COORDMOTION: LEARNING PREDICTIVE EYE-BODY COORDINATION STATES FOR HUMAN MOTION FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human motion forecasting is an important topic in the areas of computer vision, robotics, and human-aware artificial intelligence. Prior human behavioral studies have shown that eye–body coordination provides informative structure for understanding actions, intentions, and task progression. However, this relationship has not been explicitly explored for human motion forecasting. We present **CoordMotion**, a novel method that learns eye–body coordination as structured predictive behavioral states to forecast human motion. Specifically, (1) we learn structured eye–body coordination states that explicitly organize task-dependent eye–body action transitions from complementary semantic text and physical value; (2) we forecast the evolution of the learned coordination states and use the predicted coordination states to guide deterministic base prediction and horizon-dependent residual diffusion; and (3) we introduce a causal Coordination-Guided Verifier that leverages predicted coordination states to rank and select motion candidates, enabling coordination-aware test-time selection from multiple stochastic samples. We conduct extensive experiments on both deterministic and stochastic human motion forecasting tasks using the EE4D-Motion and Nymeria datasets. Experimental results show that our method outperforms state-of-the-art deterministic forecasting methods, reducing the mean per-joint position error by up to 9.7%, and produces higher-quality stochastic futures, reducing the average displacement error on EE4D-Motion by 20.2%. These results validate the significant potential of structured eye–body coordination states for human motion forecasting.

1 INTRODUCTION

Human motion forecasting aims to predict future human motion from past observations, a capability that is fundamental to computer vision, robotics, and human-aware artificial intelligence. In common human activities (Pelz et al., 2001), the human body exhibits task-dependent eye–body coordination amid continuously changing motions, and such coordination evolves dynamically as actions proceed. For instance, gaze may anticipate an upcoming motor response (Pelz et al., 2001), monitor task-critical events during object manipulation (Johansson et al., 2001), or shift toward the next task object before the current action is completed (Land & Hayhoe, 2001a).

Recent methods exploit gaze to infer intent or reduce uncertainty in motion forecasting, including SIF3D (Lou et al., 2024), GIMO (Zheng et al., 2022), and GazeMoDiff (Yan et al., 2023). UniEgoMotion (Patel et al., 2025) and Ego3DLM (Bae et al., 2026) further incorporate egocentric observations and scene context to predict future motion. Nevertheless, using gaze, motion, and scene observations as conditioning signals does not by itself capture how their relationships evolve with the underlying task. As illustrated in Fig. 1 (c), existing frameworks typically treat these signals as direct conditioning signals. This leaves a missing link between multimodal observations and the latent behavioral structure that organizes task-dependent eye–body coordination.

Building on this behavioral insight, we present **CoordMotion** to learn eye–body coordination as structured predictive behavioral states for human motion forecasting. Specifically, we first model eye–body coordination from two complementary perspectives. Explicitly, we generate eye–body coordination semantic text based on atomic-action descriptions. In parallel, we implicitly encode

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

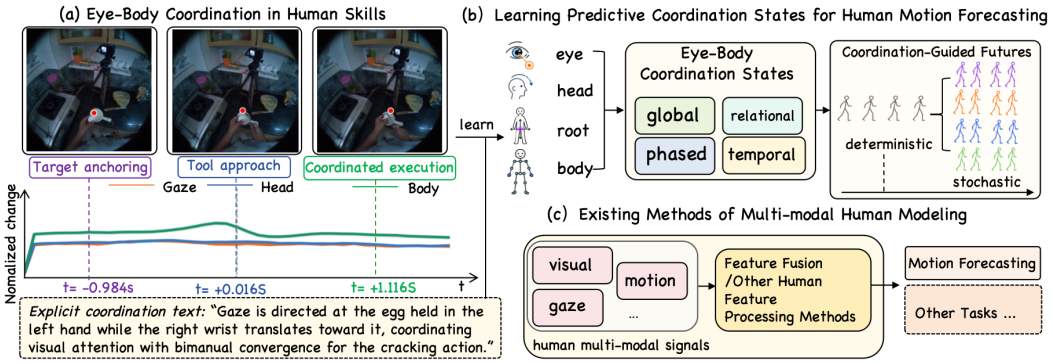


Figure 1: **Motivation for learning eye–body coordination states for human motion forecasting.** (a) An example from Ego-Exo4D dataset illustrates how gaze, head, and body evolve jointly across action phases. (b) We organize eye–body relation to predict structured coordination states that support motion forecasting. (c) Existing multimodal human modeling methods typically process different modalities independently.

fine-grained physical value of coordination from observed eye and body displacement and motion dynamics that are difficult to articulate precisely in language. Based on these complementary representations, we align these complementary cues to infer the current eye–body coordination states and forecast how they evolve across future action phases. As shown in Fig. 1 (b), we predict four complementary coordination representations: global, relational, phased and temporal eye–body coordination states.

