LEARNING INTERACTIVE COUNTERFACTUAL FUTURES FOR RISK-AWARE EMBODIED PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many existing embodied planning methods learn direct mappings from observations and instructions to actions through imitation of recorded demonstrations. Such mappings do not explicitly account for surrounding agents' responses to alternative ego actions, potentially leaving interaction risks unanticipated. We propose a framework that learns interactive counterfactual futures for risk-aware embodied planning. Given multimodal observations and language instructions, a vision-language backbone generates candidate ego trajectories. For each candidate, a learned interaction graph encodes inter-agent relationships, and a meta-action-guided response predictor models surrounding agents' speed and path behaviors before decoding them into future trajectories. The resulting candidate-specific future scenes enable explicit risk assessment and trajectory selection. Training proceeds in two stages: task module pretraining first learns interaction relationships from recorded trajectories; LoRA adaptation then aligns perceptual representations with the planning objective by supervising the risk-aware ego trajectory against the recorded ego future. On Bench2Drive, our method with a Qwen2-0.5B language backbone achieves a Driving Score of 87.47 and a Success Rate of 72.73%, exceeding the strongest compared baseline by 9.43 points and 17.64 percentage points, respectively. The full framework also improves aggregate planning scores over the variant without interaction analysis, with modest reductions in efficiency and comfort.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have enabled embodied agents to interpret multimodal observations and follow language instructions for planning Zeng et al. (2026); Yue et al. (2026); Gu et al. (2022); Zhang et al. (2026). Reliable decision-making in dynamic, safety-critical environments, however, requires anticipating the consequences of candidate actions Liu et al. (2026). Surrounding agents exhibit uncertain, evolving behaviors that depend on their interactions with the ego agent Zheng et al. (2026). Consequently, the safety of an ego trajectory depends not only on its geometric feasibility, but also on the responses it may elicit and the resulting evolution of the scene.

Existing approaches address this challenge through modular prediction and planning Mohanan & Salgoankar (2018), end-to-end learning Yang et al. (2025a), and world-model-based planning Hou et al. (2026). Modular frameworks separate prediction from decision-making, facilitating independent validation and explicit safety constraints. However, mismatched optimization objectives and the propagation of prediction errors can compromise downstream planning. End-to-end approaches enable joint optimization across perception and planning, but do not necessarily represent the consequences and interaction risks of individual candidates explicitly. World-model-based approaches incorporate predicted scene states or latent features into action generation and evaluation, bringing future information directly into decision-making. Figure 1 summarizes these approaches schematically. Our focus is on an explicit connection between each ego candidate, predicted surrounding-agent behavior, and the cost used to select a trajectory. This requires preserving candidate-specific responses through prediction and evaluation.

Counterfactual reasoning examines how consequences would differ under alternative actions in a given context. We adopt this perspective to compare future-scene hypotheses conditioned on different ego candidates before execution. In interactive environments, constructing these hypotheses

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.

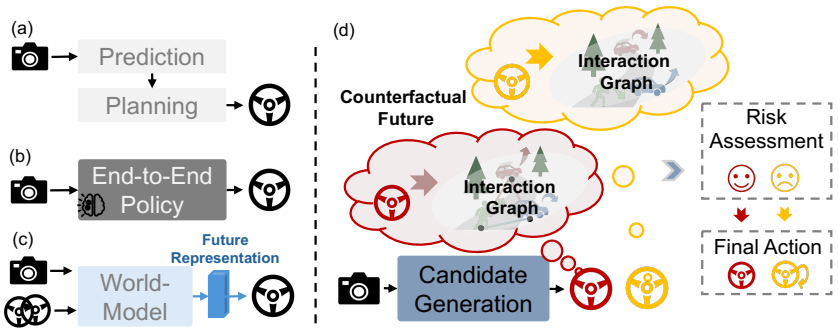


Figure 1: Schematic comparison of planning approaches: (a) modular prediction and planning; (b) an end-to-end policy; (c) world-model-based planning; and (d) proposed method, which evaluates candidate-conditioned future interactions for trajectory selection.

requires accounting for candidate-dependent relationships between the ego agent and surrounding agents, together with their possible behavioral responses. This motivates a candidate-conditioned interaction graph that represents these relationships and supports response prediction, providing an explicit connection between candidate actions, future scene evolution, and interaction risk.

To address this challenge, we propose Counterfactual-Agent, an interaction-aware counterfactual planning framework for safety-critical dynamic environments. We use counterfactual to denote future-scene hypotheses conditioned on alternative ego candidates, enabling their predicted consequences to be compared before execution. A VLM backbone encodes visual observations and language instructions into scene representations that capture surrounding-agent spatial information and environmental context and support candidate ego trajectory generation. For each candidate, Counterfactual-Agent constructs an interaction graph and predicts surrounding-agent responses to form a corresponding counterfactual future scene. The framework then evaluates each scene using a deterministic cost that combines interaction-related penalties with motion and progress terms, and selects the lowest-cost ego trajectory.

Our main contributions are:

- We propose Counterfactual-Agent, an interaction-aware counterfactual planning framework that evaluates candidate ego actions through their predicted interaction consequences. By constructing a corresponding future scene for each candidate, Counterfactual-Agent connects surrounding-agent response prediction with risk-based action selection.
- We develop a ego-conditioned behavior modeling mechanism that integrates interaction graphs with a discrete meta-action representation. This mechanism links ego-conditioned interaction modeling to continuous surrounding-agent trajectory prediction, enabling the construction of behavior-conditioned counterfactual futures.
- We evaluate Counterfactual-Agent on Bench2Drive through closed-loop comparisons and ablations, examining how candidate-conditioned response prediction and interaction modeling contribute to planning performance under a shared risk assessment framework.

2 Related Work

2.1 Safety-Critical Embodied Planning

Safety-critical embodied planning commonly quantifies collision hazards through temporal and geometric measures, such as time-to-collision (TTC) and minimum separation distances. These measures enter planning as objective penalties or safety constraints. Penalty-based formulations balance collision avoidance against progress and comfort, but depend on weight selection and do not inherently guarantee constraint satisfaction. Constraint-based approaches explicitly restrict admissible states and controls. Nguyen & Sreenath (2021) incorporate robust control barrier functions (CBFs) into quadratic programs to enforce safety under bounded model uncertainty. Xu et al. (2025) com-

bine reinforcement learning with CBF-based safety correction, while Chen et al. (2025) integrate diffusion-based trajectory generation with dynamic high-order CBFs. Although these formulations provide guarantees under specified assumptions, competing constraints and uncertainty bounds may introduce infeasibility or conservatism. Moreover, safety enforcement alone does not characterize how surrounding agents respond to alternative ego actions. Our work focuses on predicting these responses and assessing their implications for future interaction risk.

