PROMAC: LANGUAGE-GUIDED WORLD ACTION MODELING FOR SINGLE-VIEW EGOCENTRIC NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric visual navigation requires an agent to infer how to move through its surroundings toward a desired destination. We formulate this problem as predicting a feasible sequence of 2D waypoints from a *single current egocentric image* with a marked goal point. The goal specifies *where* to move, but reaching it requires reasoning about traversable space and intermediate motion while anticipating how the scene will evolve. This motivates representations that jointly capture semantic transition understanding and action-conditioned visual dynamics. We introduce **PROMAC**, a world action model for egocentric navigation that uses auxiliary chain-of-thought (CoT) reasoning reconstruction objective only at training time to guide world modeling. PROMAC trains the visual token hidden states of a large multimodal model (LMM) backbone through two complementary pathways: (i) a frozen auxiliary language decoder provides supervision that compresses CoT reasoning information into the visual hidden states; (ii) a next-frame latent prediction head uses these "CoT-conditioned" visual patch hidden states to predict future visual token representations from a pretrained vision encoder. This *language-guided world modeling* objective grounds semantic transition reasoning in predicted visual dynamics, encouraging the trajectory policy to capture both how to move and how the scene will change under that movement. The reasoning and world modeling modules are used *only during training*. At inference, the large multimodal model directly generates the 2D trajectory without explicit CoT generation or latent rollouts. On Ego4D Fut-Loc, PROMAC reduces ADE, FDE, L2@First4, and DTW by 33.4%, 59.2%, 13.2%, and 49.1%, respectively, relative to trajectory-only Qwen3-VL-4B.

1 INTRODUCTION

Egocentric visual navigation requires transforming first-person perception into spatially grounded motion through the surrounding environment. This distinction is increasingly important as camera-equipped wearable devices such as smart glasses provide persistent, hands-free perception of a user's surroundings (Pan et al., 2025; Yehia et al., 2026; Kim et al., 2026a). Such systems could support applications ranging from augmented and mixed reality to wearable robotics and assistive navigation for people who are blind or have low vision (BLV), where useful guidance must communicate not only scene content but also traversable space and actionable movement (Azzino et al., 2024; Kim et al., 2026b).

Recent large multimodal models (LMMs) have demonstrated strong understanding and narration of first-person scenes, including objects, spatial relations, and ongoing activities (Lin et al., 2022; Zhao et al., 2023; Pramanick et al., 2023; Mangalam et al., 2023; Kim et al., 2026b; Windecker et al., 2026). Yet language remains an indirect representation of physical motion: describing that "a pedestrian is ahead" or that "the walkway bends left" does not itself specify a geometrically valid path. Recent embodied evaluations similarly expose a gap between semantic scene understanding and spatially grounded navigation (Wang et al., 2025; Zhao et al., 2025; Kim et al., 2026b). Image-space trajectories, in contrast, provide a more direct interface, representing movement as an ordered sequence of 2D waypoints grounded in the current view (Lu et al., 2026; Windecker et al., 2026).

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

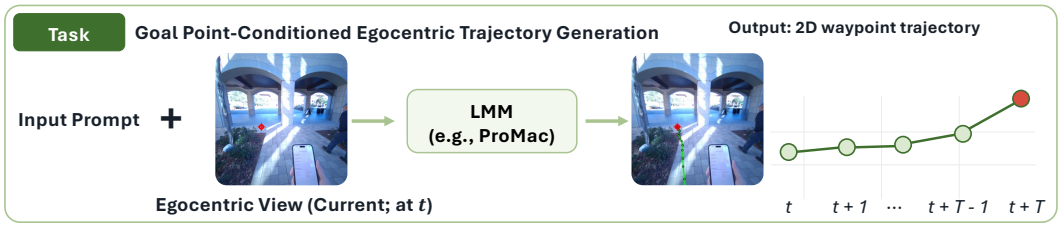


Figure 1: We study *goal-point-conditioned egocentric trajectory generation* as a local visual planning task. Given a single egocentric observation with a reachable local goal marked in the image, the model predicts the intermediate 2D waypoint trajectory needed to reach it.

In this work, we study *goal-point-conditioned egocentric trajectory generation* (Fig. 1): predicting an ordered 2D path from a single current image with a reachable local destination marked in the image. This *local visual planning* task separates goal localization from path construction: a planner, semantic grounding module, or user could supply the goal, e.g., a smartphone handling global GPS routing can use standard Visual-Inertial Odometry (VIO) and camera calibration to mathematically project the next map node directly into the 2D camera view, while a fixed language query isolates intermediate path generation from instruction interpretation. For Aria Navigation Dataset (AND) (Pan et al., 2025) and Ego4D Future Locomotion (Fut-Loc)(Grauman et al., 2022), we derive goals from the final valid visible waypoint of actual human trajectories projected into the current view (§4.1). Even with the endpoint supplied, predicting a path requires identifying traversable space, avoiding obstacles, and respecting local scene geometry during the navigation.

Recent vision-language-action (VLA) models have increasingly coupled intermediate reasoning with action or trajectory generation, including chain-based driving reasoning (Nie et al., 2025; Tian et al., 2025; Liu et al., 2025; Zeng et al., 2025), embodied chain-of-thought (CoT) reasoning (Zawalski et al., 2025), and textual reasoning for trajectory generation (Zhou et al., 2025b; Liu et al., 2025). However, explicit CoT remains an indirect representation of motion and adds autoregressive latency before action generation, motivating adaptive, latent, and multimodal alternatives (Luo et al., 2025; Tan et al., 2026; Zhang et al., 2026; Lu et al., 2026). Moreover, much of this trajectory-generation literature focuses on autonomous driving, where motion is constrained by road structure, vehicle dynamics, traffic conventions, and explicit control prompts (Ghosh et al., 2025; Dauner et al., 2024; Jia et al., 2024). Human locomotion instead requires selecting paths through less structured spaces such as sidewalks, corridors, crowds, and open areas directly from local egocentric geometry. This motivates our central question: **can language reasoning shape the predictive world representation used for egocentric navigation without requiring the model to explicitly verbalize that reasoning?**

We introduce **ProMac**, a language-guided world action model with complementary training objectives: a frozen auxiliary language decoder supervises navigation-relevant CoT information in backbone visual-token hidden states, while a lightweight head reconstructs dense future-frame features from those same states (§3). This *language-guided world modeling* grounds semantic transition reasoning in visual dynamics to shape the trajectory policy. Both auxiliary branches are used only during training; inference directly autoregressively generates a 2D trajectory from the goal-overlaid image, without explicit CoT, future-frame input, or latent rollouts. We also manually curate and open-source **2K human-annotated CoT trajectory rationales** for AND and Ego4D Future Locomotion by 33 human annotators, covering traversable space, movement directions, and intermediate movement toward the goal (§4.1), in which we demonstrate the importance and utility of CoT rationale's visual grounding in egocentric navigation.

Our contributions are as follows:

1. *Task and Benchmark:* Goal-point-conditioned human locomotion benchmarks built on AND (Pan et al., 2025) and Ego4D (Grauman et al., 2022), mapping a single goal-overlaid image to directly renderable 2D paths.
2. *Language-Guided World Action Modeling:* ProMac jointly supervises shared visual states with structured CoT and latent future-state reconstruction, only during training, for direct trajectory generation at inference.

3. *Human Reasoning Annotations and Experiments:* 2K human annotated CoT rationales; comparisons with trajectory-only, CoT-based, and latent world-modeling approaches; and analyses of trajectory quality, reasoning information, future-state prediction, and efficiency.

2 Related Work

2.1 Egocentric Human Locomotion and Visual Navigation.

Prior work forecasts ground-plane trajectories in Ego4D (Grauman et al., 2022), future wearer locations and walking paths (Yagi et al., 2018; Wang et al., 2023), and 6D head-pose trajectories in LookOut, which introduces the Aria Navigation Dataset (AND) (Pan et al., 2025). These methods infer motion from temporal history, while NaviTrace predicts language-instructed 2D navigation traces from single images (Windecker et al., 2026). Wearable accessibility systems motivate actionable guidance (Azzino et al., 2024; Kim et al., 2026b), but advances in egocentric understanding (Lin et al., 2022; Zhao et al., 2023; Pramanick et al., 2023; Mangalam et al., 2023) have not resolved gaps in spatially grounded navigation (Wang et al., 2025; Zhao et al., 2025) guided by language-based reasoning. We derive image-space paths from observed human locomotion and mark their destinations, separating goal specification from intermediate path generation in the current image.

2.2 Reasoning and Trajectory Generation for Vision-Language-Action Models.

ROADWork, NAVSIM, DriveLM, and Bench2Drive study visual planning under driving constraints (Ghosh et al., 2025; Dauner et al., 2024; Sima et al., 2025; Jia et al., 2024), while human navigation must also accommodate sidewalks, corridors, and crowds. Driving rationales (Nie et al., 2025; Tian et al., 2025; Liu et al., 2025; Zeng et al., 2025), embodied CoT (Zawalski et al., 2025), and AutoVLA (Zhou et al., 2025b) expose decision-relevant semantics but often add inference-time deliberation. Adaptive, latent, and multimodal reasoning aim to reduce this cost (Luo et al., 2025; Tan et al., 2026; Zhang et al., 2026; Lu et al., 2026). OneVL supervises compact latent representations through separate CoT and future-frame decoders (Lu et al., 2026). ProMac also omits explicit CoT at inference, but jointly supervising backbone visual-token states with reasoning and direct future-feature reconstruction for human navigation.

2.3 Latent World Models for Action and Trajectory Prediction.

LAW predicts future scene features from current representations and ego trajectories (Li et al., 2025), DINO-WM learns action-conditioned transitions in pretrained feature space (Zhou et al., 2025a), and DriveVLA-W0 adds future-state prediction to driving VLA training (Li et al., 2026). ProMac couples future-feature reconstruction with structured CoT supervision through shared backbone states, linking predictive and semantic learning while retaining trajectory-only inference.

3 ProMac: Language-Guided World Action Modeling for Trajectory Generation

3.1 Preliminaries

ProMac combines trajectory supervision, structured chain-of-thought (CoT) rationales, and future observations to shape shared backbone visual hidden states for egocentric trajectory generation.

Let X_t be the current goal-overlaid egocentric image, q the navigation prompt, and $\tau_t = [(x_1, y_1), \ldots, (x_K, y_K)]$ the target 2D trajectory, serialized as tokens $\mathbf{y}_t = (y_1, \ldots, y_M)$. Each training example also provides a bare future frame X_{t+1}^{bare} and a structured rationale $\mathbf{r}_t = (r_1, \ldots, r_J)$, giving $\mathcal{D}_t = (X_t, q, \tau_t, X_{t+1}^{\text{bare}}, \mathbf{r}_t)$. Here $t+1$ denotes the dataset-provided future observation, not necessarily the adjacent video frame.

