DR-VLA: Learning Human-Like Quadruped Navigation from Egocentric Web Videos via Domain-Invariant Scene Representations

Anonymous authors
Paper under double-blind review

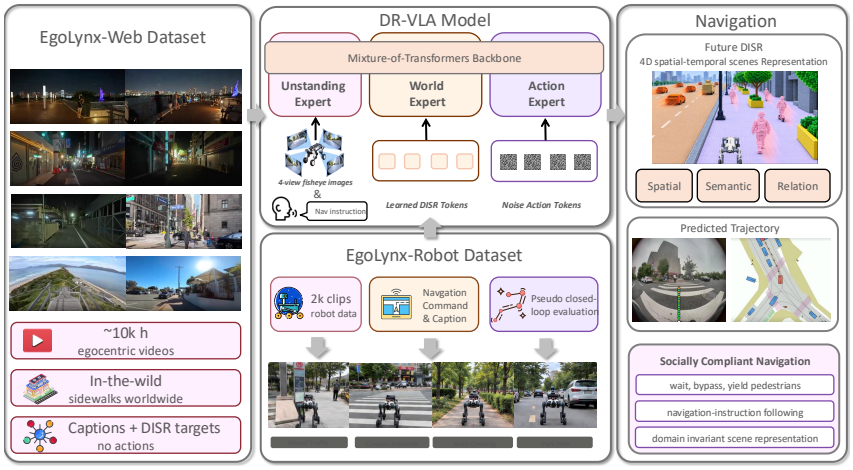


Figure 1: **Overview.** DR-VLA learns to understand and anticipate scenes from egocentric web videos (left) and to act from teleoperated quadruped clips (center bottom). Three experts (center top) map four fisheye views and a navigation instruction to a trajectory and to future domain-invariant scene representations (DISR), which encode semantic, spatial, and relational structure (right).

Abstract

Quadruped robots performing outdoor deliveries must interpret complex scenes, anticipate changes, and navigate in socially compliant ways. Robot demonstrations are costly to collect at the scale needed for generalization. Egocentric web videos offer abundant human navigation experience, but differences in viewpoint and motion make this experience difficult to transfer to robots. We introduce DR-VLA, a Mixture-of-Transformers framework that transfers human navigation experience through a domain-invariant scene representation (DISR). The framework combines vision-language understanding and latent world modeling with a flow-matching action expert. We introduce EgoLynx, comprising 10,000 hours of curated egocentric web video (EgoLynx-Web) and 2,000 clips of teleoperated robot navigation collected during real-world deliveries (EgoLynx-Robot). We pretrain the understanding and world experts on EgoLynx-Web without action labels, then adapt the model to robot observations and train the action expert on EgoLynx-Robot to predict executable trajectories. We also propose the Quadruped Navigation Score (QNS), a pseudo-closed-loop metric combining safety, instruction compliance, comfort, and progress. On the held-out EgoLynx-Robot test split, DR-VLA improves QNS by 5% and reduces open-loop final heading error (FHE) by 14.6%, each relative to the best competing result for that metric. Ablations show that web-video pretraining improves generalization, the model's key components contribute to performance, and DISR outperforms five alternative visual representations as a future-prediction target. Trained separately on NAVSIM v1 without web pretraining, DR-VLA achieves 92.1 PDMS, suggesting that the framework also applies to autonomous driving.

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

1 INTRODUCTION

Last-mile delivery in shared outdoor spaces, from city sidewalks to village paths, is a promising application for quadruped robots (Alverhed et al., 2024; Lee et al., 2024). Navigating these spaces requires more than reaching a goal without collision. A robot must understand complex scenes, anticipate how the environment will change, and navigate in socially compliant ways. The prevailing approach learns these capabilities end-to-end from robot demonstrations via imitation learning (Shah et al., 2023a;b; Sridhar et al., 2024; Hirose et al., 2026). However, collecting demonstrations requires skilled teleoperation and careful safety oversight (Liu et al., 2025). As a result, existing quadruped data are limited in both scene coverage and diversity of social interactions: even the largest public real-world datasets contain only 5–15 hours of legged-robot data each, typically from a single city or region (Karnan et al., 2022; Liu et al., 2025; Tuna et al., 2026). With such data alone, models struggle to cover the long-tail situations that these three capabilities must handle.

In contrast, egocentric videos on the Internet offer another source of data. They are easy to obtain and span cities and weather conditions worldwide. More importantly, they show how people move through varied environments, respond to other agents, and choose paths around obstacles, which is exactly what robot data lack the most. However, transferring this experience to robots presents two challenges. The first is an observation gap: head-mounted human videos and a robot's low-mounted, multi-camera fisheye observations differ in viewpoint, imaging characteristics, and camera motion. The second is an action gap: Internet videos do not provide robot action labels, and human motion cannot be directly executed by a quadruped. Prior work estimates pseudo-actions from camera motion or reconstructed scene geometry (Liu et al., 2025; Cheng et al., 2025; Kuno et al., 2026; Seneviratne et al., 2026). These methods help address missing action labels, but the resulting supervision may retain human viewpoint and motion biases. Other approaches learn from unlabeled videos by predicting future visual content (Bar et al., 2025; Zhu et al., 2025b). These methods learn scene dynamics, but they entangle embodiment-specific factors such as appearance and viewpoint with scene evolution, which hinders transfer across embodiments. Some works transfer at an embodiment-agnostic abstract level (Wang et al., 2026), but they rely on depth and pose captured with dedicated wearable devices and thus cannot scale to Internet video. Existing methods therefore face a dilemma: they either scale but transfer at an embodiment-specific level, or use abstract representations but fail to scale. These limitations motivate a central question: can we learn transferable scene understanding and future prediction directly from raw Internet videos, at an abstract level shared by humans and robots?

We address this question with a domain-invariant scene representation (DISR) designed to retain information useful for navigation across human and robot observations. We observe that scene understanding for outdoor navigation comes down to semantics, space, and relations: what is in the scene, where it is, and how entities relate to one another. We obtain semantic information from segmentation and spatial structure from depth, then fuse them with attention to represent relations among entities and to the surrounding geometry. A pretrained scene autoencoder compresses this information into a compact latent. Segmentation and depth can be estimated from RGB video using vision foundation models (Carion et al., 2026; Lin et al., 2026), allowing DISR targets to be generated without dedicated depth sensors or manual action annotation. By emphasizing scene structure over texture and imaging style, DISR reduces appearance-related differences between human and robot observations; adaptation on robot data addresses the remaining viewpoint and embodiment differences. We model scene dynamics by predicting future DISR latents rather than reconstructing future pixels. Future observations in human videos reflect the consequences of navigation decisions, such as moving around a crowd or stopping to yield. Predicting their scene-level evolution thus provides a way to learn human navigation priors without explicit action labels.

