arXiv:2512.16907v2 [cs.CV] 30 Dec 2025

Flowing from Reasoning to Motion: Learning 3D Hand Trajectory Prediction from Egocentric Human Interaction Videos

Mingfei Chen[1,2] Yifan Wang[1] Zhengqin Li[1] Homanga Bharadhwaj[1] Yujin Chen[1] Chuan Qin[1]
Ziyi Kou[1] Yuan Tian[1] Eric Whitmire[1] Rajinder Sodhi[1] Hrvoje Benko[1] Eli Shlizerman[2] Yue Liu[1]

[1]Meta [2]University of Washington

https://egoman-project.github.io/

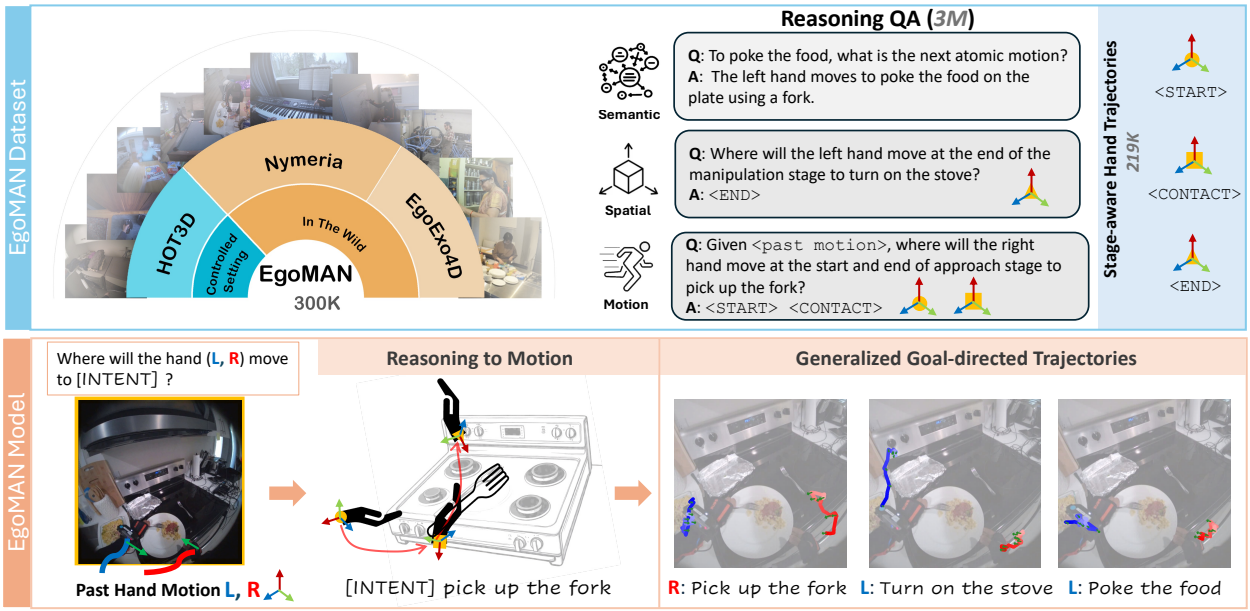

Figure 1. **EgoMAN project.** We introduce 1) the **EgoMAN dataset** (top), a large-scale egocentric dataset for interaction stage–aware 3D hand trajectory prediction with 219K 6-DoF trajectories and 3M structured QA pairs for semantic, spatial, and motion reasoning. During inference, 2) the **EgoMAN model** (bottom) takes an image, past hand motion, and an intent query as input, performs stage-aware reasoning to infer intent-specific waypoints, and then generates 6-DoF hand trajectories of distinct motions for different intent queries.

Abstract

*Prior works on 3D hand trajectory prediction are constrained by datasets that decouple motion from semantic supervision and by models that weakly link reasoning and action. To address these, we first present the **EgoMAN dataset**, a large-scale egocentric dataset for interaction stage–aware 3D hand trajectory prediction with 219K 6DoF trajectories and 3M structured QA pairs for semantic, spatial, and motion reasoning. We then introduce the **EgoMAN model**, a reasoning-to-motion framework that links vision–language reasoning and motion generation via a trajectory-token interface. Trained progressively to align reasoning with motion dynamics, our approach yields accurate and stage-aware trajectories with generalization across real-world scenes.*

1. Introduction

Predicting future 3D hand motion is essential for *in-context interaction* and proactive assistance, where a system anticipates human intent from visual, linguistic, and motion cues. Humans perform this naturally, *i.e.*, understanding the goal of an action, interpreting the scene layout, and coordinating movement based on recent dynamics. Achieving this computationally requires jointly reasoning about task semantics, spatial geometry, and temporal motion. We develop a model that predicts long-horizon 3D hand trajectories by integrating visual and motion context with language, which

conveys intent and disambiguates visually similar actions. Such capabilities enable applications in robot manipulation, language-conditioned motion synthesis, and assistive systems that respond to human intent.

A major bottleneck is the lack of large-scale, high-quality 3D trajectory data. Controlled datasets [4, 26, 30] offer accurate annotations but limited diversity, while large-scale egocentric video datasets [17, 42] contain rich real-world interactions but noisy, weakly goal-directed trajectories and little temporal structure. Crucially, they lack explicit **interaction stages**, *e.g.*, *approach* and *manipulation*, which are needed to separate purposeful motion from background and to connect trajectories to intent. Models trained on such raw videos often generalize poorly because the links between intent, spatial relations, and motion dynamics are missing.

Beyond data limitations, existing modeling approaches also fall short. Affordance-based methods [2, 11] rely on object detectors and affordance estimators, which propagate upstream detection errors and introduce additional computational overhead. End-to-end motion predictors, including those based on diffusion [20, 40], variational [5], and state-space models [6], focus on short-term dynamics with limited semantic grounding. Vision-Language-Action (VLA) systems [24, 38, 60] exhibit strong reasoning ability, but applying VLMs [3, 12, 13, 15, 34] directly to generate continuous 3D motion remains challenging, as they struggle to produce smooth, high-frequency action sequences. Bridges between VLM reasoning and motion experts [10, 28, 31, 33, 48, 49, 54, 58] typically rely on implicit tokens or lengthy reasoning chains, which limits efficiency, generalization, and interpretability when generating fine-grained, fast actions.

To address these challenges, we introduce the **EgoMAN** project, which couples a large-scale, stage-aware dataset with a modular reasoning-to-motion framework. The **EgoMAN dataset** contains over 300K egocentric clips from 1,500+ scenes, including 219K 6DoF hand trajectories annotated with interaction stages (*approach*, *manipulation*) and 3M structured vision–language–motion QA pairs. This supervision explicitly encodes *why*, *when*, and *how* hands move, enabling models to learn intent-linked, spatially grounded motion patterns at scale.

Building on this dataset, the **EgoMAN model** introduces a compact **trajectory-token interface** that connects high-level reasoning to continuous 3D hand motion. We define four trajectory tokens: one semantic token (`<ACT>`) and three stage-aware waypoint tokens (`<START>`, `<CONTACT>`, `<END>`) marking key transitions in interaction. These tokens represent wrist-centered spatio-temporal waypoints rather than object-centric affordances, providing a clear, structured interface for conditioning a flow-matching motion expert. A progressive three-stage training strategy learns (i) intent-conditioned and stage-aware reasoning over semantics, spatial and motion, (ii) motion dynamics, and (iii) their alignment through the token interface, enabling long-horizon, intent-consistent 3D trajectory prediction in diverse real-world scenes.

Our main contributions are:

- **EgoMAN dataset**: a large-scale, interaction stage–aware 6DoF hand trajectory dataset with structured semantic, spatial, and motion reasoning annotations.
- **EgoMAN model**: a modular reasoning-to-motion architecture with a trajectory-token interface and progressive training that aligns semantic intent with physically grounded motion generation.
- We achieve state-of-the-art accuracy and generalization with high efficiency in 3D hand trajectory prediction across diverse real-world egocentric scenes.

2. Related Works

Hand Trajectory Prediction. Egocentric hand forecasting aims to infer future hand motion from past observations under ego-motion and depth ambiguity. Large-scale works often predict short-horizon *2D* trajectories at low framerates [5, 19, 36, 39, 40], while curated datasets enable *3D* trajectory prediction [6, 20, 41]. Prior 3D methods generally follow either: (a) *object-centric, affordance-driven models* [2, 11, 36], which rely on detectors and affordance estimators but suffer from error propagation and additional computational efficiency cost from detection; or (b) *end-to-end motion models* predicting trajectories directly from video and past hand motion [39–41], sometimes incorporating egomotion [39–41] or 3D priors such as point clouds [41]. Given that 3D labels are often uncertain and scarce [6], generative models have become standard: VAEs [5], state-space models [6], diffusion [20, 40], and hybrid variants [39, 41]. However, these methods typically forecast short fixed horizons, focus on low-level motion, and encode intent implicitly, limiting generalization in diverse real-world egocentric scenarios. Our work instead predicts long-horizon, semantically grounded 6DoF trajectories by explicitly conditioning on intent, spatial context, and interaction stages.

Learning Interactions from Human Videos. Human videos provide rich demonstrations of hand–object interactions, driving research in reconstruction and forecasting [6, 7, 35, 36, 40]. Controlled datasets [4, 26, 30] offer precise 3D annotations but limited task diversity; robotic imitation datasets [21, 23, 53] provide structured demonstrations but remain narrow and scripted. Large-scale egocentric datasets [17, 42] capture varied daily activities with language annotations but often contain noisy trajectories and unclear interaction boundaries. We address these gaps by curating *EgoMAN-Bench*, consolidating real-world egocentric datasets into a stage-aware supervision benchmark. Our model builds on this benchmark to connect reasoning about

interaction stages with accurate, long-horizon 3D trajectory prediction, aligning with recent efforts in robot learning from human videos [2, 8, 9, 21, 23, 56].

