Ordinal Neural Collapse as a Representation Prior for Visual Navigation

E-In Son*, Jung-Taak Kim*, and Seung-Woo Seo†

Dept. of ECE & INMC, Seoul National University
{pingpang,mychoco333,seo}@snu.ac.kr,

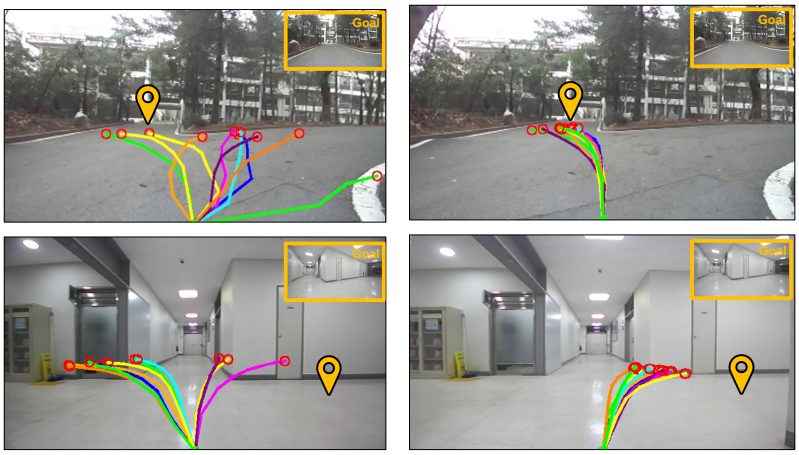
Abstract. Learning robust navigation policies directly from visual observations remains a fundamental challenge in vision-based robotic navigation. In end-to-end imitation learning approaches, the visual encoder and action decoder are jointly optimized using a single action loss, which provides only an indirect supervisory signal to the encoder. This indirect supervision frequently results in the encoder learning ambiguous, action-agnostic representations. The problem is further complicated by substantial variations in scene structure and appearance across diverse environments, as well as the prevalence of visual distractors inherent to real-world navigation settings. Such action-agnostic features cause the navigation policy to produce inconsistent actions at ambiguous decision points, leading to navigation failure. To overcome these limitations, we propose ORION (Ordinal Neural Collapse for Visual Navigation), a method that explicitly organizes the encoder's representation space according to the ordinal structure of navigation actions. In the context of goal-directed navigation, ego-centric control categories from *Far Left* to *Far Right* exhibit a natural ordinal relationship in which neighboring classes share similar visual contexts, while semantically opposing classes differ substantially in appearance. We encourage class representations to be arranged sequentially along a single discriminative axis, while suppressing off-axis variance within each class. The pretrained encoder is then integrated into a diffusion-based navigation framework, and the full pipeline is fine-tuned end-to-end. Extensive experiments in both simulation and real-world settings show that ORION consistently outperforms end-to-end and neural collapse baselines in navigation success rate and goal progress, with notable gains in visually challenging scenarios such as complex multi-way intersections.

Keywords: Goal-Conditioned Visual Navigation · Neural Collapse · Representation Geometry

1 Introduction

Recently, end-to-end imitation learning has become a dominant approach for vision-based navigation. By jointly optimizing a vision encoder and an action

* Equal contribution.
† Corresponding author.



(a) NoMaD (Baseline) **(b)** ORION (Ours)

Fig. 1: Structured representations improve action consistency at intersections. Given the same goal image, (a) the NoMaD baseline generates candidate waypoints scattered across multiple directions, while (b) our method concentrates them toward the goal direction. Each colored trajectory represents one action candidate sampled from the action decoder.

decoder from expert demonstrations, recent systems have achieved strong performance across diverse environments [2, 8, 26]. Diffusion-based action decoders have further improved performance by capturing multimodal action distributions [2, 27].

However, relying solely on a single action loss provides limited learning signals for the vision encoder. Since the objective is to imitate actions, the training loss acts on the final action output rather than supervising the intermediate representation. Prior work has shown that encoders trained this way often latch onto visually salient but task-irrelevant features [28,34]. In visual navigation, this problem is worse because training data spans diverse environments with wide variation in layout, lighting, and terrain [3, 19]. The encoder must compress this visual diversity into a shared latent space, yet receives no explicit signal to organize its representations around control-relevant structure. As a result, the learned representations lack consistent action-related semantics, leading to degraded performance in novel visual conditions. Fig. 1a illustrates this effect. At an intersection, the baseline policy produces candidate paths scattered across multiple directions rather than consistently toward the goal. This suggests that the model focuses on imitating expert actions, while the learned representation may not sufficiently capture goal-relevant visual structure.

Recent work has explored structuring visual representations to be *control-aware*, decoupling representation learning from policy learning through staged training [14, 18, 20, 28, 34]. A particularly compelling perspective comes from *neural collapse* (NC) [21], which characterizes the terminal phase of training in

deep classifiers: features of the same class converge to their class mean, and the class means form a symmetric simplex geometry. Notably, Qi *et al.* [23] observed that neural collapse also emerges in vision-based control tasks formulated with discrete control classes, and showed that explicitly encouraging this structure through encoder pretraining improves downstream policy performance. Their work raises a natural question: what is the appropriate class definition and geometric structure for a visual navigation task?

