4.3).

	Video realism			Instruction alignment			
Method	FVD ↓	Precision ↑	Recall ↑	Completeness ↑	Structural ↑	Sem. Align. ↑	Mean ↑
DOME [14]	437	0.55	0.03	N/A	N/A	N/A	N/A
COME [36]	**391**	0.56	0.04	N/A	N/A	N/A	N/A
DynamicCity [3]	401	**0.57**	**0.08**	N/A	N/A	N/A	N/A
DynamicCity + CLIP	388	0.60	0.008	3.03	3.54	2.69	3.09
DynamicCity + T5	396	**0.64**	0.02	2.84	3.34	2.73	2.97
OccDirector w/o history-prefix control	354	0.48	**0.12**	3.59	3.66	3.04	3.43
OccDirector w/ history-prefix control	**275**	0.58	**0.12**	**3.81**	**4.01**	**3.32**	**3.73**

achieves superior performance across all single-frame metrics. This confirms that our architecture effectively learns the underlying scene distribution and generates high-fidelity occupancy frames, outperforming naive adaptations of previous works.

For video-level generation (Tab. 4, left), OccDirector significantly improves temporal coherence and dynamic realism. Our full model achieves an FVD of 275, outperforming the strongest baseline (DynamicCity + CLIP, FVD 388) by a remarkable 29.1%. Furthermore, OccDirector achieves the highest video Recall (0.12) while maintaining a strong Precision (0.58). We observe an interesting phenomenon with the text-adapted baselines (DynamicCity + CLIP/T5): while they achieve relatively high Precision (e.g., 0.64 for T5), their Recall is extremely low (e.g., 0.008 for CLIP). This suggests a severe mode collapse, where standard text encoders fail to adequately capture complex text instructions, leading the model to repeatedly generate a limited set of "safe" or static scenarios. In contrast, our method generates spatio-temporal occupancy sequences that are not only realistic but also encompass a significantly wider diversity of dynamic scenarios.

Furthermore, in terms of text-to-occupancy instruction alignment (Tab. 4, right), OccDirector substantially surpasses the text-adapted baselines across all VLM-based rubric scores. For instance, our full model achieves a mean

score of 3.73. The low semantic alignment scores of DynamicCity + CLIP/T5 further corroborate our observation regarding their low Recall: standard text encoders struggle to differentiate between diverse prompts, resulting in poor alignment and limited generation diversity. This highlights the severe semantic-spatiotemporal gap in standard text encoders and proves the superiority of our VLM-driven architecture. Notably, the inclusion of the history-prefix control mechanism further boosts the structural correctness and instruction-following capabilities (improving the mean score from 3.43 to 3.73), demonstrating the effectiveness of our proposed conditioning design for complex multi-agent scenario orchestration.

Qualitative results on history-prefix control. As the teaser in Sec. 1 highlights OccDirector's diverse controllability, we further investigate its capability in long-horizon autoregressive rollouts. Fig. 4 illustrates two multi-stage narratives where the ego vehicle is denoted by a *black car icon*. In the **stop-and-go scenario** (top), the model follows the instruction to decelerate on a one-way street and stop behind a van at a red light (Stage 1). In Stage 2, responding to the "green light" prompt, the van initiates movement across the intersection edge (marked by *red dashed lines*), correctly prompting the ego to prepare to follow. In the **emergency braking scenario** (bottom), the ego initially cruises on a wide, empty straight road. Subsequently, the model successfully generates a sudden pedestrian crossing event at a T-intersection (highlighted by a *red circle*), forcing the ego to execute an emergency stop. These results confirm that history-prefix control effectively maintains spatio-temporal coherence and causal logic across generation stages.

Qualitative comparison on fine-grained text alignment. Fig. 5 compares OccDirector with text-adapted baselines on three prompts requiring precise semantic understanding. In the first column "...follows a lead car, with *no other* traffic...", baselines fail to respect the negation, generating extraneous vehicles. We highlight these hallucinations in baselines versus the correct single lead car in our result using *red circles*. In the second column "...moving ego... *stop* by a bus", baselines exhibit attribute confusion (e.g., DynamicCity + CLIP renders a completely static scene, while the T5 variant incorrectly halts the ego vehicle while moving the bus). Here, *red circles* contrast the baselines' incorrect objects with our successfully generated blocking bus. Finally, for the "cut-in" prompt, we indicate the ego's forward direction with *red dashed lines*. While baselines revert to simple lane-following, only OccDirector synthesizes the lead car's lateral intrusion across this line, demonstrating superior grasp of spatial dynamics.