2.2 Future-Aware Planning

Future-aware planning represents anticipated scene evolution through generated views, bird's-eye-view (BEV) states, or latent features. DriveWorld Min et al. (2024), LAW Li et al. (2025a), and Drive-JEPA Wang et al. (2026) use temporal supervision to improve scene representations and trajectory learning. Other methods incorporate predicted futures directly into planning: Wang et al. (2024b) evaluates action-conditioned visual futures, Li et al. (2025b) uses a BEV world model for trajectory evaluation, Zheng et al. (2025) ranks trajectories through intention-conditioned latent futures, and Hong et al. (2026) conditions trajectory generation on predicted future features. These approaches connect future prediction with decision-making through representation learning, planner conditioning, or candidate evaluation Zhong et al. (2026). However, the utility of predicted features depends on whether they preserve the consequences relevant to each candidate. Sharing or aggregating future representations without retaining candidate-dependent responses can obscure differences in interaction risk. Furthermore, action-conditioned prediction does not by itself provide an explicit representation of inter-agent relationships. Counterfactual-Agent represents these relationships through candidate-conditioned interaction graphs and uses predicted agent responses to construct future scenes for risk-based trajectory planning.

2.3 Counterfactual Reasoning in Planning

Counterfactual reasoning evaluates how consequences would differ under alternative actions or conditions. Planning methods apply this perspective to supervision, action evaluation, and environmental modeling. The methods of Glossop et al. (2026); Hamid et al. (2026) use counterfactual labels to improve instruction following, while Zhou et al. (2026) learns interaction necessity and target selection from counterfactual navigation data. Zhu et al. (2026) employs temporal counterfactual preference learning to mitigate prediction–execution mismatch in asynchronous VLAs. Before execution, Tian et al. (2026a) evaluates alternative meta-actions through structured risk reasoning, whereas the methods of Peng et al. (2026); Tian et al. (2026b) revise potentially unsafe meta-actions to guide trajectory generation. Counterfactual Planning for Generalizable Agents' Actions further introduces a structural causal model to support adaptation to stochastic environment–action interactions Fu et al. (2026b). These methods improve instruction adherence, action refinement, or adaptation, but counterfactual supervision and reflection alone do not explicitly model the interactions underlying candidate-dependent outcomes. Our work focuses on how surrounding agents respond to each ego candidate and how those responses shape scene evolution and risk, linking counterfactual reasoning to explicit interaction modeling and trajectory selection.

3 Method

3.1 Problem Formulation

We consider an embodied agent navigating a dynamic environment shared with N surrounding agents. At time t, the agent receives a multimodal observation $o_t = (I_t, s_t, M_t)$ and a language instruction L_t, where I_t denotes visual inputs, s_t the ego state, and M_t available structured environmental information. The objective is to generate an ego trajectory over a horizon of H steps. We use index 0 for the ego agent and $i \in \{1, \dots, N\}$ for surrounding agents.

Our framework evaluates counterfactual futures by modeling how surrounding agents may respond to alternative ego trajectories, as shown in Figure 2. A vision-language backbone jointly encodes the current observations and a language instruction into a scene representation z_t. We use a predefined set of $K = 7$ ego meta-action candidates $\{m_0^k\}_{k=1}^K$ corresponding to distinct speed hypotheses. Conditioned on the scene representation z_t and each ego meta-action m_0^k, the ego planner generates

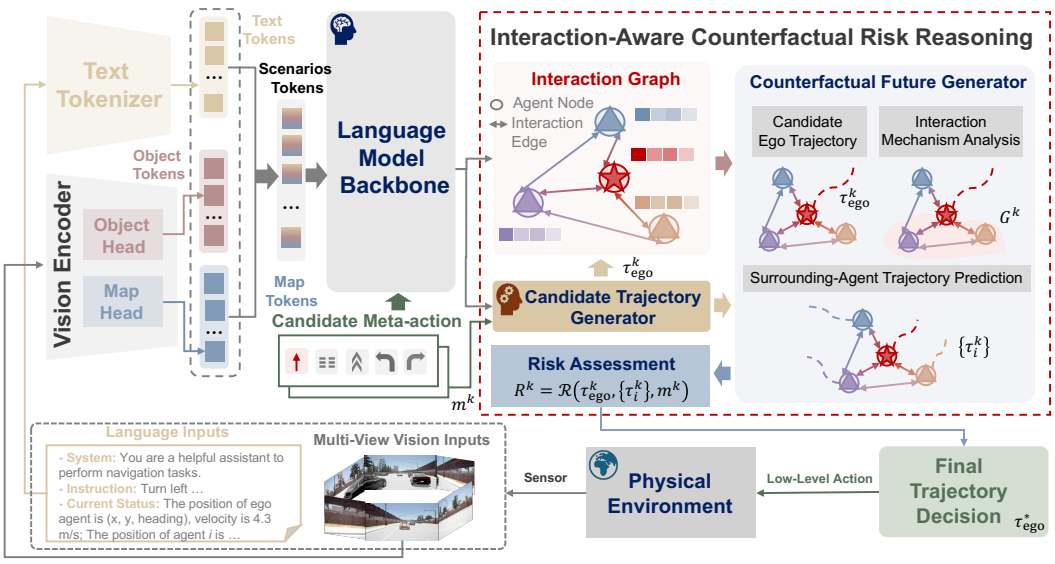


Figure 2: Overview of Counterfactual-Agent. Language and visual inputs are encoded and processed by a language model backbone. Candidate ego meta-actions m^k guide the generation of ego trajectories τ^k_{ego}. The counterfactual future generator combines candidate trajectories with interaction graph analysis to predict surrounding-agent responses $\{\tau_i^k\}$. Risk assessment scores the resulting futures to select the lowest-cost ego trajectory for execution.

a candidate trajectory τ_0^k. These ego meta-actions are predefined candidates rather than predicted surrounding-agent behaviors. For each candidate, we construct an interaction graph and predict surrounding agents' behaviors and future trajectories, forming a corresponding counterfactual future scene. These candidate-conditioned interaction representations connect proposed ego motions to anticipated agent responses, enabling risk-aware decision among the candidate trajectories.

