Seeing Beyond the Next Step: World-Model-Guided Human-Like Navigation in Multi-Agent Scenes

Seong-Eun Hong[1*] **JuYeong Hwang**[1*] **Jinhyun Kim**[1]
Hanyoung Jang[2] **HyeongYeop Kang**[1†]
[1]Korea University [2]NC AI

Abstract

Selecting a next motion in a populated scene is not just a matter of avoiding what is currently in the way: an action can be locally feasible yet drift onto the road instead of the crosswalk, cut into a stream of crossing pedestrians, or bottleneck a narrow passage. These failure modes are invisible from the current state alone, and acting on them requires reasoning about what each candidate action will cause. World models offer exactly this, namely action-conditioned predictions of the future, but rolling them out at every decision step is a cost deployable navigation systems cannot pay, and the field has largely traded foresight for speed. We argue this is a false dichotomy. The natural place for a world model in human-aware navigation is at training time, as a preference oracle that distills foresight into a fast policy, rather than at inference time, as an online planner. We pair a discrete visual action prior with two task-aligned world models, one for nearby social dynamics and one for future scene semantics, that score candidate actions by their predicted consequences and refine the policy through group-relative preference distillation. At deployment, both world models are removed and the policy acts in a single forward pass. Across five held-out evaluation scenarios, this yields navigation that is simultaneously safer, more socially and group-coherent, and more goal-directed than strong reactive and learned baselines, with naive observers rating the resulting motion as more human-like. Taken together, our findings point to a different approach for bringing world-model foresight into deployable agents: not accelerating rollouts, but amortizing them into the policy so that the cost of looking ahead is confined to training. Project page: `https://seoneun.github.io/-NeurIPS-2026-Seeing-Beyond-the-Next-Step/`.

1 Introduction

An agent can clear every person, vehicle, and obstacle around it at the current moment and still, two seconds later, end up cutting through a group already moving as one, drifting onto a stretch of road a turning car is about to reach, or blocking a doorway someone is about to step out of — each of these is locally feasible, geometrically valid, and consistent with a reasonable demonstration. None of them is good behavior. The shared property of these failure modes is that they are invisible from the current state alone: they are determined by what the chosen action will cause, not by what it currently avoids.

Acting well in such scenes therefore requires foresight, anticipating how nearby people, dynamic objects, and the navigable layout will evolve under each candidate motion. But foresight is expensive, and the field has mostly traded it away for speed. Geometric reactive controllers decide from

*Equal contribution. †Corresponding author.
Emails: `seong_eun@korea.ac.kr` `05judy02@korea.ac.kr` `rlawlsgus0623@korea.ac.kr` `ncdrjang@ncsoft.com` `siamiz_hkang@korea.ac.kr`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

instantaneous positions and velocities [18, 41], and learned social and reinforcement-learning policies fold longer-horizon trade-offs into a value function but still act on the current state at decision time [29, 11, 30]. Vision-based steering methods and recent vision-language-action (VLA) models extend this paradigm to richer visual and semantic perception [12, 7, 48, 23], but they invest that capacity in reading the present scene more carefully, not in asking what each candidate action would actually produce. World models can ask exactly that question through action-conditioned rollouts [16, 17], but they have largely been built as monolithic generators of photorealistic or latent-state futures. Rolling one out at every step, over a vocabulary of roughly a hundred plausible motions, is a cost the live decision loops of crowd simulation, embodied interactive agents, and on-robot navigation cannot afford. As a result, world models remain largely unexplored in human-aware navigation: the field has effectively been forced to choose between foresight and deployability, and with it, human-like behavior — by which we mean motion that is perceptually preferred over baselines by naive observers and aligns with demonstrated human trajectories along measurable axes.

We argue that this is a false dichotomy, and that the natural place for a world model in human-aware navigation is at training time, as a preference oracle, rather than at inference time, as an online planner. We instantiate this view in a human-aware navigation policy through three deliberate design choices. First, we train a VLA-style visual action prior over a discrete vocabulary of short-horizon motion anchors, so that candidate actions already lie on a behaviorally plausible manifold shaped by demonstration. Second, we replace the usual monolithic, photorealistic world model with two task-aligned predictors. The social branch predicts relative position and velocity with a direction term capturing who is heading toward whom. The scene branch predicts only the navigation-relevant slice of the future scene, with pedestrians excluded so the two predictors specialize cleanly on motion and layout. Third, we keep both world models entirely on the training side: their predictions are converted into a compact preference signal and distilled into the policy via Group Relative Policy Optimization (GRPO). The resulting policy inherits the look-ahead reasoning of the world models without paying their cost — at deployment, it selects an anchor with a single forward pass, and the world models are not needed.

Across five held-out evaluation scenarios, this design outperforms all baselines in goal success and social-region violation on every scenario, and achieves the lowest or tied-lowest group distance on four of the five, with violation rates reduced by up to an order of magnitude without sacrificing goal completion. A user study with naive observers further confirms that our policy's rollouts are perceived as significantly more human-like than those of all three baselines, showing that world-model-based foresight can be brought into human-aware navigation and yield human-like behavior without sacrificing real-time deployability.

Our contributions are:

- **World models as training-time preference oracles for human-aware navigation.** We frame action selection as ranking locally plausible candidates by their predicted consequences, and realize this through a visual action prior refined by GRPO-based preference distillation, yielding a policy that acts in a single forward pass with no world-model rollout at inference.
- **Dual task-aligned world models, not a monolithic generator.** Our key insight is that future prediction for human-aware navigation decomposes naturally along two axes, social dynamics and scene layout, and we instantiate each branch with a navigation-aligned target: position-velocity with a heading-aware loss, and pedestrian-excluded scene semantics. Ablations confirm that the two branches catch complementary failure modes.
- **Train-time foresight, test-time efficiency, and human-like rollouts.** With world models used only as training-time preference oracles, the resulting policy outperforms all baselines in goal success and social-region violation on all five held-out evaluation scenarios, achieves the lowest ADE against demonstrated human trajectories, and is judged significantly more human-like by naive observers.

2 Related Work

Human-like navigation and crowd simulation. Human-like navigation in this work is treated as a crowd-behavior problem: an agent must move in a way that is safe, socially plausible, and consistent with the local pedestrian flow. Human locomotion is anticipatory rather than purely reactive, and crowd self-organization often emerges from mutual prediction among nearby people [42, 19, 28].

This motivates our central question: when several motions are locally feasible, which one will remain human-like after the scene evolves?

Classical crowd models and vision-based steering methods provide efficient local interaction rules from current geometry or visual evidence [18, 41, 12]. Data-driven and learning-based crowd simulators further improve realism by learning from demonstrations, reference trajectories, or reinforcement-learning objectives [25, 24, 10, 47, 29, 11, 30]. These works are the closest baseline family for our experiments because they target plausible crowd behavior. However, most of them either react to the current state or encode future effects implicitly inside a policy or reward. Our work instead makes the future consequences of candidate motions explicit during training. Other approaches plan with model predictive control over the outputs of trajectory forecasters [33] or use conformal prediction to account for forecast uncertainty [45].

VLA models and anchor-based action priors. Recent vision-language-action models show that strong visual-semantic backbones can be adapted into policies that map rich observations to actions [7, 48, 23]. We use this idea as a visual action prior for crowd navigation, not as an end-to-end solution. In crowded scenes, the difficulty is often not proposing one plausible action, but choosing among multiple plausible actions whose downstream effects differ.

To make this comparison tractable, we represent short-horizon motion with trajectory anchors. Anchor-based prediction has been effective for multimodal future modeling because a compact set of trajectories can approximate diverse local motion modes [9, 32]. Low-rank and anchor-based descriptors likewise approximate pedestrian trajectories compactly [2, 4, 38]. Pedestrian forecasting work similarly shows that crowd motion is multimodal and shaped by social and scene context [1, 15, 34, 35, 46, 3]. Our anchors differ from forecasting outputs: they are executable candidate motions for the controlled agent. This design gives the policy a human-like action manifold from demonstrations, while leaving the final choice to future-aware preference learning.

World models for consequence-aware policy refinement. World models are attractive because they allow actions to be evaluated by their predicted consequences rather than only by the current observation. In model-based learning, latent dynamics models such as PlaNet and Dreamer use imagined rollouts for planning or policy optimization [16, 17]. More recent visual world models scale this idea toward interactive simulation, controllable visual generation, driving futures, navigation rollouts, and game-like environments [44, 8, 20, 13, 6, 14, 39, 40, 26]. Their main goal is usually to build a simulator, planner, data generator, or future evaluator.

Our contribution is different in the role, representation, and deployment use of the world model. First, the world models are used only during training as a preference oracle; they are removed at inference, so the final policy selects an anchor with a single forward pass. Second, we avoid a monolithic photorealistic future generator. Instead, we use two task-aligned predictors: a State Transition World Model for future relative position–velocity dynamics, and a Semantic Transition World Model for future navigation-oriented segmentation. This follows the idea that decision-relevant semantic futures can be more useful than raw RGB prediction for downstream action selection [27]. Third, the world models do not directly control the agent. They score candidate anchors by goal progress, smoothness, interaction risk, and semantic validity, and these relative preferences are distilled into the VLA-style anchor policy through GRPO [37].

3 Method

Our pipeline learns a deployable policy in two stages, as shown in Figure 1. We first train a *visual action prior* that imitates demonstrated motion over a discrete vocabulary of human-like motion anchors. We then refine this prior with two navigation-aligned world models that score each candidate anchor by its predicted social and semantic consequences, and distill the resulting preferences back into the policy via Group Relative Policy Optimization. Crucially, the world models are used only during refinement: at deployment, the refined policy selects an anchor with a single forward pass, and no world-model rollout is performed.

We use the following notation throughout. At simulation step t, the controlled agent receives the observation

$$o_t = \{I_t, g_t, v_t, v_{t-1}, \mathcal{N}_t, \mathcal{D}_t\}, \tag{1}$$

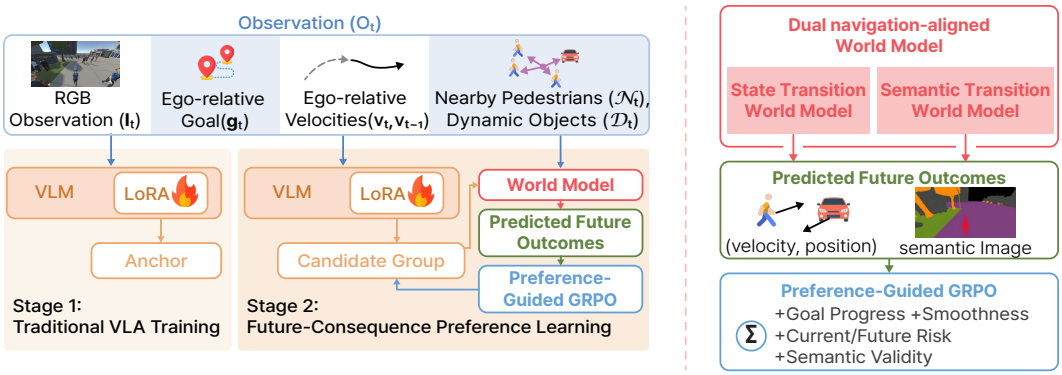


Figure 1: **Overview of the proposed training pipeline.** In Stage 1, we adapt a pretrained VLM into a VLA policy through supervised learning, grounding agent-centric visual observations and structured navigation context to executable motion-skill actions. In Stage 2, we further refine the policy with dual world models, which teaches the agent action–outcome knowledge by comparing plausible candidate actions according to their predicted social, semantic, and goal-directed consequences.

where I_t is the RGB observation, $g_t \in \mathbb{R}^2$ is the ego-relative goal, $v_t, v_{t-1} \in \mathbb{R}^2$ are the current and previous ego velocities, and $\mathcal{N}_t$, $\mathcal{D}_t$ are the sets of nearby pedestrians and dynamic objects. Each surrounding entity j stores an ego-frame relative position $p_{t,j} \in \mathbb{R}^2$ and relative velocity $u_{t,j} \in \mathbb{R}^2$. The policy is a categorical distribution $\pi_\theta(a_k \mid o_t)$ over an anchor vocabulary $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$, where each anchor $a_k = \{y_{k,0}, y_{k,1}, \ldots, y_{k,L}\}$ is a short ego-relative trajectory and $y_{k,\ell} \in \mathbb{R}^2$ is the displacement at anchor step ℓ.

3.1 Action Prior over Motion Anchors

We start from a discrete vocabulary because, over a short horizon, the space of plausible human motion is well approximated by a manageable set of motion modes. This turns action selection into classification over a finite, behaviorally meaningful vocabulary, which (i) anchors candidate actions on a human-like manifold from the outset, and (ii) makes consequence-based comparison tractable.