Vision-Language Models for Embodied AI. Modern VLMs unify perception and language [3, 13, 16, 27, 34, 43, 51], enabling broad video understanding and reasoning. Their extensions to Vision-Language-Action (VLA) systems [24, 38, 60] support manipulation and navigation via robot datasets, but direct action prediction through VLMs often struggles to produce smooth, high-frequency trajectories. To mitigate this, recent works have sought to incorporate hand trajectory prediction in VLAs either through pre-training or co-training [29, 55]. Coupled VLM–motion systems, where the VLM is linked to an action module, use implicit feature routing [10, 49, 54], which suffers from poor generalization and limited interpretability, while other approaches rely on long reasoning chains as the interface [28, 31, 58], resulting in high inference cost and low efficiency. In contrast, we introduce a trajectory-token interface that directly links high-level reasoning to continuous 3D motion using four specialized semantic and spatiotemporal waypoint tokens, enabling an efficient, interpretable interface that effectively guides the motion expert to generate smooth, accurate high-frequency trajectories.

3. EgoMAN Dataset

EgoMAN is a large-scale egocentric interaction dataset (300+ hrs, 1,500+ scenes, 220K+ 6DoF trajectories) built from Aria glasses [14] across EgoExo4D [17], Nymeria [42], and HOT3D-Aria [4]. It provides high-quality wrist-centric trajectories, structured interaction annotations, and rich QA supervision for semantic, spatial, and motion reasoning. This section summarizes dataset statistics, annotation pipeline, trajectory annotation, QA construction, and the dataset split.

Dataset Statistics. The data spans diverse interactions—scripted manipulation in HOT3D-Aria, real-world activities (bike repair, cooking) in EgoExo4D, and everyday tasks in Nymeria. Trajectories cover substantial variation: 27.8% exceed 2 s, 34.0% move over 20 cm on average, and 35.3% rotate more than 60°. We train on EgoExo4D and Nymeria and reserve HOT3D-Aria as test-only set.

Annotation Pipeline. We use GPT-4.1 [45] to extract interaction annotations for EgoExo4D and Nymeria. At each atomic action timestamp [17, 42], we crop a 5 s clip and annotate two wrist-centric interaction stages: (1) *Approach*—the hand moves toward the target manipulation region; (2) *Manipulation*—the hand performs the action with the object in hand. Detailed prompts and filters are provided in the appendix. For HOT3D, we infer interaction stages using hand–object trajectories, defining approach as 0.5–2.0 s prior to object motion (object visible and within 1 m), and the manipulation stage corresponds to period after motion onset.

EgoMAN Trajectory. The EgoMAN dataset provides 6DoF wrist trajectories for both hand wrists (3D position + 6D rotation [59]), sampled at 10 frames per second (FPS). For *EgoExo4D*, we use hand tracking data produced by Aria's Machine Perception Services (MPS) [1]. For Nymeria dataset, we use trajectories obtained from two wrist-mounted devices. For HOT3D dataset, we directly use the high-quality 6DoF hand trajectories provided by the dataset. All trajectories are aligned by transforming positions and orientations into the camera coordinate frame of the final visual frame before interaction begins.

EgoMAN QA. We generate structured question–answer pairs using GPT, covering *semantic* (21.6%), *spatial* (42.6%), and *motion* (35.8%) reasoning .

(1) *Semantic reasoning* questions target high-level intent, such as:

- "*What will be the next atomic action?*"
- "*What object will the hand interact with next?*"
- "*Why does the next action happen?*"

These questions connect language to goal-directed hand behaviors, enabling deeper understanding of the motivations and purposes behind specific actions.

(2) *Spatial reasoning* questions ground intent within metric 3D space by querying the wrist's state at key interaction stages such as approach onset, manipulation onset (approach completion), and manipulation end. These questions may target a single stage (*e.g.*, "*Where/When will the left hand complete the manipulation?*") or span multiple stages (*e.g.*, "*Where/When is the right hand at the start and end of manipulation?*"), enabling reasoning about transitions between interaction stages. Some questions explicitly reference objects and stage timestamps, supporting reasoning over object-time-space relationships that align with interaction intent.

(3) *Motion reasoning* questions probe how past motion informs both semantic and spatial understanding, supporting reasoning about the evolution of motion over time. To construct these questions, we augment a random subset of semantic and spatial questions by prepending a 0.5-second 6DoF hand trajectory sequence from before the interaction start time. (*e.g.*, "*Given the <past motion>, where will the right hand complete the approach stage?*") This approach enables analysis of how previous hand movements influence subsequent actions and spatial positions, deepening the connection between motion history and interaction intent.

Dataset Split. To support our progressive training pipeline, we split the EgoMAN dataset into 1,014 scenes for pretraining (64%), 498 for finetuning (31%), and 78 for testing (5%). The pretraining set contains lower-quality trajectory annotations—where the target object may be occluded, image quality is low, or interaction intent is ambiguous, and interacitons are generally sparse. In total, the pretrain set comprises 74K samples, 1M QA pairs. The finetune set, by contrast, provides 17K high-quality trajectory samples.

For evaluation, we introduce **EgoMAN-Bench** as our test set, which consists of two settings: (1) **EgoMAN-Unseen** includes 2,844 trajectory samples from 78 held-out EgoExo4D and Nymeria scenes with high-quality trajectories, used to evaluate generalization to new in-domain but previously unseen scenes. (2) **HOT3D-OOD** includes 990 trajectory samples from HOT3D (dataset only used in testing), designed to evaluate out-of-distribution (OOD) performance on novel subjects, objects, and environments.

4. EgoMAN Model

As illustrated in Fig. 2, the **EgoMAN** model has two components: a **Reasoning Module** that extracts cues and reasons over semantics, spatial relations, and motion to produce stage-aware waypoints, and a **Motion Expert** that generates 6DoF hand trajectories. In this section, we first formalize the prediction task in Sec. 4.1, then detail the Reasoning Module in Sec. 4.2, the Motion Expert in Sec. 4.3, and Reasoning to Motion via Trajectory-Token Interface in Sec. 4.4.

4.1. Problem Formulation

Given a single RGB frame $\mathbf{V}_t$, past wrist trajectories $\{\mathbf{L}_\tau, \mathbf{R}_\tau\}_{\tau=t-H}^{t}$, and an intent description $\mathbf{I}$ as input, the task is to predict future 6DoF trajectories $\{\tilde{\mathbf{L}}_\tau, \tilde{\mathbf{R}}_\tau\}_{\tau=t+1}^{t+T}$ across manipulation stage, *e.g.*, reaching, manipulating, or releasing. Each position vector $\mathbf{L}_\tau \in \mathbb{R}^6$ and rotation vector $\mathbf{R}_\tau \in \mathbb{R}^{12}$ represent the 3D positions and 6D rotations of both wrists. Our **EgoMAN** model acts as the function $\mathcal{F}$ that maps the inputs to future trajectories:

$$\mathcal{F} : \left(\mathbf{V}_t, \{\mathbf{L}_\tau, \mathbf{R}_\tau\}_{\tau=t-H}^{t}, \mathbf{I}\right) \mapsto \{\tilde{\mathbf{L}}_\tau, \tilde{\mathbf{R}}_\tau\}_{\tau=t+1}^{t+T}$$

4.2. Reasoning Module

To predict accurate hand trajectory that aligns with human intent and environment context, we need to understand the spatial context of the environments as well as intent semantics. Therefore, the first module of EgoMAN model is reasoning model which aligns spatial perception and motion reasoning with task intent semantics and interaction stages. Built on Qwen2.5-VL [3], it takes as input an egocentric frame $\mathbf{V}_t$, a language query with intent description $\mathbf{I}$, and past wrist trajectories $\{\mathbf{L}_\tau, \mathbf{R}_\tau\}_{\tau=t-H}^{t}$. The past-motion sequence is encoded into the same latent space as the VLM's visual and language features, and then fused with them. Depending on the query, the module outputs either (i) a natural-language answer or (ii) a set of structured *trajectory tokens* that represent key interaction semantics and waypoints.

We introduce four trajectory tokens, one action semantic token and three waypoint tokens, to explicitly capture intent semantics and key spatiotemporal transitions across interaction stages. The action semantic token<ACT> decodes an action semantic embedding corresponding to the interaction phrase (e.g., "left hand grabs the green cup"). The three waypoint tokens: <START>, <CONTACT>, and <END> denote the approach onset, manipulation onset (*i.e.*, approach completion), and maunipulation completion stages respectively. Each waypoint token is equipped with a lightweight head that predicts a timestamp, 3D wrist positions, and 6D wrist rotations. These tokens allow the module to align semantic intent with the corresponding spatiotemporal hand states.

Reasoning Pre-training. To support this dual functionality to predict text and trajectory tokens, we first pretrain the module on 1M question–answer pairs from the EgoMAN pretraining split (Sec. 3). Semantic questions requiring natural language answers are supervised with the standard next-token prediction loss ($\mathcal{L}_{\text{text}}$). In contrast, queries requiring numeric outputs (*e.g.*, timestamps, 6DoF location), such as spatial reasoning queries, append a special token <HOI_Query> to the question end, instructing the model to decode trajectory tokens. For these queries, in addition to the language modeling loss that supervises the special token as text ($\mathcal{L}_{\text{text}}$), we supervise the <ACT> token with an action-semantic loss ($\mathcal{L}_{\text{act}}$) and the waypoint tokens with a dedicated waypoint loss ($\mathcal{L}_{\text{wp}}$).

