COLLECTIVE EGO-AGENT WORLD MODELING FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

World modeling enables end-to-end autonomous driving (E2EAD) to anticipate scene evolution and learn from action consequences. However, existing methods predominantly adopt an ego-centric perspective, treating surrounding agents as scene content rather than explicit rollout subjects. This formulation limits the completeness of future scene representations and leaves predictive consistency across ego and agent perspectives underconstrained. In this paper, we propose CoDreamer, an E2EAD framework for collective ego-agent world dreaming. CoDreamer integrates an ego-centric branch and selected agent-centric branches through a shared latent dynamics model. To improve spatial completeness, each agent branch reconstructs future features over its full spatial window from a coordinate-transformed, partially observed Bird's-Eye-View (BEV) representation, learning to infer how regions beyond its spatial support evolve. To encourage cross-perspective agreement, ego and agent rollouts are conditioned on their respective logged trajectories and jointly supervised at the prediction horizon by coordinate-aligned views of the same realized future. Ego-Agent Spatial Tokens (EAST) encode relative geometry and entity type, enabling the shared predictor to distinguish the rollout subject from other participants across coordinate frames. This collective learning objective provides additional supervision for planning representations, with future world-model transitions confined to training in the evaluated configuration. On NAVSIMv1, CoDreamer achieves state-of-the-art planning performance among world-model-based methods using the same ResNet-34 backbone, attaining 92.0% PDMS. Ablation studies further demonstrate improved prediction accuracy and scene completion beyond observed regions. These findings underscore the value of agent-centric world modeling for capturing more complete and coherent scene dynamics. Code will be released.

1 INTRODUCTION

"One World, One Dream."

— Olympic Games Beijing 2008

End-to-end autonomous driving (E2EAD) directly maps raw sensor inputs to trajectories or control commands, integrating perception and decision-making within a jointly trainable framework Hu et al. (2023); Jiang et al. (2023); Sun et al. (2025). World modeling extends this paradigm by forecasting scene evolution to support representation learning Li & Cui (2025), action generation Shang et al. (2026), trajectory evaluation Li et al. (2025b), and iterative refinement Zhang et al. (2026). By explicitly predicting the consequences of candidate decisions, it provides a basis for context-aware reasoning beyond correlations learned through imitation Zheng et al. (2026).

However, as illustrated in Fig. 1(b), most world-model-based E2EAD methods formulate future prediction around the ego vehicle, treating surrounding agents as scene elements rather than rollout subjects with their own actions and reference frames. This asymmetry introduces two coupled limitations. First, *viewpoint-restricted supervision* restricts local predictive constraints and annotation utilization. An ego-centered reconstruction objective supervises agent neighborhoods only through their contribution to the global target, without explicitly requiring complete and detailed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

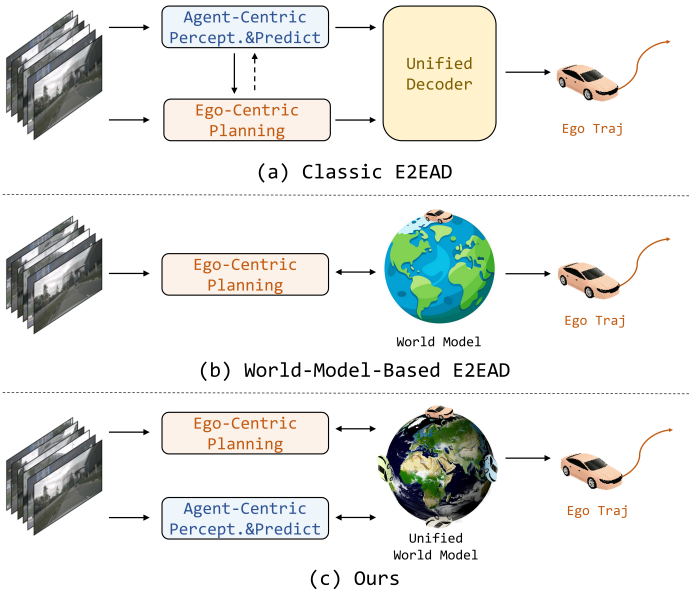

Figure 1: Comparison of E2EAD paradigms. Classic E2EAD (a) couples planning with perception and prediction. Ego-centric world-model-based E2EAD (b) predicts scene evolution around the ego vehicle and is included in our baseline. CoDreamer (c) introduces agent-centric prediction alongside ego dreaming to learn participant-centered scene dynamics. The world-model connections depict auxiliary training in the evaluated configuration.

agent-centered predictions. Consequently, the scene predictor receives no dedicated supervision for local dynamics and spatial relationships, although these details are essential to the surrounding scene evolution modeled by the ego branch. Moreover, ego-only reconstruction underutilizes scene-wide annotations by restricting supervision to a single ego-centered window. Second, *underconstrained ego-agent consistency* leaves participant-conditioned forecasts without a shared compatibility constraint. Although participants act from different perspectives, their realized trajectories belong to the same evolving scene. Forecasts conditioned on these trajectories should therefore describe compatible future states after coordinate alignment. Ego-only world modeling supervises a single action-conditioned perspective, leaving agreement across participant perspectives unregularized. Existing E2EAD methods couple ego and agent reasoning through joint perception Jia et al. (2025) or integrated motion prediction and planning Sun et al. (2025). These approaches motivate learning shared scene dynamics through compatible participant-conditioned forecasts.

