EgoMAN: Interaction-Structured Reasoning for Egocentric 3D Hand Trajectory Prediction

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/

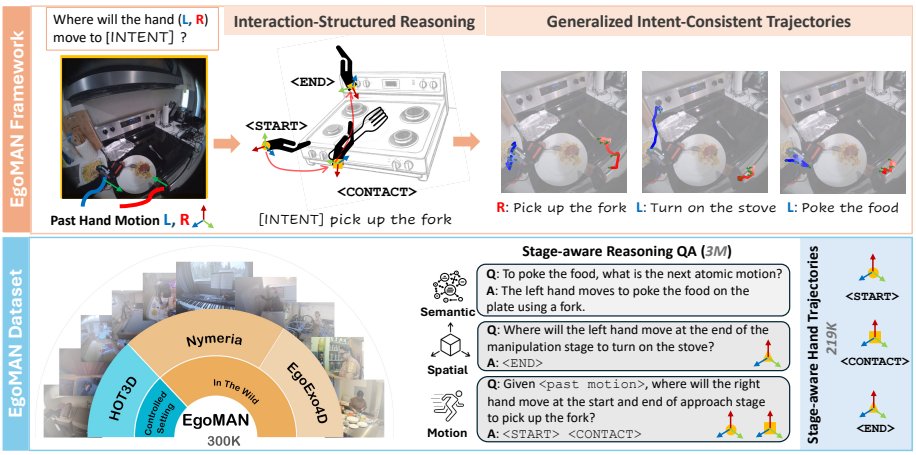


Fig. 1: EgoMAN overview. EgoMAN predicts egocentric 3D hand trajectories through *interaction-structured reasoning*. Given visual observations, past hand motion, and an intent query, the model infers key interaction stages (start, contact, end) and generates intent-consistent 6DoF trajectories. To enable this formulation, we introduce the EgoMAN dataset, a large-scale egocentric benchmark with stage-aware trajectory annotations and 3M structured QA pairs for semantic, spatial, and motion reasoning.

Abstract. Our work addresses 3D hand trajectory prediction in egocentric interaction, where future hand motion is inferred from visual observations, past motion, spatial context, and intent. Real-world actions follow stage-aware interaction structures (e.g., approach, manipulate) describing how the hand interacts with objects over time. However, prior works typically treat trajectory prediction as continuous signal regression, decoupling motion from semantic supervision and ignoring interaction structure. Without stage-aware cues to infer intent, models struggle to

* Work done during an internship at Meta.

separate purposeful motion from egocentric noise and generalize across diverse interactions. We therefore present EgoMAN, a unified framework for interaction-structured 3D hand trajectory prediction that models hand motion as stage-aware interactions between the hand and surrounding objects. EgoMAN introduces a novel Trajectory-Token Interface where a small set of tokens encodes interaction stages, temporal progression, and 6DoF pose, enabling interaction stage-aware reasoning to guide efficient long-horizon 3D trajectory generation while preserving physical interpretability. To support this formulation, we construct the EgoMAN dataset with 219K 6DoF trajectories, stage-aware annotations, and 3M semantic, spatial, and motion QA pairs. Experiments show that EgoMAN improves trajectory accuracy, smoothness, and generalization, enabling interaction-structured reasoning for egocentric hand motion prediction for applications in robotics and assistive systems.

Keywords: 3D Hand Trajectory Prediction · Egocentric Vision · Interaction Reasoning · Vision–Language–Action

1 Introduction

Predicting future 3D hand motion is essential for *in-context interaction* and proactive assistance, where systems anticipate human actions and interactions with objects using visual observations, past motion, spatial context, and task intent. Humans perform this naturally by understanding action goals, interpreting object layouts, and coordinating movements over time. Achieving this computationally requires modeling the interplay between task semantics, scene geometry, and the temporal evolution of motion. Motivated by these requirements, our work develops a framework that predicts long-horizon 3D hand trajectories by integrating visual observations, motion history, and high-level intent cues. This capability enables applications in robot manipulation, language-conditioned motion synthesis, and intent-aware assistive systems.

In real-world egocentric videos, hand motion often contains substantial transitional and exploratory movements that are weakly tied to task goals, making it difficult to distinguish purposeful interaction from incidental motion. Large-scale egocentric video datasets [15,37] reflect this challenge: while they capture diverse activities, they often contain noisy, weakly goal-directed trajectories without clear interaction stage progression. As a result, most existing 3D motion prediction methods [4,5,18,34] treat motion as a continuous homogeneous sequence focused on short-term dynamics, conflating intent-driven actions with incidental motion and limiting alignment between trajectories, intent, and spatial context and often leading to poor generalization.

To capture the underlying organization of hand motion, we define an *interaction structure* that represents hand–object interactions as stage-aware phases (e.g., approaching, manipulating). Modeling trajectories through this structure anchors motion around meaningful interaction events, helping separate purposeful actions from incidental motion.