Training proceeds in three stages: **Stage 0** learns CoT reconstruction through a projector and frozen auxiliary LLM; **Stage 1** calibrates the future-state reconstruction head on frozen backbone features; and **Stage 2** jointly optimizes trajectory generation and both auxiliary objectives.

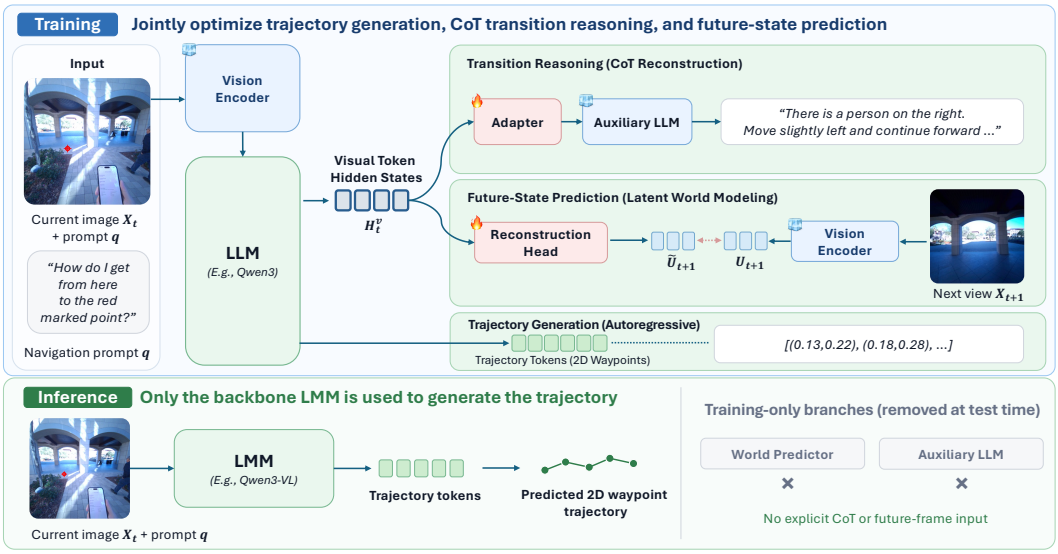


Figure 2: **Overview of ProMac.** An LMM backbone receives a goal-overlaid egocentric image and navigation prompt and autoregressively generates a 2D waypoint trajectory. During training, ProMac reconstructs structured CoT from visual hidden states and uses the reasoning-aligned representation to condition latent next-state prediction; at inference, both auxiliary branches are removed for trajectory-only generation.

3.2 Stage 0: Compress Chain-of-Thought Reasoning for Navigation

Let F_θ be the LMM backbone, e.g., Qwen3-VL, and $\mathcal{V}_t$ the positions occupied by image tokens in its multimodal input sequence. We extract the final-layer language-backbone hidden states at those positions:

$$\mathbf{H}_t^v = [\mathbf{h}_i^L]_{i \in \mathcal{V}_t} \in \mathbb{R}^{N_t \times d_h}. \tag{1}$$

These are contextual backbone hidden states, not the raw outputs of the vision encoder. $\mathbf{H}_t^v$ encode the goal-overlaid image and the preceding context, and can acquire reasoning-relevant information through training gradients (**Stage 0**). We encourage $\mathbf{H}_t^v$ to support navigation-relevant reasoning through a frozen auxiliary language model A_ω. A trainable adapter ψ maps backbone visual states into the auxiliary model's input-embedding space:

$$\tilde{\mathbf{H}}_t^v = \psi(\mathbf{H}_t^v). \tag{2}$$

The adapter consists of two linear layers with a GELU nonlinearity, preceded and followed by layer normalization. We insert its outputs at the image-token positions of the auxiliary model's input prompt, replacing the usual image embeddings with $\tilde{\mathbf{H}}_t^v$. Let q_r denote the navigation prompt adapted to request a reasoning trace. With teacher forcing, the auxiliary objective is

$$\mathcal{L}_{\text{reason}} = -\frac{1}{J} \sum_{j=1}^{J} \log p_\omega(r_j \mid r_{<j}, \tilde{\mathbf{H}}_t^v, q_r). \tag{3}$$

The auxiliary parameters ω remain frozen, but gradients pass through A_ω to the adapter and, during backbone training, to θ, encouraging navigation reasoning to be encoded in the visual hidden states.

3.3 Stage 1: Direct Future Visual-State Reconstruction

To encode the visual environment's temporal change, a lightweight reconstruction head g_ϕ predicts future visual features from the same CoT-supervised backbone hidden states $\mathbf{H}_t^v$ defined in Eq. 1:

$$\hat{\mathbf{U}}_{t+1} = g_\phi(\mathbf{H}_t^v). \tag{4}$$

The target $\bar{\mathbf{U}}_{t+1}$ is constructed from the bare future frame using the frozen vision encoder. We optimize

$$\mathcal{L}_{\text{next}} = \text{MSE}\Big(\hat{\mathbf{U}}_{t+1}, \text{sg}(\bar{\mathbf{U}}_{t+1})\Big), \tag{5}$$

where sg denotes stop-gradient. The target construction, reconstruction-head parameterization, and loss normalization are detailed in Appendix B.

Both auxiliary losses update the shared backbone hidden states, grounding language-guided representations in future visual dynamics. The reconstruction head receives neither generated rationale tokens, auxiliary-decoder hidden states, nor trajectory-token states; its predictions influence the trajectory policy through backbone supervision, not as decoder inputs.

The trajectory backbone is jointly trained with teacher-forced autoregressive likelihood over trajectory-response tokens only:

$$\mathcal{L}_{\text{traj}} = -\frac{1}{M}\sum_{j=1}^{M} \log p_\theta(y_j \mid y_{<j}, X_t, q). \tag{6}$$

The combined objective is $\mathcal{L} = \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \lambda_{\text{reason}}\mathcal{L}_{\text{reason}} + \lambda_{\text{next}}\mathcal{L}_{\text{next}}$, with $(\lambda_{\text{traj}}, \lambda_{\text{reason}}, \lambda_{\text{next}}) = (1, 0.01, 0.05)$. Joint training updates the language backbone, reconstruction head, and reasoning adapter while keeping the vision encoder and auxiliary LLM frozen. At inference, both auxiliary branches are removed: the backbone generates the trajectory from (X_t, q) without explicit CoT, future frames, or latent rollouts.

4 Experiments

We evaluate ProMac on goal-point-conditioned egocentric trajectory generation using the Aria Navigation Dataset (AND) (Pan et al., 2025) and Ego4D Future Locomotion (Grauman et al., 2022). Our experiments are designed to answer three primary questions: (i) whether ProMac improves goal-conditioned trajectory generation over direct trajectory prediction and existing reasoning/world-modeling approaches; (ii) whether the world-modeling branch learns representations that are predictive of the correct future visual state; and (iii) whether transition-reasoning supervision encourages the visual hidden states to encode navigation-relevant semantic information.

4.1 Dataset Construction

We construct goal-point-conditioned trajectory benchmarks from AND (Pan et al., 2025), which records indoor and outdoor navigation via human locomotion with Project Aria glasses, and Ego4D Future Locomotion (Fut-Loc) (Grauman et al., 2022), which provides first-person video and future wearer trajectories recovered from camera motion. We project their 3D future ego-trajectories into the current image to obtain 2D path supervision.

Table 1: **Statistics of the goal-point-conditioned egocentric trajectory-generation benchmarks.** Both datasets use a 4 s future horizon but differ in camera geometry and temporal sampling. "Waypoints" reports the median number of valid 2D trajectory points per example.

	AND	**Ego4D Fut-Loc**
Train examples	9,400	36,306
Validation examples	1,051	3,016
Image resolution	512×512	1280×720
Future horizon	4.0 s	4.0 s
Sampling interval	0.25 s	0.10 s
Median waypoints	12	40
Coordinate space	2D pixels	2D pixels

Each benchmark example consists of a single current frame X_t, an image-space goal point $\mathbf{g}_t$, and a projected 2D trajectory

$$\tau_t^{2D} = [(x_1, y_1), \ldots, (x_K, y_K)]\,. \tag{7}$$

We select the final valid projected waypoint as the destination and render it as a red marker on X_t, producing the goal-overlaid observation $\tilde{X}_t$. The model receives only $\tilde{X}_t$ together with the fixed prompt *"How do I get from here to the red marked point?"* and must generate the corresponding ordered sequence of 2D waypoint coordinates. Importantly, only the goal marker is rendered in the input; the ground-truth trajectory itself is never shown to the model and the textual instruction does not indicate the directionality of movement.

AND trajectories are constructed by projecting future Project Aria device poses into the current rectified RGB view over a 4 s horizon, while Fut-Loc trajectories are obtained by projecting Ego4D future camera centers into the current fisheye image using the provided camera calibration. Because the two datasets use different cameras and resolutions, all trajectory metrics are computed within

Table 2: **Goal-point-conditioned egocentric trajectory generation on AND and Ego4D Future Locomotion (Fut-Loc).** We report Average Displacement Error (ADE), Final Displacement Error (FDE), L2 error over the first four trajectory points (L2@First4), and Dynamic Time Warping distance (DTW). Lower is better for all metrics. All methods receive the same single goal-overlaid egocentric observation at inference time. Best and second-best distinct values are bold and underlined, respectively.

Method	Ego4D Fut-Loc				Aria Navigation Dataset (AND)			
	ADE↓	FDE↓	L2@First4↓	DTW↓	ADE↓	FDE↓	L2@First4↓	DTW↓
Zero-Shot Baselines								
Gemini-3.8-Flash	331.14	293.02	494.03	291.00	190.59	166.00	200.11	147.92
GPT-6 Astra	189.85	87.72	365.99	152.31	46.00	39.11	49.37	26.10
Claude Fable 5.1	203.00	88.21	379.50	158.94	48.21	37.29	51.66	26.31
Qwen3-VL-4B	308.00	347.81	510.16	299.54	196.49	269.36	139.25	186.11
Qwen3-VL-4B (CoT)	332.15	387.05	541.46	324.73	133.78	187.93	113.92	139.97
Qwen3-VL-8B	280.10	312.59	443.91	260.70	258.21	297.49	216.72	229.50
Qwen3-VL-8B (CoT)	299.83	329.49	465.01	278.90	228.97	254.03	197.74	190.43
Qwen3-VL-32B	280.83	322.17	467.87	264.33	221.81	246.56	192.50	190.96
Qwen3-VL-32B (CoT)	329.85	346.81	529.44	301.77	159.40	195.18	134.90	144.60
Fine-Tuned Qwen Baselines								
Qwen3-VL-4B	163.13	200.91	119.26	132.47	17.79	9.45	24.93	16.28
Qwen3-VL-4B (CoT)	141.78	174.46	112.18	112.74	<u>17.02</u>	**8.22**	<u>24.48</u>	<u>15.97</u>
Reasoning-Augmented Trajectory Generation								
AutoVLA (Zhou et al., 2025b)	151.35	216.37	**83.82**	103.62	20.39	9.32	27.42	18.98
OneVL (Lu et al., 2026)	160.30	228.79	<u>90.18</u>	104.22	28.27	9.83	28.79	18.19
World-Modeling-Based Trajectory Generation								
LAW (Li et al., 2025)	184.12	<u>174.00</u>	155.42	136.85	32.54	33.30	33.41	30.74
DriveVLA-W0 (Li et al., 2026)	<u>109.91</u>	82.06	104.43	<u>67.62</u>	17.60	<u>8.69</u>	25.21	16.32
ProMac	**108.40**	**82.01**	103.46	**67.45**	**16.83**	9.04	**23.39**	**14.42**

each dataset separately. Full projection equations, camera conventions, visibility filtering, goal-selection rules, and quality-control procedures are provided in Appendix A, where Appendices A.4–A.5 detail CoT construction, schema design, and human annotation.