Building on this insight, we propose DR-VLA, a Mixture-of-Transformers (Liang et al., 2025) framework whose three experts have separate parameters and exchange information via shared self-attention. This tightly couples understanding, prediction, and decision-making while avoiding interference among training objectives. The understanding expert is a vision-language model that takes images and navigation instructions as input and reasons about the scene under chain-of-thought-style caption supervision (Wei et al., 2022; Zawalski et al., 2024). The world expert predicts future DISR latents. Conditioned on the other experts, the action expert generates robot trajectories through flow matching (Lipman et al., 2023). To support training across human and robot observations, we curate EgoLynx from complementary data sources. EgoLynx-Web contains 10,000 hours of egocentric

Internet video selected to cover diverse environments and human navigation behaviors. EgoLynx-Robot contains 2,000 training clips (approximately 17 hours) and a held-out test split, both selected from a larger collection of human-teleoperated wheeled-legged robot navigation in real-world delivery settings. We first pretrain the understanding and world experts on EgoLynx-Web without action labels, then adapt the model to robot observations, and supervise the action expert with EgoLynx-Robot.

We evaluate DR-VLA on real-world quadruped navigation and on the NAVSIM autonomous-driving benchmark (Dauner et al., 2024). We also introduce the Quadruped Navigation Score (QNS) to evaluate navigation quality beyond agreement with a recorded trajectory. We report it alongside open-loop average and final displacement errors (ADE and FDE) on the held-out EgoLynx-Robot test split. Against eight navigation and vision-language-action baselines, DR-VLA improves QNS by 5% and reduces FHE by 14.6%, each relative to the strongest competing result for that metric. Ablations show that web-video pretraining improves performance on held-out robot data and that DISR is a more effective future-prediction target than five alternative scene representations. Trained independently on NAVSIM v1 without EgoLynx-Web pretraining, DR-VLA achieves 92.1 PDMS, and removing the world expert reduces the score to 89.5.

In summary, our contributions are fivefold: (i) We propose a domain-invariant scene representation that abstracts human and robot observations into shared semantic, spatial, and relational structure, allowing human navigation priors such as social compliance to be distilled from large-scale egocentric video and transferred across embodiments; (ii) We present DR-VLA, a Mixture-of-Transformers latent world action model whose understanding, world, and action experts decouple supervision across data sources and extend readily to diverse navigation settings; (iii) We curate EgoLynx, a dataset that pairs 10k hours of egocentric web videos with 2,000 teleoperated quadruped clips under a shared annotation format of instructions, captions, and DISR targets, enabling action-free pretraining followed by data-efficient adaptation; (iv) We introduce QNS, a pseudo-closed-loop quadruped navigation metric that evaluates safety, instruction compliance, comfort, and progress; (v) DR-VLA outperforms state-of-the-art methods by 5% in QNS on real-world quadruped navigation and achieves 92.1 PDMS on NAVSIM, corroborating the generality of the representation across embodiments.

2 Related Work

VLA and WAM. Vision-language-action (VLA) models adapt pretrained vision-language models to robot control (Brohan et al., 2023; Kim et al., 2024; Black et al., 2025b;a; NVIDIA et al., 2025b), and recent variants adopt Mixture-of-Transformers (MoT) designs whose function-specific experts share attention (Liang et al., 2025; Black et al., 2025b; Lv et al., 2025; Bi et al., 2026). World action models (WAMs) (Ye et al., 2026) additionally predict future observations: early ones forecast pixels or video latents (Du et al., 2023; Wu et al., 2024; Hu et al., 2025; Zhu et al., 2025a; Li et al., 2025a; Zhao et al., 2025b), spending capacity on visual detail irrelevant to control, whereas recent ones predict latent embeddings of future observations as training-only targets (Zheng et al., 2025; Zhang et al., 2025b; Sun et al., 2026; Lyu et al., 2026; Luo et al., 2026b). DR-VLA follows this latent route but replaces generic visual embeddings, which retain the recording embodiment's appearance, with a structured scene representation that removes appearance by construction, and partitions its experts by supervision source.

Visual and Socially Compliant Navigation. Visual navigation foundation models transfer across platforms (Shah et al., 2023a;b; Sridhar et al., 2024) and have been extended with web video (Hirose et al., 2024; Liu et al., 2025) and VLA formulations (Zhang et al., 2024; 2025a; Hirose et al., 2026), including for legged robots (Cheng et al., 2025; Ding et al., 2024; Song et al., 2024; Tong et al., 2025; Zhao et al., 2025a). Social navigation further requires following implicit human conventions (Mavrogiannis et al., 2023), which prior work encodes through hand-crafted models or reinforcement learning in simulated crowds (Helbing & Molnár, 1995; van den Berg et al., 2011; Chen et al., 2017; 2019), small-scale demonstrations (Karnan et al., 2022; Nguyen et al., 2023), or VLM-based reasoning (Song et al., 2025; Payandeh et al., 2025), yet current VLMs still trail human consensus in social scene understanding (Munje et al., 2025). DR-VLA instead learns such conventions from large-scale human video and explicitly bridges the human-to-quadruped gap that trajectory-level imitation leaves open.

Learning from Egocentric Human Videos. Egocentric datasets (Grauman et al., 2022; Hoque et al., 2026; Li et al., 2025c) have made human video a practical source of supervision. Beyond visual pretraining (Nair et al., 2022; Ma et al., 2023), prior methods recover pseudo-actions through pose retargeting (Kareer et al., 2025; Yang et al., 2026; Luo et al., 2026a) or visual odometry (Liu et al., 2025; Cheng et al., 2025; Kuno et al., 2026); learn latent actions from inter-frame change (Ye et al., 2025; Bu et al., 2025; Chen et al., 2025; 2026), which tends to entangle scene dynamics with camera motion unless explicitly separated; or align human and robot latents jointly with actions (Punamiya et al., 2025), which requires action-labeled human data. Most closely related, EgoWAM (Li et al., 2026a) shows in a controlled manipulation study that the world-prediction target governs human-to-robot transfer and that pixel targets transfer weakly, yet finds no off-the-shelf target uniformly superior. DR-VLA builds on this premise with a dedicated target combining semantic, geometric, and relational structure, inspired by mid-level representations that aid transfer in navigation and driving (Müller et al., 2018; Sax et al., 2019; Zhou et al., 2019); unlike these, it serves only as a prediction target, and it addresses navigation, where ego-motion differs on both sides of the gap.

3 TASK FORMULATION

Navigation policy. We cast outdoor navigation as instruction-conditioned trajectory prediction. At time t, the agent receives an observation $\mathbf{o}_t = \{I^v_{t-T+1}, \dots, I^v_t\}_{v=1}^{V}$ comprising the last T frames from V camera views, a natural-language navigation instruction ℓ, and its ego state s_t. The policy outputs an action chunk of N_a planar body-frame poses spaced δ apart,

$$\mathbf{a}_t \sim \pi(\cdot \mid \mathbf{o}_t, \ell, s_t), \qquad \mathbf{a}_t = \{(x_j, y_j, \theta_j)\}_{j=1}^{N_a}. \tag{1}$$

On the quadruped, $V{=}4$ surround-view fisheye cameras observe the scene, ℓ specifies the route (e.g., *follow the path*, *branch left*), and $\mathbf{a}_t$ spans 1 s with $N_a{=}10$ and $\delta{=}0.1$ s. In driving, ℓ is one of four high-level commands and $\mathbf{a}_t$ spans 4 s with $N_a{=}8$ and $\delta{=}0.5$ s. Only V, N_a, and δ depend on the embodiment. Importantly, ℓ prescribes *where* to go but not *how* to behave along the way: when to stop, yield, or detour around other agents is left unspecified and must be inferred from the scene.