In navigation, the robot's immediate control action is largely determined by where it needs to go. This intuition motivates the use of goal-relative directions as a natural basis for defining discrete control classes, which can be divided into ordered categories ranging from *Far Left* to *Far Right* [22]. Unlike categorical control classes where all pairs are treated as equally spaced, goal-relative classes are naturally more similar to neighboring classes than to distant ones. Neighboring classes correspond to nearby steering commands and share similar visual contexts, while opposing classes represent different navigational situations. We leverage this natural ordering as a prior structure for the learned representation space. Recent theoretical analysis has shown that ordinal classes naturally collapse onto a one-dimensional ordered subspace, rather than the simplex geometry assumed in standard neural collapse [17].

Building on this, we propose **ORION** (Ordinal Neural Collapse for Visual Navigation), which explicitly imposes ordinal geometric structure on a vision encoder's representation space for continuous visual control. To our knowledge, this represents the first attempt to leverage ordinal neural collapse geometry for visual navigation.

Rather than restricting the encoder's capacity, our method provides a structured initialization from which downstream end-to-end training can refine the full range of navigation-relevant features. Specifically, we develop a supervised ordinal projection that aligns class means along a learned axis, paired with a compactness objective that suppresses variance perpendicular to the axis. We build on the NoMaD [27] navigation framework, which combines a vision encoder with a Transformer context module, and a diffusion action decoder [2]. Our approach operates in two stages: we first pretrain the vision encoder using ordinal neural collapse objectives on direction-labeled navigation data, then fine-tune the entire pipeline end-to-end with the standard diffusion policy loss.

In simulation and real-world navigation experiments, ORION demonstrates improved robustness compared to the end-to-end trained baseline, particularly in novel environments and visually ambiguous scenarios such as intersections with similar visual features.

Our contributions are as follows:

- We identify that goal-relative navigation classes have inherently non-uniform inter-class similarity, forming a natural ordinal structure. Existing control-oriented methods overlook this property by assuming uniform class separation.

- We propose ORION, a pretraining method for ordinal control classes, comprising supervised ordinal projection and orthogonal variance suppression, that enforces ordinal alignment while collapsing representations toward the ordinal axis.
- We evaluate our method in simulation and real-world navigation experiments and show that ORION improves the success rate by up to +26 pp and reduces heading jerk by up to 71% over the NoMaD baseline.

2 Related Work

2.1 Representation Learning for Robot Control

While end-to-end imitation learning jointly optimizes perception and control, it often yields encoders that overfit to task-irrelevant visual distractors rather than actionable structures [6, 11, 34]. This has motivated staged training approaches that decouple representation learning from policy optimization through contrastive objectives [14, 28], task-aware pretraining on large-scale data [20], or systematic benchmarking of pretrained visual backbones for embodied tasks [18]. These works establish that control-relevant representations improve downstream performance, but they do not prescribe a specific target geometry for the representation space.

2.2 Neural Collapse and Control-Oriented Clustering

Neural Collapse in Classification. Papyan *et al.* [21] discovered that deep classifiers in their terminal training phase exhibit a striking regularity: within-class features collapse to their class mean, and the class means converge to a Simplex Equiangular Tight Frame (ETF) with uniform pairwise cosine similarity $\cos\theta = -1/(K-1)$. This geometry has been theoretically grounded [35, 36] and extended to imbalanced settings [4, 9]. Critically, NC has evolved from an observation into a prescriptive tool, where explicitly enforcing NC geometry improves robustness and generalization in downstream tasks [15, 32].

Control-Oriented Clustering. Qi *et al.* [23] brought this prescriptive perspective to visual control. They observed that NC-like clustering spontaneously arises in behavior cloning when features are labeled by action-derived classes, and proposed Relative Pose Orthants (REPOs)—a multi-dimensional binary partition yielding 2^d categorical classes—as control-oriented labels. Pretraining encoders to minimize NC metrics improved downstream policy performance by 10–35% across multiple tasks. Because REPO classes are categorical with no ordinal relationship, the Simplex ETF geometry is structurally appropriate: all class pairs *should* be equidistant.

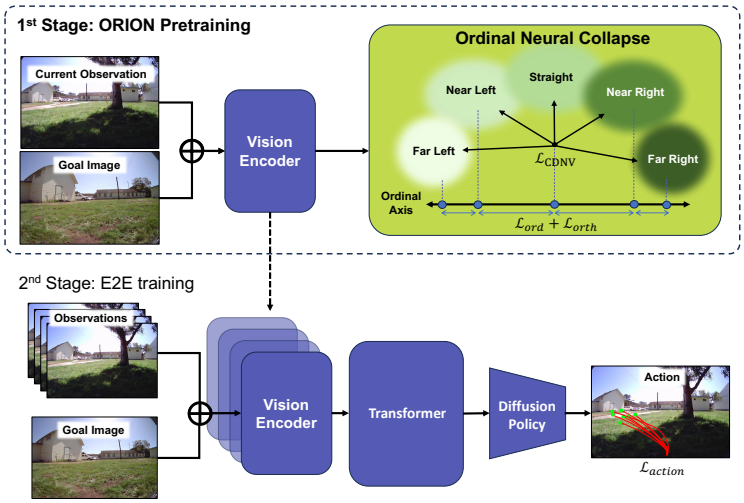


Fig. 2: Overall pipeline of ORION. In Stage 1, ORION pretraining aligns the encoder feature space with the ordinal relationships among navigation actions. In Stage 2, the pretrained encoder is plugged into the NoMaD pipeline and fine-tuned end-to-end with the diffusion policy loss.