4.2 Experimental Setup

Backbone and Architecture. We use Qwen3-VL-4B-Instruct as the main trajectory backbone in this work and a separate frozen Qwen3-VL-4B as the auxiliary language decoder. The reconstruction head linearly maps each final-layer backbone visual state to four dense pre-merger future features (Appendix B). We also provide Qwen3-VL-8B as backbone experiments in the Appendix as an additional experiment on the effect of parameter size changes.

Baselines. We compare ProMac with trajectory-only Qwen3-VL-4B, explicit reasoning followed by trajectory generation in AutoVLA (Zhou et al., 2025b), action-conditioned latent world modeling in LAW (Li et al., 2025), prepended latent token-based language and visual supervision in OneVL (Lu et al., 2026), and future-state prediction during VLA training in DriveVLA-W0 (Li et al., 2026). AutoVLA, OneVL and DriveVLA-W0 are finetuned with AND and Fut-Loc datasets on Qwen3-VL-4B backbones. We also evaluate open-source Qwen3-VL at 4B, 8B, and 32B parameters, with and without CoT reasoning, to assess how model size affects trajectory generation. We additionally test state-of-the-art closed-source Gemini-3.8-Flash, GPT-6 Astra, and Claude Fable 5.1 models. All methods receive the same goal-overlaid egocentric image and generate 2D waypoints.

Evaluation Metrics. We report four complementary trajectory errors in image-space pixels. Let $\hat{\tau} = \{\hat{\mathbf{p}}_k\}_{k=1}^{K}$ and $\tau = \{\mathbf{p}_k\}_{k=1}^{K}$ denote the aligned predicted and ground-truth waypoint sequences, truncated to their shared length K, and let $e_k = \|\hat{\mathbf{p}}_k - \mathbf{p}_k\|_2$. **Average Displacement Error (ADE)**, $\mathrm{ADE} = K^{-1}\sum_{k=1}^{K} e_k$, measures average waypoint error. **Final Displacement Error (FDE)**, $\mathrm{FDE} = e_K$, measures error at the last aligned waypoint. Since the destination is explicitly marked in the input, we also report **L2@First4**, $\mathrm{L2@First4} = K_4^{-1}\sum_{k=1}^{K_4} e_k$, $K_4 = \min(4, K)$, to emphasize near-term path accuracy. Finally, **Dynamic Time Warping (DTW)** measures path similarity while allowing flexible alignment between the original waypoint sequences. We report $\mathrm{DTW} = |\pi^\star|^{-1}\sum_{(i,j)\in\pi^\star}\|\hat{\mathbf{p}}_i - \mathbf{p}_j\|_2$, where $\pi^\star$ is the monotonic warping path minimizing total Euclidean alignment cost and $|\pi^\star|$ is its number of matched pairs. Unlike point-wise errors, DTW accommodates differences in waypoint sampling. We average per-example scores across the evaluation set; lower is better for all four metrics.

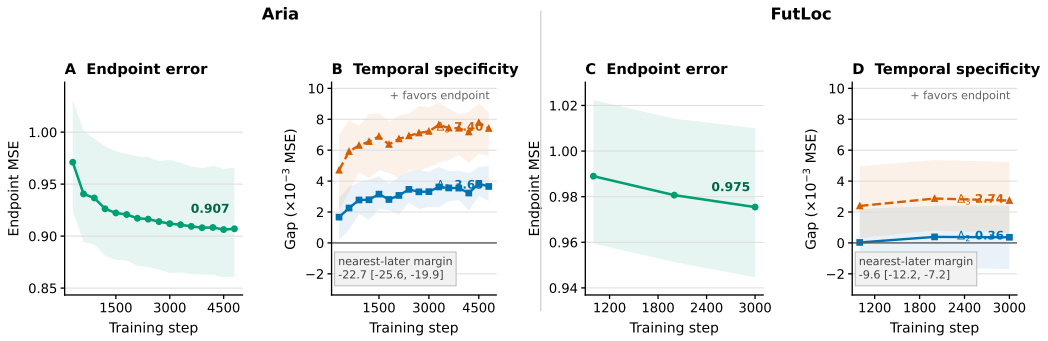


Figure 3: **Latent world modeling and temporal specificity.** A,C: supervised-endpoint MSE across checkpoints on Aria and FutLoc. B,D: paired gaps $\Delta_k = d_{t+k} - d_{t+1}$, $k \in \{2, 3\}$; positive values favor the endpoint. Curves show sample means with 95% sequence-cluster bootstrap intervals. Boxes report the mean nearest-later margin and its 95% interval; gaps and margins use 10^{-3} MSE units.

4.3 Main Results: Goal-Conditioned Egocentric Trajectory Generation

Table 2 compares ProMac with direct trajectory-generation, explicit reasoning, and world-modeling baselines under the same single-image input.

Overall Trajectory Accuracy. ProMac achieves the lowest reported ADE and DTW on both datasets without explicit CoT generation. On Fut-Loc, it reduces ADE/FDE/DTW by 33.4%/59.2%/49.1% over trajectory-only Qwen and surpasses DriveVLA-W0 on ADE, L2@First4, and DTW while matching FDE. On AND, it leads on ADE, L2@First4, and DTW and improves all four errors over trajectory-only Qwen. Table 3 analyzes the impact of ProMac's auxiliary objectives.

4.4 Ablation of Auxiliary Objectives

Table 3: **Ablation of ProMac's auxiliary training objectives.** Predictor Only uses future-state reconstruction, Aux. LLM Only uses reasoning supervision, and Full combines both with trajectory training.

Variant	Ego4D Fut-Loc				AND			
	ADE	FDE	L2@First4	DTW	ADE	FDE	L2@First4	DTW
ProMac (Predictor Only)	110.76	86.07	102.76	87.32	16.13	**7.43**	23.78	15.57
ProMac (Aux. LLM Only)	109.92	84.68	**102.16**	67.85	**16.09**	7.57	23.43	14.75
ProMac (Full)	**108.40**	**82.01**	103.46	**67.45**	16.83	9.04	**23.39**	**14.42**

Either auxiliary objective improves all four errors over trajectory-only Qwen (Table 3), consistent with prior predictive (Li et al., 2025; 2026) and reasoning supervision (Zawalski et al., 2025; Zhou et al., 2025b). Full improves Fut-Loc ADE/FDE/DTW but worsens L2@First4. On AND, reasoning alone is already competitive: Full improves L2@First4/DTW but worsens ADE/FDE, indicating a trade-off rather than uniform complementarity. Near-straight paths account for 96.8% of AND versus 51.6% of Fut-Loc targets (Appendix E). This result suggests that simpler geometry in trajectory and minimal change in temporal dynamics contributes to such competing optimization signals.

4.5 Latent World Modeling and Temporal Specificity

Figure 3 examines future-state learning using predictions conditioned on the current image and ground-truth intermediate trajectory tokens. We compare predictions with the supervised endpoint (t+1) and two later observations (t+2, t+3) using patch-level MSE over visual tokens. Endpoint MSE falls by 6.6% on Aria and 1.4% on FutLoc. Although mean errors favor t+1 over each fixed later observation, the mean per-query nearest-later margin $m = \min(d_{t+2}, d_{t+3}) - d_{t+1}$ is negative: -22.7×10^{-3} and -9.6×10^{-3} MSE. The reduction in patch-level reconstruction error demonstrates improved prediction of future visual features, while the negative nearest-later margins indicate limited discrimination between the supervised future observation and nearby later states.

4.6 Do Visual Hidden States Encode Transition Reasoning?

We probe whether navigation-relevant CoT information is accessible from the backbone's visual hidden states by decoding a rationale through the frozen auxiliary LLM. The adapter maps the final-layer visual states $\mathbf{H}_t^v$ to $\hat{\mathbf{H}}_t^v = \psi(\mathbf{H}_t^v)$, which replace the auxiliary LLM's image-token embeddings. Given these states and the reasoning prompt, the auxiliary LLM generates a CoT sequence. Vanilla Qwen's image-conditioned CoT provides a quality reference for this representation readout. Additional CoT-generation efficiency results appear in Section 4.9.

LLM-as-a-Judge Protocol and Reasoning Readout. We use frozen Qwen3-VL-235B-A22B-Instruct to judge each rationale against a reference, given the goal-overlaid image and task prompt, with source-blinded, deterministically randomized A/B labels. Grounding, trajectory consistency, safety, and clarity receive independent 1–5 scores, with Overall being their mean (Figure 4). The auxiliary readout scores 3.53 overall and 3.33 for grounding versus Qwen's 3.60 and 3.43, closely matching its rounded trajectory-consistency score (3.78), and produces valid JSON in 98.57% of cases. These results support navigation-relevant reasoning information being recoverable from adapted backbone visual states by a frozen decoder given the task prompt, despite modest gaps from direct generation. Such an observation, along with the results in Table 2, also strongly indicates that the compressed CoT related information within the visual hidden states are beneficial for improved trajectory generation.

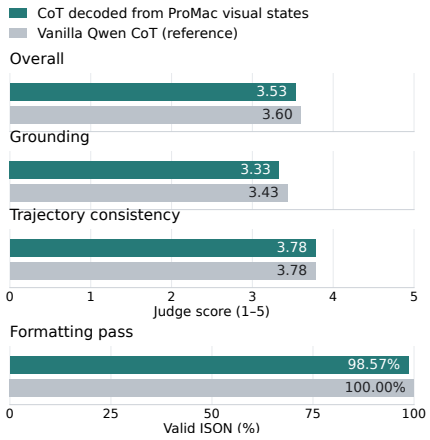


Figure 4: **CoT decoded from visual states on Aria.** Vanilla Qwen CoT serves as the upper bound quality reference.

4.7 Semantic Consistency of Human-Reviewed CoT

Complementing Section 4.6, we assess human revisions for 1,000 samples per benchmark. Annotators could revise any candidate or write custom responses; the reason we opt for pairwise cosine similarity instead of inter-annotator agreement in this work. Across all pairs, MPNet cosine similarity of final cause and critical-factor texts reaches 0.800 on Aria and 0.842 on FutLoc (Figure 5A), which demonstrates that the CoT rationales selected or revised by annotators, overall, share high semantic similarities. Final directions fit a common ± 1-hour window showing 82.0% and 98.0% in agreement (B). Shared content supports the supervision premise (§ 3.2); preceding probes examine what visual states retain (§ 4.6).