5.3 Ablation Study

To thoroughly evaluate the effectiveness of our proposed architectural designs, we conduct comprehensive ablation studies. Since the impact of the history-prefix control has already been demonstrated in Tab. 4 and Fig. 4, we establish our ablation baseline using the text-to-occupancy generation setting **without (w/o)**

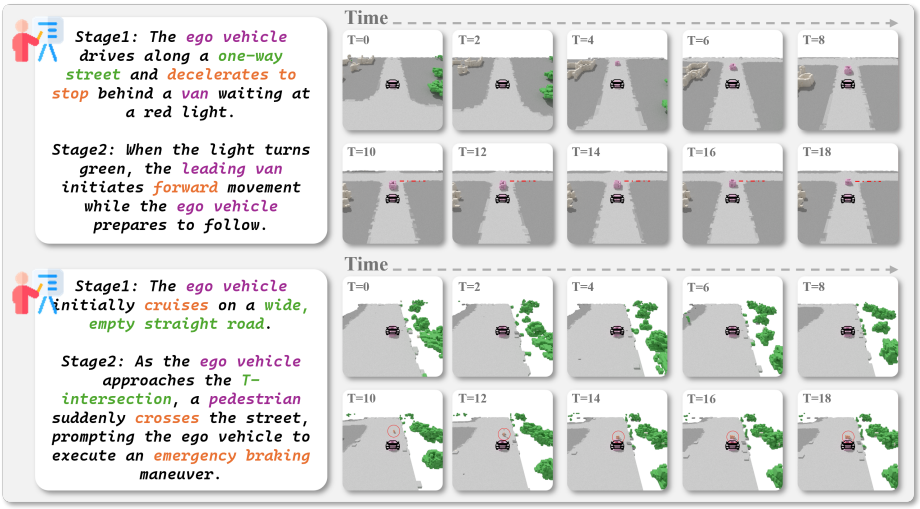

Fig. 4: Long-horizon narrative generation. We visualize multi-stage rollouts for a stop-and-go scenario (top) and an emergency braking event (bottom).

Table 5: Ablation study on model architecture. We evaluate the impact of the attention mechanism (Full vs. STSA) and the Token Refiner under the setting w/o history control.

Architecture Design		Video realism			VLM score ↑			
Backbone	Token Refiner	FVD ↓	Precision ↑	Recall ↑	Completeness ↑	Structural ↑	Instr. follow ↑	Mean ↑
Full attention	×	592	**0.48**	0.02	**2.98**	**2.86**	2.48	2.77
Full attention	✓	**532**	0.41	**0.07**	2.97	2.74	**2.63**	**2.78**
STSA (ours)	×	407	0.49	**0.14**	3.46	3.39	2.86	3.24
STSA (ours)	✓	**354**	**0.58**	0.12	**3.59**	**3.66**	**3.04**	**3.43**

history control. We focus on two key components: (i) The backbone strategy, where we compare our Spatio-Temporal Separated Attention (STSA) against a Full Attention baseline that applies standard joint attention over the entire flattened 4D token sequence. (ii) The Token Refiner, a bidirectional transformer module designed to bridge the semantic-spatial gap of VLM features. Results are summarized in Tab. 5.