A structural view of the embodiment gap. Humans, quadrupeds, and vehicles traverse similar scenes, yet they perceive and act in them differently. We model an observation from embodiment e as a rendering of an underlying scene state,

$$\mathbf{o}_t = \Gamma_e\big(\boldsymbol{\zeta}_t;\, \boldsymbol{\eta}_e\big), \tag{2}$$

where $\boldsymbol{\zeta}_t$ captures the scene structure relative to the observer, namely what is present, where it lies, and how entities relate, and $\boldsymbol{\eta}_e$ collects embodiment-specific factors such as camera height, appearance, and low-level motion characteristics. Under this view, a human egocentric recording is simply another embodiment with $V{=}1$. The gap between embodiments then arises at two levels: Γ_e renders the same scene differently, and the action spaces differ, so human motion does not define an action $\mathbf{a}_t$ in the robot's space. We assume, however, that the evolution of scene structure, $p(\boldsymbol{\zeta}_{t+\Delta:t+K\Delta} \mid \boldsymbol{\zeta}_{\le t}, \ell)$, is largely shared across embodiments. Since $\boldsymbol{\zeta}$ is expressed relative to the observer, its evolution also reflects where the observer chooses to go, and hence encodes human navigation priors such as yielding to pedestrians or steering around crowds.

Policy decomposition. This view suggests factorizing the policy into a scene-evolution predictor Ψ shared across embodiments and an embodiment-specific action generator π_e,

$$\hat{\boldsymbol{\zeta}}_{t+\Delta:t+K\Delta} = \Psi(\mathbf{o}_t, \ell), \qquad \mathbf{a}_t \sim \pi_e\big(\cdot \mid \mathbf{o}_t, \ell, s_t, \hat{\boldsymbol{\zeta}}_{t+\Delta:t+K\Delta}\big). \tag{3}$$

Because Ψ depends only on scene structure, it can be learned from observations alone, including those of embodiments whose actions are unavailable or incompatible with the robot. In contrast, π_e grounds the anticipated scene evolution in the robot's own kinematics and therefore requires action supervision from the target embodiment.

4 METHODOLOGY

4.1 DOMAIN-INVARIANT SCENE REPRESENTATION

We define the domain-invariant scene representation (DISR) as a compact spatio-temporal latent $\mathbf{z}$ that serves as a tractable approximation of the scene structure $\boldsymbol{\zeta}$. DISR characterizes a scene along

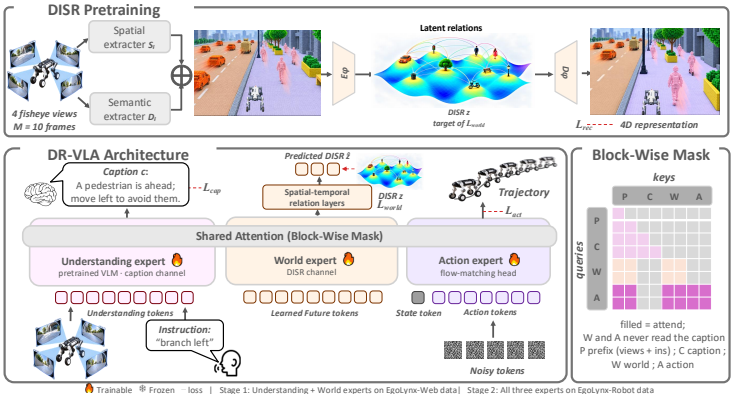


Figure 2: **DR-VLA and the DISR Pretraining.** *Top:* Combined Spatial and Semantic information are encoded by E_ϕ into the DISR latent $\mathbf{z}$ and reconstructed by D_ϕ. *Bottom left:* the understanding, world, and action experts share attention and are trained with $\mathcal{L}_{\text{cap}}$, $\mathcal{L}_{\text{world}}$ (targets from the frozen E_ϕ), and $\mathcal{L}_{\text{act}}$. *Bottom right:* block-wise attention mask.

three axes: *semantic*, *spatial*, and *relational*. Vision foundation models extract the first two explicitly from raw frames, while an encoder learns the third across frames (Fig. 2, top).

Disentangling structure from appearance. For each frame I_i of each view, two frozen teachers provide structural cues. The semantic cue S_i is produced by the concept segmentation model SAM 3 (Carion et al., 2026) over a fixed category vocabulary $\mathcal{C}_{\text{sem}}$ shared across all data (e.g., pedestrian, vehicle), and is encoded as $|\mathcal{C}_{\text{sem}}|$ channels of category masks. The spatial cue D_i is the relative depth estimated by the relative-depth model of Depth Anything 3 (Lin et al., 2026); by discarding absolute scale, it also mitigates the scale discrepancy induced by different camera heights. Together, the image-plane layout of the masks and the depth locate each entity in direction and distance relative to the observer. Both teachers operate directly on raw frames, including unrectified fisheye images. Because the encoder receives only (S_i, D_i) and never RGB, appearance factors such as texture, lighting, and imaging style are excluded by construction.

Relational encoding in a shared latent space. The DISR encoder E_ϕ and a lightweight decoder D_ϕ form an autoencoder. For $M{=}10$ consecutive frames, S_i and D_i are concatenated along the channel dimension, tokenized, and processed by L_{rel} layers of causal spatio-temporal self-attention over the tokens of all frames,

$$\{\mathbf{z}_1, \ldots, \mathbf{z}_M\} = E_\phi\big(\{(S_i, D_i)\}_{i=1}^{M}\big), \qquad \mathbf{z}_i \in \mathbb{R}^{N_z \times d_z}. \tag{4}$$

These attention layers realize the relational axis, capturing entity correspondence across frames, relative depth ordering, and co-occurrence among entities. The decoder reconstructs S_i and D_i from each $\mathbf{z}_i$ under a reconstruction loss plus a regularizer $\mathcal{R}(\mathbf{z}_i)$. Compared with discrete label maps, the continuous latent is more compact and better suited as a regression target.

We train E_ϕ on a mixture of human videos and robot data, then freeze it and apply the same weights to clips from every embodiment, mapping all observations into a common target space. DISR serves only as the training target of the world expert (Sec. 4.2) and never as a policy input, so neither the teachers nor the encoder is needed at inference.

4.2 Mixture-of-Transformers Architecture

We instantiate the decomposition in Eq. (3) as DR-VLA (Fig. 2). The understanding and world experts jointly realize Ψ: the former grounds the instruction and multi-view observations in scene semantics, and the latter predicts the DISR of K future frames. The action expert realizes π_e, translating the anticipated scene evolution into trajectories for the robot's own body. Following the Mixture-of-Transformers design (Black et al., 2025b; Liang et al., 2025; Deng et al., 2025), each expert owns its attention projections. Image tokens of view v receive a learnable view embedding

e_v added to their positional encoding, which allows the model to distinguish the four fisheye views without rectification.