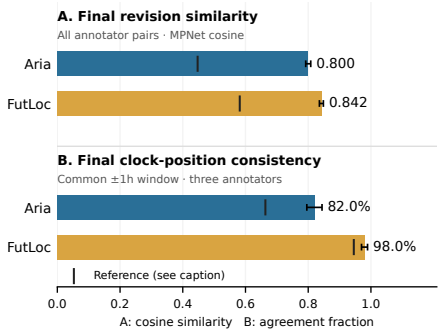


Figure 5: **Human CoT consistency.** Whiskers show 95% CIs; black ticks show reference levels defined in Appendix A.5.

4.8 Effect of Prior-Frame Context

We use $K = \{1, 3\}$ prior frames at 1 Hz throughout ProMac training and inference. All errors favor the archived $K = 0$ reference (Table 4), supporting current-view-only inputs in our currently proposed task design. The similarity between the prior and current frames (i.e., frame redundancies) could account for the drop in performance (Appendix F).

Table 4: **Prior frame-trained promac on AND.** Stage-2 errors on the same 834 validation examples; $K = 0$ is an archived reference.

K	ADE↓	FDE↓	L2@First4↓	DTW↓
0	**16.83**	**9.04**	**23.39**	**14.42**
1	19.51	11.39	26.96	18.27
3	18.95	10.84	26.35	18.13

4.9 Efficiency Gains in Reasoning Token Generation

ProMac reduces paired CoT-generation latency by 0.98% on Fut-Loc and 3.87% on 1,044 valid Aria pairs (Table 5), because during inference ProMac does not rely on generating CoT rationales to enable enhanced trajectory generation. Completion-token counts fall by 4.34% and 7.56%, respectively (Figure 6). At deployment, ProMac generates only trajectory tokens, omitting rationale decoding and both auxiliary branches.

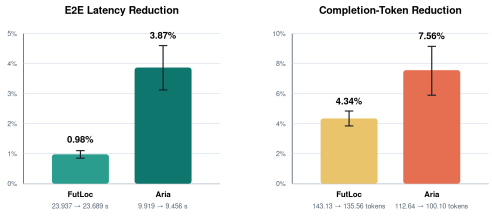


Figure 6: **CoT efficiency gains over vanilla Qwen:** latency and token reductions.

Table 5: **Absolute CoT latency and paired efficiency gains.** Latency is reported in seconds; intervals are 95% CIs.

Dataset	Vanilla Qwen (s)	ProMac (s)	Paired gain (95% CI)
Fut-Loc	23.937	23.689	0.98% [0.85%, 1.10%]
Aria	9.919	9.456	3.87% [3.12%, 4.60%]

4.10 Qualitative Trajectory Comparisons



Figure 7: **Selected trajectory comparisons on AND (Aria).** Each scene pairs ProMac (left) with vanilla Qwen3-VL answer-only (right). Green: ground truth; red: prediction; large dots: final waypoints.

Figure 7 shows two selected cases favoring ProMac. In the plaza, ADE falls from 63.94 to 17.44 pixels and DTW from 53.82 to 14.23. In the walkway, both models reach the goal, but ProMac better follows the intermediate path, illustrating the limits of endpoint accuracy alone.

5 Conclusion, Limitations, Future Works

We presented ProMac, which uses navigation reasoning and future-state prediction to improve trajectory generation without requiring either process at inference. Its two auxiliary objectives shape shared visual hidden states: language supervision emphasizes route constraints, while future-feature reconstruction connects those representations to observed visual change. We also introduce goal-point-conditioned benchmarks derived from AND and Ego4D Fut-Loc, together with human-annotated navigation rationales. ProMac achieves the lowest reported ADE and DTW on both benchmarks, reducing Fut-Loc errors by 33.4% and 49.1%, respectively, over trajectory-only Qwen. Ablations show that both objectives help individually, while their combined benefits depend on the dataset and metric. These findings support training with richer reasoning and predictive supervision while deploying a direct trajectory generator. Our single-image, open-loop evaluation does not establish dynamic-obstacle understanding or safe execution. Future work will integrate temporal learning and test closed-loop navigation, uncertainty, and efficiency on wearable devices for real-world deployment of the proposed method.

AI USE STATEMENT

Generative AI assistance, including Codex, was used to refine hypotheses, develop analysis code, and revise manuscript text, tables, and figures. These uses included the validation-trajectory geometry analysis in Appendix E. Separately, multimodal language models generated candidate navigation rationales used as training supervision (Appendix A.4), and Qwen3-VL-235B-A22B-Instruct served as an automated rationale-quality judge (Section 4.6). Human revisions of generated rationales are distinguished from automatic annotations in Appendix A.5. AI-generated suggestions and judgments are not treated as independent evidence of correctness or safety. The authors take responsibility for the final text, claims, code, analyses, and other artifacts, including those developed with AI assistance.

ETHICS STATEMENT

This work studies navigation assistance using existing egocentric recordings and human review of navigation rationales. Such recordings do not contain identifiable people and sensitive environmental information; their use and any release of derived materials must respect the source datasets' (AND, Ego4D) access conditions, licenses, and privacy protections. Appendix A.5 describes the annotation workflow and distinguishes individual human revisions from adjudicated consensus. Our evaluation measures agreement with recorded trajectories and does not establish collision avoidance, accessibility, or safety during real-world execution. Dataset coverage, camera geometry, and differences in mobility or surroundings may limit generalization. Deployment, particularly for blind and low-vision users, requires evaluation with intended users, uncertainty-aware safeguards, and appropriate human oversight.

REPRODUCIBILITY STATEMENT

Section 3 specifies the architecture and training objectives, and Section 4.2 describes the baselines and evaluation metrics. Appendix A documents trajectory projection, goal selection, dataset splits, and input-leakage controls; Appendices A.4 and A.5 describe rationale construction and human-review processing. Appendix B details reconstruction targets, head parameterization, and loss normalization. The zero-shot evaluation protocol, trajectory-geometry analysis, and prior-frame-context experiment are described in Appendices C, E, and F, respectively. These descriptions identify the preprocessing choices and evaluation conditions needed to interpret and reproduce the reported comparisons.

REFERENCES

Tommy Azzino, Marco Mezzavilla, Sundeep Rangan, Yao Wang, and John-Ross Rizzo. 5g edge vision: Wearable assistive technology for people with blindness and low vision. In *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2024. doi: 10.1109/WCNC57260.2024.10570607.

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*, volume 37, 2024.

Anurag Ghosh, Shen Zheng, Robert Tamburo, Khiem Vuong, Juan Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G. Narasimhan. ROADWork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6132–6142, 2025.

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, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jeonghwan Kim, Renjie Tao, Sanat Sharma, Jiaqi Wang, Kai Sun, Zhaojiang Lin, Seungwhan Moon, Lambert Mathias, Anuj Kumar, Heng Ji, and Xin Luna Dong. Pixel-grounded retrieval for knowledgeable large multimodal models. *arXiv preprint arXiv:2601.19060*, 2026a.

Junhyeok Kim, Jaewoo Park, Junhee Park, Sangeyl Lee, Jiwan Chung, Jisung Kim, Ji Hoon Joung, and Youngjae Yu. GuideDog: A real-world egocentric multimodal dataset for blind and low-vision accessibility-aware guidance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5545–5574. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.251.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations (ICLR)*, 2025.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-W0: World models amplify data scaling law in autonomous driving. In *International Conference on Learning Representations (ICLR)*, 2026.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7575–7586, 2022.

Xueyi Liu, Zuodong Zhong, Junli Wang, Yuxin Guo, Zhiguo Su, Qichao Zhang, Yun-Fu Liu, Yinfeng Gao, Yupeng Zheng, Qiao Lin, Huiyong Chen, and Dongbin Zhao. ReasonPlan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3051–3068. PMLR, 2025.

Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, Fang Li, Chenxu Dang, Junli Wang, Tao Xu, Jing Wu, Jianhua Wu, Xiaoshuai Hao, Wen Zhang, Tianyi Jiang, Lingfeng Zhang, Lei Zhou, Yingbo Tang, Jie Wang, Yinfeng Gao, Xizhou Bu, Haochen Tian, Yihang Qiu, Feiyang Jia, Lin Liu, Yigu Ge, Hanbing Li, Yuannan Shen, Jianwei Cui, Hongwei Xie, Bing Wang, Haiyang Sun, Jingwei Zhao, Jiahui Huang, Pei Liu, Zeyu Zhu, Yuncheng Jiang, Zibin Guo, Chuhong Gong, Hanchao Leng, Kun Ma, Naiyan Wang, Guang Chen, Kuiyuan Yang, Hangjun Ye, and Long Chen. Xiaomi OneVL: One-step latent reasoning and planning with vision-language explanation. *arXiv preprint arXiv:2604.18486*, 2026.

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. *arXiv preprint arXiv:2509.13769*, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In *Computer Vision – ECCV 2024*, pp. 292–308. Springer, 2025. doi: 10.1007/978-3-031-73347-5_17.

Boxiao Pan, Adam W. Harley, Francis Engelmann, C. Karen Liu, and Leonidas J. Guibas. LookOut: Real-world humanoid egocentric navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24977–24988, 2025.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5285–5297, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Computer Vision – ECCV 2024*, pp. 256–274. Springer, 2025. doi: 10.1007/978-3-031-72943-0_15.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krähenbühl, Marco Pavone, and Boris Ivanovic. Latent chain-of-thought world modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39724–39733, 2026.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4698–4726. PMLR, 2025.

Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via MetaVQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22453–22464, 2025.

Weizhuo Wang, Michael Raitor, Steve Collins, C. Karen Liu, and Monroe Kennedy, III. Trajectory and sway prediction towards fall prevention. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10483–10489, 2023. doi: 10.1109/ICRA48891.2023.10161361.

Tim Windecker, Manthan Patel, Moritz Reuss, Richard Schwarzkopf, Cesar Cadena, Rudolf Lioutikov, Marco Hutter, and Jonas Frey. NaviTrace: Evaluating embodied navigation of vision-language models. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7593–7602, 2018.

Ahmad Yehia, Abduallah Mohamed, Tianyi Wang, Jiseop Byeon, Kun Qian, Junfeng Jiao, and Christian Claudel. EgoTraj: Real-world egocentric human trajectory dataset for multimodal prediction. *arXiv preprint arXiv:2605.19004*, 2026.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3157–3181. PMLR, 2025.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. FutureSightDrive: Thinking visually with spatio-temporal cot for autonomous driving. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Lingjun Zhang, Yujian Yuan, Changjie Wu, Xinyuan Chang, Xin Cai, Shuang Zeng, Linzhe Shi, Sijin Wang, Hang Zhang, and Mu Xu. MindDriver: Introducing progressive multimodal reasoning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17831–17841, 2026.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. UrbanVideo-Bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32400–32423. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1558.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6586–6597, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025a.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

A Construction of the Goal-Point-Conditioned Navigation Benchmarks

We construct two goal-point-conditioned egocentric trajectory-generation benchmarks from existing human-locomotion datasets: the Aria Navigation Dataset (AND) (Pan et al., 2025) and Ego4D Future Locomotion (Fut-Loc) (Grauman et al., 2022). Both datasets contain temporally realized ego-motion, but neither directly provides the task studied in this work: predicting an image-space path from a single egocentric observation toward an explicitly marked destination.