3.2 Candidate-Conditioned Interaction Representations

The interaction relevance of a surrounding agent can change with the proposed ego trajectory. We capture this dependence through an attributed graph $\mathcal{G}_t^k = (\mathcal{V}_t^k, \mathcal{E}_t^k)$ for each candidate τ_0^k. Its nodes represent the ego agent and surrounding agents, with features

$$h_0^k = \Phi_{\text{ego}}(o_t, \tau_0^k), \qquad h_i^k = h_i = \Phi_{\text{agent}}(o_t^i), \quad i = 1, \ldots, N, \tag{1}$$

where o_t^i denotes the observed motion state of agent i. The surrounding-agent features are shared across candidates, while the ego feature explicitly encodes the proposed motion.

For each unordered pair, the relation module estimates

$$p_{ij}^k = \Phi_{\text{rel}}\left(h_i^k, h_j^k, r_{ij}^k\right), \qquad p_{ij}^k = p_{ji}^k, \tag{2}$$

where r_{ij}^k encodes relative spatial and motion information available from the current scene and the ego candidate. The predicted binary relations define the edge set $\mathcal{E}_t^k = \{\{i, j\} : \widehat{y}_{ij}^k = 1,\ 0 \le i < j \le N\}$. Edges represent interaction relevance rather than a directed causal influence.

A graph reasoning module propagates information among connected agents:

$$\{g_i^k\}_{i=0}^N = \Phi_{\text{int}}(z_t, \mathcal{G}_t^k). \tag{3}$$

The resulting features g_i^k provide candidate-conditioned context for response prediction. In particular, changes in the ego trajectory can alter ego–agent relations and the information propagated to surrounding-agent representations.

3.3 RISK-AWARE TRAJECTORY DECISION

Structured response prediction. We introduce a discrete meta-action $m_i^k = (m_i^{k,\mathrm{spd}}, m_i^{k,\mathrm{path}}) \in \mathcal{M}_{\mathrm{spd}} \times \mathcal{M}_{\mathrm{path}}$ to represent each surrounding agent's speed and path behavior. Unlike the predefined ego candidates m_0^k, these surrounding-agent meta-actions are predicted for each ego trajectory candidate before decoding the corresponding agent trajectories. This intermediate representation separates behavioral prediction from continuous trajectory decoding. Conditioned on the interaction features and ego candidate, the response policy and trajectory decoder produce

$$\begin{aligned} \widehat{m}_i^k &= \arg\max_{m \in \mathcal{M}} \pi_{\mathrm{resp}}\left(m \mid g_i^k, g_0^k, \tau_0^k\right), \\ \widetilde{\tau}_i^k &= \Psi_{\mathrm{traj}}\left(z_t, g_i^k, \widehat{m}_i^k, \tau_0^k\right), \qquad i = 1, \ldots, N, \end{aligned} \tag{4}$$

where $\mathcal{M} = \mathcal{M}_{\mathrm{spd}} \times \mathcal{M}_{\mathrm{path}}$ and $\widetilde{\tau}_i^k \in \mathbb{R}^{H \times d}$, with d denoting the position dimension. The same observed agent can therefore receive different behavior and trajectory predictions under different ego candidates.

Risk-Aware Trajectory Decision. Each ego candidate and its predicted surrounding-agent responses jointly define a counterfactual future scene $\widetilde{\mathcal{S}}_{t:t+H}^k = (\tau_0^k, \{\widetilde{\tau}_i^k\}_{i=1}^N)$. A deterministic planning cost $\mathcal{R}$ scores each scene using interaction-related penalties together with motion and progress terms. Its interaction component sums per-agent minimum-distance proximity penalties and scales the total by one plus the mean predicted relevance over valid agents. These per-agent relevance scores $r_i^{(k)}$ are obtained by applying a sigmoid to candidate-conditioned relevance logits from the interaction and response network; they are distinct in notation and role from the pairwise relation probabilities p_{ij}^k in Eq. 2. The candidate with the lowest total score is selected:

$$k^\star = \arg\min_{k \in \{1,\ldots,K\}} \mathcal{R}\left(\widetilde{\mathcal{S}}_{t:t+H}^k\right), \qquad \tau_0^\star = \tau_0^{k^\star}. \tag{5}$$

The scoring criteria and aggregation procedure are detailed in Appendix A.2. The selected trajectory is passed to the low-level controller for execution. The score ranks the available candidates under surrogate costs; it is not a calibrated collision probability or a safety guarantee.

3.4 TRAINING

We first learn the task modules from factual trajectories, then adapt the backbone using the planning error after risk-based decision.

Supervision from recorded episodes. At each sampled time index t, a training example contains (o_t, L_t) and synchronized future trajectories $\{\tau_i^{\mathrm{gt}}\}_{i=0}^N$ over H steps. We derive surrounding-agent meta-action labels m_i^{gt} from their recorded future motion. For path-conditioned ego planning, the instruction is paired with the corresponding recorded ego trajectory.

Pairwise interaction labels are constructed as

$$y_{ij} = \mathbb{I}\left[d_{ij}^{\min} \le \delta_d \ \vee \ \mathrm{TTC}_{ij}^{\min} \le \delta_{\mathrm{TTC}}\right], \tag{6}$$

where $d_{ij}^{\min} = \min_{1 \le h \le H} \|p_{i,t+h}^{\mathrm{gt}} - p_{j,t+h}^{\mathrm{gt}}\|_2$ and $\mathrm{TTC}_{ij}^{\min}$ is the minimum valid TTC over the same horizon. These labels supervise geometric interaction relevance in the recorded episode.

For interaction and response learning, τ_0^{gt} serves as the conditioning ego trajectory; superscript f denotes quantities computed under this factual condition. Surrounding-agent futures provide supervision only. At inference, generated ego candidates replace the factual ego trajectory, so predictions for unexecuted candidates rely on generalization from recorded interactions.

Stage I: Pretraining. We freeze the pretrained backbone and optimize the ego planner, interaction module, response policy, and trajectory decoder:

$$\mathcal{L}_{\mathrm{pre}} = \lambda_{\mathrm{ego}}\mathcal{L}_{\mathrm{ego}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{meta}}\mathcal{L}_{\mathrm{meta}} + \lambda_{\mathrm{traj}}\mathcal{L}_{\mathrm{traj}}, \tag{7}$$

where the λ coefficients balance the tasks. Ego and surrounding-agent trajectories are supervised by

$$\begin{aligned}\mathcal{L}_{\text{ego}} &= \frac{1}{H}\sum_{h=1}^{H}\left\|\widehat{p}_{0,t+h} - p^{\text{gt}}_{0,t+h}\right\|_2^2,\\ \mathcal{L}_{\text{traj}} &= \frac{1}{NH}\sum_{i=1}^{N}\sum_{h=1}^{H}\left\|\widetilde{p}^{\text{f}}_{i,t+h} - p^{\text{gt}}_{i,t+h}\right\|_2^2.\end{aligned} \tag{8}$$

During this stage, the surrounding-agent trajectory decoder receives the annotated meta-actions; inference uses predicted meta-actions instead. The relation and response predictors receive explicit classification supervision:

$$\begin{aligned}\mathcal{L}_{\text{rel}} &= \sum_{0\le i<j\le N}\text{BCE}(p^{\text{f}}_{ij}, y_{ij}),\\ \mathcal{L}_{\text{meta}} &= -\frac{1}{N}\sum_{i=1}^{N}\log\pi_{\text{resp}}\left(m^{\text{gt}}_i \mid g^{\text{f}}_i, g^{\text{f}}_0, \tau^{\text{gt}}_0\right),\end{aligned} \tag{9}$$

where BCE denotes binary cross-entropy. These losses train the discrete relation and behavior predictions without differentiating through their hard selection operations.