Understanding expert: grounding instructions in the scene. The understanding expert f_u is a pretrained vision-language model, Qwen3-VL-2B (Bai et al., 2025a). Images from all views are encoded by its ViT and the instruction ℓ by its text tokenizer, together forming the prefix tokens. During training, f_u generates the clip caption c by next-token prediction,

$$\mathcal{L}_{\text{cap}} = -\sum_{n} \log p_{f_u}\big(c_n \mid c_{<n}, \mathbf{o}_t, \ell\big). \tag{5}$$

The caption describes the scene layout, surrounding agents, potential hazards, and ego-motion, steering the hidden states $\mathbf{h}^u$ toward navigation-relevant content. It serves only as auxiliary supervision and is not decoded at inference.

World expert: anticipating scene evolution. The world expert f_w operates on learnable world tokens $\mathbf{w} = \{w_1, \ldots, w_K\}$, where w_k holds N_w query tokens per view, augmented with a temporal embedding for time $t+k\Delta$ and the view embedding e_v. The world tokens first read the prefix through shared attention and then pass through L_w spatio-temporal relation layers spanning all world tokens. These layers relate the predicted frames and views, extending the relational axis of DISR into the future. With $N_w = N_z$, a token-wise linear head g_w yields $\hat{\mathbf{z}}_{t+k\Delta} = g_w(\mathbf{h}^w_k) \in \mathbb{R}^{N_z \times d_z}$. The target $\mathbf{z}_{t+k\Delta}$ is computed by the frozen E_ϕ on the causal clip ending at frame $t + k\Delta$ (Sec. 4.1). Abbreviating $\hat{\mathbf{z}}_{t+k\Delta}$ and $\mathbf{z}_{t+k\Delta}$ as $\hat{\mathbf{z}}_k$ and $\mathbf{z}_k$, the world loss is

$$\mathcal{L}_{\text{world}} = \frac{1}{K}\sum_{k=1}^{K}\Big[d_{\text{SL1}}(\hat{\mathbf{z}}_k, \mathbf{z}_k) + 1 - \cos(\hat{\mathbf{z}}_k, \mathbf{z}_k)\Big], \tag{6}$$

where d_{SL1} denotes the smooth-L1 distance and $\cos$ the mean token-wise cosine similarity.

Action expert: embodiment-specific trajectory generation. The action expert f_a generates the action chunk $\mathbf{a}_t$ via conditional flow matching (Lipman et al., 2023; Black et al., 2025b). Each of the N_a waypoints is represented by one action token, and the ego state s_t occupies one additional token. Given $\tau \sim \mathcal{U}[0,1]$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{a}_t^\tau = \tau\,\mathbf{a}_t + (1-\tau)\,\boldsymbol{\epsilon}$, f_a predicts the velocity field under

$$\mathcal{L}_{\text{act}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\Big\| v_{f_a}\big(\mathbf{a}_t^\tau, \tau, s_t;\, \mathbf{h}^u, \mathbf{h}^w\big) - (\mathbf{a}_t - \boldsymbol{\epsilon})\Big\|_2^2. \tag{7}$$

The action expert attends to $\mathbf{h}^u$ and $\mathbf{h}^w$ through shared attention. Since $\hat{\mathbf{z}}$ is read out from $\mathbf{h}^w$ by g_w, conditioning on $\mathbf{h}^w$ realizes the dependence of π_e on the anticipated scene evolution in Eq. (3). At inference, $\mathbf{a}_t$ is integrated from noise in 4 Euler steps.

Structured information flow. A block-wise attention mask (Black et al., 2025b) governs how information flows among tokens (Fig. 2, bottom right). World and action tokens do not attend to the caption, so their inputs are identical at training and inference. Moreover, since action tokens are not visible to the prefix or world tokens, $\mathbf{h}^u$ and $\mathbf{h}^w$ are computed once at inference and reused across integration steps.

4.3 EGOLYNX AND TWO-STAGE TRAINING

To instantiate the decomposition in Eq. (3), we curate EgoLynx, a dataset that pairs action-free human videos with action-labeled robot demonstrations under a shared annotation format: every clip carries an instruction ℓ, a caption c, and DISR targets $\mathbf{z}$, while robot clips additionally carry actions $\mathbf{a}$. Its two parts, EgoLynx-Web and EgoLynx-Robot, drive two training stages that learn Ψ and π_e, respectively, with the DISR encoder E_ϕ kept frozen throughout.

Stage 1: learning scene evolution from EgoLynx-Web. EgoLynx-Web is collected from YouTube and covers walking, cycling, and other forms of egocentric human navigation across urban and rural areas, different times of day, and diverse weather conditions. After multi-stage cleaning, it retains about 10k hours, roughly twice the size of Sekai and three times that of Ego4D (Li et al., 2025c; Grauman et al., 2022). For each clip, Gemini 3.1 Flash-Lite (Google DeepMind, 2026) observes the full clip, including its subsequent motion, and generates the instruction ℓ and the caption c. This stage trains only f_u and f_w with $\mathcal{L}_1 = \mathcal{L}_{\text{cap}} + \lambda_w \mathcal{L}_{\text{world}}$. We deliberately avoid estimating pseudo-trajectories from the videos: human motion does not correspond to the robot's action space, and camera motion estimated on monocular, dynamic footage is noisy (Li et al., 2025d).

Table 1: Quantitative Navigation results on the EgoLynx-Robot test split. ADE/FDE: average/final displacement error; AHE/FHE: average/final heading error; NC: no collision; WAC: walkable-area compliance; TTC: time-to-collision; CL: clearance; EP: ego progress; IC: instruction compliance; QNS: Quadruped Navigation Score.

	Open-loop				*Pseudo-closed-loop*						
Model	**ADE↓**	**FDE↓**	**AHE↓**	**FHE↓**	**NC↑**	**WAC↑**	**TTC↑**	**CL↑**	**EP↑**	**IC↑**	**QNS↑**
VLN-based policies											
InternNav (Wei et al., 2026a)	0.079	0.098	0.038	0.062	0.945	0.955	0.871	0.815	0.893	0.818	0.764
NaVILA (Cheng et al., 2025)	0.030	0.059	0.028	0.048	0.968	0.975	0.912	0.838	0.941	0.893	0.843
OmniNav (Xue et al., 2026)	0.090	0.152	0.033	0.054	0.939	0.949	0.863	0.811	0.866	0.803	0.743
StreamVLN (Wei et al., 2026b)	0.036	0.067	0.032	0.053	0.964	0.972	0.905	**0.843**	0.933	0.884	0.833
VLA-based policies											
LingBot-VLA 2.0 (Wu et al., 2026)	0.044	0.068	0.038	0.054	0.958	0.966	0.893	0.829	0.926	0.863	0.809
DepthVLA (Yuan et al., 2026)	0.056	0.074	0.032	0.049	0.952	0.961	0.882	0.825	0.915	0.842	0.789
$\pi_{0.5}$ (Black et al., 2025a)	0.036	0.075	0.032	0.052	0.961	0.969	0.899	0.834	0.922	0.874	0.819
OpenVLA (Kim et al., 2024)	0.041	0.068	0.030	0.051	0.956	0.964	0.888	0.823	0.928	0.855	0.802
Ours											
DR-VLA w/o EgoLynx-Web	0.034	0.062	0.030	0.052	0.960	0.970	0.905	0.832	0.924	0.878	0.823
DR-VLA	**0.027**	**0.052**	**0.024**	**0.041**	**0.982**	**0.979**	**0.928**	0.842	**0.945**	**0.921**	**0.884**