Specifically, we calculate the action-semantic loss by projecting the hidden state of <ACT> to a semantic embedding and contrast against a CLIP-encoded [50] GT embedding. To stabilize training under varying batch sizes caused by the flexible mix of query types, with questions requiring an <ACT> answer varying in proportion across batches, we adaptively use cosine similarity or InfoNCE [44]:

$$\mathcal{L}_{\text{act}} = \begin{cases} 1 - \frac{1}{K}\sum_{i=1}^{K} \text{sim}(z_i, z_i^+), & K < \kappa, \\ -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1}^{K}\exp(\text{sim}(z_i, z_j^+)/\tau)}, & K \geq \kappa. \end{cases}$$

where z_i and z_i^+ denote normalized predicted and GT embeddings, $\text{sim}(\cdot)$ is cosine similarity, and τ is a learnable temperature parameter. When the number of valid training samples K falls below a threshold κ, we apply a cosine similarity loss to avoid unstable contrastive updates; otherwise, we use an InfoNCE-style contrastive loss.

For waypoint learning, each waypoint token is supervised with Huber losses weighted by Gaussian time windows:

$$\mathcal{L}_{\text{wp}} = \lambda_t \mathcal{L}_{\text{time}} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_r \mathcal{L}_{\text{rot6D}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}.$$

We use the continuous 6D rotation parameterization [59] with a geodesic rotation loss ($\mathcal{L}$geo), and compute the 2D loss ($\mathcal{L}2D$) by projecting predicted 3D positions into the input image frame. Only visible waypoints are supervised to avoid ambiguity.

The complete reasoning pre-training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \lambda_{\text{wp}}\mathcal{L}_{\text{wp}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}},$$

where the λ terms weight each loss component.

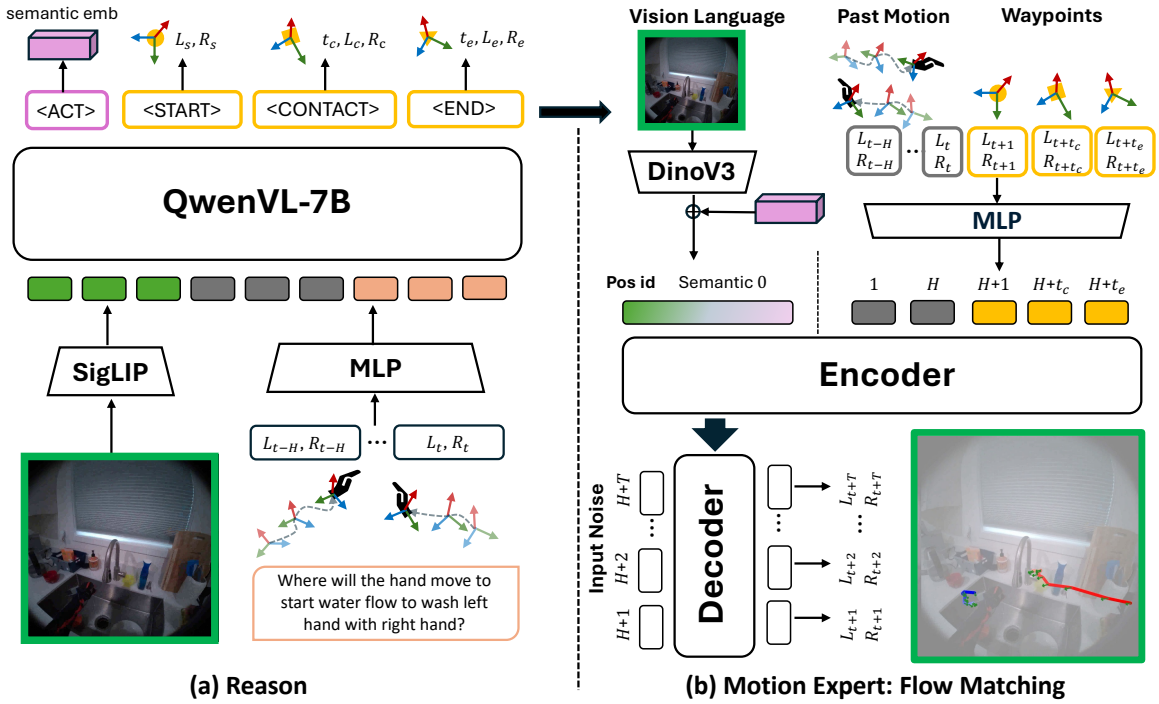

Figure 2. **Overview of the EgoMAN model.** The EgoMAN model is a modular reasoning-to-motion framework that predicts future 6DoF hand trajectories from an egocentric RGB frame, past wrist trajectories, and a language intent. The **Reasoning Module** (a), built on QwenVL-7B, extracts semantic and spatial features and outputs trajectory tokens with waypoints and intent semantic cues. The **Motion Expert** (b), using Flow Matching, predicts future trajectories based on waypoints, past motion, intent semantics and visual input. The trajectory tokens of (a) form the **Trajectory-Token Interface** which replaces semantic and waypoint condition inputs of (b) to bridge from Reasoning to Motion Expert.

4.3. Motion Expert

Given the trajectory tokens from the Reasoning Module, the Motion Expert predicts high-frequency 6DoF wrist trajectories by modeling fine-grained hand dynamics. The Motion Expert is an encoder–decoder transformer using Flow Matching (FM) [33] conditioned on past wrist motion, intent semantics, low-level visual features, and stage-aware waypoints. FM learns a conditional velocity field that yields smooth, probabilistic trajectories, with the three waypoint tokens providing structural guidance.

As shown in Fig. 2 (b), the encoder organizes all inputs into a unified sequence. Motion-related tokens lie on a unified temporal axis: H past wrist motion points occupy steps 1–H, waypoint tokens are placed at predicted timestamps offset by H, and future queries span $H{+}1$–$H{+}T$. These temporal tokens receive positional IDs based on their timestamps. In parallel, intent semantics and DINOv3 [52] visual features are added as non-temporal context tokens. The decoder then generates the T future 6DoF trajectory points by attending to this complete encoded context.

We follow the standard FM: a noisy sample x_0 is interpolated with the ground truth x_1, and the supervision target is $\hat{v} = x_1 - x_0$. The loss is a mean squared error over 3D positions and 6D rotations:

$$\mathcal{L}_{\text{FM}} = \|\hat{v} - (x_1 - x_0)\|_2^2 .$$

At test time, we sample an initial random trajectory x_0 and integrate the velocity field over N steps:

$$x_{k+1} = x_k + \Delta t \cdot \hat{v}(x_k, t_k), \quad \Delta t = \tfrac{1}{N},$$

to obtain future wrist trajectories.

Motion Pre-Training. We found joint training of the Reasoning Module and the Motion Expert is unstable due to mismatched learning objectives. To address this, we pretrained the FM model separately on the EgoMAN finetuning split (Sec. 3), using GT waypoints and action phrase semantics as conditions. This provides strong low-level motion prior that stabilizes joint training with Reasoning Module.

4.4. Reasoning to Motion

Once both the Reasoning Module and Motion Expert are pretrained, we jointly train them to connect high-level reasoning with low-level motion generation. A key challenge in this stage is the distribution mismatch. The Reasoning Module was pretrained to predict tokens based on ground-truth, while

Method	ADE (m) ↓			FDE (m) ↓			DTW (m) ↓			Rot (°) ↓			Dataset
	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	
USST* [6]	0.233	0.233	0.233	0.394	0.394	0.394	0.220	0.220	0.220	46.98	46.98	46.98	EgoMAN Unseen
MMTwin* [41]	0.213	0.208	0.206	0.261	0.257	0.256	0.211	0.206	0.204	49.53	49.12	48.98	
HandsOnVLM* [5]	0.176	0.172	0.171	0.232	0.228	0.228	0.166	0.162	0.161	35.49	35.29	35.22	
FM-base	0.188	0.166	0.160	0.265	0.236	0.229	0.171	0.150	0.144	37.92	37.27	37.00	
EgoMAN-ACT	0.162	0.146	0.141	0.225	0.210	0.204	0.148	0.132	0.127	36.24	35.28	35.03	
EgoMAN (Ours)	**0.151**	**0.130**	**0.124**	**0.206**	**0.186**	**0.179**	**0.137**	**0.117**	**0.111**	**33.88**	**33.00**	**32.75**	
USST* [6]	0.245	0.245	0.245	0.409	0.409	0.409	0.226	0.226	0.226	55.80	55.80	55.80	HOT3D OOD
MMTwin* [41]	0.214	0.210	0.209	0.263	0.260	0.259	0.212	0.208	0.207	44.75	44.46	44.37	
HandsOnVLM* [5]	0.197	0.194	0.194	0.266	0.262	0.262	0.191	0.188	0.186	38.29	38.18	38.13	
FM-base	0.176	0.165	0.161	0.252	0.241	0.237	0.161	0.151	0.147	40.23	39.64	39.47	
EgoMAN-ACT	0.172	0.158	0.153	0.247	0.232	0.228	0.159	0.145	0.141	39.60	38.68	38.42	
EgoMAN (Ours)	**0.166**	**0.147**	**0.141**	**0.246**	**0.224**	**0.217**	**0.155**	**0.137**	**0.130**	**36.11**	**35.40**	**35.09**	

Table 1. **Comparison of 6DoF hand trajectory prediction on EgoMAN-Unseen and HOT3D-OOD.** Lower is better. Best values are **bold**, second-best are underlined. Our EgoMAN model outperforms the strongest external baseline (HandsOnVLM) by 27.5% ADE on both the held-out EgoMAN-Unseen test split and the out-of-distribution HOT3D-OOD dataset.

the Motion Expert was pretrained to consume ground-truth waypoints and action phrases. At inference, however, the Motion Expert must consume the predicted (and potentially noisy) tokens from the Reasoning Module. To bridge this gap, we align the two components through joint training on the Trajectory-Token Interface.