Effectiveness of STSA and Token Refiner. First, STSA is critical for modeling 4D dynamics. Compared to the Full Attention baseline (FVD 532), STSA significantly reduces FVD to 354 and improves the VLM score (2.78 → 3.43), confirming the benefit of factorizing spatial and temporal computation. Second, the Token Refiner effectively bridges the gap between frozen text embeddings and occupancy features. Removing the refiner degrades performance (FVD 354 → 407), verifying its necessity for aligning high-dimensional text features with spatial

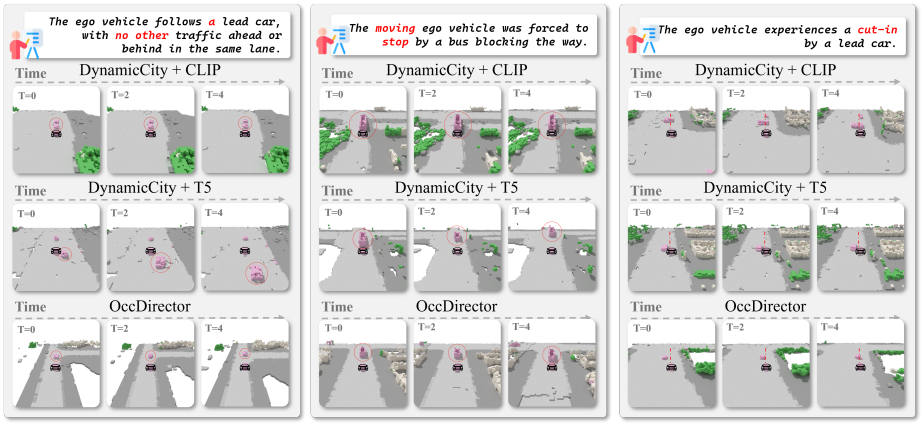

Fig. 5: Qualitative comparison of text-to-occupancy alignment. We evaluate adherence to fine-grained instructions: (Left) Negation and cardinality; (Middle) Dynamic attribute binding; (Right) Complex "cut-in" trajectory.

constraints. Notably, the Full Attention backbone fails to fully leverage the refined tokens (VLM score stagnates at 2.78), whereas our STSA architecture successfully integrates them to achieve state-of-the-art fidelity and alignment.

6 Conclusion

In this paper, we presented OccDirector, a pioneering framework capable of generating dynamic 4D occupancy scenes purely from natural language instructions. By shifting the paradigm from rigid geometric conditioning to high-level semantic orchestration, OccDirector acts as a "scenario director", enabling intuitive control over complex multi-agent behaviors and environmental interactions. To achieve this, we introduced a VLM-driven Spatio-Temporal MMDiT architecture equipped with a novel history-prefix anchoring strategy, effectively bridging the semantic-spatiotemporal gap and ensuring physically plausible, interaction-consistent rollouts. Furthermore, we contributed OccInteract-85k, a large-scale dataset with multi-level language descriptions, alongside a rigorous VLM-based evaluation benchmark. Extensive experiments demonstrate that our approach achieves state-of-the-art generation quality and unprecedented instruction-following capabilities.

References

1. Alaba, S.Y., Ball, J.E.: A survey on deep-learning-based lidar 3d object detection for autonomous driving. Sensors **22**(24), 9577 (2022)
2. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
3. Bian, H., Kong, L., Xie, H., Pan, L., Qiao, Y., Liu, Z.: Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. arXiv preprint arXiv:2410.18084 (2024)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10164–10183 (2024)
7. Chen, R., Wu, Z., Liu, Y., Guo, Y., Ni, J., Xia, H., Xia, S.: Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. arXiv preprint arXiv:2412.04842 (2024)
8. Diehl, C., Sykora, Q., Agro, B., Gilles, T., Casas, S., Urtasun, R.: Dio: Decomposable implicit 4d occupancy-flow world model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27456–27466 (2025)
9. Dong, W., Lu, S., Chen, X., Zhang, S., Liu, Q., Liu, Z., Chen, L., Wang, H., Cai, Y.: End-to-end autonomous driving: From classic paradigm to large model empowerment—a comprehensive survey. IEEE Internet of Things Journal **13**(3), 3870–3898 (2025)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
11. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28135–28144 (2025)
12. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023)
13. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
14. Gu, S., Yin, W., Jin, B., Guo, X., Wang, J., Li, H., Zhang, Q., Long, X.: Dome: Taming diffusion model into high-fidelity controllable occupancy world model. arXiv preprint arXiv:2410.10429 (2024)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)