Stage II: Risk-Aware Adaptation. We introduce LoRA adapters into the backbone while keeping its original weights frozen. The trajectory selected by the risk-based decision provides the output for imitation supervision. We supervise this final planning output using the recorded ego future:

$$\mathcal{L}_{\text{LoRA}} = \frac{1}{H}\sum_{h=1}^{H}\left\|p^{\star}_{0,t+h} - p^{\text{gt}}_{0,t+h}\right\|_2^2. \tag{10}$$

The objective supervises the selected trajectory rather than directly minimizing the risk score. In the hard-selection formulation of Eq. 5, risk determines which candidate is supervised; the loss does not provide a differentiable objective through the discrete selection index. This distinction separates imitation of the selected output from end-to-end optimization of the risk-ranking operation.

4 Experiments

4.1 Experimental Setup

Benchmark and Data We evaluate our framework on Bench2Drive Jia et al. (2024), a CARLA-based benchmark, using autonomous driving as a testbed for embodied planning in dynamic multi-agent environments. We adopt the official base set of 1,000 clips, using 950 for training and 50 for open-loop validation. Interaction labels and surrounding-agent meta-action annotations are constructed from the recorded trajectories following Section 3.4. Closed-loop evaluation is conducted on the official Bench2Drive benchmark, comprising 220 short routes across 44 interactive scenarios.

Metrics We report open-loop average L2 displacement error over a 2-second horizon at 2 Hz as a reference for trajectory accuracy. Closed-loop evaluation provides the primary measure of planning performance, using Driving Score (DS), Success Rate (SR), and Multi-Ability. DS accounts for route completion and infraction penalties, while SR measures successful route completion. Multi-Ability covers Merging, Overtaking, Emergency Brake, Give Way, and Traffic Sign, with individual scores and their benchmark-defined mean reported. We additionally report time-to-collision (TTC) and minimum ego–agent distance to assess interaction safety.

Implementation Details We use Qwen2-0.5B Wang et al. (2024a) as the language backbone and EVA-02-L as the visual encoder. The ego planner generates one candidate ego trajectory for each of seven predefined ego meta-actions representing distinct speed hypotheses. The candidate vocabulary and its role in trajectory generation are detailed in Appendix A.1. The interaction module models up to ten dynamic surrounding agents, with each agent's response represented by seven speed and six path meta-action categories. Training follows the two-stage procedure described in Section 3.4. The risk-function terms and fixed coefficients are described in Appendix A.2. All experiments are conducted on 4 NVIDIA H200 GPUs, each with 144 GB of memory.

Table 1: Open-loop, closed-loop, and multi-ability results on Bench2Drive. Avg. L2 is averaged over predictions within 2 seconds at 2 Hz. DS: Driving Score; SR: Success Rate; M: Merging; O: Overtaking; EB: Emergency Brake; GW: Give Way; TS: Traffic Sign. * denotes expert feature distillation.

Method	**LLM**	**Reference**	**Open-loop**	**Closed-loop**		**Ability (%) ↑**					
			Avg. L2 ↓	DS ↑	SR (%) ↑	M	O	EB	GW	TS	**Mean**
Traditional End-to-End Paradigm											
AD-MLP	–	–	3.64	18.05	0.00	0.00	0.00	0.00	0.00	4.35	0.87
UniAD-Tiny	–	–	0.80	40.73	13.18	8.89	9.33	20.00	20.00	15.43	14.73
UniAD-Base	–	Hu et al. (2023)	**0.73**	45.81	16.36	14.10	17.78	21.67	10.00	14.21	15.55
VAD	–	Jiang et al. (2023)	0.91	42.35	15.00	8.11	24.44	18.64	20.00	19.15	18.07
TCP*	–	–	1.70	40.70	15.00	16.18	20.00	20.00	10.00	6.99	14.63
TCP-ctrl*	–	–	–	30.47	7.27	10.29	4.44	10.00	10.00	6.45	8.23
TCP-traj*	–	Wu et al. (2022)	1.70	59.90	30.00	8.89	24.29	51.67	40.00	46.28	34.22
TCP-traj w/o distillation	–	–	1.96	49.30	20.45	17.14	6.67	40.00	50.00	28.72	28.51
ThinkTwice*	–	Jia et al. (2023b)	0.95	62.44	31.23	27.38	18.42	35.82	50.00	54.23	37.17
DriveAdapter*	–	Jia et al. (2023a)	1.01	64.22	33.08	28.82	26.38	48.76	50.00	56.43	42.08
GenAD	–	Zheng et al. (2024)	–	44.81	15.90	–	–	–	–	–	–
MomAD	–	Song et al. (2025)	–	44.54	16.71	–	–	–	–	–	–
WoTE	–	Li et al. (2025b)	–	61.71	31.36	–	–	–	–	–	–
DriveTransformer-Large	–	Jia et al. (2025)	–	63.46	35.01	17.57	35.00	48.36	40.00	52.10	38.60
DiffAD	–	Wang et al. (2025)	–	67.92	38.64	30.00	35.55	46.66	40.00	46.32	38.79
DriveDPO	–	Shang et al. (2025)	–	62.02	30.62	–	–	–	–	–	–
Raw2Drive	–	Yang et al. (2025b)	–	71.36	50.24	43.35	51.11	60.00	50.00	62.26	53.34
Vision-Language-Action Paradigm											
Driveπ0	Paligemma-3B	Yang et al. (2026)	–	60.45	30.00	29.35	36.58	48.83	40.00	54.45	41.84
DriveMoE	Paligemma-3B	Yang et al. (2026)	–	74.22	48.64	34.67	40.00	65.45	40.00	59.44	47.91
ORION	Vicuna-7B	Fu et al. (2025)	–	77.74	54.62	25.00	71.11	**78.33**	30.00	69.15	54.72
ORION	Qwen2-0.5B†	Fu et al. (2025)	–	72.89	45.83	26.32	62.22	55.55	50.00	63.30	51.37
RecogDrive	Qwen2.5-7B	Li et al. (2026)	–	71.36	45.45	29.73	20.00	69.09	20.00	71.34	42.03
MindDrive-IL	Qwen2-0.5B	–	–	75.85	49.30	30.26	66.67	60.00	33.33	56.91	49.44
MindDrive	Qwen2-0.5B	Fu et al. (2026a)	–	78.04	55.09	32.89	75.56	68.33	50.00	57.89	56.94
Ours	Qwen2-0.5B	–	1.09	**87.47**	**72.73**	**61.25**	**97.78**	**78.33**	**80.00**	**92.63**	**82.00**

4.2 Comparison with Baselines

Table 1 compares our framework with traditional end-to-end and vision-language-action methods on Bench2Drive.