Anchor construction. We build $\mathcal{A}$ offline from training trajectories, recorded at 25 FPS. Each demonstration is cut into 20-frame clips with a 10-frame stride, yielding ego-relative displacement sequences $x_n = [x_{n,1}, y_{n,1}, \ldots, x_{n,20}, y_{n,20}] \in \mathbb{R}^{40}$. We run K-means with $K = 100$ in this 40-dimensional space; each cluster center is reshaped into 20 two-dimensional points to form one anchor. We discuss the choice of K and show nearest-anchor reconstructions in Appendix D.

Supervised training. The first stage trains π_θ as an anchor classifier. We construct a supervised dataset $\mathcal{D}_{\mathrm{SFT}}$, where each observation o_t is paired with the anchor $a_t^\star \in \mathcal{A}$ closest to the demonstrated future trajectory at time t. We minimize

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(o_t, a_t^\star) \sim \mathcal{D}_{\mathrm{SFT}}} \left[\log \pi_\theta(a_t^\star \mid o_t) \right]. \tag{2}$$

We instantiate π_θ with a pretrained multimodal backbone and adapt it with LoRA [21]. This yields a strong human-like imitation prior, but imitation alone cannot distinguish locally plausible actions with different downstream consequences.

3.2 Navigation-Aligned World Models

Refining the action prior requires asking, for each candidate anchor, *what would happen if it were executed?* A natural choice would be to train a single generative world model that simulates the next observation in pixel space. We deliberately avoid this. In populated navigation, the quantities that determine whether an action is acceptable are not the visual texture of the future, but two distinct things: how nearby agents will *move*, and how the navigable scene will be *occupied*. We therefore decompose future prediction along these two axes — social dynamics and scene layout — and train a specialized predictor for each. We use the term *world model* in the broad sense of an action-conditioned future predictor [6], since each predictor is conditioned on a full anchor (an L-step

sub-trajectory) and serves as a preference oracle rather than a simulator; a single anchor-conditioned prediction is sufficient for ranking candidates, and recursive frame-level rollout is unnecessary.

Each predictor takes as input a short visual context $h_t = (I_{t-3}, I_{t-2}, I_{t-1}, I_t)$ and the candidate anchor a_k; the input is RGB so the model can read the current scene fully, but the output is restricted to a navigation-aligned target. Both predictors are pretrained, then frozen during policy refinement.

State Transition World Model. The State Transition World Model f_{state} predicts the next ego-frame relative state of each surrounding entity:

$$\Delta\hat{p}_{t,j}(a_k), \hat{u}_{t+1,j}(a_k) = f_{\text{state}}(h_t, a_k, p_{t,j}, u_{t,j}),$$

with $\hat{p}_{t+1,j}(a_k) = p_{t,j} + \Delta\hat{p}_{t,j}(a_k)$. During pretraining, the targets are $\Delta p_{t,j} = p_{t+1,j} - p_{t,j}$ and $u_{t+1,j}$, and we minimize

$$\mathcal{L}_{\text{dyn}} = \mathbb{E}_{t,j}\Big[\|\Delta\hat{p}_{t,j} - \Delta p_{t,j}\|_2^2 + \lambda_u\|\hat{u}_{t+1,j} - u_{t+1,j}\|_2^2 + \lambda_{\text{dir}}\big(1 - \cos(\hat{u}_{t+1,j}, u_{t+1,j})\big)\Big]. \quad (3)$$

We add an explicit direction term because anticipating *who is heading toward whom* is what makes this branch useful for ranking interaction risk. Hyperparameter values are reported in Appendix A.

Semantic Transition World Model. The Semantic Transition World Model f_{sem} predicts a future segmentation map under candidate anchor a_k:

$$\hat{S}_{t+1}(a_k) = f_{\text{sem}}(h_t, a_k, \mathcal{D}_t). \quad (4)$$

Two design choices distinguish this branch from a generic visual world model.

First, we predict segmentation, not RGB. The score function in Eq. (6) only needs to know which navigational classes the agent is about to occupy — we use the navigation-relevant set {sidewalk, crosswalk, road, vehicle, building, obstacle, grass}, defined in Appendix C. Predicting RGB would spend capacity on texture that the policy never reads. Predicting segmentation aligns the model's output with the quantity that actually drives semantic acceptability.

Second, we deliberately remove pedestrians from the segmentation target. Pedestrians act as transient occluders rather than stable scene structure, and their motion is already modeled by f_{state}. We therefore let pedestrians appear in the input image but assign their pixels an ignore label in the target, so f_{sem} learns to predict the underlying environmental layout — including regions temporarily occluded by people — rather than fleeting agent occupancy. This separation means that the two world models do not duplicate signal: f_{state} owns who moves where, f_{sem} owns what the scene affords.

We instantiate f_{sem} as an EDM-style conditional denoiser [22]. Given the clean target S_{t+1}, sampled noise level $\sigma > 0$, Gaussian noise ϵ_{sem}, and noisy input $\tilde{S}_{t+1} = S_{t+1} + \sigma\epsilon_{\text{sem}}$, the denoising network D_ψ produces $\hat{S}^\sigma = D_\psi(\tilde{S}_{t+1}, \sigma, c_t(a_k))$, where $c_t(a_k)$ are the conditioning tokens built from h_t, a_k, and $\mathcal{D}_t$. We train with

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_{\sigma,\epsilon_{\text{sem}}}\left[w(\sigma)\left\|\hat{S}^\sigma - S_{t+1}\right\|_2^2 + \lambda_{\text{LPIPS}}\,\text{LPIPS}\left(\hat{S}^\sigma, S_{t+1}\right)\right], \quad (5)$$

where $w(\sigma)$ is the EDM noise-dependent weight and λ_{LPIPS} is a coefficient balancing the perceptual term against the reconstruction term.

3.3 Consequence-Based Preference Refinement

The two frozen world models give us, for each candidate anchor, a prediction of what would happen next. We turn these predictions into a preference signal and distill them into the policy. Crucially, this distillation is the only place in the pipeline where the world models are used; once refinement is complete, they are discarded.

Candidate group. For each observation o_t, we form a group of ten candidates. Five are the highest-probability anchors under the current policy (exploitation of the supervised prior), and five are sampled uniformly at random from the rest of the vocabulary (exploration beyond the prior's preference). This avoids comparing only near-identical high-probability actions, which would make the refinement signal degenerate.

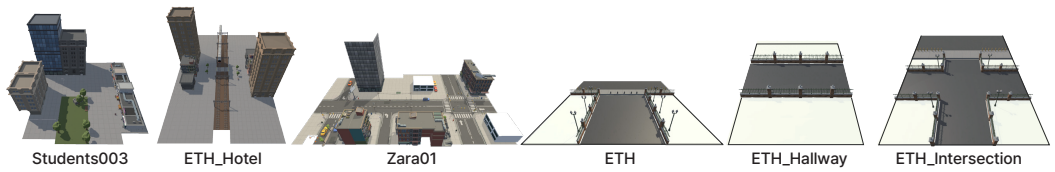


Figure 2: **Overview of the training and evaluation scenes.** We use six scenes in our experiments: Students003, ETH_Hotel, Zara01, ETH, ETH_Hallway, and ETH_Intersection.

Consequence score. Each candidate $a_t^{(i)}$ receives a scalar score

$$r_t^{(i)} = G_t^{(i)} - C_t^{(i)} - P_{\mathrm{cur},t}^{(i)} - P_{\mathrm{fut},t}^{(i)} + M_t^{(i)}. \tag{6}$$

Here $G_t^{(i)}$ measures goal progress, $C_t^{(i)}$ penalizes abrupt changes in motion, $P_{\mathrm{cur},t}^{(i)}$ measures immediate interaction risk, $P_{\mathrm{fut},t}^{(i)}$ measures predicted future interaction risk using the State Transition World Model, and $M_t^{(i)}$ measures semantic validity using the Semantic Transition World Model. All component weights are set to one and all components are scale-normalized; full formulas are in Appendix B.

Preference distillation. Within each group, we update π_θ with GRPO, which prefers candidates whose score exceeds the group mean. This treats the dual world models as a *training-time preference oracle*: rather than training the policy to imitate world-model rollouts or to plan against them online, we train it to *prefer* the actions that the world models judge to lead to better consequences.

Training stages and deployment. Training proceeds in three parts. First, the action prior π_θ is trained with $\mathcal{L}_{\mathrm{SFT}}$. Second, f_{state} and f_{sem} are pretrained with $\mathcal{L}_{\mathrm{dyn}}$ and $\mathcal{L}_{\mathrm{sem}}$. Third, the supervised policy checkpoint is refined with GRPO using the score in Eq. (6), while both world models remain frozen. At inference, both world models, the candidate-group construction, and the score function are discarded. Given o_t, the deployed policy returns a categorical distribution over $\mathcal{A}$ in a single forward pass; the selected anchor is executed directly. The look-ahead reasoning learned from the world models is amortized into π_θ's parameters, so its inference cost is independent of the world models' size, denoising depth, or rollout horizon.

4 Experiments

We compare against three navigation baselines spanning the major paradigms: GBM [12], a synthetic-vision gradient-based steering model; CCP [29], a configurable RL profile model; and CEDRL [30], an example-driven RL method. All results are averaged over three runs per scenario. Implementation details are in Appendix A.

4.1 Data and Evaluation Scenes

We train the policy and both world models on *ETH_Hotel*, *Zara01*, *ETH*, and *Students003* from the ETH/UCY pedestrian trajectory datasets [25, 31], reconstructed as Unity scenes (Figure 2). For *Zara01* and *Students003*, trajectories are split 8:1:1 into train, validation, and test, with only the training split used for optimization. We evaluate on five scenarios. *Zara01* and *Students003* reuse training scenes with held-out trajectories. *ETH_Hallway* and *ETH_Intersection* are new layouts built from the ETH scene asset, never seen during training, to test spatial generalization, with Hallway emphasizing constrained passages and Intersection emphasizing crossing flows. *Zara01_Group* augments *Zara01* with larger group configurations to test group-size generalization.

World-model inputs are short sequences of third-person RGB frames; Appendix H shows example inputs with ground-truth and predicted segmentation maps.

Table 1: **Closed-loop results on five held-out evaluation scenarios.** Higher is better for GSR, while lower is better for SRVR and GD.

Scene	Method	GSR ↑	SRVR ↓	GD ↓
Zara01	GBM [12]	55.6	21.0	**0.6**
	CCP [29]	33.3	16.0	2.6
	CEDRL [30]	50.0	15.6	1.3
	Ours	**72.2**	**7.0**	**0.6**
Zara01_Group	GBM [12]	38.9	17.5	**1.9**
	CCP [29]	27.8	19.5	4.8
	CEDRL [30]	22.2	17.4	5.9
	Ours	**77.8**	**1.2**	2.1
Students003	GBM [12]	64.0	15.2	1.4
	CCP [29]	28.0	20.0	3.4
	CEDRL [30]	37.0	27.1	1.8
	Ours	**66.0**	**7.2**	**0.7**
ETH_Intersection	GBM [12]	72.2	23.5	2.7
	CCP [29]	61.1	13.0	2.3
	CEDRL [30]	55.6	13.3	2.9
	Ours	**83.3**	**3.9**	**0.9**
ETH_Hallway	GBM [12]	68.8	9.3	1.3
	CCP [29]	62.5	7.5	1.8
	CEDRL [30]	56.3	19.1	1.5
	Ours	**87.5**	**0.9**	**0.8**

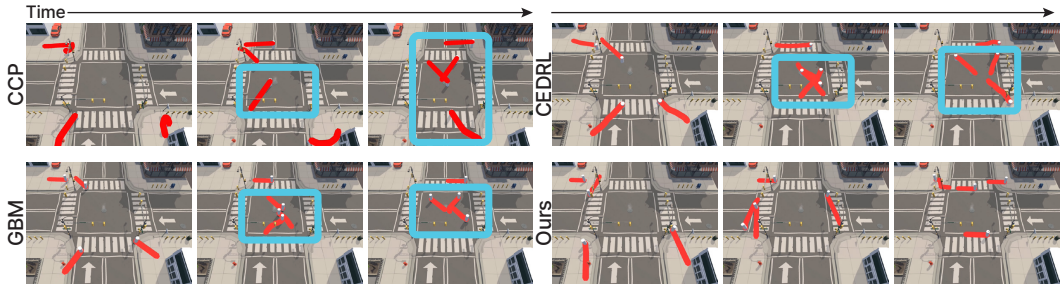


Figure 3: **Crosswalk navigation comparison.** We compare CCP, CEDRL, GBM, and our method in a road-crossing scene with a marked pedestrian crosswalk. Red curves show each agent's trajectory over the previous 60 frames. Blue boxes mark frames where the baselines leave the crosswalk and traverse road regions directly, indicating limited use of semantic scene cues. Our method, by contrast, remains on the crosswalk throughout the crossing, producing semantically compliant and pedestrian-like behavior. All trajectories are generated by the respective policies. In the second frame of ours, the agent moves along the crosswalk edge to avoid an oncoming pedestrian while staying on the crosswalk.