In the full EgoMAN model, the Reasoning Module is prompted with QA-style input, e.g., *Given the past wrist motion: {past_motion}. Where will the hands move to {intent}?*<HOI_QUERY>, and produces the structured trajectory token sequence <ACT><START><CONTACT><END>. These tokens are then decoded into Motion Expert inputs: <ACT> yields an action-semantic embedding that replaces the ground-truth phrase embedding, while <START>, <CONTACT>, and <END> decode into 6DoF waypoints and timestamps (i.e., their positional encodings), replacing ground-truth waypoints. To align the reasoning and motion components, we jointly train them on the EgoMAN finetuning dataset using two objectives: (1) a next-token prediction loss $\mathcal{L}_{\text{text}}$ over the trajectory-token sequence, and (2) the Flow Matching loss $\mathcal{L}_{\text{FM}}$ on the trajectories generated by the Motion Expert, as in Sec. 4.3. This unified setup enables efficient intent reasoning and produces physically consistent 6DoF trajectories aligned with the intent semantics.

5. Experiments

We evaluate the *EgoMAN* model thoroughly on EgoMAN-Bench to answer three core questions: (1) *Does the reasoning-to-motion pipeline improve long-horizon 6DoF prediction over state-of-the-art baselines?* (2) *How effectively does the Reasoning Module generate accurate and reliable waypoints for intent-aligned spatial prediction?* (3) *How do the progressive training strategy and the trajectory-token interface contribute to overall performance?* We further provide qualitative results showing diverse generalization and controllable intent-conditioned motion.

5.1. Evaluation Setting and Metrics

Trajectory Metrics. We evaluate all methods using standard hand–trajectory forecasting metrics, including **Average Displacement Error (ADE)**, **Final Displacement Error (FDE)**, and **Dynamic Time Warping (DTW)**, all reported in meters, as well as **Angular Rotation Error (Rot)** in degrees. To assess stochastic generative prediction, each model samples $K{=}1/5/10$ trajectories per query. Unless otherwise specified, all results are reported as **best-of-**K, which selects the trajectory with minimum error to the ground truth.

Waypoint (WP) Metrics. We evaluate the <CONTACT> and <END> waypoints predicted by our VLM. We report two metrics in meters to quantify the localization accuracy of key intent states: **Contact Distance (Contact):** The Euclidean distance between the predicted and ground-truth wrist locations at the approach-completion timestamp. and **Trajectory-Warp Distance (Traj):** The average Euclidean distance from each predicted waypoint to its nearest point on the GT trajectory.

5.2. Baselines

Hand Trajectory Predictor Baselines. We compare against five trajectory baselines. Baselines marked with (*) are adapted for fair comparison by matching the *EgoMAN* setting: using a single RGB image, an intent text embedding, and past motion as inputs to predict up to 5-second 6DoF bi-hand trajectories, with metrics computed over the ground-truth duration. 1) **USST*** [6]: an uncertainty-aware state–space transformer for egocentric 3D hand trajectory forecasting; 2) **MMTwin*** [41]: a model using twin diffusion experts and a Mamba–Transformer backbone for joint ego-motion and hand motion prediction; 3) **HandsOnVLM*** [5]: a VLM that predicts 2D trajectories via dialogue, which we adapt to 6DoF poses using a Conditional Variational

Method	WP	Detect	FPS ↑	EgoMAN		HOT3D	
				Contact ↓	Traj ↓	Contact ↓	Traj ↓
HAMSTER* [31]	2D-text	✗	0.17	0.342	0.297	0.236	0.219
VRB* [2]	3D	✓	0.03	0.300	0.271	0.216	0.224
VidBot [11]	3D	✓	0.04	0.290	0.269	0.190	0.147
EgoMAN-WP	3D	✗	**3.45**	**0.192**	**0.127**	**0.188**	**0.110**

Table 2. **Waypoint prediction results.** Lower is better for *Contact* and *Traj*; higher is better for *FPS* (averaged over 50 samples on an NVIDIA PG509-210, 80GB). *EgoMAN-WP* achieves the best accuracy, improving Contact by 33.8% and Traj by 52.8% on EgoMAN-Unseen, and runs orders of magnitude faster at 3.45 FPS.

Pretrain		WP	ADE ↓	FDE ↓	DTW ↓	Rot ↓
Reason	FM					
✗	✗	✗	0.273	0.308	0.260	51.79
✓	✗	6DoF	0.215	0.255	0.198	43.03
✗	✓	✗	0.162	0.225	0.148	36.24
✓	✓	✗	0.161	0.224	0.147	35.90
✓	✓	Emb	**0.150**	0.210	0.138	34.02
✓	✓	6DoF	0.151	**0.206**	**0.137**	**33.88**

Table 3. **Ablation on EgoMAN-Unseen** (K=1)**.** Lower is better. *Reason* and *FM* pretraining with *6DoF* waypoints yield the highest accuracy.

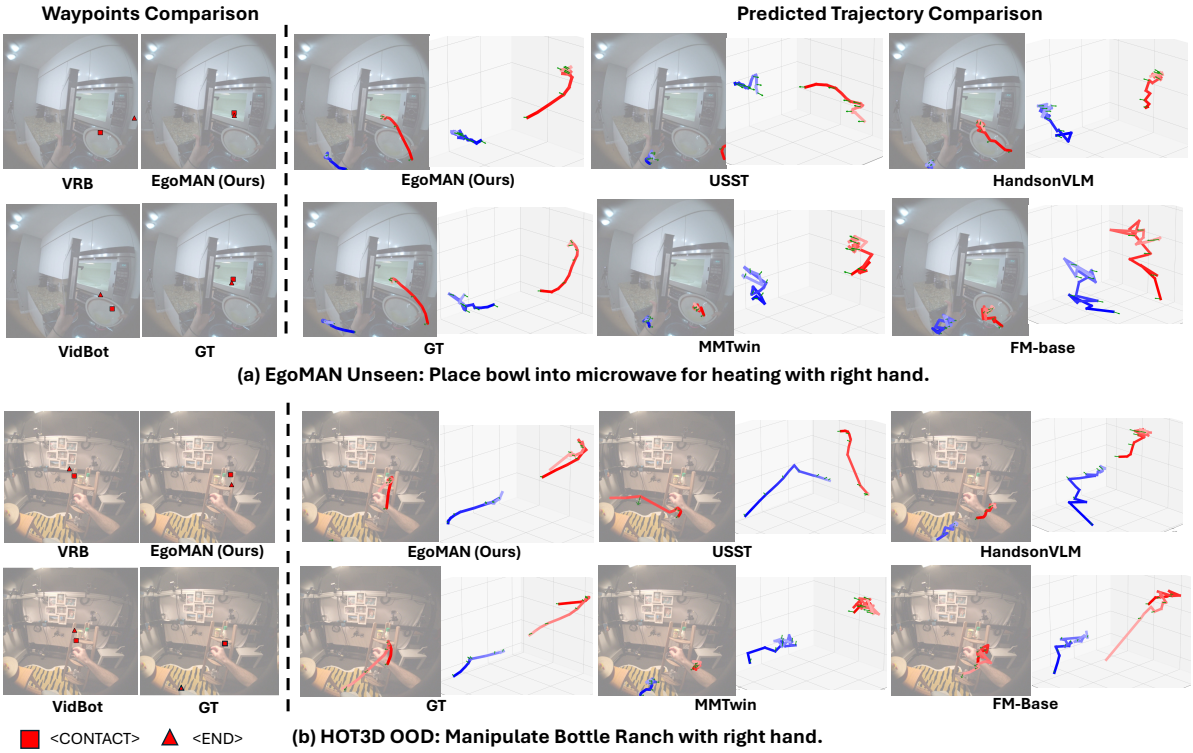

Figure 3. **Qualitative comparisons on EgoMAN-Bench.** We visualize best-of-K=10 predictions for waypoints and full trajectories. Left: <CONTACT> and <END> waypoint predictions compared with *VRB** and *VidBot*. Right: 3D hand trajectory forecasts and 2D projections compared with prior baselines. Our *EgoMAN* model produces the smoothest and closest results to ground truth.

Autoencoder (CVAE) [25] head from 50 predicted special hand tokens; We also include two ablations from our own pipeline: 4) **FM-Base**: our Flow Matching Motion Expert conditioned on image, intent, and past motion, but without VLM reasoning; and 5) **EgoMAN-ACT**: our variant that removes reasoning pre-training and waypoint supervision, conditioning the Motion Expert only on a VLM-learned <ACT> token.

Affordance-driven Baselines. We evaluate three affordance-based methods: **HAMSTER*** [31], **VRB*** [2], and **VidBot** [11], each adapted to our waypoint prediction setting for a fair comparison. All methods take the same RGB image, Metric3D depth [22], and verb–object text as input, and predict contact and goal points aligned with EgoMAN's <CONTACT> and <END> waypoints. Aria fisheye images are rectified to pinhole views using device calibration. *VRB** and *HAMSTER** produce 2D affordance points that we unproject to 3D, and for *HAMSTER** we treat the first and last predicted points as contact and goal. *VidBot* and *VRB** return object-conditioned affordance candidates; when multiple candidates appear, we select the one closest to the target object. Since these models output affordance points rather than wrist poses, we approximate wrist locations by choosing the predicted point closest to the GT wrist within 5 cm.

5.3. Results

Trajectory Evaluation. As shown in Table 1, the full *EgoMAN* model achieves the lowest ADE, FDE, DTW, and

rotation errors for all sampling budgets $K \in 1, 5, 10$. It outperforms the strongest external baseline *HandsOnVLM** by a large margin, reducing ADE at $K{=}10$ by 27.5% on the held-out EgoMAN-Unseen split and achieving a similar 27.3% reduction on the out of distribution HOT3D-OOD dataset, demonstrating strong generalization across domains.

FM-base already outperforms state-space and Mamba-based predictors (*USST**, *MMTwin**) on both test splits, showing that Flow Matching–based encoder–decoder modeling provides a stronger foundation for long-horizon 6DoF motion forecasting. Incorporating vision–language supervision further improves accuracy. *HandsOnVLM** leverages text guidance, but its CVAE decoder predicts trajectories from 50 VLM-produced hand tokens learned from noisy egocentric data, yielding higher 6DoF errors and showing little improvement even as the sample count K increases.