We present **CoDreamer**, an E2EAD framework for **Co**llective Ego-Agent World **Dream**ing. CoDreamer treats the ego vehicle and selected surrounding agents as explicit rollout subjects of a shared latent world model. For each agent, we transform ego-centric BEV features into its local frame and zero-pad regions beyond the original spatial support. Inspired by masked representation learning (Assran et al., 2025), each branch reconstructs a complete local future from partially supported scene features. An action token encoded from the logged trajectory of the corresponding participant conditions each prediction. Ego-Agent Spatial Tokens (EAST) further encode entity types and geometry in both ego and subject frames, providing explicit spatial context for participant-conditioned prediction. The shared predictor performs an action-conditioned transition in latent BEV space, and a shared semantic decoder maps the resulting latent state to a future BEV semantic map. Full-window reconstruction targets supervise both overlapping regions and areas beyond ego support, providing dedicated local constraints and exploiting additional supervision from existing annotations. All targets describe the same realized future in their respective reference frames. Aligned reconstruction targets regularize participant-conditioned forecasts toward a compatible shared future, while extending supervision to the detailed evolution of individual agent neighborhoods.

Collective world dreaming serves as training-only auxiliary supervision without additional inference latency. On NAVSIM v1, CoDreamer achieves state-of-the-art performance among methods using

a ResNet-34 backbone, reaching 92.0 PDMS. Further analyses of agent-centric supervision, spatial conditioning, and future-scene prediction support the effectiveness of agent-centric world modeling. Our main contributions are as follows:

- We propose CoDreamer, a collective world-modeling framework that jointly predicts ego- and agent-centered futures through a shared action-conditioned predictor.
- We develop full-window agent-centric reconstruction and Ego-Agent Spatial Tokens to incorporate local future supervision and participant-relative geometry into world modeling, without additional inference latency.
- CoDreamer achieves 92.0 PDMS on NAVSIM v1, setting a new state of the art among methods using a ResNet-34 backbone. Further analyses examine agent-centric supervision, spatial conditioning, and future-scene prediction.

2 Related Work

End-to-End Autonomous Driving (E2EAD) maps sensor observations directly to trajectories or control commands, jointly optimizing perception and decision-making. UniAD (Hu et al., 2023) integrates perception, prediction, and planning within a planning-oriented framework, while VAD (Jiang et al., 2023) adopts vectorized scene representations for efficient planning. SparseDrive (Sun et al., 2025) models the ego vehicle and surrounding agents as sparse instances, supporting parallel motion prediction and planning with explicit ego-agent interactions. iPad (Guo et al., 2025) organizes representation learning around trajectory proposals and iteratively refines them using planning-relevant features. DiffusionDrive (Liao et al., 2025) and GoalFlow (Xing et al., 2025) address multimodal planning through truncated diffusion and goal-conditioned flow matching, respectively. CoDreamer complements these advances by jointly learning ego- and agent-centered future scenes to enrich planning representations. Collective world dreaming serves as training-only auxiliary supervision without additional inference latency.

World modeling in E2EAD predicts scene evolution to support representation learning and planning. Generative approaches (Ren et al., 2025; Gao et al., 2025) synthesize controllable driving scenes for simulation and data generation, while other methods integrate future prediction into driving policies. Future prediction provides supervision for representation learning (Li & Cui, 2025; Li et al., 2026a), supports dynamics-informed action generation (Zhao et al., 2025; Shang et al., 2026), and enables trajectory evaluation through action-conditioned outcomes (Wang et al., 2024; Li et al., 2025b). Predicted future states can also guide iterative trajectory refinement (Zhang et al., 2025b; 2026). Despite these diverse uses, many driving world models formulate prediction in an ego-centric frame, treating surrounding agents as scene elements rather than explicit prediction subjects with their own conditioning actions and local reference frames. Related interaction-aware planners couple participants through joint representations or integrated prediction and planning (Jia et al., 2025; Sun et al., 2025), motivating a participant-centered formulation of world modeling. CoDreamer treats both the ego vehicle and surrounding agents as prediction subjects within a shared action-conditioned world model. Full-window agent-centric reconstruction extends local future supervision beyond the ego BEV support, while spatial conditioning and coordinate-aligned targets encourage compatible predictions of the same realized future across participant perspectives.

3 Method

3.1 Overall Framework

Figure 2 illustrates CoDreamer, which augments an anchor-based end-to-end planner with collective ego-agent world dreaming. A shared action-conditioned world model predicts future scenes from ego-centric and agent-centric perspectives, providing complementary supervision through full-window local reconstruction (Section 3.2). Ego-Agent Spatial Tokens (EAST) encode participant-relative geometry and entity types to support prediction across reference frames (Section 3.3). Collective world dreaming provides training-only auxiliary supervision for planning without additional inference latency (Section 3.4).