Ordinal Neural Collapse. Ma *et al.* [17] proved that in ordinal regression, class means collapse onto a one-dimensional subspace ordered by class rank, rather than forming a Simplex ETF. Their analysis relies on cumulative probabilities and threshold parameters specific to ordinal regression; we adopt the geometric insight—ordinal alignment on a 1D subspace—as an inductive bias, not the theoretical framework itself (Section 3).

2.3 Learning-Based Visual Navigation

Learning-based navigation has progressed from classical modular pipelines [30] and early visuomotor policies [3, 19, 22] toward scalable, data-driven approaches. ViNT [26] trained a Transformer-based model on diverse multi-robot data for topological goal-reaching. NoMaD [27] extended this with a diffusion policy decoder [2] that represents multimodal action distributions, unifying goal-conditioned navigation and exploration through goal masking. NoMaD architecture serves as our baseline. Subsequent work has further refined the diffusion-based action decoder through cost-guided sampling [33] and bridge models with informative priors [24].

Despite advances in policy architecture and sampling strategy, the encoder in these systems is trained end-to-end without explicit structure in the representation space. Hong *et al.* [10] addressed this partially through semantic map supervision for indoor navigation, but no prior work has organized encoder representations to reflect the ordinal structure of the navigation action space. Our approach fills this gap by pretraining the vision encoder with ordinal alignment objectives before policy learning.

3 Method

3.1 Problem Setup and Overview

We consider image-goal navigation via imitation learning: given an observation o_t and a goal image g, the agent predicts a continuous action $a_t \in \mathbb{R}^d$. We build on the NoMaD architecture [27], which decomposes the policy into a vision encoder f_θ that maps each observation and the goal to a per-frame feature $h_t = f_\theta(o_t, g) \in \mathbb{R}^D$, a Transformer context module T_ϕ that temporally aggregates the per-frame features into $z_t = T_\phi(h_{t-p:t})$, and a diffusion decoder D_ψ [2] that generates actions conditioned on z_t.

When such an architecture is trained end-to-end, the encoder is supervised only indirectly through the action denoising loss. Our approach adds a pretraining stage before this end-to-end pipeline. In Stage 1, we train only the encoder f_θ using the Ordinal Neural Collapse (ONC) objectives, which shape the geometry of the feature space to better align with the structure of visual navigation tasks. In Stage 2, the pretrained encoder is integrated into the full NoMaD pipeline and fine-tuned end-to-end with the diffusion policy loss. We emphasize that pretraining provides a structured geometric initialization; it does not constrain what the encoder learns during downstream training. The overall pipeline is illustrated in Fig. 2.

3.2 Goal-Relative Control Classes

To define supervisory structure for representation shaping, we first discretize the continuous action space into ordered classes based on the angular direction toward the goal. Let $\alpha_t \in [-\pi, \pi)$ denote the relative angular offset between the agent's heading and the goal direction. We partition the front field of view into K uniform angular bins and assign each sample to a class $y_t \in \{0, \ldots, K-1\}$; if the goal lies outside the field of view, the sample is assigned to a separate out-of-view class $y_t = K$, yielding $K + 1$ classes in total.

These classes have a natural ordinal structure: class 0 corresponds to the far right goal direction, class $K - 1$ to the far left, and intermediate classes are ordered between them. Consequently, neighboring classes share similar steering contexts, while distant classes represent different behaviors. The out-of-view class is excluded from the ordinal objectives, which assume an ordered relationship between classes. It contributes only to $\mathcal{L}_{\mathrm{CDNV}}$, which does not rely on any ordinal assumption.

3.3 ORION Pretraining Objectives

We propose ORION pretraining loss, which consists of three components: class separability, supervised ordinal projection, and orthogonal compactness. Let μ_k denote the mean feature of class k, $\bar{\mu}$ the global mean, $\tilde{\mu}_k := \mu_k - \bar{\mu}$ the centered class mean, and σ_k^2 the within-class variance.

Class Separability. To encourage compact class clusters, we minimize the class-distance normalized variance (CDNV) [5], which measures intra-class compactness relative to inter-class separation:

$$\mathcal{L}_{\text{CDNV}} = \frac{1}{|\mathcal{P}|} \sum_{(k,k')\in\mathcal{P}} \frac{\sigma_k^2 + \sigma_{k'}^2}{2\,\|\tilde{\mu}_k - \tilde{\mu}_{k'}\|_2^2}, \tag{1}$$

where $\mathcal{P} = \{(k, k') \mid k < k'\}$ is the set of all unordered class pairs.

Supervised Ordinal Projection. Our method aligns class means along a one-dimensional axis according to their ordinal order. This design is motivated by ONC theory [17], which shows that under ordinal regression, class means collapse onto a one-dimensional subspace ordered by class rank.

We assign each class an equally spaced target position $t_k = k - (K-1)/2$ for $k \in \{0, \ldots, K-1\}$, and define the supervised ordinal axis as the direction that best aligns the centered class means with the target ordering:

$$w = \frac{\tilde{M}^\top \mathbf{t}}{\|\tilde{M}^\top \mathbf{t}\|_2}, \tag{2}$$

where $\tilde{M} = [\tilde{\mu}_0, \ldots, \tilde{\mu}_{K-1}]^\top \in \mathbb{R}^{K\times D}$ is the matrix of centered class means and $\mathbf{t} = [t_0, \ldots, t_{K-1}]^\top$. The ordinal projection loss then encourages the projected class means $p_k = \tilde{\mu}_k^\top w$ to match the targets:

$$\mathcal{L}_{\text{ord}} = \frac{1}{K} \sum_{k=0}^{K-1} (p_k - t_k)^2 . \tag{3}$$

Unlike PCA-based approaches that identify the principal axis without label information, our supervised projection directly incorporates the known ordinal structure, guaranteeing zero ordinal violations by construction.