Waypoints Evaluation. Table 2 shows that *EgoMAN-WP*, which directly regresses 3D <CONTACT> and <END> wrist positions from the Reasoning Module, achieves the best performance on both EgoMAN-Unseen and HOT3D-OOD. It reduces contact error from 0.29–0.34,m to 0.19,m and lowers DTW from 0.27–0.30,m to 0.13,m on EgoMAN-Unseen, while maintaining the lowest contact error on the challenging test set HOT3D-OOD. *EgoMAN-WP* is also far more efficient, running at 3.45,FPS compared to < 0.05,FPS for *VRB** and *VidBot*, which require heavy detection and 3D post-processing. By predicting only four structured and compact trajectory tokens, our *EgoMAN* model gains both speed and robustness.

5.4. Ablation Study

As shown in Table 3, we ablate different components of our method. For pretraining, we toggle **Reasoning Pretrain** (Reason) and **Flow-Matching Pretrain** (*FM*). For **Waypoint** (*WP*) choices, we compare using no waypoints ($\mathcal{X}$), our explicit 6DoF waypoints (*6DoF*), and decoder's final hidden state as implicit embeddings (*Emb*).

Starting from a model without any pretraining or waypoints, adding only *FM* pretraining (✗/✓/✗) yields the largest single gain across all metrics, showing the importance of a motion-aware initialization. *Reason* pretraining alone (✓/✗/*6DoF*) also provides substantial improvements, but remains weaker than *FM* pretraining. Combining both pretraining signals (✓/✓/✗) further reduces error, indicating complementary benefits. Finally, with *Reason* and *FM* pretraining, adding the waypoint interface yields the strongest overall performance: the implicit waypoint variant *Emb* achieves the lowest ADE and closely approaches the full model with explicit *6DoF* waypoint design, which delivers the lowest FDE, DTW, and rotation errors. Please see more detailed analysis of the ablation results in appendix.

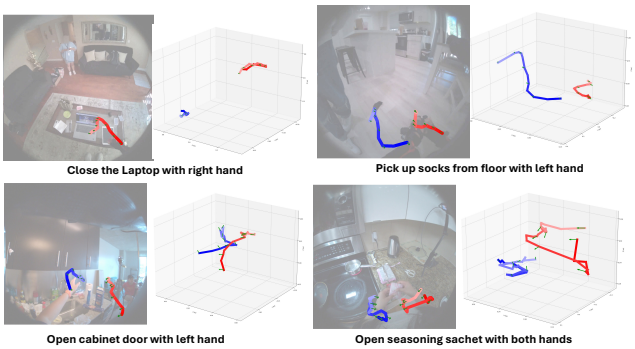

Figure 4. **Qualitative results of diverse activities.** EgoMAN generates accurate 6DoF hand trajectories for diverse activities, aligning motion with the intent description and scene spatial.

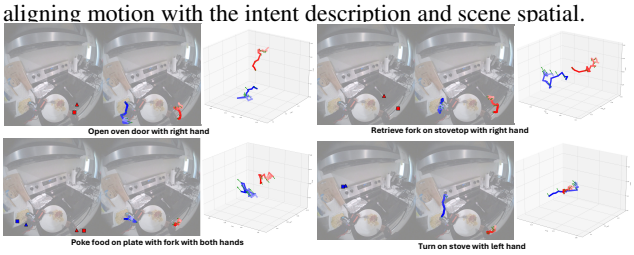

Figure 5. **Multiple intents.** With the same image and past motion, *EgoMAN* model produces distinct 6DoF trajectories for different intent queries, showing controllable intent-to-motion generation.

5.5. Qualitative Analysis

We visualize qualitative best-of-$K{=}10$ forecasts in Figure 3 on EgoMAN-Bench (EgoMAN-Unseen and HOT3D-OOD). On the left of Figure 3, we compare waypoints from affordance baselines (*VRB**, *VidBot*) with those from our *EgoMAN* model. Our predicted contact and end points align closely with the GT wrist positions, while affordance methods often miss the target surface due to detection errors or collapse toward the hand instead of the intended goal region. On Figure 3 right, we compare full 6DoF trajectories against trajectory baselines (*USST*, *HandsOnVLM*, *MMTwin*, *FM-base*). *EgoMAN* generates smoother and more accurate motions that reach the target and complete the manipulation with correct wrist orientation, while baselines often underreach, overshoot, or drift in cluttered scenes or under unfamiliar objects and intent descriptions.

Figure 4 further illustrates diverse activities such as closing a laptop, picking up socks, opening a cabinet, and opening a seasoning sachet. Across these scenarios, *EgoMAN* model consistently produces trajectories that follow the verb phrase and scene spatial contexts, demonstrating that the reasoning-to-motion pipeline generalizes well to a wide range of real-world hand–object interactions.

Finally, Figure 5 illustrates intent-conditioned motion generation under the same visual and motion context. Given different intent queries, *EgoMAN* model reasons about the action, predicts distinct waypoints, and produces correspond-

ingly different yet valid 6DoF trajectories (*e.g*. opening the oven, retrieving a fork, poking food, turning on the stove). This controllable intent-to-motion mapping enables flexible generation of diverse hand trajectories, which can support robot learning and data augmentation.

6. Conclusion

We introduced the **EgoMAN dataset**, a large-scale egocentric benchmark for interaction stage–aware 6DoF hand trajectory prediction, featuring structured QA pairs that capture semantic, spatial, and motion reasoning. We also presented the **EgoMAN model**, a modular reasoning-to-motion framework that aligns high-level intent with physically grounded 6DoF trajectories through a trajectory-token interface and progressive training. Experiments show strong gains over both motion-only and VLM baselines: Flow Matching yields smoother and more stable trajectories, VLM-driven reasoning improves semantic alignment and generalization to novel scenes and intents, and the trajectory-token interface enables efficient inference, bridging intent-conditioned stage–aware reasoning with precise low-level motion generation. Overall, **EgoMAN** offers a practical step toward in-context action prediction, supporting applications in robot manipulation, language-conditioned motion synthesis, and intent-aware assistive systems.

Acknowledgment

We thank Lingni Ma for valuable discussions and support related to the Nymeria dataset. We also thank Hanzhi Chen for assisting with adapting the affordance model for our waypoint evaluation protocol. We are grateful to the teams behind Project Aria, EgoExo4D, and HOT3D for releasing the foundational datasets that enabled this research.

References

[1] Project Aria Machine Perception Services. 3

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023. 2, 3, 7

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4, 12

[4] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 2, 3, 12

[5] Chen Bao, Jiarui Xu, Xiaolong Wang, Abhinav Gupta, and Homanga Bharadhwaj. Handsonvlm: Vision-language models for hand-object interaction prediction. *Transactions on Machine Learning Research*, 2025. 2, 6, 14

[6] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 14

[7] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6904–6911. IEEE, 2024. 2

[8] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *ECCV*, 2024. 3

[9] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *Conference on Robot Learning*, 2025. 3

[10] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision–language–action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2, 3

[11] Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. VidBot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *CVPR*, 2025. 2, 7

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[13] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2, 3

[14] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 3

[15] Aaron Grattafiori et al. The llama 3 herd of models, 2024. 2

[16] Gemma Team et al. Gemma 3 technical report, 2025. 3

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 3, 12

[18] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022. 13

[19] Masashi Hatano, Ryo Hachiuma, and Hideo Saito. Emag: Ego-motion aware and generalizable 2d hand forecasting from egocentric videos. In *European Conference on Computer Vision Workshops (ECCVW)*, 2024. 2

[20] Masashi Hatano, Zhifan Zhu, Hideo Saito, and Dima Damen. The invisible egohand: 3d hand forecasting through egobody pose estimation. *arXiv preprint arXiv:2504.08654*, 2025. 2

[21] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 2, 3

[22] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7

[23] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025. 2, 3

[24] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 3

[25] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 7

[26] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 2

[27] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 3

[28] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025. 2, 3

[29] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025. 3

[30] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[31] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, et al. Hamster: Hierarchical action models for open-world robot manipulation. *arXiv preprint arXiv:2502.05485*, 2025. 2, 3, 7

[32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. 14

[33] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 5, 12

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3

[35] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020. 2

[36] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 12

[38] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. 2, 3

[39] Junyi Ma, Xieyuanli Chen, Wentao Bao, Jingyi Xu, and Hesheng Wang. Madiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos, 2024. 2

[40] Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. *arXiv preprint arXiv:2405.04370*, 2024. 2

[41] Junyi Ma, Wentao Bao, Jingyi Xu, Guanzhong Sun, Xieyuanli Chen, and Hesheng Wang. Novel diffusion models for multimodal 3d hand trajectory prediction, 2025. 2, 6, 14

[42] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *the 18th European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 12

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4, 12

[45] OpenAI and et al. Gpt-4 technical report, 2024. 3

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002. 14

[47] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 13

[48] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. 2

[49] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision–language–action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 2, 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4, 13

[51] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[52] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 5

[53] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20270–20281, 2023. 2

[54] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025. 2, 3

[55] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 3

[56] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024. 3

[57] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 14

[58] Wenqi Zhang, Mengna Wang, Gangao Liu, Huixin Xu, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, Weiming Lu, Peng Li, and Yueting Zhuang. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025. 2, 3

[59] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 3, 4

[60] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 2, 3

A. Summary of Appendix

In this appendix, we provide:

1. A video demonstration of our system, including representative interaction cases (Sec. B).
2. Implementation details of the EgoMAN model pipeline and our progressive training strategy (Sec. C).
3. Extended analysis of ablation results presented in the main paper (Sec. D).
4. Evaluation of semantic alignment between predicted hand trajectories and action verbs, with comparisons to baselines (Sec. E).
5. Comparison across different parameter scales of the Reasoning Module, including EgoMAN-QA accuracy and trajectory prediction performance (Sec. F).
6. Representative prompt examples used in data annotation (Sec. G).
7. Limitations and future work (Sec. H).