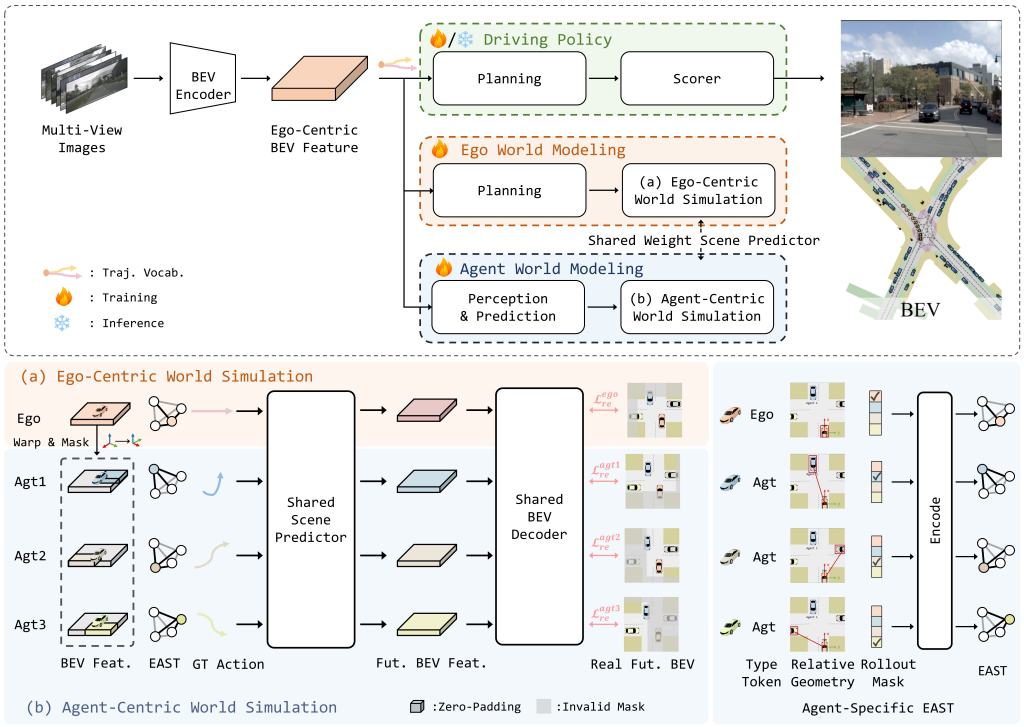

Figure 2: Overview of CoDreamer. A shared action-conditioned world model predicts future scenes from ego- and agent-centered perspectives. Ego-Agent Spatial Tokens (EAST) encode participant-relative geometry and entity types. Full-window reconstruction with coordinate-aligned targets encourages spatial completeness and cross-perspective consistency. Collective world dreaming provides training-only auxiliary supervision for planning without additional inference latency.

Scene Encoding. Given the current multi-view images $\mathcal{I}_0$, an image encoder (He et al., 2016) extracts features $\mathbf{F}_{\text{cam}} \in \mathbb{R}^{N_{\text{cam}} \times H_f \times W_f \times C_f}$, where N_{cam} denotes the number of camera views and H_f, W_f, and C_f denote the feature height, width, and channel dimension. Geometric lifting and BEV encoding (Philion & Fidler, 2020) produce an ego-centered representation $\mathbf{B}_0^{\text{ego}} \in \mathbb{R}^{H_b \times W_b \times C}$, where H_b and W_b are the BEV dimensions and C is the latent dimension. The image and BEV features support trajectory-conditioned planning and spatial scene understanding, respectively. To incorporate instance-level geometry and motion into the learned representations, we jointly train auxiliary agent detection and motion prediction branches. Detection queries encode surrounding agents under classification and localization supervision and initialize the corresponding motion queries for future trajectory prediction. This perception-to-motion pathway also provides the agent queries used for action conditioning in collective world dreaming (Section 3.2).

Trajectory Planning. Following anchor-based planning (Guo et al., 2025; Li et al., 2025b), we initialize N_{traj} trajectory anchors $\{\bar{\boldsymbol{\tau}}_{\text{ego},i}\}_{i=1}^{N_{\text{traj}}}$, where each anchor contains N_{wp} waypoints with planar positions and headings. The corresponding queries $\mathbf{Q}_{\text{ego},i}^{(0)} \in \mathbb{R}^{N_{\text{wp}} \times C}$ combine anchor, ego-status, and waypoint embeddings. At stage $r \in \{1, \ldots, N_{\text{ref}}\}$, the planner predicts anchor-relative offsets and updates the queries using image features sampled along the refined trajectory:

$$\begin{aligned} \hat{\boldsymbol{\tau}}_{\text{ego},i}^{(r)} &= \bar{\boldsymbol{\tau}}_{\text{ego},i} + \text{MLP}_{\text{reg}}\left(\mathbf{Q}_{\text{ego},i}^{(r-1)}\right), \\ \mathbf{Q}_{\text{ego},i}^{(r)} &= \text{TrajAttn}\left(\mathbf{Q}_{\text{ego},i}^{(r-1)}, \hat{\boldsymbol{\tau}}_{\text{ego},i}^{(r)}, \mathbf{F}_{\text{cam}}\right), \end{aligned} \tag{1}$$

where N_{ref} is the number of refinement stages and TrajAttn denotes trajectory-conditioned image attention. A base scorer evaluates the final proposals for trajectory selection, while collective world dreaming supplies auxiliary future-scene supervision during training.