Stage 2: grounding in the target embodiment with EgoLynx-Robot. EgoLynx-Robot comprises 2,000 training teleoperated clips (approximately 17 hours) selected from a larger collection of human-teleoperated wheeled-legged robot navigation in real-world delivery settings. Its captions are generated with the same prompt as in Stage 1, grounded in the event annotations. This stage initializes the action expert and the view embeddings of the four fisheye cameras, and fine-tunes all parameters with $\mathcal{L}_2 = \mathcal{L}_1 + \lambda_a \mathcal{L}_{\text{act}}$. Retaining $\mathcal{L}_{\text{cap}}$ and $\mathcal{L}_{\text{world}}$ guards against the erosion of the understanding and prediction capabilities acquired in Stage 1 during action learning. Meanwhile, continued training of the world expert on the DISR targets of robot clips absorbs the residual viewpoint gap caused by camera height and fisheye distortion.

5 Experiments

5.1 Experimental Setup

EgoLynx. All quadruped baselines are trained on the training split of EgoLynx-Robot, and only DR-VLA is additionally pretrained on EgoLynx-Web (Sec. 4.3). All methods are evaluated on the held-out test split of EgoLynx-Robot, which is sampled from the same teleoperation collection but shares no clips with the training split. We report open-loop displacement errors over the 1 s action chunk. We also report the proposed Quadruped Navigation Score (QNS), a pseudo-closed-loop metric inspired by NAVSIM (Dauner et al., 2024) that aggregates sub-scores covering safety, instruction compliance, comfort, and progress. Its computation is detailed in Appendix B.

NAVSIM. We follow the standard protocol of NAVSIM v1 (Dauner et al., 2024), training on navtrain and evaluating on navtest with camera inputs only, and report the PDM score (PDMS).

Implementation details. The understanding expert of DR-VLA is initialized from Qwen3-VL-2B (Bai et al., 2025a). Stage 1 and Stage 2 are trained for 20k and 5k steps, respectively, on 32 NVIDIA H800 GPUs. The driving model shares the same architecture as the quadruped model and differs only in V, N_a, and δ.

5.2 Main Results on Quadruped Navigation

To evaluate quadruped navigation, we compare DR-VLA with state-of-the-art policies on EgoLynx, comprising four VLN-based and four VLA-based models, all adapted to the same inputs and action chunk. DR-VLA achieves the best overall performance, improving QNS over the strongest competitor, NaVILA, by 5% while also producing the most accurate open-loop trajectories (Table 1). The improvement is most pronounced in instruction compliance and collision-related metrics: DR-VLA follows route-level instructions more faithfully and maintains a larger temporal margin from surrounding agents, consistent with the world expert anticipating how the scene evolves. These gains

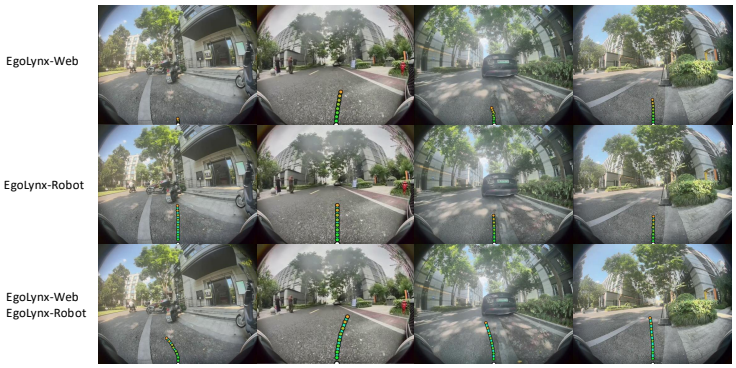


Figure 3: **Predicted trajectories in unseen scenes.** Replayed observations; rows: trained on EgoLynx-Web only, EgoLynx-Robot only, or both; columns: four scenes.

Table 2: **Ablation .** (a) Replacing DISR with alternative targets: Cosmos pixel latents (NVIDIA et al., 2025a), DINOv3 (Siméoni et al., 2025) and V-JEPA 2.1 (Mur-Labadia et al., 2026) features, depth-augmented DINO video queries (Wu et al., 2026), and a FLARE-style embedding (Zheng et al., 2025). (b) Ablating DISR components. Architecture and two-stage training are fixed

(a) Alternative targets

Target	ADE↓	FDE↓	NC↑	TTC↑	IC↑	QNS↑
No world expert	0.031	0.058	0.966	0.906	0.896	0.841
Cosmos latent	0.031	0.057	0.967	0.909	0.897	0.845
DINOv3	0.030	0.056	0.971	0.914	0.903	0.856
V-JEPA 2.1	0.029	0.055	0.973	0.916	0.906	0.861
Depth + DINO video	0.029	0.055	0.975	0.919	0.908	0.866
FLARE-style	0.030	0.057	0.970	0.912	0.905	0.853
DISR (ours)	**0.027**	**0.052**	**0.982**	**0.928**	**0.921**	**0.884**

(b) DISR components

Variant	ADE↓	FDE↓	NC↑	TTC↑	IC↑	QNS↑
No world expert	0.031	0.058	0.966	0.906	0.896	0.841
Segmentation only	0.029	0.055	0.974	0.917	0.911	0.864
Depth only	0.029	0.054	0.976	0.921	0.909	0.868
w/o relation layers	0.028	0.053	0.978	0.922	0.914	0.873
DISR (full)	**0.027**	**0.052**	**0.982**	**0.928**	**0.921**	**0.884**

in safety come without sacrificing progress or clearance, which remain on par with or above the best baseline.

To disentangle the contribution of human videos from that of the architecture, we further evaluate DR-VLA trained on EgoLynx-Robot alone. This variant already outperforms all VLA-based policies but falls short of the strongest VLN models; pretraining on EgoLynx-Web lifts it above all baselines, with consistent gains across every metric and the largest in instruction compliance and collision avoidance. The advantage of DR-VLA therefore stems mainly from knowledge transferred from human videos rather than from the architecture alone, as we analyze further in Sec. 5.3.

Figure 3 provides a qualitative comparison of web-only, robot-only, and jointly trained variants on replayed observations from unseen scenes. For the web-only variant, we obtain trajectory supervision from DPVO pseudo-trajectories, following the idea of Liu et al. (2025); no robot demonstrations are used. In the examples shown, its predictions steer around obstacles but are poorly calibrated to the quadruped's metric scale. The robot-only variant predicts more suitably scaled motion but less consistently follows socially compliant routes. The jointly trained variant exhibits both properties: in the third column, its predicted path avoids an obstacle and then returns to the sidewalk.