17. Khurana, T., Hu, P., Held, D., Ramanan, D.: Point cloud forecasting as a proxy for 4d occupancy forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1116–1124 (2023)
18. Kim, S.W., Philion, J., Torralba, A., Fidler, S.: Drivegan: Towards a controllable high-quality neural simulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5820–5829 (2021)
19. Koishigarina, D., Uselis, A., Oh, S.J.: Clip behaves like a bag-of-words model cross-modally but not uni-modally. arXiv preprint arXiv:2502.03566 (2025)
20. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
21. Lee, J., Lee, S., Jo, C., Im, W., Seon, J., Yoon, S.E.: Semcity: Semantic scene generation with triplane diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28337–28347 (2024)
22. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the computer vision and pattern recognition conference. pp. 11971–11981 (2025)
23. Li, L., Shao, W., Dong, W., Tian, Y., Zhang, Q., Yang, K., Zhang, W.: Data-centric evolution in autonomous driving: A comprehensive survey of big data system, data mining, and closed-loop technologies. arXiv preprint arXiv:2401.12888 (2024)
24. Liang, A., Kong, L., Yan, T., Liu, H., Yang, W., Huang, Z., Yin, W., Zuo, J., Hu, Y., Zhu, D., et al.: Worldlens: Full-spectrum evaluations of driving world models in real world. arXiv preprint arXiv:2512.10958 (2025)
25. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
26. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
27. Ma, B., Zong, Z., Song, G., Li, H., Liu, Y.: Exploring the role of large language models in prompt encoding for diffusion models. arXiv preprint arXiv:2406.11831 (2024)
28. Ma, J., Chen, X., Huang, J., Xu, J., Luo, Z., Xu, J., Gu, W., Ai, R., Wang, H.: Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21486–21495 (2024)
29. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
30. Mei, J., Hu, T., Yang, X., Wen, L., Yang, Y., Wei, T., Ma, Y., Dou, M., Shi, B., Liu, Y.: Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. arXiv preprint arXiv:2409.04003 (2024)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)

33. Ruan, B.K., Tsui, H.T., Li, Y.H., Shuai, H.H.: Traffic scene generation from natural language description for autonomous vehicles with large language model. arXiv preprint arXiv:2409.09575 (2024)
34. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
36. Shi, Y., Jiang, K., Meng, Q., Wang, K., Wang, J., Sun, W., Wen, T., Yang, M., Yang, D.: Come: Adding scene-centric forecasting control to occupancy world model. arXiv preprint arXiv:2506.13260 (2025)
37. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
39. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. Advances in Neural Information Processing Systems **36**, 64318–64330 (2023)
40. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
41. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8406–8415 (2023)
42. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
43. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
44. Wang, L., Zheng, W., Ren, Y., Jiang, H., Cui, Z., Yu, H., Lu, J.: Occsora: 4d occupancy generation models as world simulators for autonomous driving. arXiv preprint arXiv:2405.20337 (2024)
45. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
46. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17850–17859 (2023)
47. Wang, Y., Huang, X., Sun, X., Yan, M., Xing, S., Tu, Z., Li, J.: Uniocc: A unified benchmark for occupancy forecasting and prediction in autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25560–25570 (2025)
48. Wilson, J., Song, J., Fu, Y., Zhang, A., Capodieci, A., Jayakumar, P., Barton, K., Ghaffari, M.: Motionsc: Data set and network for real-time semantic mapping in

dynamic environments. IEEE Robotics and Automation Letters **7**(3), 8439–8446 (2022)
49. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
50. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
51. Xu, H., Chen, J., Meng, S., Wang, Y., Chau, L.P.: A survey on occupancy perception for autonomous driving: The information fusion perspective. Information Fusion **114**, 102671 (2025)
52. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
53. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
54. Yang, Y., Liang, A., Mei, J., Ma, Y., Liu, Y., Lee, G.H.: X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. arXiv preprint arXiv:2506.13558 (2025)
55. Yang, Z., Liu, Z., Lu, Y., Hou, L., Miao, C., Peng, S., Feng, B., Bai, X., Zhao, H.: Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. arXiv preprint arXiv:2512.12751 (2025)
56. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)
57. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9433–9443 (2023)
58. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
59. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)