Orthogonal Compactness. To concentrate features onto the ordinal subspace, we penalize within-class variance in the directions orthogonal to the ordinal axis:

$$\mathcal{L}_{\text{orth}} = \frac{1}{N} \sum_{n=1}^{N} \left\| (I - ww^\top)(h_n - \mu_{y_n}) \right\|^2 , \tag{4}$$

where w is the supervised ordinal axis defined in Eq. (2), and N is the number of (non-OOV) samples in the mini-batch. While $\mathcal{L}_{\text{ord}}$ positions class means along w, $\mathcal{L}_{\text{orth}}$ collapses the orthogonal directions, encouraging features to lie on the ordinal axis. This geometric constraint applies only during pretraining. Downstream end-to-end fine-tuning is free to recover orthogonal variation as needed for the full navigation task.

Algorithm 1 ORION Pretraining and Downstream Fine-tuning

Require: Dataset $\mathcal{D} = \{(o_t, g, a_t, y_t)\}$, encoder f_θ, context module T_ϕ, diffusion decoder D_ψ

Stage 1: ORION Pretraining (train f_θ only)

Initialize running class means $\{\mu_k\}$, variances $\{\sigma_k^2\}$, axis w
for each mini-batch $\mathcal{B}$ **do**
  $h_t \leftarrow f_\theta(o_t, g)$ for all $(o_t, y_t) \in \mathcal{B}$
  Update $\{\mu_k\}$, $\{\sigma_k^2\}$ via Welford's algorithm
  Compute centered class means $\tilde{\mu}_k \leftarrow \mu_k - \bar{\mu}$
  Compute $\mathcal{L}_{\text{CDNV}}$ ▷ Eq. (1)
  Compute supervised axis: $w \leftarrow \text{EMA}\Big(w,\ \tilde{M}^\top \mathbf{t} / \|\tilde{M}^\top \mathbf{t}\|\Big)$
  Compute $\mathcal{L}_{\text{ord}}$ from projections $p_k = \tilde{\mu}_k^\top w$ ▷ Eq. (3)
  Compute $\mathcal{L}_{\text{orth}} = \frac{1}{N}\sum_{n=1}^{N} \left\|(I - ww^\top)(h_n - \mu_{y_n})\right\|^2$ ▷ Eq. (4)
  Update θ by $\nabla_\theta \mathcal{L}_{\text{ORION}}$ ▷ Eq. (5)
end for

Stage 2: Fine-tune f_θ, T_ϕ, D_ψ end-to-end with diffusion policy loss [27].

Overall Objective. The full pretraining loss is:

$$\mathcal{L}_{\text{ORION}} = \lambda_{\text{cdnv}}\, \mathcal{L}_{\text{CDNV}} + \lambda_{\text{ord}}\, \mathcal{L}_{\text{ord}} + \lambda_{\text{orth}}\, \mathcal{L}_{\text{orth}}, \tag{5}$$

where $\mathcal{L}_{\text{CDNV}}$ promotes class separability, $\mathcal{L}_{\text{ord}}$ enforces ordinal alignment along a learned axis, and $\mathcal{L}_{\text{orth}}$ reduces orthogonal variation to reinforce one-dimensional structure. The full training procedure is shown in Alg. 1.

3.4 Training Considerations

Encoder-level Application. We apply ORION pretraining to the encoder features h_t rather than the context-level features z_t. The context module integrates temporal and goal information through attention. Applying geometric constraints at this stage disrupts temporal aggregation and results in unstable feature norms. Additional experimental results are included in the supplementary material.

Stable Statistics Estimation. The pretraining objectives depend on class means and variances, which can be noisy when estimated from single mini-batches. We maintain running statistics using Welford's online algorithm [31] and smooth the ordinal axis w via exponential moving average across batches. Gradient flow is preserved by computing losses from current-batch features centered with the incrementally updated statistics (details in supplementary material).

4 Experiments

4.1 Experimental Setup

Training Data and Control Classes. We use the same training dataset as NoMaD [27], which consists of trajectories collected from diverse environments and robotic platforms. The dataset includes RECON [25], SCAND [12], GoStanford [8], and SACSoN [7], totaling approximately 140 hours of navigation time. Goal images are sampled from future observations along the same trajectory, drawn 3–20 steps ahead of the current observation during training. Each sample is labeled with a goal-relative class based on the angular offset α_t between the agent's heading and the goal. We partition the front field of view $[-90°, 90°]$ into $K = 5$ uniform angular bins and assign samples outside this range to an out-of-view class, yielding six classes in total. This explicit out-of-view class captures goal-irrelevant states, allowing the encoder to convey such conditions to downstream control.

Training. In Stage 1, we pretrain the EfficientNet-B0 [29] encoder for 30 epochs using AdamW [16] with a learning rate of 10^{-5}, cosine scheduling, and batch size 256. Loss weights are $\lambda_{\text{cdnv}} = 1.0$, $\lambda_{\text{ord}} = 1.0$, $\lambda_{\text{orth}} = 1.0$. In Stage 2, we integrate the pretrained encoder into the full NoMaD pipeline and fine-tune all parameters end-to-end for 30 epochs with the diffusion policy loss, following the NoMaD defaults [27]: a 1D conditional U-Net decoder with 10 denoising steps, trajectory prediction horizon of 8, and learning rate 10^{-4}.