5.3 Ablation Study

We performed ablation studies to demonstrate the contribution of each components. All ablations are evaluated on the EgoLynx-Robot test split. More ablations like expert layout, the pretraining signals, and the data scale are detailed in Appendix A.

Without the world expert, QNS drops to 0.841. We then retain the expert and replace DISR with five alternative targets under the same training data and schedule (Table 2(a)). Cosmos latents yield 0.845 QNS, close to the no-expert baseline. Depth-augmented DINO queries perform best among the alternative targets, but the full DISR model achieves the highest QNS. The differences are especially

Table 3: **Results on NAVSIM v1 navtest (Dauner et al., 2024).** Sub-scores and PDMS of prior work are taken from the cited papers. "C" cameras, "L" LiDAR.

Method	Input	Backbone	NC	DAC	TTC	Comf.	EP	PDMS
TransFuser (Chitta et al., 2023)	C+L	ResNet-34	97.7	92.8	92.8	100	79.2	84.0
DiffusionDrive (Liao et al., 2025)	C+L	ResNet-34	98.2	96.2	94.7	100	82.2	88.1
WoTE (Li et al., 2025b)	C+L	ResNet-34	98.5	96.8	94.9	99.9	81.9	88.3
DiffusionDriveV2 (Zou et al., 2025)	C+L	ResNet-34	98.3	97.9	94.8	99.9	87.5	91.2
PRIX (Wozniak et al., 2026)	C	ResNet-34	98.1	96.3	94.1	100	82.3	87.8
iPad (Guo et al., 2026)	C	ResNet-34	98.6	**98.3**	94.9	100	**88.0**	91.7
Fine-tuned Qwen2.5-VL (Bai et al., 2025b)	C	Qwen2.5-VL-7B	97.8	92.1	92.8	100	78.3	83.3
AdaThinkDrive (Luo et al., 2026c)	C	InternVL3-8B	98.4	97.8	95.2	100	84.4	90.3
ReCogDrive (Li et al., 2026b)	C	InternVL3-2B + diffusion	97.9	97.3	94.9	100	87.3	90.8
DR-VLA w/o world expert	C	Qwen3-VL-2B	98.4	96.9	94.8	100	85.6	89.5
DR-VLA (DISR target)	C	Qwen3-VL-2B	**99.0**	98.0	**96.9**	100	87.8	**92.1**

apparent in time-to-collision metric. These results suggest that the benefit depends on what the world expert is trained to predict, rather than on adding a future-prediction objective alone.

We next isolate the contributions of DISR's prediction cues and encoder structure (Table 2(b)). Using segmentation alone or depth alone reduces performance relative to full DISR. Depth retains more of the collision-related performance, whereas segmentation retains more of the instruction-compliance performance. Removing the encoder's relation layers reduces QNS by 1.1 points, with the largest declines in time-to-collision. Together, these results support combining semantic and geometric cues with relational modeling in the world-expert target.

5.4 Generalization to Autonomous Driving

We further evaluate DR-VLA on NAVSIM to examine whether its design extends beyond quadruped navigation. Following proposal–selection planners (Li et al., 2024; 2025b), the action expert decodes 64 candidate trajectories from learnable action queries, and a scorer supervised by the PDM sub-scores selects the final trajectory; all other components remain unchanged. DR-VLA achieves 92.1 PDMS, outperforming the compared planners, including LiDAR-based methods and VLM-based planners built on larger backbones (Table 3). Removing the world expert will lower PDMS to 89.5, with a notable drop in time-to-collision, which rewards anticipating the motion of other agents. The benefit of predicting future DISR therefore carries over to an embodiment whose viewpoint and dynamics differ from those of both humans and quadrupeds.

6 Conclusion and Limitations

Conclusion. This work studies how large-scale egocentric human videos can provide transferable knowledge for outdoor quadruped navigation. Our central thesis is that human-to-robot transfer should be grounded in scene structure rather than in pixels or actions. We instantiate this view with DR-VLA, a Mixture-of-Transformers model whose world expert learns to predict a domain-invariant scene representation of semantic, spatial, and relational structure from egocentric human videos, and curate EgoLynx, which pairs 10k hours of egocentric web videos with nearly 2,000 teleoperated quadruped clips. Ablations show that the prediction target matters: a reconstruction-oriented latent target provides little benefit over removing the world expert, whereas DISR performs best. These results support structured future-scene prediction as a way to transfer useful knowledge from human videos to robot navigation.

Limitations. First, DISR is defined in the image plane, so the residual viewpoint gap still requires adaptation on robot data and precludes zero-shot cross-embodiment transfer. Second, DISR is supervised by pseudo-labels from off-the-shelf vision foundation models over a fixed category vocabulary, whose noise and coverage constrain the representation of long-tail scenes. Third, our quadruped evaluation relies on open-loop and pseudo-closed-loop metrics complemented by real-robot replay, leaving large-scale closed-loop deployment to future study. Future work will lift DISR into a 3D semantic occupancy space, explore open-vocabulary prediction targets, and further scale the web data. More broadly, we envision world models that are pretrained on human videos and operate on scene structure as a shared interface for navigation across wheel-legged robots and vehicles.

AI USE STATEMENT

In this work, we used generative AI tools for one task with required disclosure: as part of our data pipeline, a multimodal large language model (Gemini 3.1 Flash-Lite (Google DeepMind, 2026)) generates the navigation instructions and scene captions that serve as training annotations for EgoLynx, as described in Sec. 4.3. We manually inspected a random sample of N generated annotations to verify their quality. We have not used generative AI tools to propose hypotheses, design or give feedback on the methodology or experiments, implement methods, clean datasets, analyze data, or interpret results, and the remaining tasks with required disclosure (theoretical models, mathematical claims, proofs, and translation) are not applicable to this work. Additionally, we used large language model assistants (Claude and ChatGPT) for tasks with recommended disclosure: identifying relevant literature, verifying and formatting bibliographic references, polishing grammatical correction, and improving the clarity of the manuscript. We have reviewed all AI-assisted work: every suggested reference was checked against its published version, all reported numbers were checked against our experimental results, and all edits were revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses videos that capture people in public spaces. EgoLynx-Web consists of publicly available YouTube videos, and EgoLynx-Robot was collected by teleoperating quadruped robots during real-world scenarios and was manually annotated. We will release EgoLynx. Before release, we apply privacy protection to both the web videos and the robot recordings, including blurring the faces of all visible people. We respect the copyright of the original web videos and release them in a way that does not infringe the rights of their creators for non-commercial research use only, with a procedure for content owners to request removal. In addition, the prediction targets of our model are computed from segmentation masks and relative depth alone and thus carry no appearance or identity information. All robot data were collected under human teleoperation, and all policies in this paper are evaluated offline, on open-loop and pseudo-closed-loop metrics and replayed observations, without autonomous operation among the public. Deploying learned navigation policies in shared spaces carries safety risks, and our evaluation does not replace closed-loop safety validation, which should precede any deployment. Social conventions learned from web video reflect the regions and behaviors that are over-represented online and may not hold in every community. Finally, automating last-mile delivery may affect employment in logistics; we present this work as a research contribution rather than a deployable system.