However, such a structure is rarely explicit in existing datasets: controlled datasets [3, 23, 28] provide accurate scripted interactions but limited diversity, while large-scale egocentric datasets [15, 37] capture diverse activities yet often lack explicit stage-level organization.

Existing modeling approaches also struggle to exploit interaction structure efficiently, *i.e.* reasoning over stage-aware hand–object interaction phases that link task intent to motion. Vision-Language-Action (VLA) systems [22, 33, 54] show strong high-level reasoning ability but remain limited in directly generating smooth, high-frequency 3D trajectories from VLMs [2,11,12,14,30]. Several works therefore couple reasoning models with motion experts [9, 25, 27, 29, 41, 42, 48, 52], but often rely on implicit tokens or lengthy reasoning chains, resulting in inefficient reasoning-to-motion coupling with limited awareness of interaction structure. Meanwhile, affordance-based methods [1, 10] incorporate interaction cues such as contact prediction but depend on object detectors and predefined affordance estimators, propagating perception errors and limiting flexibility for general reasoning. Overall, prior approaches either lack explicit interaction stage modeling, rely on brittle perception pipelines, or couple reasoning and motion inefficiently, limiting interaction-structured 3D hand trajectory prediction. This sparse yet expressive representation leverages the inherent structure of hand-object interactions, allowing a minimal set of tokens to guide the generation of extended trajectories while maintaining both efficiency and physical interpretability.

We further introduce a progressive three-stage training strategy that learns (i) intent-conditioning and stage-aware reasoning, (ii) fine-grained motion dynamics, and (iii) their alignment through the Trajectory-Token Interface. This structured design synchronizes high-level intent with physically grounded motion generation, enabling long-horizon, stage-aware, and intent-consistent 3D trajectory prediction across diverse real-world egocentric scenes.

To support structured reasoning supervision, we construct the **EgoMAN dataset**, a large-scale egocentric benchmark containing 219K 6DoF trajectories annotated with interaction stages and over 3M structured vision–language–motion question–answer pairs. These annotations capture *why* actions occur (intent), *when* interaction stages happen, and *how* motion unfolds in 3D space, providing supervision for interaction-structured reasoning and trajectory prediction.

Our main contributions are:

- We formulate **interaction-structured 3D hand trajectory prediction**, introducing stage-aware reasoning to separate intent-driven motion from noisy egocentric dynamics.
- We introduce **EgoMAN**, a compact reasoning-to-motion framework with a Trajectory-Token Interface and progressive training for efficient 6DoF trajectory generation, achieving state-of-the-art accuracy and smoother motions.
- We construct the **EgoMAN dataset**, a large-scale egocentric dataset providing semantic, spatial, and motion reasoning supervision.

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 frame rates [4, 17, 32, 35, 36], while curated datasets enable *3D* trajectory prediction [5, 18, 34]. Existing 3D approaches typically follow two directions. *Affordance-driven models* [1, 10, 32] rely on object detectors and predefined affordances, which introduce perception errors and additional computational cost. *End-to-end motion predictors* generate trajectories directly from video and past motion, sometimes incorporating egomotion [34–36] or 3D priors [34]. Given that 3D trajectory labels are often uncertain and scarce [5], generative models have become standard: VAEs [4], state-space models [5], diffusion [18, 36], and hybrid variants [34, 35]. However, most methods focus on short-horizon dynamics and encode intent implicitly, limiting generalization in diverse egocentric scenarios. In contrast, our approach predicts long-horizon 6DoF trajectories by conditioning motion on intent, spatial context, and interaction stages.

Learning Interactions from Human Videos. Human videos provide rich demonstrations of hand–object interactions, motivating research in interaction reconstruction and forecasting [5, 7, 31, 32, 36]. Controlled datasets [3, 23, 28] offer precise 3D annotations but limited task diversity, while robotic imitation datasets [19, 21, 47] provide structured demonstrations yet remain narrow and scripted. Large-scale egocentric datasets [15, 37] capture diverse daily activities with language annotations but often contain noisy trajectories and unclear temporal boundaries of purposeful interaction stages. We address these limitations by curating the *EgoMAN dataset*, a stage-aware dataset that consolidates real-world egocentric datasets with structured interaction annotations, aligning with recent efforts in robot learning from human videos [1, 6, 8, 19, 21, 50].

Vision-Language Models for Embodied AI. Modern vision-language models (VLMs) unify perception and language [2, 12, 24, 30, 38, 44, 46], enabling strong multimodal understanding and reasoning [51]. Their extensions to Vision-Language-Action (VLA) systems [22, 33, 54] support embodied tasks such as manipulation and navigation, and recent works incorporate hand trajectory prediction via pre-training or co-training [26, 49]. However, generating smooth, high-frequency trajectories directly from VLM reasoning remains challenging. Existing approaches link VLMs to action modules via implicit feature routing [9, 42, 48] or reasoning-based interfaces [25, 27, 52], enabling VLM-guided smoother action generation but often relying on opaque routing or long reasoning chains that limit interpretability and efficiency. In contrast, we propose a compact Trajectory-Token Interface that connects interaction-structured reasoning to continuous 3D motion through four structured tokens, enabling efficient and interpretable trajectory generation.