B. Video

Our `video` provides a visual overview of the core contributions of EgoMAN, covering the dataset, the model architecture, and qualitative demonstrations. The video first introduces the EgoMAN dataset and highlights the full EgoMAN pipeline, consisting of the Reasoning Module, the Motion Expert, and the end-to-end 6-DoF trajectory generation flow (bridging reasoning to motion through the Trajectory-Token Interface).

In the dataset overview segment, we present diverse examples of hand–object interactions corresponding to the twelve most frequent verbs in EgoMAN (e.g., *Grasp*, *open*, *place*, *pour*, *stir*). The showcased clips span multiple sources such as EgoExo4D [17], Nymeria [42], and HOT3D-Aria [4], illustrating that interactions occur in realistic everyday scenes with natural noise, clutter, and challenging viewpoints. These examples demonstrate the dataset's scale, variability, and difficulty, motivating the need for robust intention-conditioned 3D hand trajectory modeling.

The qualitative results section highlights EgoMAN's ability to generate intention-guided trajectories. We first show dozens of representative cases across various scenarios: such as *stir milk* and *turn off stove* in kitchen scenes, *pick up socks* and *open door* in household scenes, *close laptop* and *grab cable* in working scenarios, and *manipulate bowl* or *manipulate ranch bottle* in HOT3D scenes. For each case, the video displays (1) the original input image, (2) the intention text, (3) intermediate waypoint predictions from the Reasoning Module, and (4) the final 6-DoF trajectory output. These predictions are visualized both on the static input image and overlaid on future ego-video frames to more clearly illustrate spatial accuracy and motion quality.

Across all demonstrations, EgoMAN consistently predicts accurate contact and end-point waypoints around target objects, with the generated 3D trajectories following realistic manipulation paths that match the intended semantics. While certain open-ended tasks (e.g., *open door*, *pick up socks*) may exhibit slight variations in final pose or timing, or minor deviations in the non-manipulating hand relative to the single ground-truth instance, the predicted trajectories for the manipulating hand remain semantically aligned with the intended goal. These results highlight EgoMAN's capability to produce reliable, intention-driven 6-DoF hand trajectories across diverse scenes and interaction types.

We further show results of *goal-directed trajectory generation*, where the same input image paired with different intention descriptions. EgoMAN model is able to predict trajectories in distinct that align with the intended goals, even in unseen environments.

Please visit our `project website` to check more trajectory prediction results in diverse interaction scenarios.

C. Implementation Details

Reasoning Module. The Reasoning Module is optimized in `bf16` using AdamW [37] with cosine learning rate decay. The vision encoder and multimodal projector are frozen. We use a base learning rate of 1×10^{-5}, a warmup ratio of 0.02, weight decay of 0.05, maximum gradient norm of 1.0, and a batch size of 256 across $8\times$NVIDIA A100 80GB GPUs. Training runs for 2 epochs on approximately 1M EgoMAN QA samples. Images use dynamic resizing with `max_pixels=50176` and `min_pixels=784`. If past motion is provided in the input question, we use the most recent 5 past points of both hands tokenized at 10 fps; otherwise a zero-initialized motion history is used. A 4-layer MLP is used to extract features from the motion, which are then fed into Qwen2.5-VL [3]. The specialized waypoint decoders are lightweight ReLU MLPs with hidden dimension 768, predicting timestamp, 3D position, and 6DoF rotation. The action semantic decoder is a single-layer MLP (dim 768). When valid samples in a batch fall below $\kappa=10$, we use cosine similarity loss; otherwise, we apply an InfoNCE loss [44]. Loss weights are set as $\lambda_{\text{wp}}=0.3$ and $\lambda_{\text{act}}=0.1$, with internal weights $\lambda_t=1.0$, $\lambda_{3D}=2.0$, $\lambda_{2D}=0.5$, $\lambda_r=0.5$, and $\lambda_{\text{geo}}=0.15$. We apply Huber loss with $\beta=0.2$ for rot6D, $\beta_{3D}=0.07$, and $\beta_{2D}=0.02$ for location terms. The geodesic rotation loss is applied only to visible waypoints. Temporal modulation is implemented using a Gaussian time window with $\sigma_{\text{time}}=3.0$.

Motion Expert. We pre-train the flow-matching (FM) [33] based motion decoder using approximately 17K trajectories. Inputs include DINOv3 image features, ground-truth action phrases, waypoint tokens, and past wrist motion. Motion sequences are sampled at 10 fps with a maximum 50-step future horizon (5 s). The FM architecture uses a hidden dimension of 768, with 6 encoder and 6 decoder transformer layers and 8 attention heads. A sinusoidal time embedding

Pretrain Reason	Pretrain FM	WP	ADE ↓	FDE ↓	DTW ↓	Rot ↓
✗	✗	✗	0.273	0.308	0.260	51.79
✓	✗	6DoF	0.215	0.255	0.198	43.03
✗	✓	✗	0.162	0.225	0.148	36.24
✓	✓	✗	0.161	0.224	0.147	35.90
✓	✓	Emb	**0.150**	0.210	0.138	34.02
✓	✓	6DoF	0.151	**0.206**	**0.137**	**33.88**

Table 4. **Ablation on EgoMAN-Unseen** (K=1)**.** Lower is better. *Reason* and *FM* pretraining with *6DoF* waypoints yield the highest accuracy.

(256-D) is mapped to FiLM-style [47](γ, β) parameters. A 2-layer self-attention block is applied before decoding, along with modality and positional embeddings. We train in FP32 using AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, a cosine schedule with 5% warmup, and a batch size of 256 on a single A100 GPU. The training objective is the sum of MSE loss on 3D positions and MSE loss on 6D rotations, with a rotation loss weight of 0.5. At inference time, we iterate for 150 steps and retain only the predicted trajectory segment beyond the length of the ground-truth target.

Joint Training of EgoMAN Model. We initialize the Reasoning Module from the reasoning pretraining checkpoint and the motion decoder from the FM pretraining weights. The training setup largely follows the reasoning pretraining configuration, but FM components are kept in FP32. We use a learning rate of 5×10^{-6} and a batch size of 128 cross 8×NVIDIA A100 80GB GPUs. The model is trained for 60 epochs on the same finetuning trajectory dataset used in motion pretraining. At inference time, we iterate for 150 steps and retain only the predicted trajectory segment beyond the ground-truth target length.

D. Detailed Analysis for Ablation Study

In this section, we provide more detailed analysis of our ablation results in main paper Sec 5.4 and Table 4 in the appendix.

Reasoning Pretraining. Removing reasoning pretraining (*EgoMAN-ACT*), which also disables WP, degrades performance (ADE $0.162\rightarrow0.215$). This pretraining uses large-scale, noisy corpora with question-answer pairs, encoding semantic, spatial, and motion-aware priors that help disambiguate intent. Learned waypoints further align trajectories with both intent and visual context. Data-efficiency results (Fig. 6) support this trend: at 20% of the training data, *EgoMAN* (ADE $\sim$ 0.13 m) remains superior to *EgoMAN-ACT* (ADE $\sim$ 0.16 m), and the gap persists even at full data (ADE $0.140\rightarrow0.125$). FDE exhibits a similar pattern. Reasoning pretraining and waypoint conditioning reduce ambiguity early, enabling stable outputs with fewer high-quality labels.

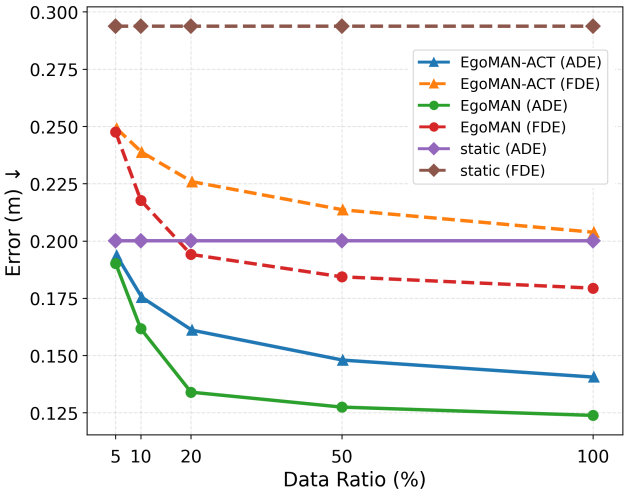

Figure 6. **Data efficiency results.** ADE/FDE (m), best-of-10. The *static* baseline repeats the last observed hand location. Without pretraining, errors of EgoMAN-ACT rise sharply under limited data, while EgoMAN maintains strong performance even at 20% data, highlighting the benefit of waypoint-based Reasoning Module and pretraining.

Trajectory-Token Interface via Waypoints (WP). WP provides a structured interface between the visual-language module and the 6DoF Motion Expert. Conditioning on predicted WP improves accuracy and stability (ADE $0.161\rightarrow0.151$, DTW $0.147\rightarrow0.137$, Rot $35.90^\circ\rightarrow33.88^\circ$). Without WP, performance drops near that of *EgoMAN-ACT*, suggesting that learned WP enhance the contribution of reasoning pretraining beyond implicit semantic embeddings. Replacing 6DoF waypoints with the decoder's final hidden state (Emb) yields only minor differences (FDE $\sim$0.4 cm worse; other metrics nearly unchanged).

Motion Pretraining (FM). Removing FM while retaining reasoning pretraining leads to clear degradation (ADE $0.150\rightarrow0.215$). Without FM, the model must learn both semantics and motion jointly through an implicit semantics interface, resulting in noisier long-horizon predictions and increased rotation error. Removing both FM and reasoning causes further decline (ADE 0.273, Rot 51.79°). While reasoning pretrain and waypoints learning help mitigate this, pretraining FM on physically plausible motion first, followed by joint fine-tuning with reasoning, produces the largest improvements across all metrics.