4.2 Evaluation Metrics

We evaluate along three complementary axes. Closed-loop navigation uses Goal Success Rate (GSR, percentage of agents reaching their destination collision-free), Social Region Violation Rate (SRVR, percentage of timesteps spent in semantically inappropriate regions such as roadways), and Group Distance (GD, mean distance to the assigned group centroid), covering goal completion, scene compliance, and group coherence respectively. Motion fidelity uses Average and Final Displacement Error (ADE, FDE) against held-out human trajectories, measuring geometric similarity to demonstrated motion. Human perception uses 7-point Likert ratings from naive observers along motion quality, naturalness, and smoothness, which together capture perceived human-likeness. Higher is better for GSR and Likert ratings; lower is better for all other metrics.

Table 2: **Ablation study on five held-out evaluation scenarios.** Higher is better for GSR, while lower is better for SRVR and GD.

Scene	Variant	GSR ↑	SRVR ↓	GD ↓
Zara01	**Ours**	**72.2**	**7.0**	**0.6**
	w/o Dual World Models	25.9	12.6	0.7
	w/o State Transition World Model	44.4	10.7	**0.6**
	w/o Semantic Transition World Model	44.4	9.3	**0.6**
	w/o VLM	48.1	10.9	0.9
	Top-10 Candidates	44.4	18.2	0.7
	Random-10 Candidates	22.2	20.2	0.7
Zara01_Group	**Ours**	**77.8**	**1.2**	2.1
	w/o Dual World Models	16.7	47.0	1.9
	w/o State Transition World Model	27.8	22.3	**1.7**
	w/o Semantic Transition World Model	38.9	17.5	2.3
	w/o VLM	29.6	7.4	**1.7**
	Top-10 Candidates	14.8	42.7	3.2
	Random-10 Candidates	18.5	18.5	2.5
Students003	**Ours**	**66.0**	**7.2**	**0.7**
	w/o Dual World Models	44.7	14.4	**0.7**
	w/o State Transition World Model	47.0	9.5	0.8
	w/o Semantic Transition World Model	63.3	8.9	0.8
	w/o VLM	60.0	10.6	0.8
	Top-10 Candidates	61.3	8.3	1.0
	Random-10 Candidates	57.3	8.5	0.8
ETH_Intersection	**Ours**	**83.3**	**3.9**	0.9
	w/o Dual World Models	40.7	13.5	1.1
	w/o State Transition World Model	51.9	10.6	1.0
	w/o Semantic Transition World Model	66.7	6.2	0.9
	w/o VLM	63.0	7.4	1.1
	Top-10 Candidates	55.6	13.9	1.0
	Random-10 Candidates	44.4	14.2	**0.8**
ETH_Hallway	**Ours**	**87.5**	**0.9**	0.8
	w/o Dual World Models	58.3	9.9	1.0
	w/o State Transition World Model	58.3	8.7	0.8
	w/o Semantic Transition World Model	70.8	13.7	**0.6**
	w/o VLM	66.7	16.9	0.9
	Top-10 Candidates	62.5	3.4	0.7
	Random-10 Candidates	70.8	12.5	1.2

4.3 Experimental Results and Discussion

We organize the results around our three contributions: (i) consequence-aware action selection improves closed-loop behavior over reactive and learning-based baselines, (ii) the two world models capture complementary failure modes, and (iii) the resulting policy is more human-like in both trajectory geometry and human perception, while requiring only a single forward pass at inference. The two world models that drive refinement are themselves accurate on held-out data — 87.0% mIoU and 99.3% pixel accuracy for the semantic branch, 0.18 m position ADE and 20.8° heading error for the state branch (Appendix E) — so the preference scores guiding GRPO rest on reliable predictions.

Closed-loop navigation across five held-out scenarios. Table 1 reports closed-loop performance on the five evaluation scenes, and Figure 3 shows a qualitative example. Across five evaluation scenes, our method achieves the best GSR and lowest SRVR in every scene, and the lowest or tied-lowest GD in four out of five scenes. Magnitudes matter: on Zara01_Group, SRVR drops from 17.4 (next-best) to 1.2, an order-of-magnitude reduction, with GSR doubling (38.9 → 77.8); ETH_Hallway shows a similar pattern (SRVR 7.5→0.9, GSR 68.8→87.5), as does the multi-flow ETH_Intersection (SRVR 13.0→3.9, GSR 72.2→83.3). Crucially, lower violations do not come from conservatism: GSR is

Table 3: **Comparison with simpler pipelines and an online world-model planner.** Trajectory predictors are combined with an MPC planner [33]. Results are averaged over the five held-out evaluation scenarios; per-scene results are in Appendix M. Higher is better for GSR, while lower is better for SRVR and GD. OnlineWM runs both world models at every decision; bold marks the best result among the remaining methods.

Method	GSR ↑	SRVR ↓	GD ↓
OnlineWM	80.6	3.2	0.9
Constant Velocity [36]	56.1	15.6	2.2
Social-LSTM [1]	51.9	15.0	2.1
Social-GAN [15]	53.2	15.0	2.2
Trajectron++ [35]	54.4	17.0	2.1
Gen-Safe-Nav [45]	42.9	14.1	2.3
Ours	**77.4**	**4.0**	**1.0**

higher than the strongest baseline on every scene. Figure 3 illustrates this qualitatively: at a marked crosswalk, the baselines leave the crosswalk and cut across the road, whereas our agent stays on it throughout the crossing. This supports our central claim that ranking locally plausible candidates by their predicted consequences yields navigation that is simultaneously safer, more compliant, and more goal-directed than baselines that act on the current state alone.

Complementary roles of the two world models. Table 2 isolates the contribution of each design choice. Removing both world models causes the largest average degradation (mean GSR 77.4 $\rightarrow$ 37.3, mean SRVR 4.0 $\rightarrow$ 19.5) and worsens both metrics on every scene, most starkly on Zara01_Group, where SRVR rises from 1.2 to 47.0 and GSR collapses from 77.8 to 16.7. This confirms that the supervised action prior alone cannot distinguish locally plausible anchors with different downstream consequences. Additional qualitative examples are provided in Appendix L.

The two world models cover different kinds of mistakes, as the single-branch ablations make clear. Both branches are themselves accurate on held-out data (87.0% semantic mIoU, 0.18 m state ADE; full diagnostics in Appendix E), so the differences below reflect *which signal* matters where, not differences in prediction quality. The State Transition World Model is most critical when interaction risk dominates: in ETH_Intersection, removing it drops GSR from 83.3 to 51.9 (a larger degradation than removing the semantic branch, which gives 66.7), reflecting that crossing flows are governed primarily by *who is heading toward whom*. Conversely, the Semantic Transition World Model is most critical when the navigable layout is the binding constraint: in ETH_Hallway, removing it raises SRVR from 0.9 to 13.7 (versus 8.7 when the state branch is removed), reflecting that constrained passages punish semantically invalid footprints more than mis-predicted neighbor trajectories. Removing either branch lowers GSR and raises SRVR in every scene, so the two branches are complementary; the state branch has the larger effect in most scenes.

The candidate-construction ablation (Top-10 vs. Random-10 vs. our top-5 + random-5) shows that the GRPO refinement also relies on a sufficiently diverse comparison set: pure exploitation collapses on Zara01_Group (GSR 14.8) while pure exploration is unstable on ETH_Intersection (GSR 44.4). The mixed strategy works because the two extremes fail in opposite ways: top-only gives candidates too similar to learn from, random-only gives candidates too irrelevant to matter. Mixing exposes the policy to alternatives that are nearby enough to be reachable yet different enough to be informative. The w/o VLM result shows that replacing the pretrained VLM with a task-trained Transformer/ViT consistently lowers GSR and raises SRVR, indicating that pretrained multimodal grounding provides useful scene-level cues beyond visual conditioning alone.

Is a world model necessary? To test whether a simpler pipeline suffices, we pair trajectory predictors [36, 1, 15, 35] with an MPC planner [33], and compare against Gen-Safe-Nav [45], which uses conformal prediction to account for forecast uncertainty. As Table 3 shows, all of these alternatives reach clearly lower goal success and higher violation rates than our method, suggesting that more accurate forecasting alone does not lead to safer navigation, consistent with [33]. We also evaluate an online planner that runs both world models at every decision (OnlineWM). Our distilled policy recovers most of its performance while requiring only a single forward pass (65 ms per

decision, compared with 1.2 s for OnlineWM). Input ablations (Appendix N) further show that visual and state inputs are complementary, and results on the standard ETH/UCY forecasting benchmark (Appendix O) show that the policy remains competitive outside our Unity reconstructions.

Human-likeness and inference cost. Three lines of evidence indicate that the refined policy produces human-like motion while retaining a single-pass inference profile. First, on motion-fidelity metrics that compare generated rollouts against held-out human trajectories (Appendix F), our method achieves the lowest ADE on both evaluated scenes (0.2064 m on Zara01, 0.1994 m on Students003) and the lowest FDE on Zara01 (0.3897 m), with Students003 FDE close to that of the best baseline. Second, in a user study with 20 naive participants rating anonymized rollouts on 7-point Likert scales (Appendix G), our method receives the highest average rating across all three perceptual axes — motion quality, naturalness, and smoothness — over GBM, CCP, and CEDRL. Third, the look-ahead reasoning is amortized at training time only: per-candidate scoring during refinement takes 0.115 s for the semantic branch and 1.9 ms for the state branch (Appendix J), but both world models are removed at deployment, leaving only the policy's single anchor-classification step at 65 ms per decision. Together, the geometric, perceptual, and computational evidence indicates that consequence-aware preference distillation produces motion that is human-aligned and naive-observer-preferred, without paying any look-ahead cost at inference.

5 Limitations

Our policy selects from a fixed vocabulary of $K = 100$ short-horizon anchors; motions outside this vocabulary can only be approximated, and longer horizons would require larger or learned vocabularies. Both world models make a single anchor-conditioned prediction, so they cannot capture long-range effects such as delayed yielding or group reformation over several seconds. When the two world models disagree, the policy follows their unit-weighted sum in Eq. (6) rather than an explicit arbitration rule; a learned or context-dependent weighting may resolve such conflicts better. All closed-loop results come from Unity reconstructions of ETH/UCY scenes, and we have not validated the policy on a real robot under sensor noise, localization error, or actuation delay. Finally, the pipeline is more complex and more costly to train than prediction-plus-planning alternatives; when scenes carry few semantic constraints or training compute is limited, a simpler predictor with MPC may be the more practical choice. Further failure cases are discussed in Appendix I.

6 Conclusion

The shape of "what to do next" in a populated scene is governed by what each action will cause, not by what it currently avoids, and the standard remedy of running a world model online at every step is a cost most deployments cannot afford. We close this gap with three intertwined moves: treating world models as training-time preference oracles rather than online planners, decomposing future prediction along the two axes that actually drive navigation, and aligning each branch with the failure mode it is best placed to catch. Together, these yield a policy that is simultaneously safer, more socially and group-coherent, more goal-directed, and perceived as more human-like, while paying no foresight cost at inference.

We see this as one instance of a broader direction. As embodied agents are built on large pretrained backbones, the question is no longer whether models can simulate the future, but how foresight can be made cheap and targeted enough to act on. Amortizing rollouts into the policy is one move; restricting prediction to the decision-relevant slice of the future is another. Natural next steps include learning the motion vocabulary jointly with the policy and extending to scenarios involving deferred coordination, where hierarchical world models may help. More broadly, the same principle should extend to settings where actions are rich and consequences are not visible from the current observation, and where the world models themselves keep improving without the deployed agent ever having to wait for them.

Acknowledgments and Disclosure of Funding

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2026-RS-2025-25441313, Contribution: 10%) grant funded by the Korea government (MSIT); by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201819, Contribution: 45%); by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2026-25615490, AI Star Fellowship Support Program, Contribution: 45%).

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[2] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9983–9995. IEEE, 2023.

[3] Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. Can language beat numerical regression? language-based multimodal trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[4] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17890–17901. IEEE, 2024.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[6] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4603–4623. PMLR, 2024.

[9] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019.

[10] Panayiotis Charalambous and Yiorgos Chrysanthou. The pag crowd: A graph based approach for efficient data-driven crowd simulation. *Computer Graphics Forum*, 33(8):95–108, 2014.

[11] Panayiotis Charalambous, Julien Pettré, Vassilis Vassiliades, Yiorgos Chrysanthou, and Nuria Pelechano. Greil-crowds: Crowd simulation with deep reinforcement learning and examples. *ACM Transactions on Graphics*, 42(4):1–15, 2023.

[12] Teófilo Bezerra Dutra, Ricardo Marques, Joaquim B Cavalcante-Neto, Creto Augusto Vidal, and Julien Pettré. Gradient-based steering for vision-based crowd simulation algorithms. In *Computer graphics forum*, volume 36, pages 337–348. Wiley Online Library, 2017.