Closed-loop planning performance. Counterfactual-Agent achieves a Driving Score of 87.47 and a Success Rate of 72.73%, the highest values among the methods in Table 1. Relative to MindDrive, these correspond to gains of 9.43 points and 17.64 percentage points, respectively. Compared with MindDrive-IL, which also uses Qwen2-0.5B, the gains are 11.62 points and 23.43 percentage points. Counterfactual-Agent also exceeds ORION with Vicuna-7B on both metrics while using a smaller language backbone. These comparisons describe overall planning performance and do not isolate individual module contributions or establish a computational-efficiency advantage.

Multi-ability evaluation. Counterfactual-Agent achieves a mean ability score of 82.00%, exceeding MindDrive by 25.06 percentage points. It obtains the highest scores among the compared methods in Merging (61.25%), Overtaking (97.78%), Give Way (80.00%), and Traffic Sign (92.63%). Merging and Give Way exceed the strongest respective baselines by 17.90 and 30.00 percentage points. Emergency Brake reaches 78.33%, matching ORION with Vicuna-7B. The improvements therefore span four ability categories, with a tie on emergency braking rather than an improvement across all five categories.

Open-loop versus closed-loop performance. Our method obtains an average L2 error of 1.09, compared with the lowest reported error of 0.73 from UniAD-Base. Despite this higher open-loop error, our framework achieves substantially better closed-loop outcomes. The different rankings indicate that open-loop imitation accuracy and closed-loop planning performance capture different aspects of behavior. This comparison supports reporting both types of evaluation, but does not by itself identify the mechanism responsible for the closed-loop gains.

Table 2: Ablation study on Bench2Drive. Model 0 is MindDrive; models 1 and 2 use candidate-conditioned responses without and with interaction analysis, respectively. Missing values are denoted by –. DS: Driving Score; SR: Success Rate; Eff.: Efficiency (relative-speed percentage); Comf.: Comfort. M: Merging; O: Overtaking; EB: Emergency Brake; GW: Give Way; TS: Traffic Sign.

ID	Components		Closed-loop				Ability (%) ↑					
	Candidate-conditioned response	Interaction analysis	DS ↑	SR (%) ↑	Eff. (%) ↑	Comf. (%) ↑	M	O	EB	GW	TS	**Mean**
0	×	×	78.04	55.09	**154.89**	**12.61**	32.89	75.56	68.33	50.00	57.89	56.94
1	✓	×	86.26	70.45	140.60	10.94	56.25	95.56	76.67	**80.00**	85.26	78.75
2	✓	✓	**87.47**	**72.73**	139.15	10.51	**61.25**	**97.78**	**78.33**	**80.00**	**92.63**	**82.00**

Table 3: Safety indicators at safety-critical scenarios.

Method	**Mean**		**Critical Case**	
	TTC (s) ↑	Min. dist. (m) ↑	TTC (s) ↑	Min. dist. (m) ↑
MindDrive	0.932	2.60	0.80	0.30
Ours	**0.935**	**2.77**	**1.32**	**3.35**

4.3 Ablation Studies

Table 2 compares baseline (ID 0), candidate-conditioned response prediction without interaction analysis (ID 1), and the full framework (ID 2). Model 1 outperforms baseline by 8.22 DS points and 15.36 percentage points in SR, although this comparison does not isolate the effect of response prediction. Adding interaction analysis further improves DS by 1.21 points, SR by 2.28 percentage points, and mean ability by 3.25 percentage points, with slight decreases in efficiency and comfort. Overall, model 2 achieves the highest DS, SR, and mean ability, exceeding MindDrive by 9.43 points, 17.64 percentage points, and 25.06 percentage points, respectively.

Safety indicators. Table 3 compares Counterfactual-Agent with baseline. The mean minimum ego–agent distance increases from 2.60 to 2.77 m, while mean TTC is nearly unchanged (0.932 vs. 0.935 s). In the reported safety-critical case, TTC increases from 0.80 to 1.32 s and minimum distance from 0.30 to 3.35 m. The distance measurements indicate greater separation on average and in this case; the nearly unchanged mean TTC does not indicate a substantial average improvement in temporal safety margin. The case-specific measurements should not be interpreted as an aggregate result over all safety-critical scenarios.

4.4 Qualitative Results

Figure 3 visualizes a planning step in a nighttime traffic scene. The interaction representation displays surrounding-agent relevance, while the seven future hypotheses show candidate ego motions and associated predicted agent trajectories. The component cost profiles summarize how these hypotheses are evaluated. The selected trajectory corresponds to slowing down while maintaining the forward path and is shown in bird's-eye and front views. This visualization exposes the intermediate predictions and scores used in decision-making; it does not independently verify the accuracy of responses to unexecuted candidates. The component labeled "Rule" is the heuristic lateral-motion cost in Appendix A.2, rather than an explicit traffic-rule check.

Figure 4 compares two outcomes in a merging scenario. In (a), the highlighted vehicles approach each other between $t = 0.8$ s and $t = 1.1$ s, with a collision marked at $t = 1.25$ s. In (b), the blue-boxed vehicle remains behind the merging area during the corresponding interval, leaving space for the pink-boxed vehicle to proceed. By $t = 2.50$ s, both vehicles have progressed along the road with spatial separation. The sequence illustrates successful conflict avoidance by Counterfactual-Agent in this case and the importance of relative entry timing at a potential conflict region. It does not isolate the contribution of interaction modeling or establish a general safety improvement on its own.