Baselines. We compare our method with the following baselines:

- **ViNT** [26]: A goal-conditioned policy trained end-to-end using a single action regression loss.
- **NoMaD** [27]: The diffusion-based policy trained end-to-end without encoder pretraining.
- **NC-ETF**: Encoder pretraining based on the Simplex ETF geometry proposed by Qi *et al.* [23]. We adapt this formulation to the visual navigation setting, enforcing three standard NC metrics (CDNV, STD Norm, STD Angle).

Evaluation Metrics. We report success rate, progress, collision, and heading jerk.

- **Success Rate (SR):** the fraction of trials in which the agent reaches the goal. The goal-reaching distance tolerance is set to 1.0 m for indoor, and 5.0 m for outdoor. The time limit is set to 300 s per episode.
- **Progress (Prog.):** the ratio of visited visual subgoal nodes to the total number of subgoal nodes along the route, capturing partial task completion in failed episodes.

Table 1: Navigation performance in simulated environments. We evaluate in Indoor (2D-3D-S) and Outdoor (Gazebo Citysim) environments. **Bold** denotes the best result per route.

Environment	Route	Method	*Task Performance*			*Traj. Quality*
			SR (%) ↑	**Prog.** (%) ↑	**Collision** ↓	**H-Jerk (med.)** ↓
Indoor (2D-3D-S)	Basic	ViNT [26]	76	74.58	0.24	4.12
		NoMaD [27]	92	95.17	0.02	4.61
		NC-ETF	86	89.32	0.14	4.33
		ORION	**96**	**96.68**	**0.04**	**1.58**
	Intersection	ViNT [26]	0	36.12	1.00	4.99
		NoMaD [27]	17	32.07	0.81	15.48
		NC-ETF	12	25.48	0.88	12.13
		ORION	**91**	**93.82**	**0.09**	**4.99**
Outdoor (Citysim)	Basic	ViNT [26]	45	73.58	1.48	1.22
		NoMaD [27]	88	89.85	0.12	5.08
		NC-ETF	89	91.18	0.21	2.38
		ORION	**94**	**94.68**	**0.06**	**1.09**
	Multi Intersection	ViNT [26]	16	44.24	2.16	**2.34**
		NoMaD [27]	48	70.30	1.80	34.49
		NC-ETF	40	68.57	1.39	11.98
		ORION	**74**	**86.39**	**0.45**	10.13

- **Collision:** the mean number of physical contacts with obstacles per episode, indicating navigation safety.
- **Heading Jerk (H-Jerk):** the median absolute change in heading angle between consecutive time steps. This metric captures short-horizon directional stability and is particularly diagnostic at intersections, where temporally inconsistent encoder features manifest as rapid heading oscillations.

4.2 Simulation Results

We evaluate our method in two simulated environments: (i) Stanford 2D-3D-S [1], an indoor environment with narrow corridors, and (ii) a Gazebo-based urban simulation (Citysim) [13]. Each environment includes two test routes with $N = 100$ trials per route. All methods are deployed on a Clearpath Husky UGV.

Overall Performance. Across both simulation environments, ORION consistently outperforms the baselines in terms of success rate and progress. (Tab. 1) On *Basic*, which is mostly straight with a single intersection, all diffusion-based methods achieve high success rates, with ORION showing moderate improvement over NoMaD (e.g., 94 vs. 88 SR on Citysim *Basic*). Performance differences become more pronounced on *Multi-Intersection* spanning approximately 200 m and requiring four consecutive directional decisions, where ORION outperforms

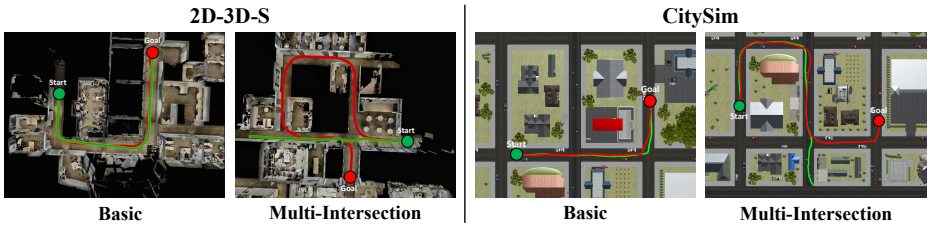


Fig. 3: Qualitative comparison in simulation. Green and red dots denote the start and goal positions, respectively. Trajectories are color-coded as NoMaD (green) and ORION (red).

NoMaD in success rate by a margin of +26 pp (48% to 74%). ViNT performs worse than both diffusion-based methods across settings.

Intersection-Level Behavior. Fig. 3 compares top-view trajectories of NoMaD and ORION at an intersection. NoMaD exhibits temporal flickering: the diffusion decoder alternates between competing directions across successive time steps. ORION instead concentrates on candidate paths toward a single direction, producing more consistent trajectories and reducing heading jerk by 71% over NoMaD (34.49 → 10.13 on *Multi-Intersection*).

Geometric Interpretation. NC-ETF, which enforces Simplex ETF geometry during pretraining, performs comparably to or below the NoMaD baseline despite using the same two-stage training procedure as ORION. This indicates that representation pretraining with a mismatched geometric structure does not improve navigation and can even degrade performance.