Our construction converts the original 3D future ego-motion into an ordered sequence of 2D image-space waypoints in the *current* camera view. The final valid projected point is selected as the goal and rendered as a red marker onto the current image. Thus, every example follows the common interface

$$\left(\tilde{X}_t,\, q\right) \to \tau_t^{2D}, \tag{8}$$

where $\tilde{X}_t$ is the current image containing only the goal overlay, q is a fixed navigation prompt, and

$$\tau_t^{2D} = [(x_1, y_1), \ldots, (x_K, y_K)] \tag{9}$$

is the projected future ego-trajectory.

Both datasets use the same prompt:

"How do I get from here to the red marked point?"

and expose a common record structure containing the current observation, goal point, projected trajectory, future observation, and diagnostic visualization metadata. The datasets nevertheless differ substantially in their camera models, source geometry, and temporal sampling, summarized in Table 6.

Table 6: **Construction details for AND and Ego4D Fut-Loc.**

	AND	**Ego4D Fut-Loc**
Source	LookOut / Project Aria	Ego4D Future Locomotion
Camera	Rectified Aria RGB	Raw egocentric fisheye
Resolution	512×512	1280×720
3D trajectory source	Future device poses	Future camera centers
Projection model	Pinhole	FOV fisheye
Future horizon	4.0 s	4.0 s
Temporal interval	0.25 s	0.10 s
Maximum trajectory size	17	40
Accepted examples	10,451	39,322
Train / validation	9,400 / 1,051	36,306 / 3,016

A.1 Aria Navigation Dataset

Source Ego-Motion. AND is derived from the Project Aria recordings released with LookOut (Pan et al., 2025). For each RGB frame, the dataset provides temporally aligned closed-loop 6-DoF device poses together with camera intrinsics and the static transformation between the device and RGB camera.

Let

$$\mathbf{T}_d^w(t) \in SE(3) \tag{10}$$

denote the device pose in the world coordinate system at time t, and let $\mathbf{T}_c^d$ denote the fixed transformation from the RGB camera to the device coordinate system. We use the closed-loop pose trajectory because accumulated drift in open-loop odometry can otherwise distort the projected four-second future path.

The future device translation corresponds approximately to the camera position rather than the wearer's ground contact point. We therefore estimate the future ground position as

$$\mathbf{p}^w(t) = \mathbf{T}_d^w(t)[:3, 3] - h\,\mathbf{e}_z, \qquad h = 1.72\ \mathrm{m}, \tag{11}$$

where h approximates the camera height.

Projection into the Current View. Each future ground point is transformed into the coordinate frame of the *current* observation at timestamp t_0:

$$\mathbf{p}^{c_t}(t) = \left(\mathbf{T}_c^d\right)^{-1} \left(\mathbf{T}_d^w(t_0)\right)^{-1} \tilde{\mathbf{p}}^w(t). \tag{12}$$

Writing

$$\mathbf{p}^{c_t}(t) = (x, y, z), \tag{13}$$

we project the point with the provided pinhole calibration:

$$u = f_x \frac{x}{z} + c_x, \qquad v = f_y \frac{y}{z} + c_y. \tag{14}$$

Points behind the current camera or with invalid coordinates are discarded. No distortion model is applied because the source AND RGB images are already rectified.

Display-Frame Rotation. Project Aria RGB images are stored in a sensor orientation that differs from the upright view used by the model. We therefore rotate both the image and projected coordinates clockwise by 90°. For a source image of height H,

$$(x, y) \mapsto (H - y, x). \tag{15}$$

All AND trajectory coordinates reported in this work therefore correspond to the upright 512×512 model input.

Temporal Sampling. For each current observation, we sample the future ego-motion over

$$t_0 + k\Delta t, \qquad k = 0, \ldots, 16, \qquad \Delta t = 0.25\ \mathrm{s}, \tag{16}$$

yielding a maximum of 17 trajectory points across a 4 s horizon. Each target timestamp is matched to the nearest available closed-loop device pose. Duplicate poses introduced by missing temporal samples are removed.

Table 7: **Construction yield for AND.**

Stage	Examples
Candidate observations	11,781
Fewer than 3 visible points	−1,225
Missing valid future frame	−88
Missing current image	−17
Accepted examples	**10,451**
Training	9,400
Validation	1,051

Visibility Filtering. Projected points are retained when they remain in front of the current camera and lie within a 64-pixel margin surrounding the image. The margin permits trajectories to briefly extend beyond the exact frame boundary while remaining associated with the visible local scene. Samples containing fewer than three valid projected waypoints are removed.

For reproducibility, we retain both an unfiltered trajectory that preserves temporal alignment with the original future-pose sequence and a filtered 2D trajectory used as the trajectory-generation target.

Goal Selection. The destination is selected as the final valid future waypoint satisfying three conditions:

1. the point is projectable from the current observation;
2. its temporal offset lies between 1 s and 4 s; and
3. the corresponding future position is at least 1 m from the current position.

The displacement threshold prevents nearly stationary future states from being selected as navigation goals.

The selected goal $\mathbf{g}_t = (u_g, v_g)$ is rendered directly onto the current frame as a red ring and crosshair. Importantly, no portion of the trajectory itself is rendered into the model input.

Future Observation. For the world-modeling objective, we additionally associate each example with the RGB observation temporally closest to the selected goal point. This image serves only as privileged supervision for predicting the future latent state and is never provided to the trajectory policy at inference.

Sampling and Splits. Candidate current observations are sampled every 20 RGB frames. Across all usable sessions, this yields 11,781 candidate examples. After removing examples with insufficient visible waypoints, missing future images, or missing current frames, 10,451 examples remain.

The archived AND split contains 9,400 training and 1,051 validation records with disjoint recording sessions. Validation uses Mainquad_jan10 (606 records), SSC2_jan17 (171), and SSC4_jan17 (274). Checkpoint selection uses validation data, so these are development results rather than an untouched test evaluation.

A.2 EGO4D FUTURE LOCOMOTION

Source Ego-Motion. Ego4D Future Locomotion (Grauman et al., 2022) provides future camera trajectories directly in the coordinate system of the current egocentric frame. Each annotation contains a sequence of future camera centers

$$\mathbf{C}_i \in \mathbb{R}^3 \tag{17}$$

together with a ground-plane normal scaled approximately by the camera height, denoted $\mathbf{u}$.

Because the future camera centers are already represented relative to the current camera, no world-coordinate transformation is required. We approximate the future ground-contact position as

$$\mathbf{p}_i = \mathbf{C}_i - \mathbf{u}. \tag{18}$$

Table 8: **Construction yield for Ego4D Fut-Loc.**

Stage	Train	Val	Total
Original annotations	47,846	3,972	51,818
Invalid future endpoint/image	−11,540	−956	−12,496
Accepted examples	**36,306**	**3,016**	**39,322**

Fisheye Projection. Unlike AND, Ego4D Fut-Loc uses the original fisheye camera geometry. We therefore apply the provided FOV distortion model rather than a pinhole projection.

For normalized image-plane coordinates

$$x_n = \frac{x}{z}, \qquad y_n = \frac{y}{z}, \tag{19}$$

define

$$r_u = \sqrt{x_n^2 + y_n^2}. \tag{20}$$

Using the provided distortion parameter ω, the distorted radius is

$$r_d = \frac{1}{\omega} \arctan\left(2 r_u \tan\frac{\omega}{2}\right). \tag{21}$$

The final image location is

$$u = f_x \frac{r_d}{r_u} x_n + p_x, \qquad v = f_y \frac{r_d}{r_u} y_n + p_y. \tag{22}$$

We use the per-segment camera intrinsics supplied by Ego4D. The resulting coordinates correspond directly to the original distorted 1280×720 fisheye image.

Temporal Sampling. Fut-Loc trajectories are sampled over a 4 s future horizon at approximately 10 Hz, yielding up to 40 waypoints per example:

$$\Delta t = 0.1 \text{ s}, \qquad K \leq 40. \tag{23}$$

Because some future offsets are missing in the upstream annotations, the realized number of projected waypoints can be smaller than 40.

Visibility and Goal Selection. As with AND, points must lie in front of the current camera and within a 64-pixel margin of the image. Starting from the end of the four-second horizon, we search backward for the latest future waypoint that is both projectable and associated with an available future RGB observation. This point defines the image-space goal.

The trajectory is then truncated to terminate at the selected goal, ensuring that

$$\mathbf{p}_K = \mathbf{g}_t. \tag{24}$$

We render only this goal point onto the current image. Diagnostic images containing the complete ground-truth trajectory are never used as model inputs.

Future Observation. For Fut-Loc, the world-modeling target is the future RGB frame corresponding directly to the selected goal waypoint. Thus, the future observation and trajectory endpoint are temporally aligned by construction.

Sampling and Splits. Unlike AND, Fut-Loc annotations are densely sampled, with approximately one example per video frame. We preserve the official Ego4D train and validation partitions rather than re-splitting the data, since adjacent examples have heavily overlapping four-second future horizons.

Starting from 51,818 annotated examples, 12,496 are removed because no projectable endpoint with an available future image can be found. The final benchmark contains 39,322 examples.

Table 9: **Two-dimensional trajectory statistics.** Pixel path lengths should not be compared directly across datasets because AND and Fut-Loc use different camera models and resolutions.

Dataset	Split	Min	Median	Mean	P90	Max
AND	train	3	12	11.7	13	17
AND	val	3	12	11.5	13	16
Fut-Loc	train	1	40	36.5	40	40
Fut-Loc	val	3	40	37.0	40	40

A.3 Benchmark Statistics

Table 9 summarizes the trajectory statistics of the two constructed benchmarks. AND trajectories contain fewer waypoints, with 0.25 s sampling versus 0.1 s for Fut-Loc.

In metric space, the median observed future path is approximately 5.1 m for AND and 4.9 m for Fut-Loc training. We observe a small number of extreme reconstruction outliers in the upstream Fut-Loc training annotations; these are discussed below.

A.4 Transition-Reasoning Supervision

In addition to trajectories and future observations, we associate training examples with structured transition rationales used by the reasoning branch of ProMac. These descriptions summarize navigation-relevant information such as visible traversable space, obstacles, ego-motion direction, and how the scene changes between the current and future observations.

Reasoning annotations are generated from the *bare* current and future frames rather than from trajectory-overlay visualizations. Consequently, the reasoning model cannot simply recover the target path from a rendered ground-truth trajectory. The goal location is supplied separately, while scene-object cues are grounded using navigation-relevant detections.