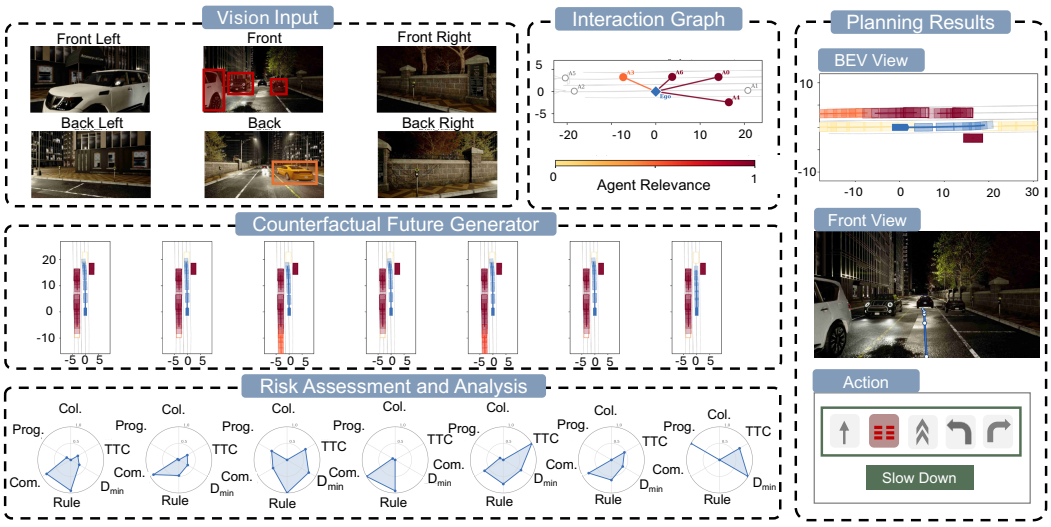


Figure 3: Planning visualization in a safety-critical scene: multi-view observations, interaction relevance, seven candidate-conditioned future predictions, component cost profiles, and the selected slow-down trajectory. The future trajectories are model predictions.



Figure 4: Temporal comparison in a merging scenario: (a) baseline, collision happened at $t = 1.25$ s; (b) Counterfactual-Agent, with no collision observed in the displayed sequence through $t = 2.50$ s.

5 Conclusion

We presented Counterfactual-Agent, an interaction-aware planning framework that predicts candidate-conditioned surrounding-agent responses to guide ego trajectory selection. On Bench2Drive, Counterfactual-Agent achieves a Driving Score of 87.47 and a Success Rate of 72.73%, with the highest aggregate planning and ability scores among the evaluated variants, at a modest cost to efficiency and comfort relative to model 1.

However, responses to unexecuted candidates remain unvalidated, response uncertainty is not explicitly aggregated, and surrogate planning costs provide no formal safety guarantee. Evaluation is limited to autonomous driving. Future work will validate candidate-dependent predictions, incorporate uncertainty, and explore broader embodied navigation tasks.

DISCLOSURE OF GENERATIVE AI USE

We used generative AI tools to assist with manuscript drafting and revision. For each section, the authors provided the intended structure, key points, and technical content, and used AI to organize and refine the language. We also used AI to review drafts and suggest revisions, which the authors evaluated and selectively incorporated. Additionally, AI assisted with literature summaries and reference formatting. The authors reviewed the AI-assisted content for accuracy and consistency with the research and take full responsibility for the final manuscript, including all AI-assisted text and claims.

REFERENCES

Di Chen, Ruiguo Zhong, Kehua Chen, Zhiwei Shang, Meixin Zhu, and Edward Chung. Dynamic high-order control barrier functions with diffuser for safety-critical trajectory planning at signal-free intersections. *IEEE Transactions on Intelligent Transportation Systems*, 26(9):14011–14024, 2025.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24823–24834, 2025.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Hongwei Xie, Bing Wang, Guang Chen, Hangjun Ye, Dingkang Liang, and Xiang Bai. Minddrive: A vision-language-action model for autonomous driving via online reinforcement learning. In *European Conference on Computer Vision*, pp. 503–522. Springer, 2026a.

Jiarun Fu, Lizhong Ding, Qiuning Wei, Yuhan Guo, Yurong Cheng, and Junyu Zhang. Counterfactual planning for generalizable agents' actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29432–29440, 2026b.

Catherine Glossop, William Chen, Arjun Bhorkar, Dhruv Shah, and Sergey Levine. Cast: Counterfactual labels improve instruction following in vision-language-action models. *IEEE Robotics and Automation Letters*, 2026.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7606–7623, 2022.

Kaiser Hamid, Can Cui, and Nade Liang. Icr-drive: Instruction counterfactual robustness for end-to-end language-driven autonomous driving. *arXiv preprint arXiv:2604.05378*, 2026.

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. *arXiv preprint arXiv:2605.09701*, 2026.

Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, et al. World model for robot learning: A comprehensive survey. *arXiv preprint arXiv:2605.00080*, 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7919–7929. IEEE, 2023a.

Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21983–21994. IEEE, 2023b.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified transformer for scalable end-to-end autonomous driving. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a7afc9957f1190223763b6ea93218f98-Abstract-Conference.html.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316. IEEE, 2023.

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, 2025a. URL https://arxiv.org/abs/2406.08481.

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*, pp. 27137–27146, 2025b.

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 *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JoXwhGbuMi.

Yushen Liu, Yin-Jen Chen, Ziyi Chen, Tao Wang, Heng Huang, Xugui Zhou, and Yanfu Zhang. Action-conditioned risk gating for safety-critical control under partial observability. *arXiv preprint arXiv:2605.14246*, 2026.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15522–15533. IEEE, 2024.

MG Mohanan and Ambuja Salgoankar. A survey of robotic motion planning in dynamic environments. *Robotics and Autonomous Systems*, 100:171–185, 2018.

Quan Nguyen and Koushil Sreenath. Robust safety-critical control for dynamic robotics. *IEEE Transactions on Automatic Control*, 67(3):1073–1088, 2021.

Zhenghao Peng, Wenhao Ding, Yurong You, Yuxiao Chen, Wenjie Luo, Thomas Tian, Yulong Cao, Apoorva Sharma, Danfei Xu, Boris Ivanovic, et al. Counterfactual vla: Self-reflective vision-language-action model with adaptive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4022–4031, 2026.

Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and Zhao-Xiang Zhang. DriveDPO: Policy learning via safety DPO for end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 38, pp. 81565–81585, 2025. doi: 10.52202/085713-2724.

Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22432–22441. IEEE, 2025.

Kefei Tian, Yuansheng Lian, Kai Yang, Xiangdong Chen, and Shen Li. C-cot: Counterfactual chain-of-thought with vision-language models for safe autonomous driving. *arXiv preprint arXiv:2605.10744*, 2026a.