Failure Analysis. ORION failures on intersection routes fall into two categories: (i) premature turning, where the agent initiates a turn too early and collides with obstacles at the intersection boundary; and (ii) branch misjudgment, where the agent selects an incorrect direction. Premature turning is also observed in NoMaD and mainly arises from imprecise turn timing rather than representation quality. Branch misjudgment is the dominant failure mode of NoMaD; ORION reduces its frequency but does not eliminate it.

4.3 Real-World Results

We evaluate our method in three complex real-world environments: *Indoor, Campus, and Field.* We conduct 10 evaluation episodes per route on *Indoor* and *Campus*. On *Field* environment, due to longer travel distances and reduced controllability of environmental factors, we conduct 6 trials per route. All policies are deployed on a Clearpath Husky UGV with a ZED 2i stereo camera (left RGB stream only) and an NVIDIA Jetson Orin for onboard inference.

Table 2: Navigation performance in real-world environments. We evaluate in Indoor, Campus, and Field environments. **Bold** denotes the best result per route.

	Indoor			Campus			Field		
Method	SR↑	Prog.↑	H-Jerk↓	SR↑	Prog.↑	H-Jerk↓	SR↑	Prog.↑	H-Jerk↓
ViNT	20	56.97	**2.34**	40	52.15	**7.76**	0	36.70	**8.51**
NoMaD	40	57.19	5.09	60	73.95	13.67	16.67	37.41	20.77
ORION	**60**	**92.12**	4.64	**70**	**80.61**	12.17	**33.33**	**48.99**	13.45

Overall Performance. Across all environments, our method achieves higher success rates and progress compared to the baselines in real-world experiments (Tab. 2).

Indoor is particularly challenging due to visually repetitive corridors where all choices appear nearly identical. Here, ORION achieves a higher success rate than both baselines, better distinguishing similar visual features.

On *Campus*, the same intersection instability observed in the simulation also appears. Both ViNT and NoMaD often make inconsistent directional decisions at visually ambiguous junctions. In contrast, ORION exhibits more stable behavior at intersections, leading to more reliable navigation outcomes.

Field environment involves long-range navigation (≈200m) over uneven terrain with sparse visual features. In this setting, all methods show relatively low success rates. Nevertheless, ORION consistently achieves higher progress toward the goal, indicating more stable long-horizon navigation without drifting away from the intended direction.

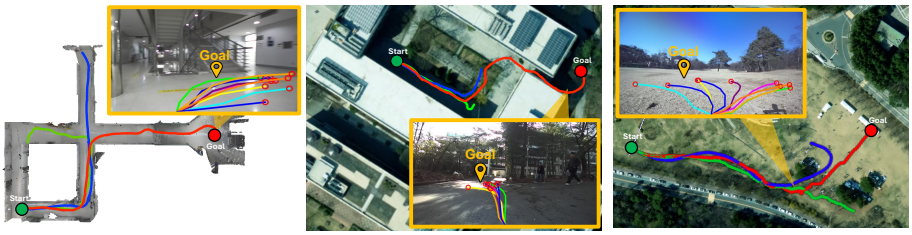
Qualitative Analysis. Fig. 4 shows qualitative comparisons across the three environments.

On *Indoor*, our policy frequently attempts to avoid the staircase obstacle, occasionally failing to reach the goal as a result. This obstacle avoidance behavior, which is not explicitly supervised during pretraining, emerges from the end-to-end fine-tuning stage. The robot generally follows the subgoal sequence and achieves high progress, with failures caused by physical constraints rather than navigation errors.

On *Campus*, baseline policies show temporal flickering near intersections, consistent with simulation observations. ORION trajectories show more consistent path selection at these junctions (See first row of Fig. 1).

Field presents a large open space where many feasible directions exist, making it easy for the robot to drift from the intended route. ORION is able to recover and return toward the route even after temporary deviations, a behavior not observed with the baselines.

Notably, ViNT achieves the lowest H-Jerk across all environments. However, this mainly reflects the limitation of its regression-based policy. By averaging over multimodal action distributions, the policy avoids decisive turns. As a result, ViNT produces smooth but less responsive trajectories, which leads to the



(a) Indoor **(b) Campus** **(c) Field**

Fig. 4: Real-world qualitative results. Green and red dots denote start and goal positions, respectively. Trajectories are color-coded as ViNT (blue), NoMaD (green), and ORION (red). The orange frames show example ego images and their corresponding sampled paths in each trajectory.

lowest success rates. In contrast, ORION achieves lower H-Jerk than NoMaD despite using the same diffusion-based decoder. This suggests that the pretrained encoder helps the policy make clearer goal-relevant decisions, leading to more decisive actions. Additional qualitative results are provided in the supplementary material.

4.4 Ablation Studies

We ablate three key design choices underlying ORION: its representation geometry, the individual loss terms, and the two-stage training structure.

Geometry vs. Direction Labels. To isolate whether ORION's gain stems from its ordinal-collapse geometry or merely from adding a direction-aware pretraining stage, we compare against CE-only pretraining: cross-entropy classification over the same K=5 directional bins, followed by the identical Stage-2 protocol. As shown in Tab. 3a, CE-only reaches only 16% SR on *Multi-Intersection*, below the no-pretraining NoMaD baseline (48%), and yields the highest heading jerk. Direction supervision here is not merely insufficient but actively harmful, as forcing directional labels through a mismatched classification objective degrades the representation rather than structuring it. ORION's gains therefore require its task-matched ordinal-collapse objectives, not direction supervision alone.