[13] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

[14] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

[15] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[16] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2555–2565, 2019.

[17] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[18] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995.

[19] Takahiro Higuchi. Visuomotor control of human adaptive locomotion: Understanding the anticipatory nature. *Frontiers in Psychology*, 4:277, 2013.

[20] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577, 2022.

[23] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[24] Kang Hoon Lee, Myung Geol Choi, Qyoun Hong, and Jehee Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 109–118, 2007.

[25] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[26] Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. Wildworld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative arpg. *arXiv preprint arXiv:2603.23497*, 2026.

[27] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 648–657, 2017.

[28] Hisashi Murakami, Claudio Feliciani, Yuta Nishiyama, and Katsuhiro Nishinari. Mutual anticipation can contribute to self-organization in human crowds. *Science Advances*, 7(12): eabe7758, 2021.

[29] Andreas Panayiotou, Theodoros Kyriakou, Marilena Lemonari, Yiorgos Chrysanthou, and Panayiotis Charalambous. Ccp: Configurable crowd profiles. In *SIGGRAPH Conference Papers*, pages 1–10, 2022.

[30] Andreas Panayiotou, Andreas Aristidou, and Panayiotis Charalambous. Cedrl: Simulating diverse crowds with example-driven deep reinforcement learning. *Computer Graphics Forum*, 44(2):e70015, 2025.

[31] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009.

[32] Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, and Eric M. Wolff. Covernet: Multimodal behavior prediction using trajectory sets. *arXiv preprint arXiv:1911.10298*, 2019.

[33] Sriyash Poddar, Christoforos Mavrogiannis, and Siddhartha S Srinivasa. From crowd motion prediction to robot navigation in crowds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6765–6772. IEEE, 2023.

[34] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[35] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision – ECCV 2020*, pages 683–700. Springer, 2020.

[36] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

[37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[38] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9641–9650. IEEE, 2023.

[39] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[40] Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

[41] Jur van den Berg, Stephen J. Guy, Ming C. Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics Research*, pages 3–19. Springer, 2011.

[42] Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

[43] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1420. IEEE, 2023.

[44] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024.

[45] Jianpeng Yao, Xiaopan Zhang, Yu Xia, Zejin Wang, Amit K Roy-Chowdhury, and Jiachen Li. Towards generalizable safety in crowd navigation via conformal uncertainty handling. *arXiv preprint arXiv:2508.05634*, 2025.

[46] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.

[47] Mingbi Zhao, Wentong Cai, and Stephen J. Turner. Clust: Simulating realistic crowd behaviour by mining pattern from crowd videos. *Computer Graphics Forum*, 37(1):184–201, 2018.

[48] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023. URL `https://proceedings.mlr.press/v229/zitkovich23a.html`.

A Implementation Details

We use Qwen3-VL-2B-Instruct [5] as the policy backbone, frozen except for LoRA adapters (rank 64, $\alpha = 16$, dropout 0.05) — chosen so that the visual prior does not drift during anchor classification. We train the anchor classifier with cross-entropy for 10 epochs at learning rate 3×10^{-5}, batch size 32 (accumulation 4), cosine schedule with 5% warmup, weight decay 0.02, and gradient checkpointing. GRPO refines the supervised checkpoint for 10 epochs at learning rate 2×10^{-5} with both world models frozen. Each candidate group contains 10 anchors: the top five under the current policy and five sampled uniformly from the remainder. The semantic transition model runs 25 denoising steps per candidate. World-model context uses three history frames; the state branch uses $\lambda_u = 1.0$, $\lambda_{\text{dir}} = 2.0$, and the semantic branch uses $\lambda_{\text{LPIPS}} = 0.1$. All five score components in Eq. (6) are summed with unit weights, deliberately avoiding scenario-specific reward tuning. For training, we used eight NVIDIA A100 GPUs, while all inference experiments were conducted on a single NVIDIA RTX 6000 Ada GPU.

B Score Function Details

This appendix gives the detailed score terms used in Eq. (6). For candidate anchor $a_t^{(i)} = \{y_{i,0}, \ldots, y_{i,L}\}$, $y_{i,L}$ is the final displacement point, $b_i = y_{i,1}$ is the first executed displacement, and $\bar{v}_i = y_{i,1} - y_{i,0}$ is the first-step velocity implied by the anchor. The operator $\text{clip}(x, l, u)$ clamps scalar x into $[l, u]$.

Goal progress and smoothness. The goal-progress score is the normalized reduction in distance to the ego-relative goal g_t:

$$G_t^{(i)} = \text{clip}\left(\frac{\|g_t\|_2 - \|g_t - y_{i,L}\|_2}{\|g_t\|_2 + \epsilon}, -1, 1\right). \tag{7}$$

The smoothness cost penalizes abrupt changes from the previous ego velocity v_{t-1}:

$$C_t^{(i)} = \text{clip}\left(\frac{\|\bar{v}_i - v_{t-1}\|_2^2}{2(\|\bar{v}_i\|_2^2 + \|v_{t-1}\|_2^2) + \epsilon}, 0, 1\right), \tag{8}$$

with $C_t^{(i)} = 0$ when both velocity magnitudes are numerically zero.

Immediate and predicted future states. For a surrounding entity with current relative position $p_{t,j}$ and relative velocity $u_{t,j}$, the immediate state after applying candidate $a_t^{(i)}$ is approximated as

$$\tilde{p}_{t,j}^{(i)} = p_{t,j} + (u_{t,j} + v_t) - b_i, \qquad \tilde{u}_{t,j}^{(i)} = (u_{t,j} + v_t) - \bar{v}_i. \tag{9}$$

The predicted future state is obtained from the State Transition World Model:

$$p_{t+1,j}^{(i)} = p_{t,j} + \Delta\hat{p}_{t,j}(a_t^{(i)}), \qquad u_{t+1,j}^{(i)} = \hat{u}_{t+1,j}(a_t^{(i)}). \tag{10}$$

Interaction costs. The interaction score uses three costs: crowd pressure, proxemic cost, and collision risk. For pedestrian positions $P = \{p_j\}_{j=1}^n$ and pedestrian velocities $U = \{u_j\}_{j=1}^n$, crowd pressure is

$$\mathcal{I}(P, U) = \text{clip}\left(\frac{\sum_{j=1}^n \exp\left(-\|p_j\|_2^2/(2\sigma_\rho^2)\right)}{N_{\max}}, 0, 1\right) \text{clip}\left(\frac{\frac{1}{n}\sum_{j=1}^n \|u_j - \bar{u}\|_2^2}{\frac{1}{n}\sum_{j=1}^n \|u_j\|_2^2 + \epsilon}, 0, 1\right), \tag{11}$$

where $\sigma_\rho = 2.5$, $N_{\max} = 15$, and $\bar{u} = n^{-1}\sum_{j=1}^n u_j$. If no pedestrian is present, the pressure term is set to zero.

The proxemic cost uses the same pedestrian positions and group indicators $q_j \in \{0, 1\}$:

$$\mathcal{C}(P, q) = \text{clip}\left(\frac{1}{N_{\max}}\sum_{j=1}^n w_j \exp\left(-\|p_j\|_2^2/(2\sigma_p^2)\right), 0, 1\right), \qquad w_j = \begin{cases} 0.5, & q_j = 1, \\ 1, & q_j = 0, \end{cases} \tag{12}$$

where $\sigma_p = 1.25$. The smaller weight for $q_j = 1$ reduces the penalty for pedestrians in the same group as the controlled agent.

Collision risk is computed for both pedestrians and dynamic objects. For positions $P = \{p_j\}_{j=1}^n$ and velocities $U = \{u_j\}_{j=1}^n$, the time-to-collision proxy is

$$T_j = -\frac{p_j^\top u_j}{\|u_j\|_2^2 + \epsilon}. \tag{13}$$

The raw collision score with entity-type scale s is

$$\mathcal{R}_{\text{raw}}(P, U; s) = \sum_{j:T_j>0} s \exp(-T_j/\tau), \tag{14}$$

where $\tau = 2.0$. Pedestrians use $s = 1$, and dynamic objects use $s = s_{\text{obj}} = 1.2$. The normalized collision risk is

$$\mathcal{R}(P^{\text{ped}}, U^{\text{ped}}, P^{\text{obj}}, U^{\text{obj}}) = \text{clip}\left(\frac{\mathcal{R}_{\text{raw}}(P^{\text{ped}}, U^{\text{ped}}; 1) + \mathcal{R}_{\text{raw}}(P^{\text{obj}}, U^{\text{obj}}; s_{\text{obj}})}{N_{\max} + s_{\text{obj}} M_{\max} + \epsilon}, 0, 1\right), \tag{15}$$

where $M_{\max} = 10$ is the maximum number of dynamic objects.

The immediate interaction penalty is

$$P_{\text{cur},t}^{(i)} = I_{\text{cur},t}^{(i)} + C_{\text{cur},t}^{(i)} + R_{\text{cur},t}^{(i)}, \tag{16}$$

where $I_{\text{cur},t}^{(i)}$, $C_{\text{cur},t}^{(i)}$, and $R_{\text{cur},t}^{(i)}$ evaluate $\mathcal{I}$, $\mathcal{C}$, and $\mathcal{R}$ on the immediate states $\tilde{p}_{t,j}^{(i)}$ and $\tilde{u}_{t,j}^{(i)}$. The future interaction penalty is

$$P_{\text{fut},t}^{(i)} = I_{\text{fut},t}^{(i)} + C_{\text{fut},t}^{(i)} + R_{\text{fut},t}^{(i)}, \tag{17}$$

where $I_{\text{fut},t}^{(i)}$, $C_{\text{fut},t}^{(i)}$, and $R_{\text{fut},t}^{(i)}$ evaluate the same costs on the predicted future states $p_{t+1,j}^{(i)}$ and $u_{t+1,j}^{(i)}$.

Semantic validity. The semantic score is computed from the predicted segmentation $\hat{S}_{t+1}(a_t^{(i)})$. A triangular local footprint is placed in the image, and $\rho_c^{(i)}$ denotes the percentage of footprint pixels assigned to semantic class c. Define $\phi_{\geq}(x, \tau, a, b)$ to return a if $x \geq \tau$ and b otherwise. The raw semantic score is

$$\begin{aligned} M_{\text{raw},t}^{(i)} =& \phi_{\geq}(\rho_{\text{sidewalk}}^{(i)} + \rho_{\text{crosswalk}}^{(i)}, 70, 1, -1) + \phi_{\geq}(\rho_{\text{road}}^{(i)}, 70, -2, 1) \\ &+ \phi_{\geq}(\rho_{\text{grass}}^{(i)}, 25, -1, 1) + \phi_{\geq}(\rho_{\text{vehicle}}^{(i)}, 5, -10, 1) \\ &+ \phi_{\geq}(\rho_{\text{building}}^{(i)}, 5, -10, 1) + \phi_{\geq}(\rho_{\text{obstacle}}^{(i)}, 5, -10, 1). \end{aligned} \tag{18}$$

The normalized semantic score is

$$M_t^{(i)} = \text{clip}\left(\frac{M_{\text{raw},t}^{(i)}}{34}, -1, 1\right), \tag{19}$$

where 34 is the maximum absolute rule score.

The reward and penalty values in the semantic score were designed according to the relative safety criticality of each semantic class. In particular, vehicle, building, and obstacle regions represent physical objects or structures for which contact must be strictly avoided. Since these classes may occupy a relatively small portion of the local footprint compared with terrain-related classes such as grass, road, sidewalk, and crosswalk, even a small overlap can indicate an unsafe or infeasible motion. Therefore, they are assigned the largest penalties. Road regions are also treated as invalid for pedestrian navigation, but they do not necessarily imply direct physical collision with a static or dynamic object; thus, they are assigned an intermediate penalty. Grass, sidewalk, and crosswalk are given lower penalty or reward magnitudes because they mainly encode terrain preference and pedestrian-validity rather than hard collision constraints. Accordingly, the penalty severity follows the hierarchy

$$\text{vehicle}, \text{building}, \text{obstacle} > \text{road} > \text{grass}, \text{sidewalk}, \text{crosswalk}.$$

The remaining scalar values, including class-specific thresholds and reward/penalty magnitudes, were determined empirically through hyperparameter tuning while preserving this relative ordering of semantic penalty severity.

Table 4: **Segmentation classes used by the Semantic Transition World Model.** Each class is identified by a semantic tag and a fixed Hex color code. The score role describes how the class is treated by the semantic-validity score in Appendix B.