Kefei Tian, Yuansheng Lian, Kai Yang, Xiangdong Chen, and Shen Li. Safealign-vla: A negative-enhanced safe alignment framework for risk-aware autonomous driving. *arXiv preprint arXiv:2605.19524*, 2026b.

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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Tao Wang, Cong Zhang, Xingguang Qu, Kun Li, Weiwei Liu, and Chang Huang. DiffAD: A unified diffusion modeling approach for autonomous driving. *arXiv preprint arXiv:2503.12170*, 2025. URL https://arxiv.org/abs/2503.12170.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14749–14759. IEEE, 2024b.

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022.

Zhenyu Xu, Ke Wang, Chaoxu Mu, and Tie Qiu. Safety-critical path planning for obstacle avoidance based on reinforcement learning and control barrier functions. *IEEE Internet of Things Journal*, 2025.

Pengxuan Yang, Yupeng Zheng, Qichao Zhang, Kefei Zhu, Zebin Xing, Qiao Lin, Yun-Fu Liu, Zhiguo Su, and Dongbin Zhao. Uncad: Towards safe end-to-end autonomous driving via online map uncertainty. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6408–6415. IEEE, 2025a.

Zhenjie Yang, Xiaosong Jia, Qifeng Li, Xue Yang, Maoqing Yao, and Junchi Yan. Raw2Drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in CARLA v2). In *Advances in Neural Information Processing Systems*, volume 38, pp. 134122–134147, 2025b. doi: 10.52202/085713-4471.

Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. DriveMoE: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10678–10688, 2026.

Lu Yue, Yue Fan, Shiwei Lian, Yu Zhao, Jiaxin Yu, Liang Xie, and Feitian Zhang. Spatial-vln: Zero-shot vision-and-language navigation with explicit spatial perception and exploration. *arXiv preprint arXiv:2601.12766*, 2026.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 33001–33026, 2026.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026.

Lei Zheng, Luyao Zhang, Peiqi Yu, Yifan Sun, Sergio Grammatico, Jun Ma, and Changliu Liu. Contingency planning for safety-critical autonomous vehicles: A review and perspectives. *arXiv preprint arXiv:2601.14880*, 2026.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 87–104. Springer, 2024.

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

Ruiguo Zhong, Benshan Ma, Xiaolong Chen, Lang Zhang, Mingyue Feng, Yaonong Wang, Pei Liu, and Jun Ma. Da-wam: Decision-aligned future latents for driving world models. *arXiv preprint arXiv:2608.19085*, 2026.

Kangjie Zhou, Zhejia Wen, Zhiyong Zhuo, Zike Yan, Pengying Wu, Shuaiyang Li, Han Gao, Kang Ding, Wenhan Cao, Wei Pan, et al. Coins: Counterfactual interactive navigation via skill-aware vlm. *arXiv preprint arXiv:2601.03956*, 2026.

Yixiang Zhu, Yonghao Chen, Zijie Yang, Yusong Hu, and Xinyu Chen. Deflect: Temporal counterfactual preference learning for delay-robust asynchronous vlas. *arXiv preprint arXiv:2605.19294*, 2026.

A Appendix

A.1 Predefined Ego Meta-Action Candidates

We construct seven ego trajectory candidates from a predefined vocabulary of speed meta-actions, listed in Table 4. The candidate index $k \in \{1, \ldots, 7\}$ identifies the ego meta-action m_0^k and its corresponding trajectory τ_0^k. The meta-actions express qualitative speed hypotheses; they are not themselves continuous trajectories or low-level control commands.

Table 4: The seven predefined ego meta-action candidates. Descriptions specify the intended qualitative speed behavior rather than a numerical speed or acceleration constraint.

k	Ego meta-action	Intended speed behavior
1	Maintain moderate speed	Maintain a moderate-speed motion profile.
2	Stop	Bring the ego vehicle to a stop or remain stopped.
3	Maintain slow speed	Maintain a low-speed motion profile.
4	Speed up	Increase the ego vehicle's speed.
5	Slow down	Reduce the ego vehicle's speed.
6	Maintain fast speed	Maintain a high-speed motion profile.
7	Slow down rapidly	Reduce speed more sharply than the slow-down hypothesis.

From meta-actions to candidate trajectories. For each predefined ego meta-action m_0^k, the ego planner generates a continuous trajectory τ_0^k conditioned on that meta-action and the scene representation described in Section 3.1. Thus, the seven candidates correspond to seven speed hypotheses, rather than seven different path instructions. The ego meta-action is a conditioning input to trajectory generation, not an additional predicted response variable.

Distinction from surrounding-agent meta-actions. For each generated ego trajectory τ_0^k, the response model predicts the surrounding-agent meta-actions $\widehat{m}_i^k$, including speed and path behaviors, and decodes the associated response trajectories as specified in Eq. 4. The predefined ego candidates m_0^k and the predicted surrounding-agent responses $\widehat{m}_i^k$ therefore have different roles: the former define the ego alternatives to evaluate, while the latter describe the model's anticipated responses to each alternative. The Risk-Aware Trajectory Decision module evaluates the resulting future scenes and returns a candidate ego trajectory according to Eq. 5.

A.2 Risk Assessment

We evaluate each candidate ego trajectory together with its associated predicted surrounding-agent trajectories using a deterministic scoring function. The score combines geometric proximity, a

time-to-collision proxy, motion smoothness, progress shortfall and heuristic lateral-motion penalties. These components are surrogate costs rather than calibrated collision probabilities.

Notation and inputs. Let $\boldsymbol{\tau}_0^{(k)} = (\mathbf{p}_{0,t}^{(k)})_{t=1}^{H}$ denote the k-th candidate ego trajectory, and let $\boldsymbol{\tau}_i^{(k)} = (\mathbf{p}_{i,t}^{(k)})_{t=1}^{H_a}$ denote the associated future trajectory of valid surrounding agent i. Positions are expressed in a shared ego-centric coordinate system, with $\mathbf{p} = (x, y)$, where x and y represent lateral and longitudinal positions, respectively. We denote the number of valid surrounding agents by N, their candidate-conditioned interaction relevance scores by $r_i^{(k)} \in [0, 1]$, and the reference ego trajectory by $\bar{\boldsymbol{\tau}}_0$. We use $[z]_+ = \max(z, 0)$.

At inference time, invalid agent predictions are removed before scoring. The relevance scores are obtained by applying a sigmoid to the interaction-response predictor's relevance logits. When candidate-conditioned agent responses are enabled, the scorer uses the resulting realized agent trajectories; otherwise, it uses the original agent forecasts. The reference ego trajectory is taken from the original planner output when available, and otherwise from the first candidate.