Loss Components. Tab. 3b reports ablation results with individual loss terms removed. With $\mathcal{L}_{\text{CDNV}}$ alone, the model achieves only 33% SR. Adding $\mathcal{L}_{\text{ord}}$ improves SR to 45% by ordering class means along a 1D axis. The full model reaches 74% SR, as $\mathcal{L}_{\text{orth}}$ tightens within-class distributions around the axis that $\mathcal{L}_{\text{ord}}$ defines. $\mathcal{L}_{\text{orth}}$ requires $\mathcal{L}_{\text{ord}}$ to be present, as it reduces orthogonal variance relative to the axis that $\mathcal{L}_{\text{ord}}$ constructs.

Table 3: Pretraining design ablations. We evaluate on Citysim *Multi-Intersection*, reporting a mean over 100 episodes.

(a) Geometry vs. direction labels.

Method	SR↑	Prog.↑	H-Jerk↓
NoMaD	48	70.30	34.49
CE-only	16	38.35	54.63
ORION	74	86.39	10.13

(b) Loss component contributions.

Variant	SR↑	Prog.↑	H-Jerk↓
$\mathcal{L}_{\text{CDNV}}$ only	33	55.54	19.41
$\mathcal{L}_{\text{CDNV}} + \mathcal{L}_{\text{ord}}$	45	69.16	42.98
ORION (full)	74	86.39	10.13

Two-Stage vs. Joint Training. Tab. 4 compares two-stage ORION pretraining against NC-Joint, which applies ORION losses as auxiliary objectives during end-to-end training from the start. NC-Joint fails entirely, achieving 0% success rate with H-Jerk an order of magnitude higher than NoMaD (147.95 vs. 5.08). Qualitatively, the NC-Joint policy generates paths that spread radially regardless of goal direction, failing to concentrate toward the target. This indicates that the ORION and diffusion objectives interfere destructively when optimized simultaneously: neither the encoder's ordinal structure nor the decoder's action prediction converges properly under conflicting gradients. Two-stage pretraining resolves this by separating the two objectives temporally, allowing the encoder to establish geometric structure in Stage 1 before the diffusion objective is introduced in Stage 2. (See supplementary material for further details)

Table 4: Two-stage vs. joint training. We evaluate on the Citysim *Basic*, reporting a mean over 100 episodes.

Variant	SR ↑	Prog. ↑	H-Jerk ↓
NoMaD	88	89.85	5.08
NC-Joint	0	10.34	147.95
ORION (Two-Stage)	94	94.68	1.09

4.5 Representation Analysis

Section 4.4 shows that joint training fails to establish ordinal structure. A natural question is whether Stage 2 fine-tuning preserves the structure established during pretraining. Fig. 5 addresses this question by comparing class distributions projected onto the supervised ordinal axis after full training (Stage 1 + Stage 2). Both models achieve correct ordinal ordering by mode, indicating that end-to-end training itself induces partial ordinal structure. However, the degree of within-class concentration differs substantially. In NoMaD (Fig. 5b), class distributions are broad with heavy cross-class overlap near the center of the axis. In ORION (Fig. 5c), each class occupies a more compact region with Far–Near

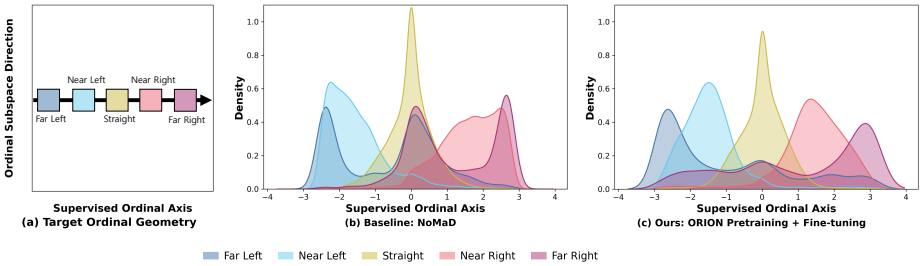


Fig. 5: Class distributions on the supervised ordinal axis after Stage 2 fine-tuning. (a) Target ordinal geometry. (b) NoMaD: Far and Near classes nearly overlap at both extremes. (c) ORION: all five classes maintain distinct modes. Projections are normalized to zero mean and unit variance. Distributions are computed on the held-out RECON [25] test split using Gaussian kernel density estimation (bandwidth $h = 0.15$). Residual overlap near boundaries reflects bin-boundary ambiguity, where visually similar samples across an angular bin boundary receive different labels.

mode separations $6-9\times$ larger than in NoMaD (0.37–0.53 vs. < 0.06 in normalized units). This confirms that the ordinal structure from Stage 1 persists through 30 epochs of end-to-end fine-tuning, validating the two-stage design.

5 Conclusion

We present ORION (Ordinal Neural Collapse for Visual Navigation) pretraining framework that embeds ordinal geometric structure into the vision encoder's representation space for image-goal navigation. ORION pretraining shapes the encoder feature space to reflect the ordinal similarity of steering decisions, providing a structured initialization for the downstream diffusion policy. Experiments in both simulation and real-world deployments show that ORION consistently improves navigation success rates over end-to-end and NC baselines, with particular gains at visually ambiguous intersections. Our results highlight the importance of structured representations for interpreting visual observations in end-to-end navigation models. We believe this principle can extend to other control tasks where actions exhibit inherent structure.