Semantic tag	Navigation meaning	Hex	Role in semantic score
`sidewalk`	Pedestrian sidewalk or primary outdoor walking surface.	`#804080`	Positive. A high sidewalk ratio indicates that the candidate anchor remains on a pedestrian-compatible walking region.
`crosswalk`	Marked pedestrian crossing region.	`#FFFFFF`	Positive. Treated as a valid walking region and grouped with sidewalk for walkability scoring.
`road`	Vehicle-oriented road surface.	`#808080`	Negative. A large road ratio indicates unsafe or semantically inappropriate pedestrian motion unless the region is explicitly labeled as crosswalk.
`obstacle`	Static obstacle such as a pole, barrier, fence, furniture, or other non-traversable object.	`#FF8C00`	Strongly negative. Footprint overlap indicates physically invalid or unsafe navigation.
`building`	Building, wall, or other non-traversable static structure.	`#464646`	Strongly negative. Overlap indicates an invalid trajectory through static structure.
`vehicle`	Moving or parked vehicle.	`#00008E`	Strongly negative. Even a small vehicle overlap indicates high collision or scene-validity risk.
`grass`	Grass or off-path terrain.	`#6B8E23`	Mildly negative. Penalized when it occupies a substantial fraction of the local footprint because it is not the primary pedestrian walking surface.
`unknown`	Undefined object, unlabeled pixel, or background.	`#000000`	Neutral. Ignored in semantic-validity scoring unless explicitly replaced by a known non-walkable class.

C Segmentation Classes

The Semantic Transition World Model predicts a compact navigation-oriented segmentation map rather than a photorealistic RGB frame. Each semantic region is represented by a discrete tag and a fixed Hex color code. The Hex code is used only as a deterministic label encoding and for visualization; during training and scoring, segmentation maps are converted into semantic class indices or one-hot masks.

Class vocabulary. We map all semantic annotations into the class set shown in Table 4. The vocabulary covers pedestrian-walkable areas, vehicle-oriented regions, and static structures. Unlike generic scene-parsing taxonomies, this vocabulary is chosen to match navigation-relevant distinctions: whether a candidate motion remains on pedestrian-compatible space, approaches a dynamic or static obstacle, or enters a road-like danger zone.

D Anchor Vocabulary Reconstruction Analysis

The purpose of the anchor vocabulary is not to reproduce every continuous trajectory exactly, but to provide a compact discrete action space that still preserves the main short-horizon motion patterns observed in real pedestrian data. In all experiments, we use $K = 100$ anchors, where each anchor represents a 20-frame short-horizon motion primitive extracted from the ETH/UCY trajectory data.

To examine whether this vocabulary is sufficiently expressive, we randomly sample 20 trajectory clips from the dataset and compare each real trajectory with its nearest anchor reconstruction. Figure 4 shows the comparison. The blue solid curves indicate the original real trajectory clips, and the red

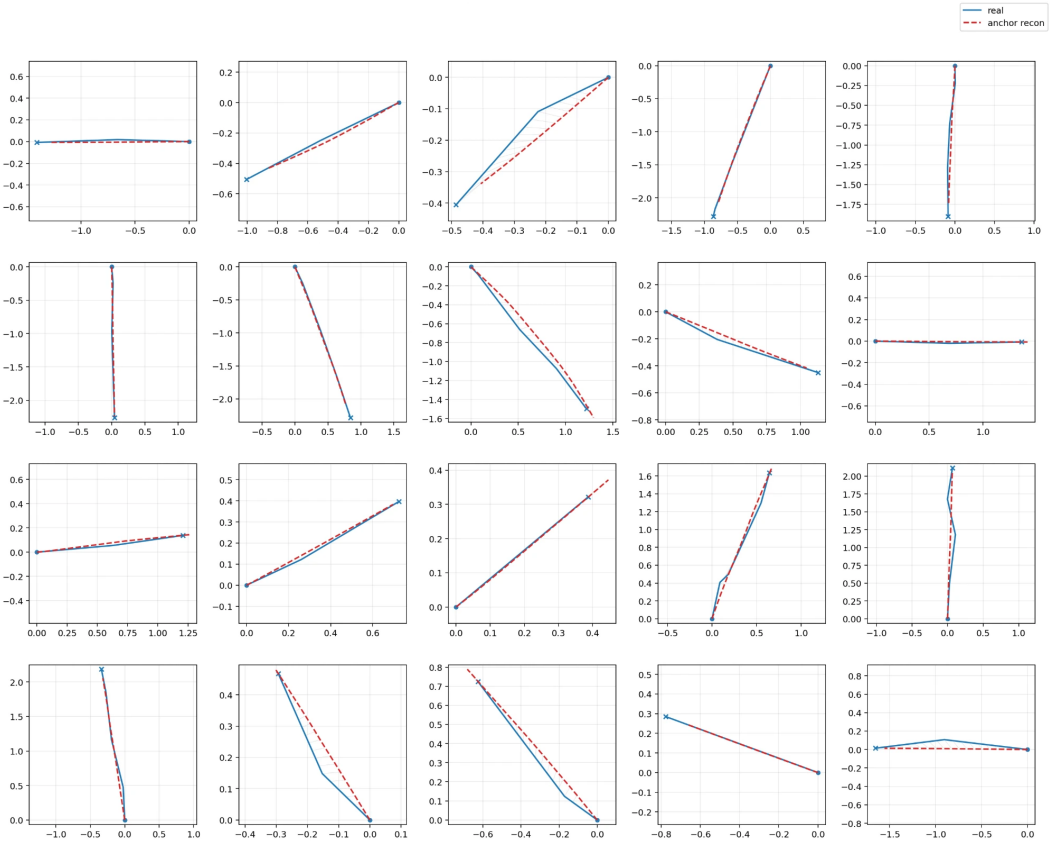


Figure 4: **Real trajectory clips and nearest-anchor reconstructions.** We randomly sample 20 trajectory clips and visualize each real trajectory against its nearest anchor reconstruction from the $K = 100$ vocabulary. Blue solid lines denote real trajectory clips, and red dashed lines denote anchor reconstructions. The examples show that the 100-anchor vocabulary preserves the main direction, scale, and endpoint tendency of short-horizon pedestrian motion, supporting its use as a compact discrete action space.

dashed curves indicate the corresponding anchor reconstructions. Across the sampled examples, the reconstructed anchors closely follow the direction, scale, and endpoint tendency of the real trajectories. Although small deviations appear in some clips, especially for rare or abrupt motions, the nearest anchors generally preserve the dominant movement pattern needed for short-horizon action selection.

This result supports the use of $K = 100$ as a practical vocabulary size. A smaller vocabulary would make the action space more compact, but it would merge distinct behaviors such as straight walking, lateral movement, turning, and speed variation into overly coarse motion primitives. On the other hand, a much larger vocabulary would increase the number of action classes, make supervised anchor classification more sparse, and increase the cost of candidate comparison during world-model-guided refinement. The reconstruction examples in Figure 4 indicate that $K = 100$ provides a good balance: it is expressive enough to approximate diverse short-horizon pedestrian motions, while remaining compact enough for efficient policy learning and candidate scoring.

Importantly, the anchor representation is used as an action abstraction rather than as a final trajectory predictor. During policy learning, the model only needs to select a plausible short-horizon motion primitive from the current observation. During refinement, the dual world models compare candidate anchors according to their predicted social and semantic consequences. Therefore, exact point-wise reconstruction is less important than preserving behaviorally meaningful motion modes. The close agreement between real clips and anchor reconstructions in the random samples suggests that the 100-anchor vocabulary satisfies this requirement.

Table 5: **Semantic prediction performance.** Class-aware test metrics for the Semantic Transition World Model.

Method	Pixel Acc. ↑	mIoU ↑	Dice/F1 ↑	FWIoU ↑
EDM	99.2848%	87.0020%	89.0814%	98.6099%
DDPM	97.7863%	86.5421%	88.2132%	95.4289%

E World Model Performance and Diagnostics

This appendix reports diagnostic test performance for the two frozen consequence predictors used during policy refinement: the State Transition World Model and the Semantic Transition World Model. The goal of this evaluation is not to benchmark a general-purpose simulator, but to verify that each branch is accurate on the decision-relevant target it contributes to candidate-anchor scoring.

Diagnostic rationale for the dual design. We evaluate the two world models separately because they answer different questions during preference construction. The State Transition World Model asks *who will move where* after a candidate anchor is executed. It therefore predicts future relative positions, velocities, speeds, and headings of nearby pedestrians and dynamic objects, which are used to estimate interaction risk. The Semantic Transition World Model asks *what navigational affordance the agent's future footprint will occupy*. It therefore predicts a navigation-oriented semantic layout, which is used to evaluate whether a candidate anchor remains on pedestrian-compatible regions or moves into invalid regions such as roads, vehicles, buildings, or obstacles.

This separation is especially important in populated scenes. Pedestrians are dynamic agents and transient occluders: they may temporarily cover a sidewalk or crosswalk, but they do not change the underlying walkability of that region. If pedestrians were included as ordinary segmentation targets, the semantic branch would be forced to solve two different problems at once: predicting persistent scene layout and forecasting every future pedestrian mask. This would duplicate the role of the state branch and introduce noisy supervision, because small errors in future pedestrian placement can dominate segmentation loss without improving semantic footprint scoring. We therefore let pedestrians appear in the RGB input, but remove them from the segmentation target by assigning their pixels to an ignore label. The semantic branch learns the underlying environmental layout, including walkable regions temporarily occluded by people, while the state branch models dynamic pedestrian motion.

The following diagnostics are therefore branch-specific. For the semantic branch, we report class-aware segmentation metrics that measure layout fidelity. For the state branch, we report position, velocity, speed, and heading errors that measure short-horizon interaction prediction quality.

Semantic Transition World Model. The Semantic Transition World Model predicts the next navigation-oriented segmentation map. Since this model is used to evaluate whether a candidate anchor moves the agent toward semantically valid or invalid regions, we report class-aware segmentation metrics rather than RGB reconstruction or perceptual losses. The reported metrics include pixel accuracy, mean IoU, Dice/F1, and frequency-weighted IoU. Pixel accuracy measures overall pixel-wise correctness, mIoU evaluates class-balanced overlap quality, Dice/F1 measures region overlap, and frequency-weighted IoU summarizes IoU while accounting for class frequency.

As shown in Table 5, the EDM-style semantic transition model achieves the best overall performance, with 99.2848% pixel accuracy, 87.0020% mIoU, 89.0814% Dice/F1, and 98.6099% frequency-weighted IoU. Compared with the DDPM variant, EDM improves pixel accuracy and frequency-weighted IoU while also slightly improving mIoU and Dice/F1. These results indicate that the semantic branch preserves both dominant layout structure and class-level region overlap, supporting its use as a footprint-based semantic consequence estimator.

State Transition World Model. The State Transition World Model predicts the next relative position and velocity of nearby pedestrians and dynamic objects. This prediction is used to estimate future interaction risk, including possible collisions, crowd pressure, and proxemic discomfort. We therefore evaluate the model with position, velocity, speed, and heading errors that directly reflect short-horizon state prediction quality.

Table 6: **State prediction performance.** Test metrics for future relative position and velocity prediction.

Pos. ADE ↓	Pos. RMSE ↓	Vel. MAE ↓	Vel. RMSE ↓	Speed MAE ↓	Head. Err. ↓
0.1846	0.2481	0.3354	0.4414	0.2096	20.77

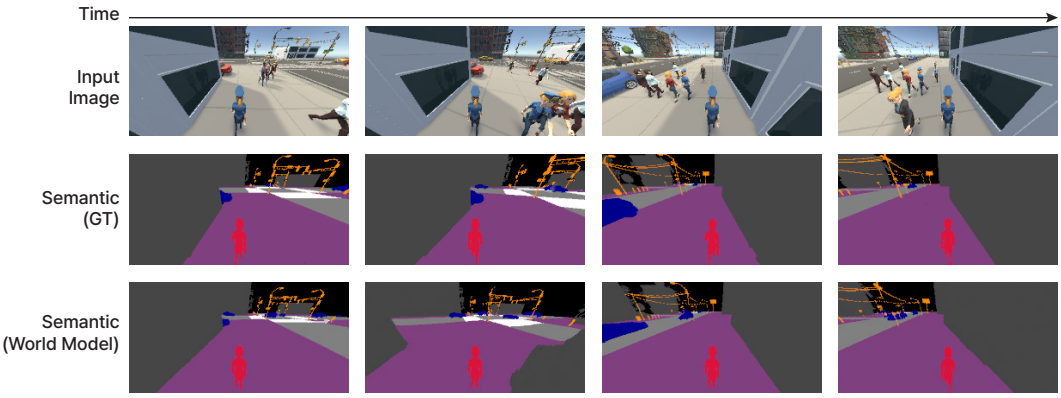


Figure 5: **Segmentation image sequence from Zara01.** Columns are ordered in time. Rows show the third-person RGB input, the ground-truth segmentation, and the segmentation predicted by the Semantic Transition World Model. The red silhouette marks the controlled agent; other pedestrians are excluded from the segmentation target.