Geometric proximity costs. We first compute the minimum synchronized distance between the ego vehicle and each surrounding agent over the common prediction horizon $T = \min(H, H_a)$:

$$d_i^{(k)} = \min_{1 \le t \le T} \left\| \mathbf{p}_{i,t}^{(k)} - \mathbf{p}_{0,t}^{(k)} \right\|_2 .$$

The collision-proximity cost is

$$C_{\text{col}}^{(k)} = \sum_{i=1}^{N} \left[d_{\text{col}} - d_i^{(k)} \right]_+ , \qquad d_{\text{col}} = 2.0.$$

This cost uses trajectory-point distances rather than explicit vehicle-footprint overlap.

A broader proximity cost is defined using $d_{\text{near}} = 5.0$:

$$C_{\text{near}}^{(k)} = \sum_{i=1}^{N} \left[d_{\text{near}} - d_i^{(k)} \right]_+ .$$

The interaction cost scales this aggregate by the mean relevance:

$$\bar{r}^{(k)} = \frac{1}{N} \sum_{i=1}^{N} r_i^{(k)}, \qquad C_{\text{int}}^{(k)} = \left(1 + \bar{r}^{(k)}\right) C_{\text{near}}^{(k)}.$$

The minimum is taken over synchronized future positions separately for each agent, after which the penalties are summed over valid agents; the cost does not accumulate penalties over time. Relevance modulates the overall proximity cost rather than weighting each agent's distance penalty individually. When relevance is unavailable, $C_{\text{int}}^{(k)} = C_{\text{near}}^{(k)}$. When no valid surrounding agent is present ($N = 0$), the interaction cost is defined as zero and the mean relevance is not evaluated. For $N > 0$, zero relevance leaves the proximity penalty unchanged rather than removing it. These proximity costs are summed over agents without normalization by the agent count. Because the distances are computed between trajectory positions rather than vehicle footprints, this term is a geometric proximity surrogate, not a collision probability.

Time-to-collision proxy. For $T \ge 2$, we compute the mean relative displacement magnitude between adjacent prediction steps:

$$u_i^{(k)} = \max\left(10^{-3}, \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \left(\mathbf{p}_{0,t+1}^{(k)} - \mathbf{p}_{0,t}^{(k)} \right) - \left(\mathbf{p}_{i,t+1}^{(k)} - \mathbf{p}_{i,t}^{(k)} \right) \right\|_2 \right).$$

We then define

$$q_i^{(k)} = \min\left(\frac{d_i^{(k)}}{u_i^{(k)}}, 10 \right), \qquad C_{\text{ttc}}^{(k)} = \sum_{i=1}^{N} \left[3 - q_i^{(k)} \right]_+ .$$

For $T = 1$, the implementation sets $u_i^{(k)} = 1$.

This quantity is a heuristic TTC proxy: relative displacements are not divided by the sampling interval, and the denominator does not isolate the closing-speed component. Consequently, $q_i^{(k)}$ is not a physical TTC measured in seconds. The scorer's stored time-step parameter does not enter this calculation.

Motion-smoothness cost. We define ego displacements and their first differences as

$$\Delta \mathbf{p}_t^{(k)} = \mathbf{p}_{0,t}^{(k)} - \mathbf{p}_{0,t-1}^{(k)}, \qquad \mathbf{a}_t^{(k)} = \Delta \mathbf{p}_t^{(k)} - \Delta \mathbf{p}_{t-1}^{(k)},$$

with the initialization

$$\mathbf{p}_{0,0}^{(k)} = \mathbf{0}, \qquad \Delta \mathbf{p}_0^{(k)} = \mathbf{0}.$$

The comfort cost is

$$C_{\text{comf}}^{(k)} = \frac{1}{H} \sum_{t=1}^{H} \|\mathbf{a}_t^{(k)}\|_2.$$

This term penalizes second differences of the trajectory positions. It is not normalized by the squared sampling interval and therefore is not an acceleration cost in physical units. Its initial term also reflects the zero-displacement initialization.

Progress shortfall. Let $\bar{y}_{\text{end}}$ be the longitudinal coordinate of the reference trajectory's final point. We penalize insufficient terminal forward displacement:

$$C_{\text{prog}}^{(k)} = \left[[\bar{y}_{\text{end}}]_+ - [y_H^{(k)}]_+ \right]_+ .$$

This is a progress-shortfall cost: matching or exceeding the reference's terminal longitudinal displacement incurs zero penalty. It does not measure path length or progress along a road centerline. If the scorer is called without a reference trajectory, it uses the candidate's detached terminal longitudinal position as the target, yielding a zero progress cost.

Heuristic lateral-motion cost. Define

$$L_{\max}^{(k)} = \max_{1 \le t \le H} |x_t^{(k)}|, \qquad L_{\text{end}}^{(k)} = |x_H^{(k)}|, \qquad S^{(k)} = [5 - [y_H^{(k)}]_+]_+.$$

The implementation's map-rule term is

$$C_{\text{map}}^{(k)} = 3[L_{\max}^{(k)} - 3.7]_+ + 0.3 L_{\text{end}}^{(k)} + 0.2 S^{(k)} L_{\text{end}}^{(k)}.$$

The three terms penalize excessive lateral excursion, residual terminal lateral displacement, and terminal lateral displacement combined with limited forward progress, respectively. Although the scorer accepts a map-context argument, this expression does not use map geometry. It therefore represents a coordinate-based heuristic rather than an explicit check of lane boundaries, drivable areas, or traffic rules.

Aggregate score and decision. For scenes containing at least one valid surrounding agent, the final score is

$$\begin{aligned} J^{(k)} = {} & 8 C_{\text{col}}^{(k)} + C_{\text{ttc}}^{(k)} + 0.2 C_{\text{int}}^{(k)} + 0.2 C_{\text{comf}}^{(k)} \\ & + 2 C_{\text{prog}}^{(k)} + C_{\text{map}}^{(k)}. \end{aligned}$$

These coefficients are fixed constants in the implementation. The candidate with the smallest score is selected:

$$k^\star = \arg\min_k J^{(k)}, \qquad \boldsymbol{\tau}_0^\star = \boldsymbol{\tau}_0^{(k^\star)}.$$

The returned per-component scores are unweighted; the coefficients above are applied only when forming the aggregate score.

The scoring procedure uses candidate trajectories, predicted surrounding-agent futures, predicted interaction relevance, and the reference ego plan. It does not require ground-truth future trajectories at inference time. Its role is to rank the available candidates under these surrogate costs; it does not establish a formal safety guarantee or identify causal effects.