E. Motion-to-Text Alignment

We evaluate semantic alignment between trajectories and action verbs by training a motion encoder to map hand trajectories to a pre-computed verb text embedding space using a CLIP-style contrastive loss [18, 50]. We report **Recall@3**

Method	R@3 ↑	FID ↓
USST* [6]	15.0	0.22
MMTwin* [41]	22.9	0.86
HandsonVLM* [5]	27.9	0.10
FM-Base	39.7	0.05
EgoMAN	**43.9**	**0.04**

Table 5. **Motion-to-Verb Text Retrieval.** Train one encoder; evaluate verb text-motion relevance over 239 verb candidates.

(fraction of samples where the GT caption is retrieved in the top 3 over 239 verbs in the test samples) and **Fréchet Inception Distance (FID)** between predicted and ground-truth motion embeddings. To account for generative diversity, we report best-of-K (K=10) retrieval results.

Table 5 shows that *EgoMAN* achieves the strongest semantic alignment between motion and verb phrases, with the highest R@3 (43.9%) and lowest FID (0.04). Generative baselines such as *USST* and *MMTwin* produce smooth trajectories but exhibit weaker text alignment, while *HandsOnVLM* benefits from language conditioning yet suffers from noisy CVAE decoding. *FM-Base* already improves verb specificity, indicating that Flow Matching promotes more structured motion. Adding VLM reasoning and waypoint constraints in *EgoMAN* further reduces ambiguity, tightening the trajectory–verb correspondence and producing a motion embedding distribution closer to ground truth.

F. Reasoning Module Scale Analysis

In this section, we analyze how scaling the Reasoning Module affects both high-level semantic reasoning and downstream motion prediction. We evaluate multiple model sizes from Qwen2.5-VL and Qwen3-VL families, using identical training data and identical Trajectory-Token Interface settings. Our analysis focuses on two components: (i) EgoMAN QA: measuring semantic and spatial reasoning, and (ii) trajectory prediction on EgoMAN-Bench: measuring the effect of Reasoning Module scale on 6-DoF hand trajectory generation.

F.1. EgoMAN QA

Evaluation Metrics. We evaluate three complementary aspects of reasoning quality:

- **Waypoint Spatial Reasoning**: We evaluate 3D waypoint accuracy for <CONTACT> and <END> using three metrics. **Location error (Loc)** is the Euclidean distance (in meters) between the predicted and ground-truth waypoints' positions. **Time error (Time)** measures the temporal accuracy of the predicted interaction stages. Since <START> (approach onset) is always aligned to time 0, we compute the L1 difference (in seconds) only for the predicted <CONTACT> (manipulation onset) and <END> (manipulation completion) timestamps.
  Rotation error (Rot) is the geodesic distance (in degrees) between predicted and ground-truth wrist orientations, computed from the relative rotation matrix. These metrics quantify spatial, temporal, and rotational grounding of interaction-stage waypoints.
- **Semantic Embedding Alignment**: we compute **Recall@3 (R@3)** between predicted and 2844 ground-truth action embeddings, as well as the **mean Pearson correlation (Pearson)** across embedding dimensions, which reflects how well the learned embedding space preserves semantic similarity.
- **Semantic Text QA**: We measure the quality of generated answers using three complementary NLP metrics. **BERTScore (BERT)** [57] computes semantic similarity using contextualized token embeddings from a pretrained BERT model, capturing paraphrases and fine-grained meaning. **BLEU** [46] evaluates n-gram precision between predictions and references, reflecting lexical overlap. **ROUGE-L (ROUGE)** [32] measures the longest common subsequence between texts, capturing phrase-level recall. Together, these metrics assess both semantic fidelity and surface-form similarity between predicted answers and ground-truth explanations.

Results Analysis. As shown in Table 6, the models achieve strong textual QA performance (BERTScore ≈0.92, ROUGE≈0.49) and moderate but meaningful semantic alignment (R@3 up to 11% and Pearson up to 0.26), providing reliable semantic grounding despite the large action-embedding space consists of 2844 samples.

Table 6 also shows that increasing reasoning-module capacity improves both waypoint grounding and semantic understanding. For spatial waypoint prediction, Qwen3-VL 4B achieves the best overall accuracy, obtaining the lowest averaged location error (0.223 m) and the lowest rotation error (40.89°) across all models. Qwen3-VL 8B further stabilizes spatial performance, achieving similarly strong location and rotation errors, indicating that spatial grounding saturates around the 4B–8B scale for Qwen3-VL. In contrast, semantic embedding alignment exhibits a different scaling trend: Qwen2.5-VL 7B reaches the highest R@3 and Pearson correlation, demonstrating the strongest alignment between reasoning tokens and action semantics. Smaller Qwen3-VL models (2B and 4B) lag in semantic alignment despite strong spatial accuracy, suggesting that *fine-grained action–semantic grounding is more capacity-dependent than geometric waypoint prediction*. Overall, scaling improves all metrics, but spatial accuracy peaks earlier (at 4B), whereas semantic–action alignment continues improving with larger reasoning capacity.

Model	Params	Spatial Reasoning (Waypoints)			Semantic Embedding		Semantic Text QA		
		Loc ↓	Time ↓	Rot° ↓	R@3 (%) ↑	Pearson ↑	BERT ↑	BLEU ↑	ROUGE ↑
Qwen2.5-VL	3B	0.229	0.483	41.36	6.26	0.239	0.914	0.155	0.469
Qwen2.5-VL	7B	0.225	**0.474**	41.27	**11.08**	**0.256**	0.916	0.165	0.481
Qwen3-VL	2B	0.244	0.495	41.88	1.62	0.107	**0.919**	0.171	0.487
Qwen3-VL	4B	**0.223**	0.481	**40.89**	1.69	0.134	**0.919**	0.171	**0.494**
Qwen3-VL	8B	0.228	0.477	41.24	5.66	0.205	0.917	**0.172**	0.488

Table 6. **Effect of model scale on spatial reasoning, semantic alignment, and text QA on EgoMAN Unseen benchmark.** We evaluate (i) waypoint spatial reasoning via 3D location, time, and rotation errors, (ii) semantic embedding alignment using R@3 (computed over 2,844 GT action-embedding candidates) and mean Pearson correlation, and (iii) semantic text QA using BERTScore, BLEU, and ROUGE. Best values are **bolded**; second-best are underlined. Spatial reasoning performance saturates early, and models larger than 2B/3B provide consistently stronger performance. Semantic alignment benefits from larger models, with Qwen2.5-VL outperforming Qwen3-VL, while text QA remains relatively stable across scales, with Qwen3-VL slightly outperforming Qwen2.5-VL.

Model	Params	EgoMAN Unseen				HOT3D OOD			
		ADE↓	FDE↓	DTW↓	Rot° ↓	ADE↓	FDE↓	DTW↓	Rot° ↓
Qwen2.5-VL	3B	0.128	0.184	0.115	33.16	0.146	0.221	0.135	35.82
Qwen2.5-VL	7B	0.124	0.179	0.111	32.75	0.141	0.217	0.130	35.09
Qwen3-VL	2B	0.130	0.186	0.118	33.48	0.142	0.216	0.132	35.62
Qwen3-VL	4B	0.123	0.178	0.111	32.63	**0.139**	**0.212**	**0.128**	34.65
Qwen3-VL	8B	**0.122**	**0.177**	**0.110**	**32.31**	0.140	0.214	0.129	**34.62**

Table 7. **Effect of Reasoning Module scale on trajectory prediction.** Best results are **bolded** and second-best are underlined. Larger reasoning models produce consistently more accurate 6-DoF trajectories on both EgoMAN Unseen and HOT3D OOD, with Qwen3-VL scaling smoothly and the 4B model offering an excellent speed–accuracy trade-off.

F.2. Trajectory Prediction on EgoMAN-Bench

Evaluation Metrics. We measure stage-aware 6-DoF trajectory prediction using four metrics that are consistent with the metrics we use in the main paper:

- **ADE**: **Average Displacement Error (ADE)** is the mean Euclidean distance between the predicted and ground-truth 3D wrist positions over all future timesteps.
- **FDE**: **Final Displacement Error (FDE)** measures this distance only at the final prediction timestep. Both are computed in meters and evaluate the overall spatial accuracy and final-state consistency of the trajectory.
- **DTW**: **Dynamic Time Warping (DTW)** measures the minimum alignment cost between the predicted and ground-truth trajectories after allowing temporal stretching or compression. It captures discrepancies in both spatial shape and temporal progression, making it sensitive to timing errors such as early or late motion onset.
- **Rotation Error (Rot)**: The mean geodesic rotation error computed from the relative rotation matrix between predicted and ground-truth wrist orientations. We use the standard geodesic distance in degrees, which measures the smallest 3D rotational difference between two orientations.

All results use best-of-K (sampling $K = 10$) on EgoMAN Unseen and HOT3D OOD)

Results. Table 7 shows that scaling the Reasoning Module leads to consistent improvements in 6-DoF trajectory prediction across both EgoMAN Unseen and HOT3D OOD. Within each model family, larger variants reduce ADE/FDE and DTW, indicating more accurate and temporally aligned motion forecasts. Qwen3-VL 4B and 8B achieve the strongest overall performance, obtaining the lowest ADE, FDE, and rotation errors on EgoMAN Unseen, and competitive or best results on HOT3D OOD. Although Qwen2.5-VL 7B maintains solid performance, the Qwen3-VL models benefit more directly from scale, suggesting that *trajectory prediction, unlike semantic embedding alignment, scales smoothly and saturates later in the Qwen3 family*.

Overall, increasing reasoning-module capacity strengthens stage-aware 6-DoF trajectory prediction. From a speed–performance perspective, Qwen3-VL 4B provides an excellent balance between efficiency and accuracy, while the 7B–8B models offer the strongest overall trajectory quality at higher computational cost.