3 EgoMAN Dataset

EgoMAN is a large-scale egocentric interaction dataset (300+ hrs, 1,500+ scenes, 220K+ 6DoF trajectories) built from Aria glasses [13] across EgoExo4D [15], Nymeria [37], and HOT3D-Aria [3]. 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 2s, 34.0% move over 20cm 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 [40] to extract interaction annotations for EgoExo4D and Nymeria. At each atomic action timestamp [15, 37], we crop a 5s clip and annotate two wrist-centric stages: *Approach*, where the hand moves toward the target manipulation region, and *Manipulation*, where the hand interacts with the object. Most trajectories follow this simple structure (approach + manipulation, or manipulation-only when starting in contact). The EgoMAN dataset provides 6DoF wrist trajectories (3D position + 6D rotation [53]) at 10 FPS. All trajectories are transformed into the camera coordinate frame of the final pre-interaction frame. To ensure annotation quality, we apply a four-stage curation pipeline that removes irrelevant segments and enforces physically plausible, clearly visible hand–object interactions involving a single unambiguous object. Implementation details are provided in the supplementary material.

EgoMAN QA. We generate structured question–answer pairs using GPT, covering *semantic* (21.6%), *spatial* (42.6%), and *motion* (35.8%) reasoning. *Semantic reasoning* questions capture high-level intent (e.g., next action, object, or purpose). *Spatial reasoning* questions ground intent in metric 3D space by querying the wrist state at key interaction stages, including approach onset, manipulation onset (approach completion), and manipulation end. *Motion reasoning* conditions these queries on a short preceding 6DoF trajectory, explicitly linking motion history to future semantic and spatial outcomes. Detailed construction procedures and examples are provided in the supplementary material.

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)

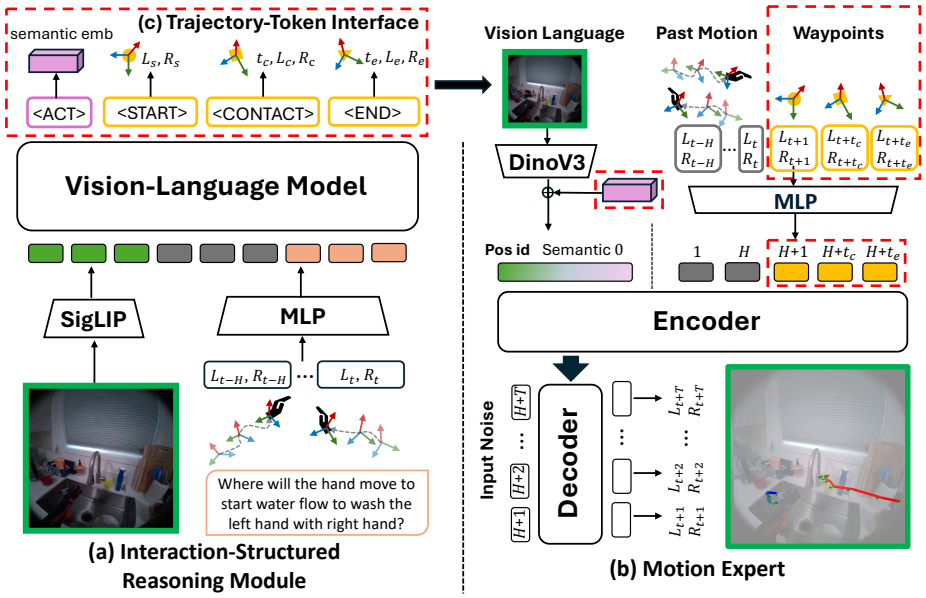


Fig. 2: Overview of EgoMAN. EgoMAN is a modular reasoning-to-motion framework that predicts future 6DoF hand trajectories from an egocentric RGB frame, past wrist motion, and a language intent. The **Interaction-Structured Reasoning Module** (a) extracts semantic and spatial cues to produce stage-aware trajectory tokens. The **Motion Expert** (b) generates future trajectories conditioned on these tokens, past motion, and visual input. The trajectory tokens form the **Trajectory-Token Interface** (c), explicitly bridging reasoning and motion generation.