3.2 COLLECTIVE EGO-AGENT WORLD DREAMING

Masked predictive learning encourages representations to capture spatiotemporal structure by predicting missing information from incomplete context. To extend this predictive learning principle to driving scenes, we formulate collective world dreaming as action-conditioned future prediction across ego- and agent-centered views. A shared latent predictor maps partially supported local BEV features to full-window future semantic targets, coupling spatial completion with dynamic modeling.

Agent-Centered Scene Representation. Let $\mathcal{A}$ index the selected surrounding vehicles and $\mathcal{S} = \{\text{ego}\} \cup \{(\text{agt}, k) \mid k \in \mathcal{A}\}$ denote the prediction subjects. For agent k, the current pose defines a local frame whose rotation $\mathbf{R}_k$ and translation $\mathbf{t}_k$ map local coordinates to the ego frame. The agent-centered representation is obtained by bilinear resampling of the ego BEV:

$$\mathbf{B}_0^{\text{agt},k} = \text{Sample}(\mathbf{B}_0^{\text{ego}}, \mathbf{R}_k\mathbf{p} + \mathbf{t}_k)\,, \tag{2}$$

where $\mathbf{p} \in \mathbb{R}^2$ denotes a local BEV position. Each local frame remains fixed at the current pose throughout prediction. To couple spatial completion with dynamic modeling, we zero-pad regions outside the input BEV support within $\mathbf{B}_0^{\text{agt},k}$ and reconstruct future semantics over the full agent-centric perspective. Following the principle of masked predictive learning, this formulation encourages inference of missing future content from available context and conditioning actions. Targets are rendered from map geometry and future annotations in the fixed local frame, with the ego vehicle assigned a distinct semantic class.

Participant-Conditioned Action Representation. To align action representations across participant types, we adopt structurally symmetric ego and agent action encoders with separate parameters. The ego encoder uses planning queries, while agent encoder uses motion queries initialized from detection queries and refined through motion prediction. Both encode the corresponding ground-truth trajectories through trajectory-conditioned image attention, followed by pooling and projection:

$$\begin{aligned} \mathbf{Q}_{\text{act}}^s &= \text{TrajAttn}\big(\mathbf{Q}_{\text{init}}^s, \boldsymbol{\tau}_{\text{gt}}^s, \mathbf{F}_{\text{cam}}\big)\,, \\ \mathbf{a}^s &= \text{Proj}(\text{Pool}(\mathbf{Q}_{\text{act}}^s))\,, \end{aligned} \tag{3}$$

where $\mathbf{Q}_{\text{init}}^s$ denotes the ego planning or refined agent motion queries, and $\boldsymbol{\tau}_{\text{gt}}^s$ is the corresponding ego or agent ground-truth trajectory expressed in ego coordinates. The resulting action token $\mathbf{a}^s \in \mathbb{R}^C$ conditions world prediction on the same recorded future as the semantic target. Agent detection and motion prediction retain their regular outputs and supervised objectives, jointly learning instance- and motion-aware representations.

Shared Future-Scene Prediction. For each subject $s \in \mathcal{S}$, a shared latent rollout integrates the local BEV $\mathbf{B}_0^s$, action token $\mathbf{a}^s$, and EAST tokens $\mathbf{Z}^s$ (section 3.3) to predict future scene features. A shared semantic decoder then reconstructs the terminal BEV:

$$\begin{aligned} \hat{\mathbf{B}}_T^s &= \mathcal{R}_\theta(\mathbf{B}_0^s, \mathbf{a}^s, \mathbf{Z}^s)\,, \\ \hat{\mathbf{Y}}_T^s &= \text{Dec}_{\text{sem}}\Big(\hat{\mathbf{B}}_T^s\Big)\,, \end{aligned} \tag{4}$$

where $\mathcal{R}_\theta$ is a transformer-based predictor that performs a single latent transition to horizon T. Action features are spatially grounded at the current and terminal positions, and predicted scene tokens are reshaped into BEV features for decoding. The decoder also supervises the current ego BEV. Shared parameters and coordinate-aligned targets encourage compatible predictions across participant perspectives without an explicit consistency loss.