For each example, we generate multiple candidate transition descriptions and retain structured outputs satisfying the required schema. The resulting reasoning annotations are used as privileged training supervision and are not available to ProMac during inference.

Our full automatically annotated corpora contain reasoning supervision for all accepted AND and Fut-Loc examples. We additionally analyze human review of 2,000 source samples (1,000 per benchmark), yielding 5,666 non-skipped final responses after deduplication, as detailed below.

A.5 Human CoT Review and Consistency Metrics

Protocol and inclusion. We analyze the Aria and FutLoc portions of the human-annotation export `think-steps-annotations-v4` (revision `43dd748`). Each source sample received three submissions. Annotators could select any generated candidate, revise its contents, write a custom response, or skip. The review interface supplied goal-overlay, trajectory-overlay, and next-frame visualizations; these are privileged annotation context, distinct from the bare frames used for automatic rationale generation and from the goal-only image used as model input. Final responses contain `cause`, `critical_factors`, and a structured `decision`. We identify source samples by dataset and stable sample ID, and keep the latest submission per source sample and annotator (output ID breaks timestamp ties). This removes one repeated FutLoc submission. Custom responses are included; skipped responses are excluded from content and direction metrics. Table 10 reports the resulting coverage. The export's audit statuses are pending; these are individual human revisions, not an adjudicated consensus.

Final-revision similarity (panel A). Candidate selection is a process variable: two annotators can start from different candidates and revise them toward similar content. We instead embed each final `cause` and `critical_factors` field separately using the frozen Sentence Transformers `all-mpnet-base-v2` encoder, following sentence-embedding cosine comparison (Reimers

Table 10: **Coverage of the human CoT analysis.** Pairwise text similarity and three-annotator clock consistency have different eligibility criteria.

Quantity	Aria	FutLoc
Source samples	1,000	1,000
Submissions after deduplication	3,000	2,999
Non-skipped final responses	2,944	2,722
Samples with at least two final responses	995	914
Unordered final-response pairs (panel A)	2,893	2,580
Complete three-annotator clock samples (panel B)	949	833
Soft-consistent clock triples	778	816

& Gurevych, 2019).[1] Attention-mask-aware mean pooling and ℓ_2 normalization produce a 768-dimensional unit embedding $e(x)$; no input field is truncated at the 384-token limit. Structured field values, if present, are serialized as JSON with sorted keys. For responses a, b on sample i, define

$$s_{iab} = \frac{1}{2} \sum_{f \in \{\text{cause,critical factors}\}} e(x_{iaf})^\top e(x_{ibf}), \qquad S = \frac{\sum_i \sum_{a<b} s_{iab}}{\sum_i \binom{m_i}{2}}, \tag{25}$$

where $m_i \geq 2$ is the number of non-skipped distinct annotators. Every eligible unordered pair receives equal weight, irrespective of selected candidate; a sample with three responses contributes three pairs and one with two contributes one. The structured decision is excluded from this prose score and its clock direction is evaluated separately. The resulting S is 0.800 (95% CI: 0.792–0.808) on Aria and 0.842 (0.836–0.848) on FutLoc. These continuous cosine scores are neither percentages nor calibrated semantic-equivalence judgments.

For a descriptive reference, we take the first retained final response per sample in deterministic export order, sort samples by the SHA-256 hash of their stable ID, and cyclically pair each with the next sample within the same dataset. Applying the same score gives 0.447 on Aria (1,000 pairs) and 0.582 on FutLoc (975 pairs; 25 samples have no final response). This reference captures some shared vocabulary and template similarity; it is not a chance-corrected reliability coefficient. Shared candidate text can also raise within-sample similarity. Pairs choosing different candidates have mean similarities of 0.696 (1,555 pairs) and 0.770 (1,513 pairs), respectively.

Final direction and soft consistency (panel B). Direction is read from the human's final `decision.clock_position`, including any revision, rather than from the selected candidate. We parse the annotated clock labels into hours $h \in \{0, \ldots, 11\}$, with 12 o'clock represented by 0. These are egocentric direction labels (12 forward, 3 right, 6 backward, 9 left), not angles recomputed from trajectory coordinates. Define circular distance in hours as

$$d(h, k) = \min\{|h - k|, 12 - |h - k|\}. \tag{26}$$

For a complete triple $H_i = (h_{i1}, h_{i2}, h_{i3})$, the common-center soft consistency indicator and dataset rate are

$$g(H_i) = \mathbf{1}\left[\exists c \in \{0, \ldots, 11\} : \max_{j \in \{1,2,3\}} d(h_{ij}, c) \leq 1\right], \qquad G = \frac{1}{N_3} \sum_{i=1}^{N_3} g(H_i). \tag{27}$$

Thus $\{11, 12, 1\}$ and $\{11, 11, 1\}$ both agree around 12, even though the endpoint pair is two hours apart and the center need not be selected. The triple $\{10, 12, 1\}$ fails. This is a common-window criterion, not pairwise distance at most one or a majority vote. The separate `stationary` category agrees only when all three responses are stationary; mixed stationary/hour triples fail. Only samples with three distinct, non-skipped, valid final clock labels enter G. Aria has $778/949 = 82.0\%$ soft consistency (95% CI: 79.6–84.4%); FutLoc has $816/833 = 98.0\%$ (97.0–98.9%). Relative to all 1,000 source samples per dataset, coverage by soft-consistent complete triples is 77.8% and 81.6%, respectively. Missing ratings are excluded, not counted as disagreements.

[1] Checkpoint: `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`, revision `e8c3b32`.

Directional reference and uncertainty. Let $p(h)$ be the empirical clock distribution among complete triples, including stationary as a thirteenth category. We enumerate all label triples to compute the independent-draw reference

$$G_{\text{ref}} = \sum_{a,b,c} p(a)p(b)p(c)\, g(a,b,c). \tag{28}$$

It is 66.3% for Aria and 94.5% for FutLoc. In FutLoc, 95.2% of labels in complete triples fall in $\{11, 12, 1\}$, explaining much of its high raw soft consistency. Figure 5 shows this reference alongside G, without treating G as chance-corrected agreement. All plotted 95% intervals are percentile intervals from 4,000 bootstrap resamples of eligible source samples (seed 20260920). Each resampled sample retains its full set of annotator pairs; we recompute the pair-weighted S or the triple mean G. Annotators are not resampled, and temporal dependence between source samples is not captured by these intervals.

Interpretation and training scope. The two panels quantify shared prose content and coarse action direction. Embedding similarity can miss negation, hazard contradictions, or grounding errors, while a tolerant clock window can hide fine directional differences. Accordingly, neither panel establishes correctness or causal faithfulness. This annotation analysis is separate from the model readout evaluation in Section 4.6 and does not report a human-CoT training ablation. For training mixtures, source IDs are matched to the original benchmark splits: 907 Aria and 894 FutLoc training samples have usable human revisions. Other training samples retain generated rationales; validation annotations are excluded from the training mixture.

A.6 Input-Leakage Prevention

Because trajectory visualizations are useful during dataset inspection, our generation pipeline produces two types of rendered image:

1. a *goal overlay*, containing only the red destination marker; and
2. a *trajectory overlay*, containing the complete ground-truth path for visualization and quality assurance.

Only the goal overlay is permitted as model input. We explicitly verify that all training and evaluation message payloads reference the goal-overlay image.

An early Fut-Loc intermediate artifact inherited a trajectory-overlay image in its message field. Before any model training, we regenerated the input mapping so that the corresponding fields reference the goal-overlay image while leaving the trajectory, geometry, and reasoning annotations unchanged. We validated this correction across all 39,322 examples. Importantly, transition-rationale generation itself was unaffected because it used bare current and future frames rather than rendered trajectory overlays.

A.7 Dataset Limitations

The constructed benchmarks inherit several limitations from their upstream datasets and from the projection procedure.

Validation-Based Model Selection. The archived AND split has disjoint recording sessions, but checkpoints are selected using validation performance. Fut-Loc also uses validation-based checkpoint selection. Neither evaluation is an untouched test, and session separation alone does not establish generalization to new environments.

Temporal Redundancy in Fut-Loc. Fut-Loc annotations are densely sampled at approximately 10 Hz, and consecutive examples share most of their four-second future horizon. We therefore preserve the official Ego4D partition and do not perform random record-level re-splitting.

Fut-Loc Reconstruction Outliers. A small number of Fut-Loc training trajectories contain physically implausible metric-scale displacements inherited from the upstream reconstruction. Since

distant 3D points can nevertheless project to valid image coordinates, these outliers are less apparent in pixel-space statistics. Metric-space analyses should therefore apply additional sanity filtering.

Degenerate Trajectories. A small number of Fut-Loc examples contain very short trajectories because of missing future frames or endpoint backoff. Experiments that require a fixed number of waypoints resample or filter these trajectories consistently across all compared methods.

Different Camera Coordinate Systems. AND trajectories are defined in a rectified, clockwise-rotated 512×512 coordinate system, whereas Fut-Loc trajectories are defined in the original distorted 1280×720 fisheye frame. Pixel-space errors are therefore reported separately for the two datasets and are not pooled without normalization.

Open-Loop Evaluation. Finally, the projected trajectory represents realized future human motion observed in the original recording. Our benchmark evaluates open-loop image-space trajectory generation from a single current observation and does not itself validate whether predicted trajectories remain safe under closed-loop execution.

B Implementation Details of Direct Visual-State Reconstruction

Future-state targets. We encode the bare future frame X_{t+1}^{bare} using the backbone's frozen vision encoder, capturing features immediately before its main spatial merger:

$$\mathbf{U}_{t+1} = E_{\text{ViT}}^{\text{pre}}(X_{t+1}^{\text{bare}}) \in \mathbb{R}^{4N_t \times d_v}. \tag{29}$$

Qwen's spatial merger combines each 2×2 group of vision patches into one backbone visual token. Thus, N_t backbone visual tokens correspond to $4N_t$ pre-merger target features. We require each current–future pair to have the same processed spatial grid. This establishes a grid-indexed target, not a correspondence between physical objects across frames.

We standardize target channels using fixed statistics estimated from training frames:

$$\bar{U}_{t+1,n,c} = \frac{U_{t+1,n,c} - \mu_c}{\max(\sigma_c, \epsilon)}, \qquad \epsilon = 10^{-6}. \tag{30}$$

The vision encoder and statistics remain fixed during calibration and joint training. Future frames are encoded without running the language backbone, and target features are detached from the optimization graph.