REPRODUCIBILITY STATEMENT

Sec. 3 formalizes the task, and Sec. 4 and Fig. 2 describe the DISR encoder, the model architecture, and all training objectives. Sec. 4.3 details the composition of EgoLynx and the two-stage training recipe, and Sec. 5.1 specifies the evaluation protocol, backbone initialization, training schedule, and compute. Appendix B defines QNS and its sub-scores, and Appendix A reports the complete ablation results. NAVSIM results use the official navtrain/navtest splits and evaluation code (Dauner et al., 2024). pon acceptance, we plan to release the data and code to support reproducible comparison with future methods.

REFERENCES

Elin Alverhed, Simon Hellgren, Hanna Isaksson, Lisa Olsson, Hanna Palmqvist, and Jonas Flodén. Autonomous last-mile delivery robots: a literature review. *European Transport Research Review*, 16(1):4, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang,

Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Conference on Robot Learning (CoRL)*, 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems (RSS)*, 2025b.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Robotics: Science and Systems (RSS)*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-X: Enhancing latent action modeling in vision-language-action models. In *International Conference on Learning Representations (ICLR)*, 2026.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Yu Fan Chen, Miao Liu, Michael Everett, and Jonathan P. How. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems (RSS)*, 2025.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12878–12895, 2023.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Pengxiang Ding, Han Zhao, Wenjie Zhang, Wenxuan Song, Min Zhang, Siteng Huang, Ningxi Yang, and Donglin Wang. QUAR-VLA: Vision-language-action model for quadruped robots. In *European Conference on Computer Vision (ECCV)*, 2024.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Google DeepMind. Gemini 3.1 Flash-Lite model card. `https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/`, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ke Guo, Haochen Liu, Xiaojun Wu, Jia Pan, and Chen Lv. iPad: Iterative proposal-centric end-to-end autonomous driving. *IEEE Robotics and Automation Letters*, 11(10):11142–11149, 2026.

Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51 (5):4282–4286, 1995.

Noriaki Hirose, Catherine Glossop, Ajay Sridhar, Oier Mees, and Sergey Levine. LeLaN: Learning a language-conditioned navigation policy from in-the-wild videos. In *Conference on Robot Learning (CoRL)*, 2024.

Noriaki Hirose, Catherine Glossop, Dhruv Shah, and Sergey Levine. OmniVLA: An omni-modal vision-language-action model for robot navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *International Conference on Machine Learning (ICML)*, 2025.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially CompliAnt navigation dataset (SCAND): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

Shoya Kuno, Yumo Ouchi, and Kanata Suzuki. Co-training with ego-centric video and demonstration for robot navigation task. *arXiv preprint arXiv:2606.01951*, 2026.

Joonho Lee, Marko Bjelonic, Alexander Reske, Lorenz Wellhausen, Takahiro Miki, and Marco Hutter. Learning robust autonomous navigation and locomotion for wheeled-legged robots. *Science Robotics*, 9(89):eadi9641, 2024.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data. In *Conference on Robot Learning (CoRL)*, 2026a.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Robotics: Science and Systems (RSS)*, 2025a.

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

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations (ICLR)*, 2026b.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025c.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast, and robust structure and motion from casual dynamic videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025d.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, Yu-Gang Jiang, and Jose M. Alvarez. Hydra-MDP: End-to-end multimodal planning with multi-target Hydra-Distillation. *arXiv preprint arXiv:2406.06978*, 2024.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025.

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

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. CityWalker: Learning embodied urban navigation from web-scale videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Haiweng Xu, Qin Jin, and Zongqing Lu. Vision-language-action pretraining from large-scale human videos. In *International Conference on Machine Learning (ICML)*, 2026a.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-H0.7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026b.

Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, Long Chen, Bing Wang, and Zhi-Xin Yang. AdaThinkDrive: Adaptive thinking via reinforcement learning for autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026c.

Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, Wenbo Cui, Senmao Qi, Shuo Wang, Yixin Zheng, Mi Yan, Xuesong Shi, Haoran Li, Dongbin Zhao, Ming-Yu Liu, Zhizheng Zhang, Li Yi, Yizhou Wang, and He Wang. LDA-1B: Scaling latent dynamics action model via universal embodied data ingestion. In *Robotics: Science and Systems (RSS)*, 2026.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations (ICLR)*, 2023.

Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):36, 2023.

Matthias Müller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. Driving policy transfer via modularity and abstraction. In *Conference on Robot Learning (CoRL)*, 2018.

Michael Joseph Munje, Chen Tang, Shuijing Liu, Zichao Hu, Yifeng Zhu, Jiaxun Cui, Garrett Warnell, Joydeep Biswas, and Peter Stone. SocialNav-SUB: Benchmarking VLMs for scene understanding in social robot navigation. In *Conference on Robot Learning (CoRL)*, 2025.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision (ECCV)*, 2026.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

Duc M. Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025a.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025b.

Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dinesh Manocha, and Xuesu Xiao. Social-LLaVA: Enhancing social robot navigation through human-language reasoning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y. Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. EgoBridge: Domain adaptation for generalizable imitation from egocentric human data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alexander Sax, Jeffrey O. Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors. In *Conference on Robot Learning (CoRL)*, 2019.

Gershom Seneviratne, Yohan Abeysinghe, Jianyu An, Vaibhav Shende, Rahul Kumar, and Dinesh Manocha. VEGA: Learning navigation VLAs from in-the-wild egocentric video with geometric trajectory supervision. *arXiv preprint arXiv:2606.18426*, 2026.

Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A general navigation model to drive any robot. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023a.

Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In *Conference on Robot Learning (CoRL)*, 2023b.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 10(1):508–515, 2025.

Wenxuan Song, Han Zhao, Pengxiang Ding, Can Cui, Shangke Lyu, Yaning Fan, and Donglin Wang. GeRM: A generalist robotic model with mixture-of-experts for quadruped robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. NoMaD: Goal masked diffusion policies for navigation and exploration. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. In *European Conference on Computer Vision (ECCV)*, 2026.

Xinyang Tong, Pengxiang Ding, Yiguo Fan, Donglin Wang, Wenjie Zhang, Can Cui, Mingyang Sun, Han Zhao, Hongyin Zhang, Yonghao Dang, Siteng Huang, and Shangke Lyu. QUART-Online: Latency-free multimodal large language model for quadruped robot learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Turcan Tuna, Jonas Frey, Frank Fu, Katharine Patterson, Tianao Xu, Maurice Fallon, Cesar Cadena, and Marco Hutter. GrandTour: A legged robotics dataset in the wild for multi-modal perception and state estimation. *arXiv preprint arXiv:2602.18164*, 2026.

Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics Research: The 14th International Symposium ISRR*, pp. 3–19. Springer, 2011.

Weizhuo Wang, Yanjie Ze, C. Karen Liu, and Monroe Kennedy III. Learning humanoid navigation from human data. *IEEE Robotics and Automation Letters*, 11(9):10449–10456, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations (ICLR)*, 2026a.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. StreamVLN: Streaming vision-and-language navigation via SlowFast context modeling. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026b.