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 Framework

The **EgoMAN** framework consists of three components (Fig. 2): (a) an **Interaction-Structured Reasoning** module that reasons over semantics, spatial context, and motion to produce stage-aware trajectory waypoints; (b) a **Motion Expert** that generates 6DoF hand trajectories; and (c) a **Trajectory-Token Interface** that bridges reasoning and motion. We first formalize the task in Sec. 4.1, then describe the Interaction-Structured Reasoning (Sec. 4.2), the Motion Expert (Sec. 4.3), and their integration via the Trajectory-Token Interface (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. *EgoMAN* acts as the function $\mathcal{F}$ that maps the inputs to future trajectories:

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

4.2 Interaction-Structured Reasoning

To predict accurate hand trajectories aligned with human intent and scene context, we need to understand both spatial relationships in the environment and the intended action. Therefore, the first module of EgoMAN is a reasoning model that aligns spatial perception and motion reasoning with task intent and interaction stages of hand–object manipulation. Built on Qwen2.5-VL [2], 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 trajectory 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 trajectory tokens,`<START>`, `<CONTACT>`, and `<END>` denote the approach onset, manipulation onset (*i.e.*, approach completion), and manipulation completion, respectively. Each trajectory token is equipped with a lightweight head that predicts a timestamp, 3D wrist positions, and 6D wrist rotations.

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 trajectory 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 [43] 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 [39]:

$$\mathcal{L}_{\text{act}} = \begin{cases} 1 - \frac{1}{K}\sum_{i=1}^{K} \text{sim}(z_i, z_i^+), & K < \kappa, \\ -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(z_i, z_j^+)/\tau)}, & K \geq \kappa. \end{cases} \tag{2}$$

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 trajectory token is supervised with Huber losses weighted by Gaussian time windows:

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

We use the continuous 6D rotation parameterization [53] 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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \lambda_{\text{wp}}\mathcal{L}_{\text{wp}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}}, \tag{4}$$

where the λ terms weight each loss component.

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) [29] 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 trajectory 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, trajectory 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 [45] visual features are added as 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:

$$\mathcal{L}_{\text{FM}} = \|\hat{v} - (x_1 - x_0)\|_2^2 . \tag{5}$$

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

$$x_{k+1} = x_k + \Delta t \cdot \hat{v}(x_k, t_k), \quad \Delta t = \tfrac{1}{N}, \tag{6}$$

to obtain future wrist trajectories.

Motion Pre-Training. We found joint training of reasoning and the Motion Expert is unstable due to mismatched learning objectives. To address this, we pre-trained 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.

4.4 Trajectory-Token Interface

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 the Motion Expert was pretrained to consume ground-truth waypoints and action phrases. At inference, however, the Motion Expert must consume the predicted 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 framework, the Reasoning Module receives the past wrist motion and an intent query, and predicts a 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, 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 introduced in Sec. 4.3. This unified setup enables efficient intent reasoning and produces physically consistent trajectories aligned with intent semantics.

Design Discussion. The Trajectory-Token Interface bridges interaction reasoning and motion generation by distilling interaction dynamics into a small set of four tokens that serve as spatiotemporal anchors. Supervising these anchors with key interaction stages provides the Motion Expert with a structural prior for trajectory generation. Rather than enforcing strict segmentation, the anchors act as soft guidance, improving robustness to spatiotemporal deviations in predicted anchors and annotation noise. This compact interface efficiently transfers semantic intent into precise 3D trajectory forecasts while scaling to longer horizons and unseen scenarios. We validate this design through comprehensive experiments.

5 Experiments

We evaluate *EgoMAN* thoroughly on EgoMAN-Bench to answer three questions: (1)Does EgoMAN improve long-horizon 6DoF prediction, waypoint accuracy, and semantic alignment over prior methods? (2)How does Interaction-Structured Reasoning with the Trajectory-Token Interface enable stable intent-conditioned motion? (3)How do the progressive training strategy and structured interface support effective reasoning-to-motion transfer? We also present qualitative results demonstrating 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 using two localization metrics (in meters): **Contact Distance (Contact):** Euclidean distance between predicted and GT wrist positions at the contact timestamp. **Trajectory-Warp Distance (Traj):** Average Euclidean distance from each predicted waypoint to its nearest point on the GT trajectory.
Motion-to-Text Alignment Metrics. We evaluate semantic alignment between trajectories and action verbs by training a motion encoder to map hand trajectories to a pre-computed verb embedding space using a CLIP-style contrastive loss [16, 43]. We report **Recall@3** (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. Considering generative diversity, we report best-of-10 retrieval results.

5.2 Baselines

Hand Trajectory Predictors. We compare against three representative trajectory models: *USST** [5], *MMTwin** [34], and *HandsOnVLM** [4]. Baselines marked with (*) are adapted for fair comparison under the EgoMAN setting: a single RGB image, intent text embedding, and past motion are used to predict up to 5-second 6DoF bi-hand trajectories, evaluated over the GT duration. We additionally report two internal variants: *FM-Base* (Motion Expert only) and *EgoMAN-ACT* (w/o. Interaction-Structured Reasoning).
Affordance-based Methods. We also evaluate *HAMSTER** [27], *VRB** [1], and *VidBot* [10], adapted to our waypoint prediction protocol. All methods use the same RGB image, Metric3D depth [20], and verb–object text input, and predict contact and goal points aligned with EgoMAN's `<CONTACT>` and `<END>` waypoints. Additional implementation details and adaptation procedures are provided in the supplementary material.