Reconstruction-head parameterization. The head applies a shared affine projection to each backbone visual hidden state $\mathbf{h} \in \mathbb{R}^{d_h}$:

$$g_\phi^{\text{token}}(\mathbf{h}) = \text{reshape}_{4 \times d_v}(W_g \mathbf{h} + \mathbf{b}_g), \qquad W_g \in \mathbb{R}^{4d_v \times d_h}. \tag{31}$$

For Qwen3-VL-4B, $d_h = 2560$ and $d_v = 1024$, yielding

$$\begin{aligned} [B, N_t, 2560] &\xrightarrow{\text{linear}} [B, N_t, 4096] \\ &\xrightarrow{\text{unpack}} [B, N_t, 4, 1024] \\ &\xrightarrow{\text{pack}} [B, 4N_t, 1024]. \end{aligned} \tag{32}$$

The four output slots represent the top-left, top-right, bottom-left, and bottom-right subpatches within each merger block. Predictions and targets follow the same merge-block-major order. The reshaping operates on the learned projection's outputs, not directly on the original backbone channels.

The projection is shared across positions, so its parameter dimensions do not depend on image resolution or token count. The head introduces no attention layers or explicit action conditioning; it relies on the contextual information already encoded in the backbone visual states.

Table 11: **Natural-language versus structured CoT: trajectory accuracy, reasoning length, and navigation-content coverage.** N: natural-language; S: structured. Bold marks the better format within each dataset (ties are unbolded). Panel (a) reports the matched CoT-format experiment, a separate run set from Table 2. In (b), tokens are mean rationale lengths; coverage (%) uses deterministic, reference-conditioned lexical matching, not human-judged semantic correctness. Object/hazard coverage includes only applicable references; complete coverage requires all applicable components.

(a) Trajectory error (↓)

	AND		Fut-Loc	
Metric	**N**	**S**	**N**	**S**
ADE	19.50	**15.38**	142.97	**141.78**
FDE	8.92	**6.99**	**174.03**	174.46
L2@First4	28.08	**22.80**	113.26	**112.18**
DTW	17.77	**14.07**	115.07	**112.74**

(b) Reasoning length and coverage

	AND		Fut-Loc	
Measure	**N**	**S**	**N**	**S**
Tokens ↓	**108.0**	116.6	**73.7**	143.6
Meta-action ↑	100.0	100.0	100.0	100.0
Object/hazard ↑	98.5	**100.0**	87.9	**94.7**
Path grounding ↑	100.0	100.0	100.0	100.0
Complete ↑	98.6	**100.0**	89.9	**95.5**

Loss normalization and padding. For one image, reconstruction loss averages squared error over all valid target patches and feature channels:

$$\mathcal{L}_{\text{next}} = \frac{1}{4N_t d_v} \sum_{n=1}^{4N_t} \left\| \hat{\mathbf{U}}_{t+1}^{(n)} - \text{sg}(\bar{\mathbf{U}}_{t+1}^{(n)}) \right\|_2^2. \tag{33}$$

For variable-length batches, padding is excluded and this normalized loss is computed separately for each image before averaging across images. Consequently, images receive equal weight regardless of their number of visual patches.

B.1 Natural-Language CoT vs. Structured CoT

We compare free-form natural-language rationales with structured CoT that organizes navigation information into explicit semantic components, including motion actions and relevant objects or hazards. The matched format experiment uses the same Qwen3-VL-4B backbone, training data, trajectory targets, and inference configuration. Table 11(a) shows that structured CoT reduces all four trajectory errors by 18.8-21.7% on AND. Gains on Fut-Loc are smaller: ADE, L2@First4, and DTW fall by 0.83%, 0.96%, and 2.03%, while FDE increases by 0.25%. Table 11(b) shows higher object/hazard and complete-content coverage, but longer rationales: 116.6 versus 108.0 tokens on AND and 143.6 versus 73.7 on Fut-Loc. Thus, structured CoT improves lexical coverage of navigation-relevant content rather than brevity, with dataset-dependent trajectory benefits.

C Zero-Shot Qwen Baseline Protocol

We define two inference-only prompting conditions for each of the original Qwen3-VL-4B-, 8B-, and 32B-Instruct checkpoints without task-specific fine-tuning: answer-only and explicit CoT followed by an answer. The CoT condition uses prompted reasoning with the same Instruct checkpoint, not a separate Thinking model; prompts and decoding settings are matched across model sizes, and weights are loaded in BF16 without quantization. Both receive only the current goal-overlaid image and identical dataset-level instructions describing egocentric walking, the red destination marker, chronological waypoint order, and original-image pixel coordinates (top-left origin; x rightward and y downward). The visible coordinate ranges are $[0, 511]^2$ for AND and $[0, 1279] \times [0, 719]$ for Fut-Loc; off-image projected ground points are explicitly permitted and are not clipped. The prompt specifies a horizon of up to four seconds, with approximately 0.25 s sampling and at most 17 waypoints for AND, or 0.10 s sampling and at most 40 waypoints for Fut-Loc, without revealing any example's reference length, trajectory, rationale, or numerical goal coordinates.

Answer-only generation requests a JSON array of numeric coordinate pairs inside `<answer>` tags. The CoT variant additionally requests brief scene-grounded reasoning about the goal, walkable space, obstacles, and route inside `<think>` tags before the same answer format. Neither prompt

contains demonstrations. The protocol uses greedy decoding with a shared budget of 1,536 new tokens on the same full validation splits as the fine-tuned baselines. Complete, correctly formatted responses are scored using the existing shared-prefix ADE/FDE/L2@First4 and full-path DTW conventions; format failures, token-limit truncations, and valid/total counts are reported separately, since errors averaged only over valid outputs must not be interpreted as unconditional performance. The implementation also reports a paired comparison on IDs valid under both prompting modes. These evaluations are pending in Table 2.

D Qualitative Trajectory Comparison on Ego4D Fut-Loc

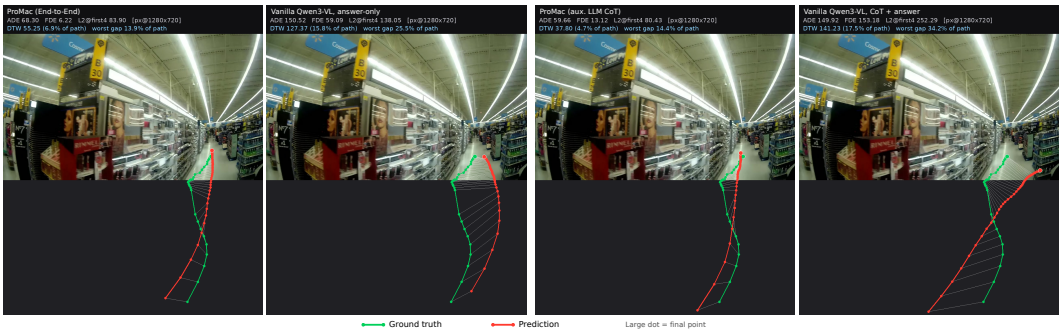


Figure 8: **Qualitative trajectory comparison on Ego4D Fut-Loc.** From left to right: ProMac (end-to-end), vanilla Qwen3-VL (answer-only), ProMac (aux. LLM CoT), and vanilla Qwen3-VL (CoT + answer), all for the same grocery-aisle observation. Green denotes the ground-truth trajectory and red the generated trajectory; large dots mark the final waypoints. The canvas is extended below the image to retain all off-image waypoints at a shared scale. Predictions contain 39–40 waypoints versus 33 reference waypoints: ADE, FDE, and L2@First4 use the shared-prefix evaluation, whereas DTW uses the complete paths. This is a selected strong-win example from a partial evaluation snapshot, not an aggregate performance comparison.

Figure 8 illustrates a case where both ProMac variants more closely follow the reference path than the vanilla baselines. End-to-end ProMac reduces ADE to 68.30 pixels and DTW to 55.25, compared with 150.52/127.37 for answer-only Qwen3-VL and 149.92/141.23 for its CoT + answer counterpart. The auxiliary-LLM CoT variant achieves still lower ADE/DTW (59.66/37.80), while the end-to-end variant has lower shared-prefix FDE (6.22 vs. 13.12 pixels). Thus, this example highlights improved path alignment over vanilla generation, but does not establish a uniform advantage of joint training over CoT compression alone.

E Validation Trajectory Geometry and Auxiliary-Objective Ablations

E.1 Hypothesis and Scope

Section 4.4 shows that combining the auxiliary objectives improves Fut-Loc ADE, FDE, and DTW over either objective alone, but slightly worsens L2@First4. On AND, the full model improves L2@First4 and DTW while worsening ADE and FDE. We investigate whether differences in trajectory geometry could help explain this pattern. Nearly straight paths may require less complementary supervision, whereas paths with bends may benefit more from combining semantic route constraints with predictive visual features. The analysis below measures ground-truth geometry; it does not directly measure how model gains depend on curvature.

E.2 Geometry Measures and Sampling

We analyze all 1,051 AND and 3,016 Fut-Loc validation target trajectories. Each 2D path is divided by its image diagonal, preserving angles and length ratios, and resampled to $N = 10$ points at equal

arc-length intervals. This gives both datasets the same number of evaluation points despite their different annotation densities. For resampled points $\mathbf{p}_1, \ldots, \mathbf{p}_N$, we compute

$$R = \frac{\sum_{i=1}^{N-1} \|\mathbf{p}_{i+1} - \mathbf{p}_i\|_2}{\|\mathbf{p}_N - \mathbf{p}_1\|_2}. \tag{34}$$

A straight path has $R = 1$; larger values indicate a longer route relative to endpoint displacement. We exclude displacements below 1% of the image diagonal to avoid unstable ratios; no 2D validation targets are excluded. We use descriptive bins of near-straight ($R \leq 1.05$), moderate deviation ($1.05 < R \leq 1.20$), and large deviation ($R > 1.20$). These thresholds are analysis choices, not dataset-provided labels.

We additionally sum the unsigned angles between consecutive nonzero path segments. This accumulated direction change includes small zigzags and is neither a net heading change nor a measurement of camera or head rotation.

Table 12: **Validation trajectory geometry.** Both datasets use 10-point arc-length resampling. Larger R and accumulated turn indicate greater path deviation. Ground-plane results use a matched temporal window; percentages describe the retained examples.

Measure	**AND / Aria**	**Fut-Loc**
2D target trajectories		
Evaluated paths	1,051	3,016
Median R	1.007	1.048
90th-percentile R	1.022	1.186
Near-straight (%)	96.8	51.6
Moderate deviation (%)	3.0	39.6
Large deviation (%)	0.3	8.9
Median accumulated turn (°)	50.3	122.2
90th-percentile accumulated turn (°)	72.8	241.4
Ground-plane paths, 0.5–4.0 s		
Evaluated paths	1,044	2,559
Median R	1.003	1.019
90th-percentile R	1.013	1.063
Near-straight (%)	97.9	85.0
Median accumulated turn (°)	44.2	74.8

E.3 Results and Robustness

Table 12 shows that AND's image-space targets are substantially straighter: 96.8% satisfy $R \leq 1.05$, compared with 51.6% on Fut-Loc. Fut-Loc also has higher median and upper-tail ratios and accumulated turning. The dataset ordering persists with 5 and 20 resampled points: AND's near-straight fractions are 97.4% and 96.1%, versus Fut-Loc's 72.8% and 37.8%. The exact fractions are therefore resolution-sensitive, particularly for Fut-Loc; short-scale zigzags and annotation noise can contribute alongside broader bends.