Maciej Wozniak, Lianhang Liu, Yixi Cai, and Patric Jensfelt. PRIX: Learning to plan from raw pixels for end-to-end autonomous driving. *IEEE Robotics and Automation Letters*, 11(5):6400–6407, 2026.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2024.

Wei Wu, Fangjing Wang, Fan Lu, He Sun, Shi Liu, Yunnan Wang, Yibin Yan, Yong Wang, Shuailei Ma, Xinyang Wang, Yibin Liu, Shuai Yang, Tianxiang Zhou, Kejia Zhang, Lei Zhou, Cheng Su, Nan Xue, Bin Tan, Han Zhang, Youchao Zhang, Fei Liao, Xing Zhu, Yujun Shen, and Kecheng Zheng. From foundation to application: Improving VLA models in practice. *arXiv preprint arXiv:2607.06403*, 2026.

Xinda Xue, Junjun Hu, Minghua Luo, Shichao Xie, Jintao Chen, Zixun Xie, Kuichen Quan, Wei Guo, Zedong Chu, Mu Xu, and Zhengzhou Zhu. OmniNav: A unified framework for prospective exploration and visual-language navigation. In *International Conference on Learning Representations (ICLR)*, 2026.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. EgoVLA: Learning vision-language-action models from egocentric human videos. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Zhuoguang Chen, Tao Jiang, and Hang Zhao. DepthVLA: Enhancing vision-language-action models with depth-aware spatial reasoning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Conference on Robot Learning (CoRL)*, 2024.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. NaVid: Video-based VLM plans the next step for vision-and-language navigation. In *Robotics: Science and Systems (RSS)*, 2024.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Robotics: Science and Systems (RSS)*, 2025a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Han Zhao, Wenxuan Song, Donglin Wang, Xinyang Tong, Pengxiang Ding, Xuelian Cheng, and Zongyuan Ge. MoRE: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025a.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *Conference on Robot Learning (CoRL)*, 2025.

Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 4(30):eaaw6661, 2019.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Robotics: Science and Systems (RSS)*, 2025a.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. DiffusionDriveV2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

A Additional Ablation Results

This appendix reports the ablations of the expert layout, the Stage 1 pretraining signals, and the data scale summarized in Sec. 5.3, evaluated on the EgoLynx-Robot test split under the same protocol.

Table 4: **Ablation on the expert layout on the EgoLynx-Robot test split.** Metrics as in Table 1.

Expert layout	ADE↓	FDE↓	NC↑	TTC↑	IC↑	QNS↑
Full model	**0.027**	**0.052**	**0.982**	**0.928**	**0.921**	**0.884**
Dense backbone	0.029	0.055	0.974	0.918	0.910	0.864
Detached world loss	0.029	0.054	0.977	0.921	0.913	0.870

Table 5: Ablation on pretraining signals and data scale with Stage 2 fixed. The first block varies what EgoLynx-Web supervises; the last two blocks vary the amount of web and robot data. Metrics as in Table 1.

Pretraining	Web (h)	Robot	ADE↓	FDE↓	QNS↑
None	0	2,000	0.034	0.062	0.823
Captions only	10k	2,000	0.031	0.058	0.843
DISR only	10k	2,000	0.029	0.055	0.866
Captions + DISR	10k	2,000	**0.027**	**0.052**	**0.884**
Captions + DISR	1k	2,000	0.031	0.058	0.845
Captions + DISR	3k	2,000	0.030	0.056	0.862
Captions + DISR	10k	500	0.036	0.068	0.829
Captions + DISR	10k	1,000	0.031	0.059	0.856

Expert layout. We next examine two choices in the expert architecture (Table 4): separating expert parameters and allowing the world objective to update the understanding expert. Replacing the Mixture-of-Transformers with a parameter-matched dense backbone reduces QNS by 2 points. Blocking gradients from the world loss to the understanding expert reduces QNS by 1.4 points. Both comparisons favor the full layout, supporting the use of separate experts and world-guided training of the understanding representation.

Pretraining signals. We examine which Stage 1 signals from EgoLynx-Web contribute to downstream navigation (Table 5). Relative to no pretraining, captions alone improve QNS by 2 points, DISR alone by 4.3 points, and their combination by 6.1 points. DISR provides the larger individual gain, while captions add a further 0.018 QNS when combined with DISR.

Data scaling. Finally, we vary the amounts of web video and robot data. Performance improves monotonically with the number of web hours and shows no sign of saturation at 10k hours, suggesting further gains from larger corpora. Web pretraining also markedly improves data efficiency: with EgoLynx-Web, DR-VLA trained on only 500 robot clips already surpasses the variant trained on all 2,000 clips without pretraining (0.829 vs. 0.823 QNS).

B Quadruped Navigation Metric

Trajectory imitation metrics. Given predicted and reference trajectories, ADE and FDE measure the mean and final Euclidean position errors, respectively. AHE and FHE measure the corresponding mean and final absolute heading errors, with angular differences wrapped to $[-\pi, \pi)$.

Pseudo-closed-loop replay.

Annotated agents are replayed against the predicted ego trajectory and six interpretable sub-scores measure No Collision (NC), Walkable-Area Compliance (WAC), Time-to-Collision(TTC), Clearance (CL), Ego Progress (EP), and Instruction Compliance (IC). They are combined as

$$\mathrm{QNS} = \mathrm{NC} \times \mathrm{WAC} \times \frac{w_{\mathrm{EP}}\mathrm{EP} + w_{\mathrm{TTC}}\mathrm{TTC} + w_{\mathrm{CL}}\mathrm{CL} + w_{\mathrm{IC}}\mathrm{IC}}{w_{\mathrm{EP}} + w_{\mathrm{TTC}} + w_{\mathrm{CL}} + w_{\mathrm{IC}}}. \tag{8}$$

The replay is non-reactive.

- **No Collision (NC):** 0 for an at-fault collision with a person, rider, vehicle, animal, or robot; 0.5 for an inanimate object; and 1 otherwise. A collision is not at fault when the ego is stationary and another object moves into it.

- **Walkable-Area Compliance (WAC):** 0 if the ego footprint crosses a curb or a road–sidewalk dividing line outside a crosswalk, unless the recorded trajectory crosses the same boundary within the horizon; 1 otherwise.
- **Time-to-Collision (TTC):** 0 if 1-s constant-velocity propagation from any predicted step intersects a replayed agent; 1 otherwise.
- **Clearance (CL):** $\min(1, d_{\min}/0.5\,\mathrm{m})$, where $d_{\min}$ is the minimum distance to any person, rider, or animal.
- **Ego Progress (EP):** predicted progress projected onto the recorded path, normalized by the recorded-path length and clipped to $[0, 1]$; set to 1 when the recorded motion is below 0.2 m.
- **Instruction Compliance (IC):** checks instruction-specific trajectory geometry: at least 20° signed heading change for branches, U-turns, and in-place rotation; signed dominant lateral velocity for lateral or lane moves; negative longitudinal velocity for reversing; displacement below 0.2 m for stationary or waiting; passage on the instructed side for obstacle avoidance; and continued motion when the operator moved for path following or continuing straight.