Table 1: Results on EgoMAN-Bench. Lower is better. Best results are **bold**; second-best are underlined. *FM-Base* denotes the Motion Expert only variant, and *EgoMAN-ACT* removes Interaction-Structured Reasoning. EgoMAN reduces ADE by 27.5% over the strongest external baseline (HandsOnVLM) on EgoMAN-Unseen and achieves similar gains on HOT3D-OOD, demonstrating strong cross-domain generalization.

Method	ADE ↓	FDE ↓	DTW ↓	Rot (°) ↓	Dataset
USST* [5]	0.233	0.394	0.220	46.98	EgoMAN Unseen
MMTwin* [34]	0.206	0.256	0.204	48.98	
HandsOnVLM* [4]	0.171	0.228	0.161	35.22	
FM-base	0.160	0.229	0.144	37.00	
EgoMAN-ACT	0.141	0.204	0.127	35.03	
EgoMAN (Ours)	**0.124**	**0.179**	**0.111**	**32.75**	
USST* [5]	0.245	0.409	0.226	55.80	HOT3D OOD
MMTwin* [34]	0.209	0.259	0.207	44.37	
HandsOnVLM* [4]	0.194	0.262	0.186	38.13	
FM-base	0.161	0.237	0.147	39.47	
EgoMAN-ACT	0.153	0.228	0.141	38.42	
EgoMAN (Ours)	**0.141**	**0.217**	**0.130**	**35.09**	

5.3 Results

Trajectory Evaluation. Table 1 shows that the full *EgoMAN* model achieves the lowest ADE, FDE, DTW, and rotation errors, outperforming the strongest external baseline *HandsOnVLM** by 27.5% ADE on EgoMAN-Unseen and 27.3% on HOT3D-OOD. These gains indicate that explicitly modeling interaction structure and stage-aware reasoning substantially improves long-horizon 6DoF prediction and cross-domain generalization.

Motion Expert only (FM-base) already surpasses state-space and Mamba-based predictors (*USST**, *MMTwin**), showing that Flow Matching provides a strong foundation for physically consistent motion forecasting. Adding structured vision–language reasoning further improves performance. Although *HandsOnVLM** uses text guidance, its CVAE decoder relies on 50 implicit hand tokens learned from noisy data, lacking explicit stage-aware grounding and yielding higher 6DoF errors. In contrast, EgoMAN via compact Trajectory-Token Interface explicitly encodes interaction stages, enabling more stable and accurate reasoning-conditioned generation. All results use sampling budget K=10; results for K=1, 5 (in supplementary material) confirm consistent gains under reduced sampling.

Waypoints Evaluation. Table 3 evaluates the stage-aware trajectory tokens predicted by the Reasoning Module. By directly regressing 3D `<CONTACT>` and `<END>` wrist positions as structured interaction states, *EgoMAN* achieves the best performance. It reduces contact error from 0.29–0.34m to 0.19m and lowers DTW from 0.27–0.30m to 0.13m on EgoMAN-Unseen, while maintaining the lowest contact error on the challenging test set HOT3D-OOD. *EgoMAN* is also far more efficient, running at 3.45FPS (averaged over 50 samples on an NVIDIA PG509-

Table 2: Ablation on EgoMAN-Unseen (K=1). R: Interaction-Structured Reasoning; M: Motion Expert pretraining, WP: Trajectory-Token Interface. Lower is better; R+M with 6DoF WP performs best.

R	M	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 3: Waypoint prediction results. Lower is better for *Contact* and *Traj*; higher is better for *FPS*. *EgoMAN* achieves the best accuracy, improving Contact by 33.8% and Traj by 52.8%, and runs faster at 3.45 FPS.

Method	WP	FPS	Contact ↓	Traj ↓
HAMSTER*	2D	0.17	0.342	0.297
VRB*	3D	0.03	0.300	0.271
VidBot	3D	0.04	0.290	0.269
EgoMAN	3D	**3.45**	**0.192**	**0.127**

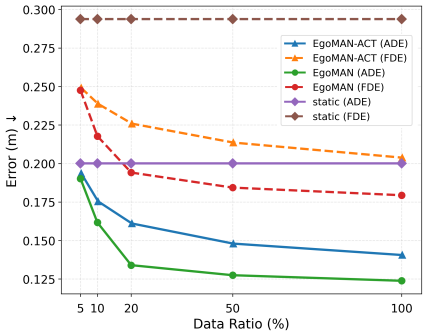


Fig. 3: Data efficiency. Best-of-10. *static*: repeats last position. EgoMAN-ACT degrades without pretraining, while EgoMAN remains robust at 20% data.

Table 4: Motion-to-Verb Retrieval. EgoMAN achieves the best alignment.