3.3 EGO-AGENT SPATIAL TOKENS (EAST)

To relate participant-centered views to the original ego observation, we introduce EAST to encode entity poses in both reference frames, together with dimensions and types. For entity j, let $\mathbf{p}_j^s$ and $\mathbf{p}_j^{\text{ego}}$ denote its current poses in the subject and ego frames. A pose $\mathbf{p} = (x, y, \theta)$ is encoded as $\psi(\mathbf{p}) = [x/d_x, y/d_y, \sin\theta, \cos\theta]$, where d_x and d_y normalize spatial coordinates. Given dimen-

sions (ℓ_j, w_j), validity indicator $m_j \in \{0, 1\}$, and type c_j, the spatial tokens are:

$$\begin{aligned} \mathbf{g}_j^s &= \mathrm{Concat}\big(\psi(\mathbf{p}_j^s), \psi(\mathbf{p}_j^{\mathrm{ego}}), \ell_j/d_x, w_j/d_y, m_j\big), \\ \mathbf{z}_j^s &= m_j \,\mathrm{LN}\big(\mathrm{Emb}(c_j) + \mathrm{MLP}(\mathbf{g}_j^s)\big)\,, \\ \mathbf{Z}^s &= \Big[\mathbf{z}_{\mathrm{ego}}^s, \mathbf{z}_{\mathrm{agt},1}^s, \ldots, \mathbf{z}_{\mathrm{agt},N_{\mathrm{agt}}}^s\Big]\,, \end{aligned} \tag{5}$$

where $\mathbf{Z}^s \in \mathbb{R}^{(N_{\mathrm{agt}}+1)\times C}$ and N_{agt} is the number of agent slots. Entity types distinguish the ego vehicle, surrounding vehicles, and padding. All views preserve the entity order, with the prediction subject at the local origin and invalid slots masked during attention. EAST is learned through future semantic supervision using only current-scene geometry.

3.4 Joint Training and Inference

Training. To obtain agent-centered future targets, we crop the global semantic map at time T according to the ground-truth pose and local BEV window of each agent:

$$\mathbf{Y}_T^{\mathrm{agt},k} = \mathrm{Crop}\Big(\mathbf{Y}_T^{\mathrm{global}}, \mathbf{p}_{0,\mathrm{gt}}^{\mathrm{agt},k}, \Omega_{\mathrm{agt}}\Big)\,, \tag{6}$$

where $\mathbf{Y}_T^{\mathrm{global}}$ combines map geometry with future annotations, $\mathbf{p}_{0,\mathrm{gt}}^{\mathrm{agt},k}$ defines the local frame at the current ground-truth pose, and Ω_{agt} specifies the full local window. We retain the planning and detection objectives of iPad (Guo et al., 2025), denoted jointly by $\mathcal{L}_{\mathrm{iPad}} = \mathcal{L}_{\mathrm{plan}} + \mathcal{L}_{\mathrm{det}}$, and introduce auxiliary motion prediction and semantic reconstruction. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{iPad}} + \mathcal{L}_{\mathrm{mot}} + \lambda_0 \mathcal{L}_{\mathrm{sem},0}^{\mathrm{ego}} + \lambda_{\mathrm{ego}} \mathcal{L}_{\mathrm{sem},T}^{\mathrm{ego}} + \lambda_{\mathrm{agt}} \mathcal{L}_{\mathrm{sem},T}^{\mathrm{agt}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{mot}}$ supervises agent motion prediction, and the semantic terms apply focal loss to current ego, future ego, and future agent BEVs, respectively. The agent term averages full-window reconstruction losses over valid subjects. Loss weights and training settings are provided in section 4.1.

Inference. Given scoring logits s_i, the base planner selects the highest-scoring final proposal:

$$\begin{aligned} i^\star &= \underset{1\le i\le N_{\mathrm{traj}}}{\arg\max}\; s_i, \\ \boldsymbol{\tau}_{\mathrm{ego}}^\star &= \hat{\boldsymbol{\tau}}_{\mathrm{ego},i^\star}^{(N_{\mathrm{ref}})}. \end{aligned} \tag{8}$$

Collective world dreaming provides auxiliary supervision without additional inference latency.

4 Experiments

4.1 Dataset and Implementation Details

Benchmark and Metrics. We evaluate CoDreamer on the NAVSIMv1 split using the PDM score (PDMS). The score combines no-at-fault collision (NC), drivable-area compliance (DAC), time-to-collision (TTC), comfort (Comf.), and ego progress (EP). Higher scores indicate better planning performance. We additionally evaluate future BEV semantic prediction using mean intersection-over-union (mIoU), comparing the learned world model (WM) with a copy-current (CP) baseline.

Implementation Details. CoDreamer processes four current camera views using a ResNet-34 image encoder (He et al., 2016). Following VeteranAD (Zhang et al., 2025a), an LSS-based BEV encoder (Philion & Fidler, 2020) produces an $8\times8\times256$ representation covering $[0, 32]\times[-32, 32]$ m. K-means clustering of training trajectories yields 64 ego trajectory anchors and six agent-motion anchors, each comprising eight poses sampled at 0.5 s intervals over a 4 s horizon. We select up to 15 agents in ascending order of Euclidean distance from the ego origin, requiring valid motion targets, a minimum distance of 1 m, and a BEV overlap threshold of 0.15. EAST uses 16 slots including the ego vehicle, with spatial normalization $d_x = d_y = 32$ m. The trajectory-attention planner follows VeteranAD (Zhang et al., 2025a) with $N_{\mathrm{ref}} = 4$ refinement stages. The agent-motion branch adopts the same architecture with separate parameters and is trained with a winner-takes-all (WTA)

Table 1: End-to-end planning performance on NAVSIM *navtest*. C, L, and T denote camera, LiDAR, and temporal inputs, respectively. Results across methods use their respective sensor and settings.