As shown in Table 6, the State Transition World Model achieves a position ADE of 0.1846 m and a position RMSE of 0.2481 m, showing accurate short-horizon relative position prediction. For motion prediction, it obtains a velocity MAE of 0.3354 m/s, a velocity RMSE of 0.4414 m/s, and a speed MAE of 0.2096 m/s. The heading error is 20.77 degrees, indicating that the model also captures the directional tendency of nearby moving entities. These results show that the state branch provides sufficiently accurate future relative-state estimates for interaction-aware candidate scoring, complementing the semantic branch rather than duplicating it.

Together, these diagnostics support the dual-world-model decomposition. The semantic branch provides reliable predictions of persistent navigation affordances, including walkable regions that may be temporarily occluded by pedestrians, while the state branch provides reliable short-horizon estimates of nearby dynamic entities. This division gives the preference score two non-overlapping sources of foresight: semantic validity from predicted layout and interaction risk from predicted motion. The main ablation results further test the behavioral necessity of using both signals during policy refinement.

F Additional Experiments for Motion Fidelity

In addition to the closed-loop navigation metrics reported in the main paper, we evaluate motion fidelity with Average Displacement Error (ADE) and Final Displacement Error (FDE). These metrics compare generated trajectories against reference human trajectories, where ADE measures the average point-wise trajectory deviation and FDE measures the endpoint deviation. Lower values indicate closer agreement with human motion.

Table 7 reports results on held-out test trajectories from two scenes, Zara01 and Students003. Our method achieves the lowest ADE in both scenes, indicating that the generated trajectories remain closer to human demonstrations over the rollout horizon. For FDE, our method obtains the best result on Zara01 and the second-best result on Students003, where it remains close to the best baseline. These results support the trajectory-fidelity aspect of our method: world-model-guided preference refinement improves not only closed-loop navigation behavior, but also the geometric similarity of generated motion to demonstrated human trajectories.

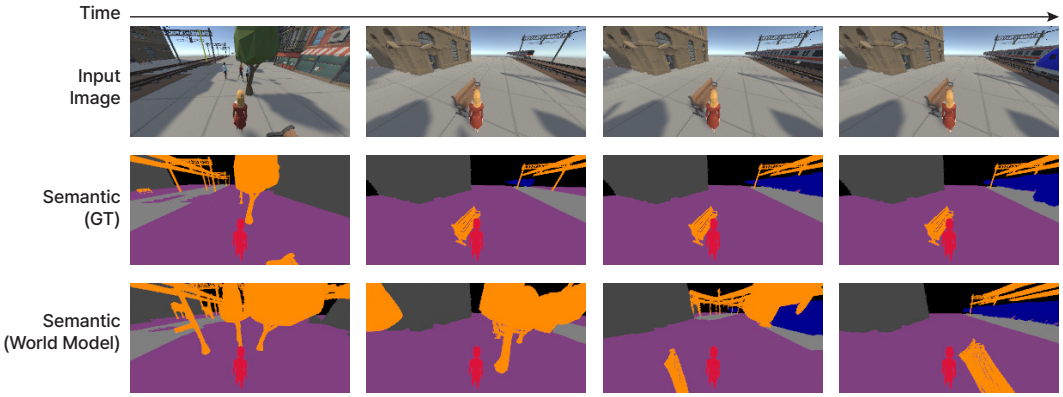


Figure 6: **Segmentation image sequence from ETH_Hotel.** Columns are ordered in time; rows are as in Figure 5.

Table 7: **Motion fidelity on held-out trajectories.** ADE and FDE (in meters) compare generated trajectories against reference human trajectories. Lower is better for both metrics; bold and underline mark the best and second-best results. Values are reported with 95% confidence intervals.

Scene	Method	ADE ↓	FDE ↓
Zara01	GBM [12]	0.3057 ± 0.0345	0.5203 ± 0.0509
	CCP [29]	$\underline{0.2406 \pm 0.0005}$	$\underline{0.4524 \pm 0.0005}$
	CEDRL [30]	0.3629 ± 0.0071	0.6817 ± 0.0134
	Ours	$\mathbf{0.2064 \pm 0.0069}$	$\mathbf{0.3897 \pm 0.0141}$
Students003	GBM [12]	0.2147 ± 0.0026	0.4194 ± 0.0056
	CCP [29]	$\underline{0.2053 \pm 0.0031}$	$\mathbf{0.3873 \pm 0.0060}$
	CEDRL [30]	0.2338 ± 0.0030	0.4382 ± 0.0058
	Ours	$\mathbf{0.1994 \pm 0.0026}$	$\underline{0.3948 \pm 0.0052}$

G User Study

We conducted a user study to evaluate the perceived quality of the generated crowd navigation behaviors. The study was administered through Google Forms, and a screenshot of the actual survey interface is shown in Figure 13. A total of 20 participants took part in the study, with a balanced gender distribution (10 male and 10 female; age 27.50 ± 6.19 years). The study was approved by the Korea University Institutional Review Board (KUIRB-2026-0115-01), and all participants provided informed consent before participating. Each participant received USD $7 as compensation for completing the study.

Each participant evaluated the results of four methods: GBM, CCP, CEDRL, and Ours. We prepared 40 evaluation scenarios in total, consisting of 10 scenarios for each method. The scenarios were evenly sampled from five evaluation scenes: Zara01, Zara01_Group, Students003, ETH_Intersection, and ETH_Hallway. For each scenario, participants watched an anonymized rollout video and rated the resulting motion using three 7-point Likert-scale questions: motion quality, naturalness, and smoothness. Higher scores indicate better perceived motion quality. The method identities were hidden from participants, and the presentation order was randomized to reduce ordering bias.

For statistical analysis, we first averaged scores at the participant level for each method and each question, rather than treating individual scenario ratings as independent samples. This avoids pseudo-replication and respects the repeated-measures structure of the study. Since all participants rated all methods and the scores are ordinal Likert ratings, we used the Friedman test to assess the overall effect of method. For post-hoc comparisons, we used paired Wilcoxon signed-rank tests comparing our method against each baseline, with Holm correction for multiple comparisons. We report statistical significance at two levels, $p < .05$ and $p < .01$.

As shown in Figure 7, our method received the highest mean score across all perceptual criteria. Our method achieved an overall score of 4.94 ± 0.44, compared to 3.19 ± 0.34 for GBM, 3.54 ± 0.34 for

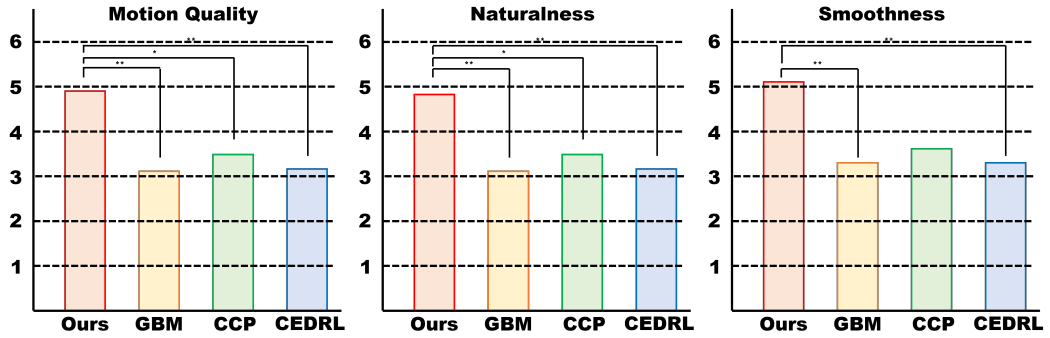


Figure 7: **User study results for motion quality, naturalness, and smoothness.** Bars show participant-level mean ratings. Brackets mark Holm-corrected paired Wilcoxon signed-rank tests comparing our method against each baseline (* $p < .05$, ** $p < .01$).

CCP, and 3.14 ± 0.34 for CEDRL. For the individual criteria, our method obtained 4.91 ± 0.44 in motion quality, 4.88 ± 0.47 in naturalness, and 5.05 ± 0.44 in smoothness. The baselines received lower scores across the three criteria: GBM achieved 3.14 ± 0.32, 3.13 ± 0.35, and 3.32 ± 0.38; CCP achieved 3.48 ± 0.32, 3.45 ± 0.35, and 3.68 ± 0.36; and CEDRL achieved 3.04 ± 0.33, 3.07 ± 0.35, and 3.31 ± 0.36 for motion quality, naturalness, and smoothness, respectively. These results suggest that participants perceived our generated navigation behaviors as more natural, higher-quality, and smoother than those of the baseline methods.

The Friedman test showed a significant overall effect of method for the aggregated score ($\chi^2(3) = 20.84$, $p < .01$). The effect was also significant for each individual criterion: motion quality ($\chi^2(3) = 21.28$, $p < .01$), naturalness ($\chi^2(3) = 16.18$, $p < .01$), and smoothness ($\chi^2(3) = 14.36$, $p < .01$). These results indicate that the perceived quality differences among methods are statistically reliable across all evaluation questions.

Post-hoc comparisons further show that our method was preferred over the baselines. For the overall score, our method significantly outperformed GBM and CEDRL at the $p < .01$ level, and CCP at the $p < .05$ level. For motion quality, our method significantly outperformed GBM and CEDRL at the $p < .01$ level, and CCP at the $p < .05$ level. For naturalness, our method also significantly outperformed GBM and CEDRL at the $p < .01$ level, and CCP at the $p < .05$ level. For smoothness, our method significantly outperformed GBM and CEDRL at the $p < .01$ level, while the difference against CCP was not statistically significant after correction.

Overall, the user study demonstrates that our method produces crowd navigation behaviors that are perceived as higher-quality and more natural than those of existing baselines. The improvements are consistent across motion quality and naturalness, where our method significantly outperforms all three baselines. For smoothness, our method still obtains the highest mean score, but its difference from CCP does not reach statistical significance after correction. This suggests that CCP can produce comparably smooth motion in some cases, while our method provides stronger overall perceptual quality by better balancing naturalness, smoothness, and human-like scene-compliant navigation.

H Qualitative Visualization of Segmentation Image Sequences

In addition to the quantitative semantic results in Appendix E, we visualize segmentation image sequences predicted by the Semantic Transition World Model. These examples are used to inspect whether the model preserves navigation-relevant semantic structure over time. The goal is not photorealistic image generation, but temporally consistent prediction of semantic regions such as walkable space, road regions, buildings, obstacles, and vehicles.

As shown in Figures 5, 6 and 8, the predicted segmentation sequences remain visually coherent across consecutive frames. Walkable regions largely preserve their spatial extent and the overall scene layout changes smoothly over time, although thin obstacles are sometimes over-extended (e.g., the orange regions in Figure 6). This temporal consistency is important because the semantic prediction is used during policy refinement to evaluate whether each candidate anchor leads to a valid or invalid region.

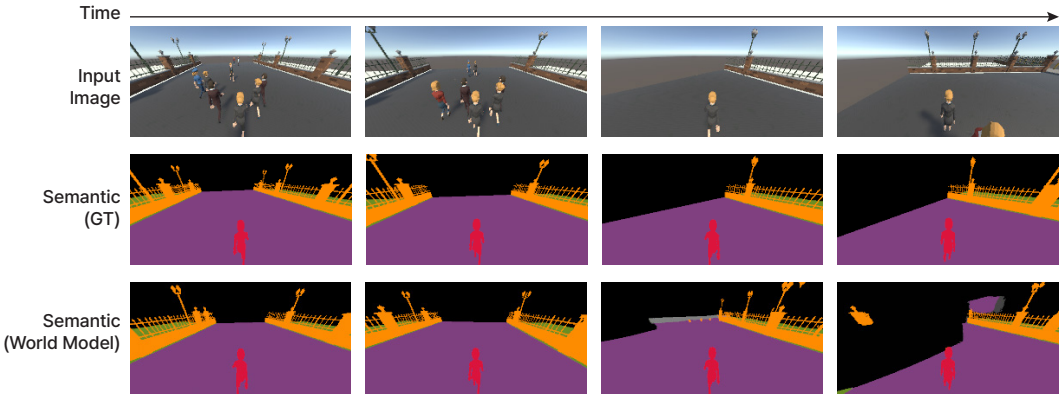


Figure 8: **Segmentation image sequence from ETH.** Columns are ordered in time; rows are as in Figure 5.