Method	R@3 ↑	FID ↓
USST*	15.0	0.22
MMTwin*	22.9	0.86
HandsOnVLM*	27.9	0.10
EgoMAN	**43.9**	**0.04**

210, 80GB) compared to < 0.05 FPS for *VRB** and *VidBot*, which require heavy detection and 3D post-processing. By predicting only four compact, stage-aware trajectory tokens, *EgoMAN* gains both speed and robustness.

Motion-to-Text Alignment. Table 4 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 lack structured semantic grounding, resulting in weaker verb alignment. *HandsOnVLM* benefits from language conditioning yet suffers from noisy CVAE decoding without explicit stage-aware constraints. These results highlight the importance of explicitly structured reasoning-to-motion transfer for intent-aligned 6DoF trajectory generation.

5.4 Ablation Study

As shown in Table 2, we ablate Interaction-Structured Reasoning (R), Motion Expert pretraining (M), and the Trajectory-Token Interface (WP). Overall, motion pretraining provides the strongest standalone gain, while Interaction-Structured Reasoning and the stage-aware Trajectory-Token Interface offer complementary improvements essential for accurate and data-efficient 6DoF trajectory prediction. In particular, effective reasoning transfer critically depends on the proposed

Trajectory-Token Interface; without it, performance approaches the no-reasoning variant. Together, these components are jointly necessary to achieve stable and high-precision intent-conditioned trajectory generation.

Interaction-Structured Reasoning (R). Removing Interaction-Structured Reasoning (*EgoMAN-ACT*) degrades performance (ADE 0.162→0.215), highlighting the importance of explicitly modeling interaction stages. Interaction-structured reasoning is pretrained on large-scale QA corpora that organizes intent through stage-aware progression (e.g., approach and manipulation). This structured reasoning reduces ambiguity by constraining motion generation to semantically meaningful interaction phases rather than treating trajectories as flat continuous signals. Learned trajectory tokens further align stage transitions with visual context and intent semantics. Data-efficiency results (Fig. 3) reinforce this effect: at 20% of training data, *EgoMAN* (ADE $\sim$ 0.13m) remains superior to *EgoMAN-ACT* (ADE $\sim$ 0.16m), and the gap persists even at full data (ADE 0.140→0.125). FDE follows a similar trend. These results indicate that interaction-structured reasoning reduces intent ambiguity early, enabling more stable and generalizable trajectory prediction even with fewer high-quality labels.

Trajectory-Token Interface (WP). Trajectory-Token Interface serves as the explicit structured interface between the reasoning module and the 6DoF Motion Expert. Conditioning on predicted stage-aware waypoints improves both accuracy and stability (ADE 0.161→0.151, DTW 0.147→0.137, Rot 35.90°→33.88°). Without WP, performance drops near that of *EgoMAN-ACT*, indicating that Interaction-Structured Reasoning is effective only when grounded through the Trajectory-Token Interface rather than implicit embeddings. Replacing the 6DoF trajectory tokens with the decoder's final hidden state (Emb) causes only minor differences (FDE $\sim$0.4cm worse), suggesting that while 6DoF waypoints perform best, results are generally stable across formats. The key factor is the stage-aware interaction structure encoded by the trajectory tokens.

Motion Expert (M). Removing Motion Expert pretraining while retaining reasoning pretraining leads to clear degradation (ADE 0.150→0.215). Without Motion Expert pretraining, the model must jointly learn semantic reasoning and motion dynamics, leading to noisier long-horizon trajectories, and unstable optimization from competing learning objectives. Removing both *M* pretraining and reasoning causes further decline (ADE 0.273, Rot 51.79°). While reasoning with trajectory tokens partially mitigate this effect, pretraining the Motion Expert on physically plausible dynamics followed by joint fine-tuning with Interaction-Structured reasoning yields the strongest performance across all metrics.

5.5 Qualitative Analysis

We visualize best-of-K=10 forecasts in Fig. 4 on EgoMAN-Bench. Left: stage-aware waypoints from affordance baselines (*VRB**, *VidBot*) often miss target surfaces or collapse toward the hand, while *EgoMAN* predicts contact and end states closely aligned with GT. Right: full 6DoF trajectories show that EgoMAN generates smoother, stage-consistent motions with correct wrist orientation, whereas baselines underreach or drift in cluttered and unseen scenes. Fig. 5a