Method	Backbone	Input	PDMS↑	NC↑	DAC↑	TTC↑	Comf.↑	EP↑
Classic E2E								
UniAD Hu et al. (2023)	ResNet-34	C	83.4	97.8	91.9	92.9	100.0	78.8
TransFuser Chitta et al. (2023)	ResNet-34	C & L	84.0	97.7	92.8	92.8	100.0	79.2
PARA-Drive Weng et al. (2024)	ResNet-34	C	84.0	97.9	92.4	93.0	99.8	79.3
Hydra-MDP Li et al. (2024)	ResNet-34	C & L	86.5	98.3	96.0	94.6	100.0	78.7
DiffusionDrive Liao et al. (2025)	ResNet-34	C & L	88.1	98.2	96.2	94.7	100.0	82.2
VeteranAD Zhang et al. (2025a)	ResNet-34	C	90.2	99.1	98.3	96.1	100.0	83.1
VLA-Based								
ExploreVLA Sheng et al. (2026)	Show-o	C	90.4	98.8	98.4	96.5	99.9	83.5
ReCogDrive Li et al. (2026b)	InternViT	C	90.8	97.9	97.3	94.9	100.0	87.3
ELF-VLA Luo et al. (2026)	InternViT	C	91.0	98.9	98.1	96.0	100.0	85.3
DynVLA Shang et al. (2026)	Emu3	C & T	91.7	98.6	98.7	95.5	100.0	86.8
DriveVLA-W0 Li et al. (2026a)	Emu3	C	93.0	99.3	97.4	97.0	99.9	88.3
World-Model-Based								
LAW Li et al. (2025a)	ResNet-34	C	84.6	96.4	95.4	88.7	99.9	81.7
World4Drive Zheng et al. (2025)	ResNet-34	C	85.1	97.4	94.3	92.8	100.0	79.9
WorldRFT Yang et al. (2026b)	ResNet-34	C	87.8	97.8	96.8	94.0	100.0	81.7
WoTE Li et al. (2025b)	ResNet-34	C & L	88.3	98.5	96.8	94.9	99.9	81.9
SeerDrive Zhang et al. (2025b)	ResNet-34	C & L	88.9	98.4	97.0	94.9	99.9	83.2
ResWorld Zhang et al. (2026)	ResNet-34	C & L & T	89.0	98.9	96.5	95.6	100.0	83.1
Drive-JEPA Wang et al. (2026)	ResNet-34	C	91.5	98.2	98.0	94.2	99.9	88.8
SafeDrive Kim et al. (2026)	ResNet-34	C & L	91.6	99.5	99.0	97.2	100.0	84.3
PWM Zhao et al. (2025)	Show-o	C & T	88.1	98.6	95.9	95.4	100.0	81.8
SeerDrive Zhang et al. (2025b)	V2-99	C & L	90.7	98.8	98.6	95.8	100.0	84.2
DriveFuture Hong et al. (2026)	V2-99	C	90.7	98.8	99.1	95.4	100.0	84.2
Auto-JEPA Yang et al. (2026a)	V-JEPA2	C	91.3	98.4	98.3	95.0	100.0	87.1
SimWAM Zhao et al. (2026)	Wan2.2-5B	C	91.7	98.4	98.7	95.5	100.0	86.4
Drive-JEPA Wang et al. (2026)	ViT-L	C	93.7	99.1	98.2	96.2	99.9	91.4
CoDreamer (Ours)	ResNet-34	C	92.0	98.3	98.3	94.8	99.9	89.1

objective (Sun et al., 2025). The shared Q-Former scene predictor uses two attention layers, a feed-forward dimension of 512, and dropout of 0.1 to predict the terminal scene in a single transition. The shared future-BEV decoder follows WoTE (Li et al., 2025b) and predicts eight semantic classes at a resolution of 128×256. Ego trajectories are derived from logged poses, and agent trajectories from annotated tracks. Agent-centered BEV features are constructed by warping and masking the ego BEV, while future semantic targets are independently rasterized from map geometry and future annotations in each agent-centered frame. We train for 20 epochs on four GPUs with a total batch size of 32 using Adam, with constant learning rates of 10^{-4} for non-world-model parameters and 2.5×10^{-5} for trainable world-model parameters. Semantic supervision uses focal loss with $\alpha = 0.5$, $\gamma = 2$, $\lambda_0 = 10$, and $\lambda_{\text{ego}} = \lambda_{\text{agt}} = 0.1$. Trajectory, scoring, and classification losses have unit weights, while auxiliary-box and detection-box losses are weighted by 0.1, the area loss by 2, and agent-motion classification and regression losses by 0.2 each.