To check whether this difference is solely an image-projection effect, we compare physical ground-plane paths over a common 0.5–4.0 s interval. We interpolate positions at 0.25 s intervals before arc-length resampling. AND uses world-plane XY coordinates; Fut-Loc camera centers are projected onto the plane orthogonal to the supplied up vector. Seven AND and 457 Fut-Loc examples lack the full interval and are excluded; no additional paths fail the 0.25 m minimum endpoint-displacement criterion. The ground-plane difference is smaller but remains in the same direction: median R is 1.003 versus 1.019, with 97.9% versus 85.0% near-straight paths.

E.4 Interpretation and Limitations

These measurements support the premise that Fut-Loc contains more geometrically complex validation paths, consistent with greater potential benefit from complementary reasoning and future-state supervision. They do not establish that complexity causes the ablation differences. A direct test

would compare each example's full-model gain over each ablation within shared geometry bins, controlling for horizon, path length, and recording, and assess variation across training seeds.

The reported statistics weight examples equally. AND validation contains three sessions and Fut-Loc contains 18 segments, with correlated neighboring examples. Equal weighting of recording groups gives 97.2% versus 60.0% near-straight 2D paths, preserving the ordering. Nevertheless, camera geometry, target visibility, truncation, and reconstruction noise remain possible contributors. This is a descriptive analysis of the available validation sets, not a significance test or evidence of greater head rotation in Fut-Loc.

F Prior-Frame Context: Protocol and Interpretation

We evaluate frozen joint checkpoints trained on single frames (AND: update 882; Fut-Loc: update 1250), prepending $K \in \{1, 3\}$ unannotated frames at 1 Hz, oldest first. The current goal image, prompt, future target, per-frame resolution, and greedy decoder remain fixed. All K share complete-history cohorts of 834 AND and 517 Fut-Loc examples for normalized future-feature MSE. ADE, measured in pixels, uses fixed, source-group-balanced subsets of 64 examples per dataset. These cohorts cover only three AND sessions and four Fut-Loc videos.

Table 4 favors the current frame alone for ADE on AND. One prior frame improves Fut-Loc MSE by 2.31% but worsens ADE by 25.92%; all history lengths worsen both AND measures. With $K = 5$, the model processes six times as many visual tokens, while ADE increases by 43.45% on AND and 73.22% on Fut-Loc. These results support single-frame deployment for these checkpoints. They measure inference-time history conditioning on models trained with single frames.

G Illustrative Human-Reviewed CoT Samples

Figures 9 and 10 illustrate two Aria and two FutLoc samples from the human-annotation analysis in Appendix A.5. Each pairs the original trajectory-overlay image with one complete final human response. Green denotes the projected observed trajectory, and red marks its goal. These overlays provide privileged annotation context; the trajectory policy receives only the goal-overlay image at inference. The figures retain the original annotation values, including imperfect wording and inconsistencies, while adjusting line wrapping and field layout. The revision notes describe changes relative to the selected generated candidate.

Selection and structure. We purposively select four samples with visible overlays and different navigation contexts, rather than estimating annotation quality from these examples. For each sample, we use the existing representative intact response: minimum total circular clock distance to the other final responses, followed by maximum mean MPNet cause/critical-factor similarity, with output ID breaking ties. This selects an individual response, not an adjudicated consensus. A1, A2, and F2 come from the validation split; F1 comes from the training split. These are qualitative illustrations, not an independent evaluation set. The accompanying `human_cot_examples.json` records full source IDs, output IDs, original candidates, final responses, field-level changes, and image checksums.

The `cause` field explains the navigation choice, while `critical_factors` describes scene evidence and constraints. Aria's `decision` additionally separates movement, direction, clock position, hazards, path constraints, and landmarks. FutLoc uses longitudinal action, lateral action, and clock position. The examples show why prose similarity, structured-field consistency, and visual grounding address different aspects of annotation quality.

A1: Campus path and a partial direction revision. In `Mainquad_jan10_9304`, the projected path extends along the road edge toward the marked goal, and the annotation identifies a parked car on the right. The annotator changes the clock position from 1 to 12 o'clock, but leaves the separate direction field as "turn right." The cause and critical factors describe forward progress. This illustrates how a local revision can leave tension between fields: agreement on a clock label does not establish consistency of the entire structured response.

A2: Storefront walkway and landmark enrichment. In `SSC4_jan17_5243`, pedestrians occupy the walkway ahead and a glass facade borders the right side. The response combines forward movement with a 1-o'clock direction, identifies a person as a hazard, and describes following the person ahead as a path constraint. The revision adds "Madewell sign" and "red shaded area" to landmark guidance while retaining the cause and critical-factor prose. This is an example of enriching the structured scene references without rewriting the navigation explanation.

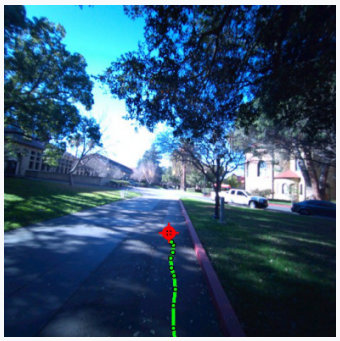
F1: Right turn beside snow and a stronger direction label. Training frame `image0010069` in FutLoc segment `e177fa1d-5620-40f8-9f8e-af0c737fbc36` shows a wet road with puddles and snow along its boundaries. The projected trajectory bends visibly to the right toward the marked goal. The final cause and critical factors describe this right turn, while the decision pairs forward walking with "turn right" and a 2-o'clock clock position. The annotator removes "slight" from the turn descriptions and changes the clock from 1 to 2 o'clock. This example illustrates a coordinated refinement of turn magnitude across the prose and clock label, with a visible directional change in the overlay. The stated concern about slippery conditions is the annotator's interpretation of the scene, not a measured safety outcome.

F2: Store aisle and more specific scene evidence. Frame `image0000314` in segment `442d7f6c-0ae2-4082-9977-f1fe1498259c_3` shows a projected path through the aisle between merchandise on the left and a floor display on the right. The revised critical factors explicitly identify the right-side display and its "Everyday Low Price $5" sign, alongside the nearby person. The cause and decision continue to specify forward walking at 12 o'clock. This illustrates a revision to supporting evidence that leaves the chosen action unchanged. Together, the four cases motivate retaining final text and structured decisions rather than using candidate-choice identity as a substitute for reasoning agreement.

Aria: human-reviewed CoT examples

Green: observed projected trajectory | Red: marked goal | Source split shown per sample

A1 Campus path



Mainquad_jan10_9304 [val]

Revision record
Clock revised from 1 to 12 o'clock; the direction field still says "turn right".

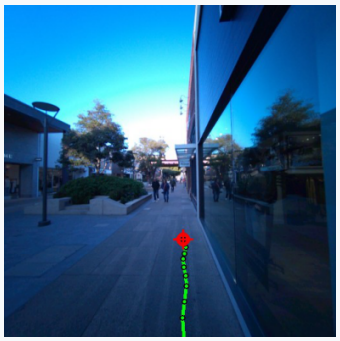
CAUSE
The path is unobstructed and leads toward the goal point, with a parked car on the right side of the road.

CRITICAL FACTORS
The path is clear and leads forward, with a parked car on the right side of the road.

DECISION

```
hazards: ["parked car"]
movement: "move forward"
direction: "turn right"
clock_position: "12 o'clock"
path_constraints: ["fence"]
landmark_guidance: ["building", "doorway"]
```

A2 Storefront walkway



SSC4_jan17_5243 [val]

Revision record
The revision adds "Madewell sign" and "red shaded area" as landmarks.

CAUSE
To maintain the walking path along the building and avoid obstacles, the user should follow the direction of the pedestrians and continue forward, keeping the building on the right.

CRITICAL FACTORS
The sidewalk is clear ahead, with pedestrians walking in the same direction, and a building with a glass facade is on the right side, indicating the path continues forward along the building.

DECISION

```
hazards: ["person"]
movement: "move forward"
direction: "right"
clock_position: "1 o'clock"
path_constraints: ["follow the person walking ahead"]
landmark_guidance: ["Madewell sign", "red shaded area"]
```

Final annotation values are reproduced verbatim; line wrapping and field layout are adjusted.

Figure 9: **Two human-reviewed Aria CoT samples.** Each panel pairs the source trajectory overlay with the complete final cause, critical factors, and decision fields. Revision notes summarize the recorded human edits. A1 illustrates a partial direction revision; A2 illustrates added landmark guidance. These selected individual responses are not adjudicated ground truth.

FutLoc: human-reviewed CoT examples

Green: observed projected trajectory | Red: marked goal | Source split shown per sample

F1 Right turn beside snow



e177fa1d... / image0010069 [train]

Revision record

The revision removes "slight" from the turn description and changes the clock from 1 to 2 o'clock.

CAUSE

The wet road and snow-covered ground require careful navigation, and the right curve in the road necessitates a turning motion to maintain the correct path.

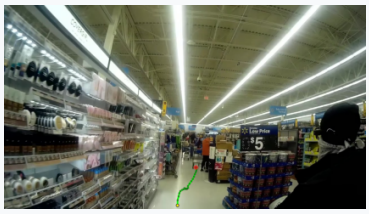
CRITICAL FACTORS

The road is wet and has puddles, indicating recent melting snow. Snow covers the ground on both sides of the road, and the path ahead is clear. A sign and a utility pole are visible on the left, while a parked car and a tree are on the right. The road curves to the right, and the sidewalk is passable. The overall environment suggests a need to navigate carefully due to slippery conditions.

DECISION

```
lateral: "turn right"
longitudinal: "walk forward"
clock_position: "2 o'clock"
```

F2 Store aisle



442d7f6c... / image0000314 [val]

Revision record

Critical factors revised to describe the floor display and "Everyday Low Price $5" sign.

CAUSE

The path is clear and the aisle is wide enough to walk forward without obstruction, with the trajectory leading towards the lower center of the frame.

CRITICAL FACTORS

The aisle is wide enough to accommodate walking with sufficient space to navigate around the person in the black jacket and the floor display with a sign indicating "Everyday Low Price $5" on the right side of the aisle. There is also a display of products on the left side. The path is clear of large obstacles, and the trajectory moves forward along the aisle towards the lower center of the frame.

DECISION

```
lateral: "straight ahead towards the lower center of
the aisle"
longitudinal: "continue walking"
clock_position: "12 o'clock"
```

Final annotation values are reproduced verbatim; line wrapping and field layout are adjusted.

Figure 10: **Two human-reviewed FutLoc CoT samples.** F1 illustrates a refined right-turn label on a wet, snow-bordered road; F2 illustrates revised scene evidence in a store aisle. The original overlays and all final annotation values are retained. Green trajectories are observed projections, rather than model predictions or verified safe routes. Full sample identities and interpretation are given in Appendix G.