Acknowledgements

This work was supported by the Challengeable Future Defense Technology Research and Development Program through the Agency for Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) (No. 915108201), and the BK21 FOUR Program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2026.

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. In: Proceedings of Robotics: Science and Systems (RSS) (2023)
3. Codevilla, F., Müller, M., López, A., Koltun, V., Dosovitskiy, A.: End-to-end driving via conditional imitation learning. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 4693–4700. IEEE (2018)
4. Fang, C., He, H., Long, Q., Su, W.J.: Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. Proceedings of the National Academy of Sciences **118**(43), e2103091118 (2021)
5. Galanti, T., György, A., Hutter, M.: On the role of neural collapse in transfer learning. In: International Conference on Learning Representations (2021)
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
7. Hirose, N., Shah, D., Sridhar, A., Levine, S.: Sacson: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters **9**(1), 49–56 (2023)
8. Hirose, N., Xia, F., Martín-Martín, R., Sadeghian, A., Savarese, S.: Deep visual mpc-policy learning for navigation. IEEE Robotics and Automation Letters **4**(4), 3184–3191 (2019)
9. Hong, W., Ling, S.: Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. Journal of Machine Learning Research **25**(192), 1–48 (2024)
10. Hong, Y., Zhou, Y., Zhang, R., Dernoncourt, F., Bui, T., Gould, S., Tan, H.: Learning navigational visual representations with semantic map supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3055–3067 (2023)
11. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. Advances in neural information processing systems **32** (2019)
12. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters **7**(4), 11807–11814 (2022)
13. Koenig, N., Howard, A.: Design and use paradigms for gazebo, an open-source multi-robot simulator. In: 2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566). vol. 3, pp. 2149–2154. Ieee (2004)
14. Laskin, M., Srinivas, A., Abbeel, P.: Curl: Contrastive unsupervised representations for reinforcement learning. In: International conference on machine learning. pp. 5639–5650. PMLR (2020)
15. Li, X., Liu, S., Zhou, J., Lu, X., Fernandez-Granda, C., Zhu, Z., Qu, Q.: Understanding and improving transfer learning of deep models via neural collapse. Transactions on Machine Learning Research (2024)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), `https://openreview.net/forum?id=Bkg6RiCqY7`

17. Ma, C., Obuchi, T., Tanaka, T.: Neural collapse in cumulative link models for ordinal regression: An analysis with unconstrained feature model. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
18. Majumdar, A., Yadav, K., Arnaud, S., Ma, Y.J., Chen, C., Silwal, S., Jain, A., Berges, V.P., Wu, T., Vakil, J., Abbeel, P., Malik, J., Batra, D., Lin, Y., Maksymets, O., Rajeswaran, A., Meier, F.: Where are we in the search for an artificial visual cortex for embodied intelligence? In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
19. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., Hadsell, R.: Learning to navigate in complex environments. In: International Conference on Learning Representations (2017)
20. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. In: Conference on Robot Learning. pp. 892–909. PMLR (2023)
21. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)
22. Pomerleau, D.A.: Alvinn: An autonomous land vehicle in a neural network. Advances in neural information processing systems **1** (1988)
23. Qi, H., Yin, H., Yang, H.: Control-oriented clustering of visual latent representation. In: The Thirteenth International Conference on Learning Representations (2025)
24. Ren, H., Zeng, Y., Bi, Z., Wan, Z., Huang, J., Cheng, H.: Prior does matter: Visual navigation via denoising diffusion bridge models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
25. Shah, D., Eysenbach, B., Rhinehart, N., Levine, S.: Rapid Exploration for Open-World Navigation with Latent Goal Models. In: 5th Annual Conference on Robot Learning (2021), https://openreview.net/forum?id=d_SWJhyKfVw
26. Shah, D., Sridhar, A.K., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: Vint: A foundation model for visual navigation. In: Conference on Robot Learning (2023), https://api.semanticscholar.org/CorpusID:259262396
27. Sridhar, A., Shah, D., Glossop, C., Levine, S.: Nomad: Goal masked diffusion policies for navigation and exploration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 63–70. IEEE (2024)
28. Stooke, A., Lee, K., Abbeel, P., Laskin, M.: Decoupling representation learning from reinforcement learning. In: International conference on machine learning. pp. 9870–9879. PMLR (2021)
29. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
30. Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics. MIT press (2005)
31. Welford, B.P.: Note on a method for calculating corrected sums of squares and products. Technometrics **4**(3), 419–420 (1962)
32. Yang, Y., Chen, S., Li, X., Xie, L., Lin, Z., Tao, D.: Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In: Advances in Neural Information Processing Systems. vol. 35, pp. 37991–38002 (2022)
33. Zeng, Y., Ren, H., Wang, S., Huang, J., Cheng, H.: Navidiffusor: Cost-guided diffusion model for visual navigation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 11994–12001. IEEE (2025)

34. Zhang, A., McAllister, R.T., Calandra, R., Gal, Y., Levine, S.: Learning invariant representations for reinforcement learning without reconstruction. In: International Conference on Learning Representations (2021)
35. Zhou, J., Li, X., Ding, T., You, C., Qu, Q., Zhu, Z.: On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In: International Conference on Machine Learning. pp. 27179–27202. PMLR (2022)
36. Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., Qu, Q.: A geometric analysis of neural collapse with unconstrained features. In: Advances in Neural Information Processing Systems. vol. 34, pp. 29820–29834 (2021)