4.2 Comparative Evaluation

Planning Performance. As shown in Table 1, CoDreamer achieves a state-of-the-art PDMS of 92.0 among the compared methods using camera inputs and a ResNet-34 backbone. Within the same backbone family, CoDreamer surpasses the reported PDMS of WoTE and SeerDrive by 3.7 and 3.1 points, respectively. It also achieves an ego-progress score of 89.1, drivable-area compliance of 98.3, and comfort of 99.9. All results follow the original settings of each method without a common retraining protocol. CoDreamer also differs in how future prediction supports planning. WoTE (Li et al., 2025b) evaluates trajectories using predicted future states, while SeerDrive (Zhang et al., 2025b) incorporates future scene features into trajectory generation. CoDreamer learns from

Modules			Performance					
Baseline	Agent-WM	EAST	PDMS↑	NC↑	DAC↑	TTC↑	Comf.↑	EP↑
✓	×	×	91.1	98.2	97.6	94.5	**99.9**	88.2
✓	✓	×	91.2	98.0	97.7	94.1	**99.9**	88.8
✓	✓	✓	**92.0**	**98.3**	**98.3**	**94.8**	**99.9**	**89.1**

Table 2: Component ablation of CoDreamer on NAVSIM. The baseline is the classic planner paired with ego-centric world modeling. Agent-WM denotes agent-centric world modeling, and EAST denotes Ego-Agent Spatial Tokens.

Agent number N_{agt}	Performance					
	PDMS↑	NC↑	DAC↑	TTC↑	Comf.↑	EP↑
4	91.0	97.9	97.8	94.3	**100.0**	88.1
10	91.4	97.8	97.9	93.9	99.9	**89.3**
15	**92.0**	**98.3**	98.3	**94.8**	99.9	89.1
20	91.6	98.0	**98.4**	94.3	99.9	88.5

Table 3: Effect of the maximum number of selected agents N_{agt} on NAVSIM planning performance.

participant-centered futures during training and selects trajectories with the base scorer at inference. These results demonstrate the effectiveness of collective world modeling as training-only supervision without online future-scene prediction.

4.3 Component Analysis

We examine collective supervision and spatial conditioning, the agent budget, and the predictive ability of the learned world model.

Collective Supervision and Spatial Conditioning. Table 2 compares the ego-only world-modeling baseline with configurations incorporating agent-centered supervision and EAST. The Agent-WM configuration without EAST obtains 91.2 PDMS, compared with 91.1 for the baseline. The complete CoDreamer model reaches 92.0 PDMS, exceeding the baseline by 0.9 points and the configuration without EAST by 0.8 points. It also improves NC, DAC, TTC, and EP over the baseline while maintaining a comfort score of 99.9. These configuration-level comparisons favor combining agent-centered future supervision with explicit participant-relative spatial context.

Number of Agent Subjects. Table 3 examines the maximum number of selected agents N_{agt}. PDMS increases from 91.0 to 91.4 and 92.0 as the budget grows from 4 to 10 and 15 agents, respectively, but decreases to 91.6 with twenty agents. This non-monotonic trend suggests diminishing returns from additional participant-centered supervision. We adopt $N_{\mathrm{agt}} = 15$, which yields the highest PDMS among the evaluated budgets.

Prediction Beyond the Current Scene. Table 4 compares learned future prediction with the copy-current baseline (CP), which directly uses the current semantic prediction as the future estimate. CoDreamer achieves 31.8 future mIoU, exceeding CP at 26.7 by 5.1 points. This gain shows that the learned transition provides a more accurate estimate of future scene semantics than retaining the current prediction.

5 Conclusion

We presented CoDreamer, a collective world-modeling framework that treats the ego vehicle and surrounding agents as explicit subjects of a shared action-conditioned predictor. Full-window agent-centric reconstruction extends future supervision beyond the ego BEV support, while EAST and

Table 4: Future BEV semantic prediction and planning performance. CP copies the current semantic prediction, while WM predicts the future semantic scene.

Configuration	Future CP mIoU↑	Future WM mIoU↑	PDMS↑
CoDreamer	26.7	31.8	92.0

coordinate-aligned targets encourage compatible predictions across participant perspectives. Collective world dreaming provides training-only auxiliary supervision without additional inference latency. On NAVSIM v1, CoDreamer achieves 92.0 PDMS, setting a new state of the art among the compared methods using a ResNet-34 backbone. It also improves future scene prediction performance. These results highlight the value of participant-centered future prediction for learning planning representations in end-to-end autonomous driving.

AI USE STATEMENT

In this work, we used ChatGPT solely to improve the language and readability of the manuscript. All AI-assisted text was reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *Pattern Analysis and Machine Intelligence (PAMI)*, 2023.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. 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.