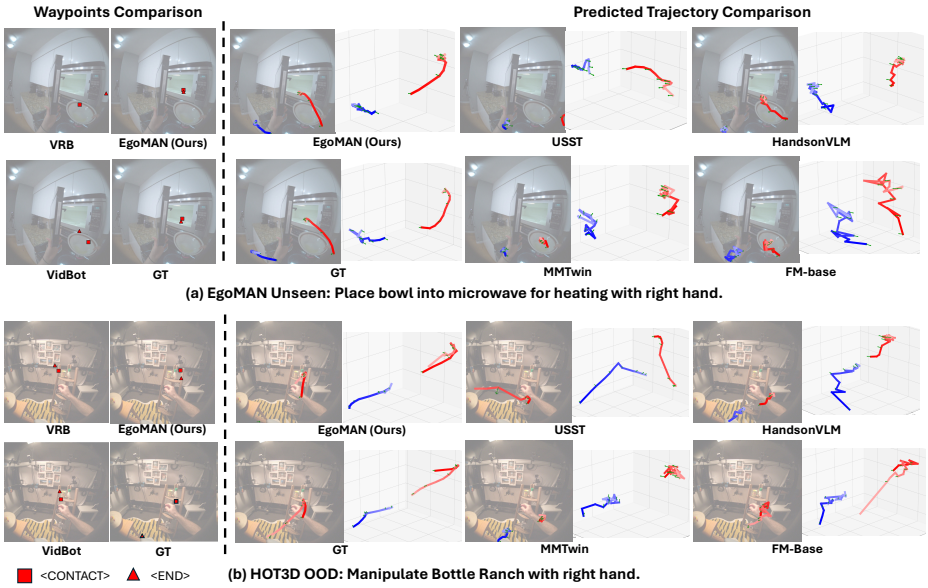
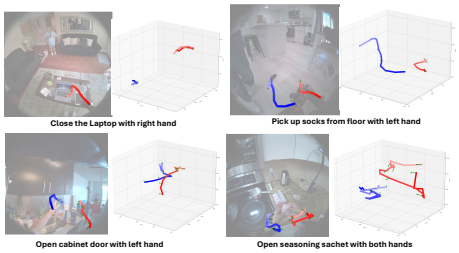


Fig. 4: Qualitative comparisons on EgoMAN-Bench. We visualize best-of-K=10 predictions for waypoints and full trajectories. Left: `<CONTACT>` and `<END>` waypoint predictions compared with *VRB** and *VidBot*. Right: 3D hand trajectory forecasts and 2D projections compared with prior baselines. Our *EgoMAN* model produces the smoothest and closest results to ground truth.

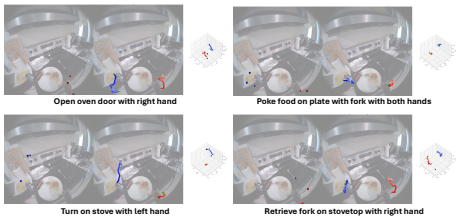
highlights diverse activities (e.g., closing a laptop, picking up socks, opening a cabinet, and opening a seasoning sachet), demonstrating that Interaction-Structured Reasoning produces trajectories consistent with both verb semantics and spatial context. Figure 5b shows intent-conditioned generation under identical visual and motion inputs. Given different intent queries, *EgoMAN* predicts distinct stage-aware waypoints and produces corresponding valid 6DoF trajectories, enabling controllable and diverse motion generation.

6 Conclusion

We presented *EgoMAN*, a unified framework for interaction-structured 3D hand trajectory prediction in egocentric settings. EgoMAN organizes motion into interaction stages and introduces a compact Trajectory-Token Interface that encodes these stages using a minimal set of tokens as sparse motion anchors, bridging intent-driven reasoning with continuous 6DoF trajectory generation. We further proposed a progressive training strategy that aligns high-level reasoning with physically grounded motion. To support this formulation, we also proposed the EgoMAN dataset, a large-scale stage-aware egocentric dataset with semantic, spatial, and motion reasoning supervision. Together, they improve trajectory



(a) Qualitative results of diverse activities. EgoMAN generates accurate 6DoF hand trajectories for diverse activities, aligning motion with intent description and scene spatial structure.



(b) Multiple intents. With the same image and past motion, EgoMAN produces distinct 6DoF trajectories for different intent queries, showing controllable intent-to-motion generation.

Fig. 5: Qualitative visualization of EgoMAN.

accuracy and generalization in the cross-domain over existing baselines while remaining efficient. Overall, *EgoMAN* establishes interaction-structured reasoning as a principled approach to egocentric 3D hand motion prediction.

Limitations and Future Directions. EgoMAN models wrist-level motion with coarse interaction stages and does not capture fine-grained dexterous behaviors. Extending to full hand articulation, richer contact semantics, and evaluating transfer to AR and robotic manipulation systems are important future directions.