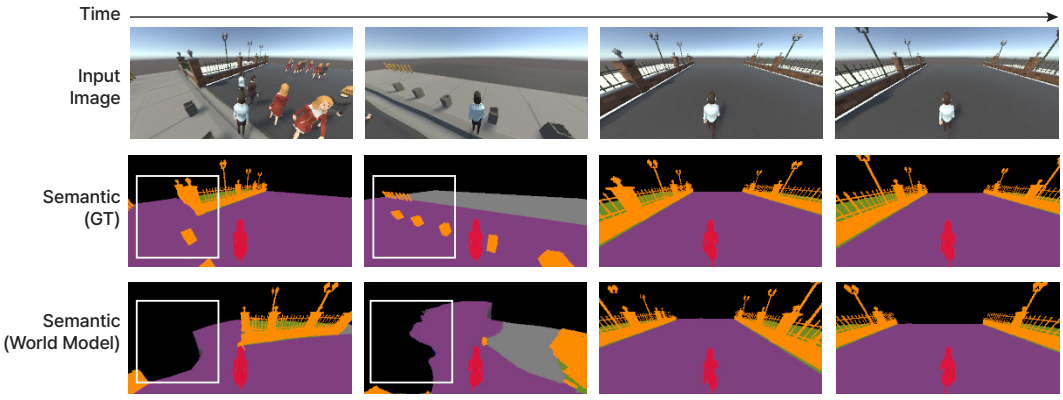


Figure 9: **Failure case of a segmentation image sequence from ETH.** Rows are as in Figure 5. White boxes mark regions where the predicted segmentation deviates from the ground truth and loses temporal consistency.

The qualitative visualizations are consistent with the semantic prediction performance reported in Appendix E. The model achieves high pixel accuracy, mIoU, Dice/F1, and frequency-weighted IoU on the test split, and the visualized sequences further show that these predictions remain stable over time. Thus, the Semantic Transition World Model provides a reliable navigation-oriented representation for semantic candidate scoring.

I Failure Cases and Limitations

This appendix summarizes the main failure cases and limitations of the current method. Although the proposed policy benefits from the State Transition World Model and the Semantic Transition World Model during refinement, the world-model predictions are not always temporally perfect.

As shown in Figure 9, the Semantic Transition World Model can fail in complex ETH scenes. While the model captures the overall layout of the scene, some semantic regions change unnaturally across time or are not reconstructed consistently. For example, parts of the sidewalk, road boundary, obstacle region, or agent area may shift or disappear between consecutive predicted frames. This type of error can affect the semantic-validity score because the policy refinement relies on the predicted segmentation map to judge whether a candidate anchor moves toward a valid region.

Another limitation is that the current world models predict short-horizon consequences. This is sufficient for comparing immediate candidate anchors, but it may not capture long-range interaction patterns such as delayed yielding, dense crowd negotiation, or group reformation over several seconds.

Table 8: **World model and VLA compute and runtime.** Runtime is measured per candidate for the world models and per decision for the deployed policy.

Component	Compute	Runtime
Semantic Transition World Model	31.1 TFLOPs	0.115 s
State Transition World Model	6.8 GFLOPs	1.904 ms
Deployed policy (Qwen3-VL-2B + LoRA, anchor classification)	23.6 TFLOPs	65 ms

Thus, some actions that appear safe in the one-step prediction may still lead to undesirable behavior later in the rollout.

The policy also uses a fixed anchor vocabulary with $K = 100$. This makes action selection efficient, but rare or unusual motions may be approximated imperfectly when they are not well covered by the anchor set. In addition, the semantic score depends on predefined segmentation labels and rule-based footprint thresholds, so transferring the method to new environments may require adjusting the semantic class mapping or validity rules.

Finally, all closed-loop experiments are conducted in reconstructed Unity scenes. Although the world models are removed at inference time and the deployed policy is efficient, real-world deployment would require additional validation under sensor noise, localization error, actuation delay, and stricter safety constraints. Therefore, the current results should be interpreted as evidence that world-model-guided refinement improves consequence-aware navigation, rather than as a complete real-robot safety guarantee.

J Runtime and Compute

This appendix summarizes the compute and runtime cost of the world models used during policy refinement. The key property of our framework is that the State Transition World Model and the Semantic Transition World Model are used only during training-time candidate scoring. At inference time, the deployed policy directly selects an anchor from the current observation, and no world-model prediction or counterfactual scoring is performed. Table 8 summarizes the compute and runtime of each component.

The Semantic Transition World Model is the dominant source of training-time compute. This is expected because it predicts a dense segmentation map through a denoising-based semantic transition model. In contrast, the State Transition World Model is lightweight because it predicts low-dimensional relative positions and velocities of nearby pedestrians and dynamic objects.

These costs are paid only when generating preference supervision during refinement. During deployment, the world models are removed and the policy performs a single anchor-selection step. Therefore, the additional computation introduced by consequence prediction improves training-time supervision without increasing test-time inference latency.

K Prompt Template

The VLA policy receives a third-person RGB observation together with structured local navigation states and predicts one anchor index from the discrete anchor vocabulary. To make the input format consistent across supervised training, GRPO refinement, and inference, we use a fixed anchor-selection prompt template (Table 9). The prompt describes the coordinate convention, the ego state, the goal location, nearby pedestrians, nearby dynamic objects, and the available anchor-index range. All coordinates are expressed in the ego-centered local frame, where the controlled agent is located at $(0, 0)$. Nearby pedestrians are listed within 7 meters, with same-group agents placed first and then non-group agents, each ordered from nearest to farthest. Nearby dynamic objects are listed within 15 meters and are also ordered by distance.

The output space is constrained to a single integer. This prevents the model from producing free-form language and makes the VLA policy directly compatible with the anchor classifier used in the rest of the method.

Table 9: **Anchor-selection prompt template.** The prompt asks the policy to choose the safest and most natural anchor for crowd simulation based on the current image, ego state, goal, nearby agents, and nearby dynamic objects.

Prompt Template

Question: The current image is your third-person view.
All coordinates below are relative to you, with your current position fixed at $(0, 0)$.
Your own velocity values are local motion values, and each nearby entity velocity is written as the other entity's velocity minus your current velocity.
Based on the scene and the state below, choose the anchor that is safest, goal-directed, collision-aware, socially compliant, smooth, and natural for crowd simulation.

Current local velocity: `{velocity}`.
Previous local velocity: `{prev_velocity}`.
Goal position: `{goal_pos}`.

Nearby agents within 7 meters (same-group agents first, then non-group agents; each part ordered from nearest to farthest):
`{nearby_agents_text}`

Nearby dynamic objects within 15 meters (ordered from nearest to farthest):
`{nearby_dynamic_text}`

Options: 0 to `{num_classes - 1}`.
Answer with only one integer from 0 to `{num_classes - 1}`.

L Additional Qualitative Results

In this section, we provide additional qualitative results that complement the quantitative evaluation in the main paper. While scalar metrics summarize overall performance, they do not fully reveal how the agent interacts with surrounding pedestrians, vehicles, group members, and semantic scene regions over time. We therefore visualize representative closed-loop rollouts to better illustrate the behavioral differences between the baselines, ablated variants, and our full method.

Figure 10 shows qualitative examples for group-coherent navigation. The goal of this setting is not only to reach the target while avoiding nearby pedestrians, but also to preserve a socially plausible relation with assigned group members. An action that is feasible for an individual agent can still be undesirable if it separates the agent from its group, produces unsafe interactions with vehicles, or stretches the group formation toward road regions.

Across the shown examples, the baseline methods often generate trajectories that are individually plausible but less consistent at the group level. In some cases, the controlled agent drifts away from its assigned group. In others, it fails to anticipate nearby vehicle motion, resulting in unsafe interactions. We also observe cases where the distance among group members becomes excessively large, causing the group formation to expand into semantically inappropriate road regions. These examples suggest that the baselines do not consistently balance goal-directed navigation, interaction safety, group coherence, and scene compliance.

By contrast, our method keeps the controlled agent closer to the group flow while still making progress toward the goal. The resulting trajectories remain more compact, avoid unsafe interactions with vehicles, and reduce the tendency of group members to spread into road regions. These qualitative results support the GD improvements reported in the main paper and show that the proposed consequence-based preference refinement improves not only numerical group-distance error, but also visually coherent and semantically safer group behavior.

Figure 11 further visualizes the effect of the dual world model. Without this component, the agent tends to rely more heavily on short-term local decisions, which can make the resulting behavior less stable in complex scenes. In particular, the ablated model may react too late to surrounding agents

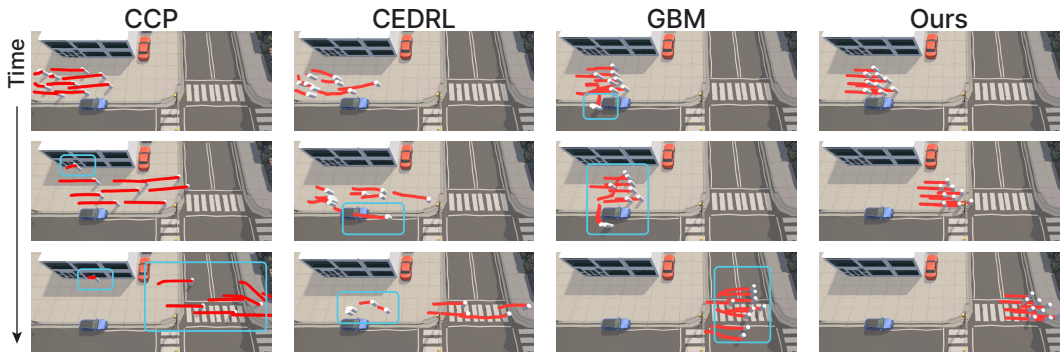


Figure 10: **Additional qualitative comparison on group-coherent navigation.** We compare CCP, CEDRL, GBM, and our method in a representative group-navigation scene. Each row shows a later timestep in the same rollout, and trajectory traces indicate recent motion over the previous 60 frames. The highlighted regions show representative failure modes of the baseline methods, including group separation, collision with a vehicle, and excessive group spread that causes group members to move into road regions. By contrast, our method maintains a more compact and coherent relation among group members while avoiding vehicles and road regions, producing trajectories that better preserve group-level motion and scene compliance.

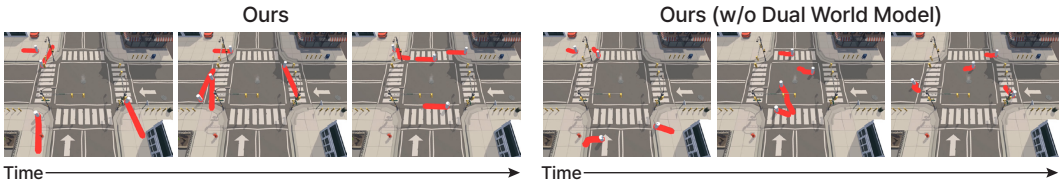


Figure 11: **Qualitative ablation on the dual world models.** We compare our full method with the variant without the dual world models in a representative closed-loop navigation scene; time runs from left to right. Without the dual world models, the agent shows typical failure cases such as less anticipatory interaction behavior, unstable trajectory adjustment, or reduced compliance with scene semantics. In contrast, the full model, refined with world-model-predicted consequences, produces more stable, safe, and scene-aware trajectories.

or dynamic objects, choose trajectories that are less compatible with the scene layout, or produce motion that appears less smooth over time.

The full method alleviates these issues by explicitly considering the predicted consequences of candidate actions during preference refinement. This allows the policy to better distinguish actions that are locally feasible from actions that remain safe and semantically appropriate over longer closed-loop rollouts. As a result, the full model produces trajectories that are more consistent with pedestrian motion, avoid unsafe interactions, and better respect scene constraints.

Together, these additional qualitative examples demonstrate that the proposed framework improves behavior beyond what is captured by individual metrics. The group-coherence examples show improved preservation of socially plausible group structure, while the ablation examples highlight the importance of the dual world model for anticipatory and scene-compliant navigation.

Figure 12 further shows that the proposed method can handle constrained interactions with dynamic vehicles. In this narrow-alley scenario, the baselines and the ablated variant often choose actions that appear locally plausible but enter the vehicle's future path. Our method instead anticipates this occupancy conflict and selects a safer side region, resulting in collision-free navigation through the alley.

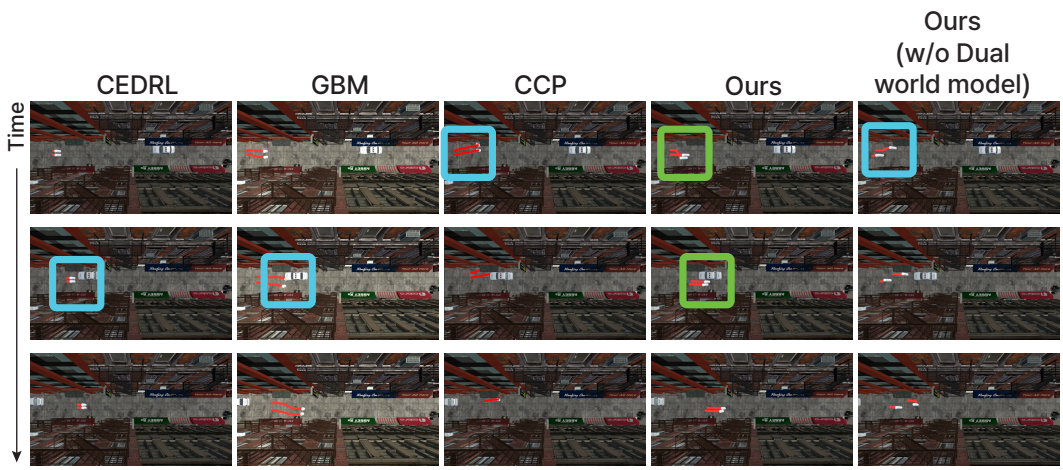


Figure 12: **Narrow-alley interaction with an approaching vehicle.** Blue highlights indicate collision-prone trajectories produced by existing methods and the ablated variant, while green highlights show the collision-free trajectory produced by our method.