Ke Guo, Haochen Liu, Xiaojun Wu, Jia Pan, and Chen Lv. ipad: Iterative proposal-centric end-to-end autonomous driving. *arXiv preprint arXiv:2505.15111*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yufeng Hong, Xiaotian Zhou, Yingyan Li, Xiangpo Zhou, Lin Liu, Yadan Luo, Shaoqing Xu, Lei Yang, and Ziying Song. Drivefuture: Future-aware latent world models for autonomous driving, 2026. URL https://arxiv.org/abs/2605.09701.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *arXiv preprint arXiv:2503.07656*, 2025.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8350, October 2023.

Jungho Kim, Jiyong Oh, Seunghoon Yu, Hongjae Shin, Donghyuk Kwak, and Jun Won Choi. Safedrive: Fine-grained safety reasoning for end-to-end driving in a sparse world. *arXiv preprint arXiv:2602.18887*, 2026.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 29118–29134, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/47cf6de08432fb54ffd6b44fc1260413-Paper-Conference.pdf.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 42942–42959, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/6aa4967920e495e90aeeaa3acf18d019-Paper-Conference.pdf.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27137–27146, October 2025b.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, AnYasong, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-w0: World models amplify data scaling law in autonomous driving. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=plrGn3RdzN.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, BING WANG, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=JoXwhGbuMi.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, Yu-Gang Jiang, and Jose M. Alvarez. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation, 2024. URL https://arxiv.org/abs/2406.06978.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047, 2025. doi: 10.1109/CVPR52734.2025.01124.

Yuechen Luo, Fang Li, Qimao Chen, Shaoqing Xu, Jiaxin Liu, Ziying Song, Zhi-xin Yang, and Fuxi Wen. Unleashing vla potentials in autonomous driving via explicit learning from failures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24833–24842, June 2026.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Shuyao Shang, Bing Zhan, Yunfei Yan, Yuqi Wang, Yingyan Li, Yasong An, Xiaoman Wang, Jierui Liu, Lu Hou, Lue Fan, Zhaoxiang Zhang, and Tieniu Tan. Dynvla: Learning world dynamics for action reasoning in autonomous driving, 2026. URL https://arxiv.org/abs/2603.11041.

Zihao Sheng, Xin Ye, Jingru Luo, Sikai Chen, and Liu Ren. Explorevla: Dense world modeling and exploration for end-to-end autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2026.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8795–8801, 2025. doi: 10.1109/ICRA55743.2025.11128800.

Linhan Wang, Zichong Yang, Chen Bai, Guoxiang Zhang, Xiaotong Liu, Xiaoyin Zheng, Xiao-Xiao Long, Chang-Tien Lu, and Cheng Lu. Drive-jepa: Video jepa meets multimodal trajectory distillation for end-to-end driving. *arXiv preprint arXiv:2601.22032*, 2026.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang 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*, pp. 14749–14759, 2024.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15449–15458, 2024. doi: 10.1109/CVPR52733.2024.01463.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. GoalFlow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1602–1611, 2025.

Jiwei Yang, Zhengxian Chen, Chaosheng Huang, and Jun Li. Auto-jepa: A latent world model of continuous intent for end-to-end autonomous driving, 2026a. URL https://arxiv.org/abs/2607.29031.

Pengxuan Yang, Ben Lu, Zhongpu Xia, Chao Han, Yinfeng Gao, Teng Zhang, Kun Zhan, Xianpeng Lang, Yupeng Zheng, and Qichao Zhang. Worldrft: Latent world model planning with reinforcement fine-tuning for autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11649–11657, Mar. 2026b. doi: 10.1609/aaai.v40i14.38149. URL https://ojs.aaai.org/index.php/AAAI/article/view/38149.

Bozhou Zhang, Jingyu Li, Nan Song, and Li Zhang. Perception in plan: Coupled perception and planning for end-to-end autonomous driving, 2025a. URL https://arxiv.org/abs/2508.11488.

Bozhou Zhang, Nan Song, jingyu li, Xiatian Zhu, Jiankang Deng, and Li Zhang. Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 10204–10229. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0ee633a6ade45eab4276352b3ee79c7a-Paper-Conference.pdf.

Jinqing Zhang, Zehua Fu, zelinxu, wenying.dai, Qingjie Liu, and Yunhong Wang. Resworld: Temporal residual world model for end-to-end autonomous driving. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ptGmMFGWmk.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From forecasting to planning: Policy world model for collaborative state-action prediction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 134585–134611. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c3b3c5f9297ad0012cfe7188c34cea0e-Paper-Conference.pdf.

Zongchuang Zhao, Xin Zhou, Tianyang Xu, Zhengyang Sun, Kaixuan Zhou, Honglin Li, Dingkang Liang, and Xiang Bai. Simwam: A simple world action model for end-to-end autonomous driving. *arXiv preprint arXiv:2608.07468*, 2026.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, XianPeng Lang, and Dongbin Zhao. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642, October 2025.

Zhiyu Zheng, Shaoyu Chen, Haoran Yin, Xinbang Zhang, Jialv Zou, Xinggang Wang, Qian Zhang, and Lefei Zhang. Resad: Normalized residual trajectory modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3729–3739, 2026.