G. Prompt Examples

In this section, we detail the LLM-based prompting pipeline used to construct the EgoMAN benchmark. Our pipeline consists of four stages: (i) extracting fine-grained hand–object interaction segments with temporal structure, (ii) filtering invalid or irrelevant interactions and canonicalizing intention goals, (iii) generating diverse QA pairs for reasoning pretraining, and (iv) filtering trajectory phrases to retain only visually grounded, unambiguous interaction samples. The corresponding prompts are shown in Figs. 7–10.

G.1. EgoMAN Interaction Annotation

We first extract temporally localized, atomic hand–object interactions from continuous egocentric video. Given reference narrations and timestamps, the LLM is prompted to decompose an interaction into structured *approach* and *manipulation* stages, each annotated with start/end times, coarse trajectory attributes (start/end locations and shape), and a natural-language atomic description and reasoning. The output is serialized as JSON and forms the core interaction representation used in later stages (Fig. 7).

G.2. Valid Interaction Annotation Filtering

Not all extracted segments correspond to clean, usable hand–object interactions. We therefore perform a second LLM pass to filter invalid or noisy annotations. Given a candidate atomic description and the reference annotations, the model judges whether the interaction is (i) relevant to the underlying sequence, and (ii) a true hand–object interaction performed by the main subject, rather than background motion or non-manipulative activities. It also summarizes the high-level intention goal into a short phrase, which we later use as a canonical intent label and conditioning token (Fig. 8).

G.3. EgoMAN QA Generation

For each valid interaction, we generate a set of diverse, intent-aware QA pairs used to train the Reasoning Module. The prompt in Fig. 9 guides the LLM to produce 8–12 short question–answer pairs that cover complementary aspects of the next interaction: intention goal, which hand will be used, upcoming action and object, spatial trajectory, temporal onset and completion, atomic motion description, and causal reasoning ("why" the action occurs). The prompt enforces that all answers must be grounded strictly in the provided interaction data and that the intention goal is injected in multiple phrasings to encourage robust semantic alignment.

G.4. Trajectory Filtering

We apply an image-conditioned filtering step to ensure that the interaction phrases used for trajectory prediction are visually grounded and unambiguous. As shown in Fig. 10, the LLM is asked to verify that (i) the described interaction is physically realistic, (ii) the target object is clearly visible in the egocentric frame, (iii) the image quality is sufficient, and (iv) the phrase refers to a single, unambiguous object. The model outputs a binary validity flag and a short failure reason when rejected. This step prunes low-quality or ambiguous samples and improves the reliability of EgoMAN-Bench trajectory supervision.

H. Limitations and Future Work

While EgoMAN demonstrates strong intention-conditioned 6-DoF trajectory prediction, several limitations remain. First, our modeling focuses primarily on wrist-level 6-DoF motion and considers only coarse interaction stages (`<START>`, `<CONTACT>`, `<END>`). More fine-grained sub-stages—such as pre-contact adjustment, micro-corrections during manipulation, or multi-step object reorientation—are not explicitly modeled, limiting the system's ability to capture high-resolution dexterous behavior. Second, although our dataset is large-scale and diverse, it inevitably contains sensor noise, imperfect annotations, and no human verification loop; higher-quality 3D trajectories and cleaner supervision would further benefit learning.

Future work includes extending the representation from wrist trajectories to full hand pose and articulation, enabling more fine-grained reasoning about object manipulation and grasp dynamics. Incorporating multi-stage interaction parsing and richer contact semantics would further improve temporal grounding. Improving dataset quality through higher-fidelity 3D annotations or curated human-verified demonstrations could significantly enhance supervision for fine-grained manipulation learning. Finally, deploying EgoMAN-derived policies on real robotic systems presents an exciting direction for evaluating how intention-grounded 6-DoF predictions transfer to embodied manipulation performance.

Prompt: Hand–Object Interaction Extraction

System Instruction: Extract hand–object interactions from video frames.

Reference Atomic Description with Timestamps: `{ref_annos}`

Output Format (JSON):

```
{
  "intent": "<action_goal>",
  "interactions": [
    {
      "approach": {
        "start_time": <float>,
        "end_time": <float>,
        "trajectory": {
          "start_point": "<location>",
          "end_point": "<location>",
          "shape": "<linear|curved|arc>"
        }
      },
      "manipulation": {
        "start_time": <float>,
        "end_time": <float>,
        "verb": "<action>",
        "object": "<object with short appearance description not mention
hand>",
        "hand": "<left|right|both>",
        "trajectory": {
          "start_point": "<location>",
          "end_point": "<location>",
          "shape": "<linear|curved|arc>"
        }
      },
      "atomic_description": "<interaction description>",
      "reasoning": "<why the action serves the goal and the trajectory
pattern>"
    }
  ]
}
```

Rules:

- The `approach` stage exists when the hand moves to reach the manipulation location; there is no contact until the object is touched.
- If the hand is already in contact at the manipulation location, skip the approach stage and start with `manipulation`.
- Each stage's trajectory must include three keys: `start_point`, `end_point`, and `shape` (one-word movement pattern).
- **Short reasoning:** Explain both (1) why the action serves the overall intent and (2) why the trajectory follows this pattern.
- Use precise timestamps derived from video frames.

Figure 7. Prompt used to generate fine-grained hand–object interaction annotations.

Prompt: Interaction Validity Judgment

System Instruction: Judge whether a proposed hand–object interaction is valid and relevant to the reference atomic descriptions, and summarize the high-level intention goal.

Inputs:
`interact_data.atomic_description` (free-text for the candidate interaction)
`{ref_annos}` (reference atomic descriptions with timestamps)

Decision Criteria:

- **Relevance:** The reference atomic descriptions summarize the whole action sequence. If the candidate interaction could plausibly be a sub-stage or part of this sequence, consider it relevant unless there is an obvious contradiction.
- **Validity:** The interaction must involve an actual hand–object interaction by the subject C (not just moving in air, not other participants, not whole-body locomotion, not looking-only).
- **Intention Goal:** Provide a concise, high-level phrase describing the goal of the interaction. Do not use parentheses or slashes.

Output Format (JSON):

```
{
  "valid": "<valid|invalid>",
  "intention_goal": "<short_high_level_goal_phrase>"
}
```

System Prompt (for LLMs):

```
You are an expert in analyzing hand-object interactions. Given an interaction
description and reference atomic descriptions, judge: 1) relevance to the
reference, 2) whether it is a real hand-object interaction of subject C, and 3)
summarize the high-level intention goal.
Return ONLY a JSON object with keys "valid" and "intention_goal".
```

Figure 8. Prompt used to judge interaction validity and summarize the intention goal.

Prompt: QA Generation

System Instruction: Generate short, diverse question–answer pairs about the next hand–object interaction using **only** the provided data.

Input: `{interact_data}` (represents the next interaction to predict)

Rules:

- Answers must be short natural phrases. **Only use information from the provided data** — do not fabricate details such as timing.
- **Do not use parentheses in questions or answers.**
- Generate **8–12** QA pairs covering:
  1. One question asking about the current intention goal.
  2. For other questions, inject the intention goal in diverse formats, e.g.:
     "Given the intention to . . . ", "To achieve . . . , . . . ?", "While pursuing . . . , . . . ?",
     "In order to . . . , . . . ?", "When attempting to . . . , . . . ?", "For the purpose of . . . , . . . ?",
     "As part of . . . , . . . ?", ". . . to accomplish . . . ?"
  3. Which hand will be used next.
  4. What action will occur next.
  5. What object will be manipulated next.
  6. What trajectory the hands will follow next for manipulation.
  7. When the next manipulation will start/end or start and end (use manipulation timestamps).
  8. Where the next manipulation will start/end or start and end.
  9. What is the next atomic motion (the atomic description of the next interaction).
  10. Why the next action will happen (reasoning).
  11. **If an approach stage exists before manipulation:** when the approach ends; where the approach starts/ends; what the approach trajectory is like.

Output Format (JSON Array):

```
[
  { "q":  "<question_text>", "a":  "<short_answer>" },
  { "q":  "<question_text>", "a":  "<short_answer>" },
  ...
]
```

System Prompt (for LLMs):

```
You are a data annotator.  Generate diverse QA pairs about the provided
hand-object interaction (the next interaction to predict).  Follow the Rules
exactly and use only the provided data.  Output a JSON array of objects with
keys "q" and "a".
```

Figure 9. Prompt used to generate diverse QA pairs for the future hand–object interaction.

Prompt: Interaction Trajectory Quality Filtering

System Instruction: Validate whether a proposed interaction phrase is realistic, unambiguous, and visually grounded in the given image.

Inputs:
`image` (single egocentric frame)
`phrase` (interaction phrase)

Validation Criteria (ALL must be satisfied):

- **Realism:** The phrase must describe a physically plausible hand–object interaction.
- **Object Visibility:** The described target object must be clearly visible in the image (hand visibility not required).
- **Image Quality:** The image must be sufficiently clear to identify the target object.
- **Unambiguous Target:** The phrase must specify a single object without ambiguity.

Output Format (JSON Only):

```
{
  "valid": true
}
```

or

```
{
  "valid": false,
  "reason": "<failed_criterion>"
}
```

System Prompt (for LLMs):

```
Analyze the image together with the interaction phrase: "<PHRASE>".
Check whether all four criteria are satisfied: (1) realistic interaction, (2)
target object visible, (3) image quality sufficient, (4) target unambiguous.
Return ONLY valid JSON:
{"valid": true} if all pass; otherwise {"valid": false, "reason": "<which
criterion failed>"}.
Example reasons: "object not visible", "ambiguous target", "poor image
quality", "unrealistic interaction".
```

Figure 10. Prompt used to filter out high-quality interaction trajectory samples.