M Comparison with Simpler Pipelines and an Online World-Model Planner

Table 10 reports the per-scene results summarized in Table 3. Each trajectory predictor is combined with an MPC planner following [33], Gen-Safe-Nav [45] integrates adaptive conformal inference into constrained reinforcement learning, and OnlineWM runs both world models at every decision. Our method achieves higher goal success and lower violation rates than all prediction-plus-planning pipelines and Gen-Safe-Nav in every scene, and stays close to OnlineWM across scenes.

N Input Ablation

To separate the contribution of the visual input from that of the hand-crafted state features, we train two variants: *State only* removes the image, and *Vision only* keeps the image and the goal but removes the other state features. As shown in Table 11, including vision generally improves GSR and SRVR, while the state features mainly help group coherence. Neither variant matches the full model, indicating that the two inputs are complementary.

O Standard ETH/UCY Forecasting Benchmark

To test whether our framework is tied to the Unity reconstructions, we retrain and evaluate it under the standard ETH/UCY forecasting protocol with leave-one-scene-out evaluation. Social-LSTM and Social-GAN numbers are taken from [15], and Trajectron++ numbers are those obtained after correcting an evaluation bug in its original implementation [43]. As shown in Table 12, our policy performs comparably to Trajectron++. Our goal is not state-of-the-art forecasting accuracy; the result indicates that the framework remains competitive outside our simulation.

Table 10: **Per-scene comparison with simpler pipelines and an online world-model planner.** Trajectory predictors are combined with an MPC planner [33]. Higher is better for GSR, while lower is better for SRVR and GD. Bold marks the best result in each scene among methods other than OnlineWM.

Scene	Method	GSR ↑	SRVR ↓	GD ↓
Zara01	OnlineWM	78.9	5.7	0.6
	Constant Velocity [36]	51.8	15.1	0.7
	Social-LSTM [1]	44.4	14.1	0.7
	Social-GAN [15]	44.4	15.0	1.0
	Trajectron++ [35]	48.2	13.8	0.8
	Gen-Safe-Nav [45]	37.0	19.1	1.3
	Ours	**72.2**	**7.0**	**0.6**
Zara01_Group	OnlineWM	77.8	0.8	1.7
	Constant Velocity [36]	25.9	18.1	6.7
	Social-LSTM [1]	33.3	19.3	5.7
	Social-GAN [15]	37.0	14.8	6.3
	Trajectron++ [35]	25.9	19.0	5.9
	Gen-Safe-Nav [45]	22.2	12.7	5.5
	Ours	**77.8**	**1.2**	**2.1**
Students003	OnlineWM	70.0	5.6	0.7
	Constant Velocity [36]	61.2	20.2	1.0
	Social-LSTM [1]	54.0	14.6	1.1
	Social-GAN [15]	62.7	18.7	1.0
	Trajectron++ [35]	60.4	20.7	1.0
	Gen-Safe-Nav [45]	32.7	11.6	1.3
	Ours	**66.0**	**7.2**	**0.7**
ETH_Intersection	OnlineWM	88.9	3.6	0.8
	Constant Velocity [36]	66.7	11.5	1.5
	Social-LSTM [1]	59.3	14.6	1.6
	Social-GAN [15]	59.3	12.5	1.5
	Trajectron++ [35]	66.7	12.8	1.5
	Gen-Safe-Nav [45]	51.9	17.3	2.2
	Ours	**83.3**	**3.9**	**0.9**
ETH_Hallway	OnlineWM	87.5	0.4	0.7
	Constant Velocity [36]	75.0	12.9	1.3
	Social-LSTM [1]	68.6	12.3	1.4
	Social-GAN [15]	62.5	14.1	1.3
	Trajectron++ [35]	70.8	18.8	1.2
	Gen-Safe-Nav [45]	70.8	9.9	1.4
	Ours	**87.5**	**0.9**	**0.8**

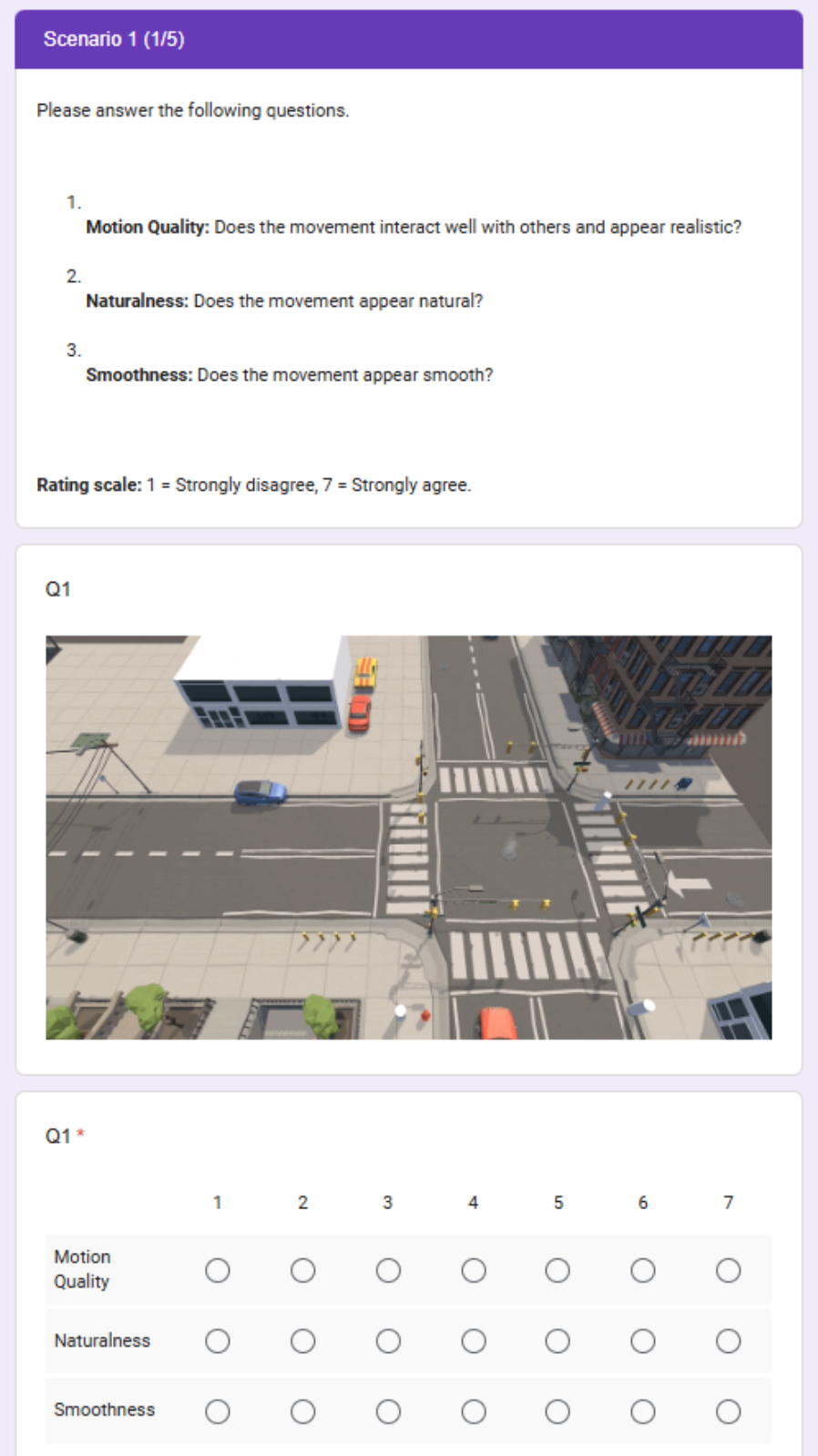


Figure 13: **Screenshot of the Google Form used for the user study.** The study was administered through Google Forms. Participants watched anonymized rollout videos and answered three 7-point Likert-scale questions evaluating motion quality, naturalness, and smoothness.

Table 11: **Input ablation on five held-out evaluation scenarios.** State only removes the visual input; Vision only removes the hand-crafted state features except the goal. Higher is better for GSR, while lower is better for SRVR and GD.

Scene	Variant	GSR ↑	SRVR ↓	GD ↓
Zara01	State only	44.4	13.1	**0.6**
	Vision only	61.2	10.9	0.8
	Ours	**72.2**	**7.0**	**0.6**
Zara01_Group	State only	44.4	12.3	2.2
	Vision only	44.4	6.4	2.3
	Ours	**77.8**	**1.2**	**2.1**
Students003	State only	50.0	12.9	**0.7**
	Vision only	54.0	8.7	0.8
	Ours	**66.0**	**7.2**	**0.7**
ETH_Intersection	State only	55.6	10.3	1.1
	Vision only	66.7	10.0	2.1
	Ours	**83.3**	**3.9**	**0.9**
ETH_Hallway	State only	62.5	1.6	1.0
	Vision only	68.8	1.1	1.6
	Ours	**87.5**	**0.9**	**0.8**

Table 12: **ETH/UCY forecasting benchmark (ADE/FDE in meters, lower is better).** Trajectron++ numbers are those after the evaluation bug fix [43].

	Most-likely			Best-of-20		
Dataset	Social-LSTM [1]	Trajectron++ [35]	Ours	Social-GAN [15]	Trajectron++ [35]	Ours
ETH	1.09/2.35	1.02/2.00	1.05/2.08	0.81/1.52	0.67/1.18	0.65/1.08
Hotel	0.79/1.76	0.33/0.62	0.37/0.68	0.72/1.61	0.18/0.28	0.25/0.30
Univ	0.67/1.40	0.53/1.19	0.40/1.02	0.60/1.26	0.30/0.54	0.23/0.50
Zara1	0.47/1.00	0.44/0.99	0.38/0.90	0.34/0.69	0.25/0.41	0.22/0.40
Zara2	0.56/1.17	0.32/0.73	0.35/0.79	0.42/0.84	0.18/0.32	0.20/0.38
Average	0.72/1.54	0.53/1.11	0.51/1.10	0.58/1.18	0.31/0.54	0.31/0.53

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by the proposed method, the closed-loop evaluation, the ablation study, and the user study. We explicitly evaluate goal success, semantic/social violations, group coherence, motion fidelity, and perceived human-likeness.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 5 and failure cases in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not introduce formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide sufficient details to reproduce the method, including the observation format, anchor vocabulary construction, supervised training objective, world-model objectives, GRPO refinement procedure, score function, hyperparameters, evaluation scenes, and metrics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The project page is publicly available at `https://seoneun.github.io/-NeurIPS-2026-Seeing-Beyond-the-Next-Step/`. The reconstructed Unity scenes are built on commercially licensed assets that cannot be redistributed, so they are not released; the ETH/UCY trajectories are available from their original sources.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the experimental setup, datasets, evaluation scenarios, baselines, metrics, training hyperparameters, and implementation details. We also provide additional appendix details for the score function, segmentation classes, world-model diagnostics, and runtime/compute cost.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report results over multiple runs and include uncertainty estimates where applicable. Motion fidelity results are reported with 95% confidence intervals, and the closed-loop experiments are conducted across multiple evaluation scenarios and repeated runs.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report implementation and compute-related details, including model backbone, LoRA settings, training hyperparameters, denoising steps, world-model runtime, and compute cost. We also clarify that world models are used only during training-time refinement and are removed at inference.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The current version does not include a dedicated broader-impact discussion. The work focuses on human-like navigation in reconstructed simulation environments and does not make claims about direct deployment in real-world safety-critical systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk foundation model, autonomous real-world control system, or other artifact that requires additional safeguards beyond the reported simulation-based evaluation.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the datasets, pretrained models, and baseline methods used in the experiments. We also respect licensing constraints for external assets and do not redistribute assets that cannot be publicly released.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not release new third-party assets. The reconstructed Unity scenes rely on external assets whose licenses do not permit redistribution, so the assets themselves will not be released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We describe the user study protocol in Appendix G, including the number of participants, anonymized rollout videos, randomized presentation order, and 7-point Likert-scale questions for motion quality, naturalness, and smoothness. A screenshot of the survey interface is shown in Figure 13, and each participant received USD $7 as compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The user study was approved by the Korea University Institutional Review Board (KUIRB-2026-0115-01), as stated in Appendix G. The study involved only watching simulation videos and answering rating questions, posing minimal risk to participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The method uses a pretrained vision-language model as the policy backbone, which is adapted into a vision-language-action policy through supervised training and GRPO-based refinement. We clearly describe how this model is used as part of the proposed method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.