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. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: CVPR (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., et al.: Introducing hot3d: An egocentric dataset for 3d hand and object tracking. arXiv preprint arXiv:2406.09598 (2024)
4. Bao, C., Xu, J., Wang, X., Gupta, A., Bharadhwaj, H.: Handsonvlm: Vision-language models for hand-object interaction prediction. Transactions on Machine Learning Research (2025)

5. Bao, W., Chen, L., Zeng, L., Li, Z., Xu, Y., Yuan, J., Kong, Y.: Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In: International Conference on Computer Vision (ICCV) (2023)
6. Bharadhwaj, H., Dwibedi, D., Gupta, A., Tulsiani, S., Doersch, C., Xiao, T., Shah, D., Xia, F., Sadigh, D., Kirmani, S.: Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. Conference on Robot Learning (2025)
7. Bharadhwaj, H., Gupta, A., Kumar, V., Tulsiani, S.: Towards generalizable zero-shot manipulation via translating human interaction plans. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6904–6911. IEEE (2024)
8. Bharadhwaj, H., Mottaghi, R., Gupta, A., Tulsiani, S.: Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In: ECCV (2024)
9. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Smith, L., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0: A vision–language–action flow model for general robot control. Robotics: Science and Systems XXI (2025)
10. Chen, H., Sun, B., Zhang, A., Pollefeys, M., Leutenegger, S.: VidBot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In: CVPR (2025)
11. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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. pp. 24185–24198 (2024)
12. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: an embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
13. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023)
14. Grattafiori, A., et al.: The llama 3 herd of models (2024)
15. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., 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. pp. 19383–19400 (2024)
16. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
17. Hatano, M., Hachiuma, R., Saito, H.: Emag: Ego-motion aware and generalizable 2d hand forecasting from egocentric videos. In: European Conference on Computer Vision Workshops (ECCVW) (2024)
18. Hatano, M., Zhu, Z., Saito, H., Damen, D.: The invisible egohand: 3d hand forecasting through egobody pose estimation. arXiv preprint arXiv:2504.08654 (2025)
19. Hoque, R., Huang, P., Yoon, D.J., Sivapurapu, M., Zhang, J.: Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709 (2025)

20. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: 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)
21. Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., Hoffman, J., Xu, D.: Egomimic: Scaling imitation learning via egocentric video. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 13226–13233. IEEE (2025)
22. Kim, M., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
23. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10138–10148 (October 2021)
24. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
25. Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y.R., Lee, S., et al.: Molmoact: Action reasoning models that can reason in space. arXiv preprint arXiv:2508.07917 (2025)
26. Li, Q., Deng, Y., Liang, Y., Luo, L., Zhou, L., Yao, C., Zeng, L., Feng, Z., Liang, H., Xu, S., et al.: Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. arXiv preprint arXiv:2510.21571 (2025)
27. Li, Y., Deng, Y., Zhang, J., Jang, J., Memmel, M., Yu, R., Garrett, C.R., Ramos, F., Fox, D., Li, A., et al.: Hamster: Hierarchical action models for open-world robot manipulation. arXiv preprint arXiv:2502.05485 (2025)
28. Li, Y., Cao, Z., Liang, A., Liang, B., Chen, L., Zhao, H., Feng, C.: Egocentric prediction of action target in 3d. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
29. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
31. Liu, M., Tang, S., Li, Y., Rehg, J.M.: 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. pp. 704–721. Springer (2020)
32. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: 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)
33. Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., Lu, Z.: Being-h0: vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597 (2025)
34. Ma, J., Bao, W., Xu, J., Sun, G., Chen, X., Wang, H.: Novel diffusion models for multimodal 3d hand trajectory prediction. 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 2408–2415 (2025)

35. Ma, J., Chen, X., Bao, W., Xu, J., Wang, H.: Madiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos (2024)
36. Ma, J., Xu, J., Chen, X., Wang, H.: Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 4291–4298 (2024)
37. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., Bailey, K., Fosas, D.S., Liu, C.K., Liu, Z., Engel, J., Nardi, R.D., Newcombe, R.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: the 18th European Conference on Computer Vision (ECCV) (2024)
38. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
39. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
40. OpenAI, et al.: Gpt-4 technical report (2024)
41. Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., Levine, S.: Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747 (2025)
42. Physical Intelligence, Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: A vision–language–action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
44. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
45. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025)
46. Team, G., et al.: Gemma 3 technical report (2025)
47. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., Joshi, N., Pollefeys, M.: 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). pp. 20270–20281 (October 2023)
48. Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., Feng, F.: Dexvla: Vision-language model with plug-in diffusion expert for general robot control. arXiv preprint arXiv:2502.05855 (2025)
49. Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.C., Zou, X., Fang, Y., Cheng, X., Qiu, R.Z., et al.: Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440 (2025)
50. Yuan, C., Wen, C., Zhang, T., Gao, Y.: General flow as foundation affordance for scalable robot learning. arXiv preprint arXiv:2401.11439 (2024)
51. Zhang, H., Chen, M., He, S., Xu, Z., Shen, Y., Huang, Y., Lu, J., Li, Y., She, Y., Fu, Y.: A survey of physical ai: A history from chatgpt to world models and embodied

agents. Preprints (June 2026). https://doi.org/10.20944/preprints202606.0173.v1, https://doi.org/10.20944/preprints202606.0173.v1
52. Zhang, W., Wang, M., Liu, G., Xu, H., Jiang, Y., Shen, Y., Hou, G., Zheng, Z., Zhang, H., Li, X., Lu, W., Li, P., Zhuang, Y.: Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. arXiv preprint arXiv:2503.21696 (2025)
53. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019)
54. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)