Based on these predicted coordination states, CoordMotion improves human motion forecasting through two key stages: future motion forecasting and a verifier for stochastic candidates. Specifically, we first use the predicted coordination states to guide a hybrid deterministic–stochastic predictor, where a deterministic predictor captures the shared motion tendency, and residual diffusion generates diverse long-term variations. We then introduce a Coordination-Guided Verifier that ranks and selects the generated candidates according to coordination consistency, physical plausibility, and temporal continuity, retaining diverse yet plausible future motions.

We conduct experiments on the EE4D-Motion (Patel et al., 2025) and Nymeria (Ma et al., 2024) datasets. On EE4D-Motion, CoordMotion reduces deterministic $J_{0\text{–}2}$ by 9.7% over the strongest reproduced baseline and reduces expected stochastic ADE by 20.2% over 2–4 s and 19.0% over 4–6 s. On Nymeria, it improves JPE by 8.6% over Ego3DLM. These results demonstrate that predictive eye–body coordination states support accurate short-term forecasting and higher-quality long-term stochastic prediction across datasets. Our contributions are summarized as follows:

- We learn structured eye–body coordination states that capture task-dependent relations within atomic actions and across action transitions, using complementary semantic text and physical value.
- We forecast the evolution of the learned coordination states through global, relational, phased, and temporal state, and use these states to guide human motion forecasting.
- We use the predicted coordination states as a causal criterion for ranking and selecting motion candidates, improving stochastic future quality and plausibility.
- We extensively validate the effectiveness of the learned eye–body coordination states that improve both deterministic and stochastic forecasting on EE4D-Motion and Nymeria.

2 Related Work

2.1 Human Motion Forecasting

Human motion forecasting predicts future poses from observed motion. Early deterministic methods modeled temporal dynamics with recurrent architectures (Martinez et al., 2017), motion attention (Mao et al., 2021), graph convolutions (Sofianos et al., 2021), and MLP-based architec-

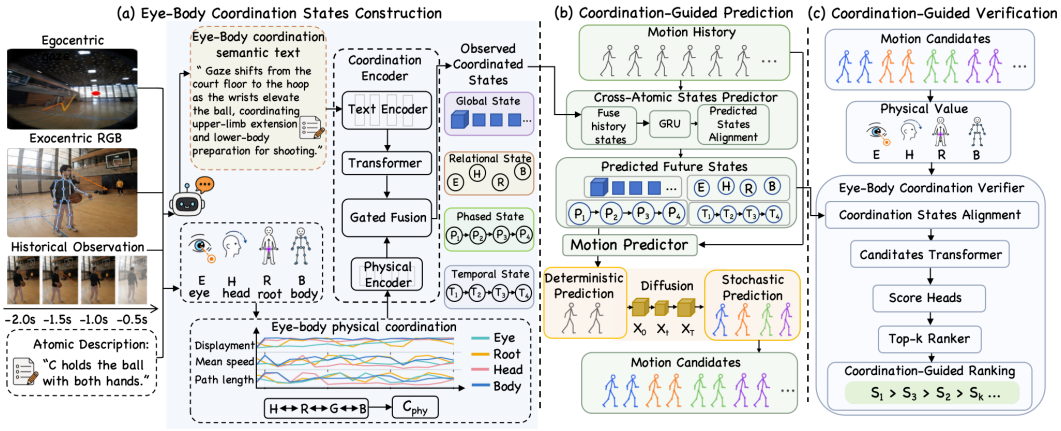


Figure 2: **Overview of CoordMotion.** (a) (Sec. 3.1) **Eye-Body Coordination States Construction**: We learn eye–body coordination states from complementary semantic text and physical value. (b) (Sec. 3.2) **Coordination-Guided Prediction**: We predict future coordination states to guide deterministic and stochastic motion forecasting. (c) (Sec. 3.3) **Coordination-Guided Verification**: We rank and select motion candidates according to their consistency with the predicted coordination states and physical constraints.

tures (Bouazizi et al., 2022). Because one motion history may admit multiple valid continuations, later methods generated diverse futures through masked motion completion (Chen et al., 2023) or diffusion-based forecasting (Yan et al., 2023; Lu et al., 2026).

Subsequent methods enriched motion forecasting with multimodal context. Yuan and Kitani (Yuan & Kitani, 2019) used first-person observations for future pose prediction. GIMO (Zheng et al., 2022) incorporated gaze and 3D scene context, while GazeMotion (Hu et al., 2024) combined predicted gaze with motion history. UniEgoMotion (Patel et al., 2025) conditioned motion diffusion on egocentric visual observations and wearer trajectories, and Ego3DLM (Bae et al., 2026) jointly forecast motion and language using explicit 3D scene context.

Although these methods enriched motion history with complementary observations, they primarily treated each modality as a direct prediction condition, without explicitly modeling the task-dependent coordination between eye and body motion. In contrast, we learn this eye–body coordination as structured predictive behavioral states, forecast their future progression, and use them to guide motion forecasting and stochastic candidate selection.

2.2 Eye–Body Coordination

Goal-directed human behavior relies on coordinated eye and body movements, through which visual attention anticipates and guides ongoing actions (Land & Hayhoe, 2001b). Gaze is coupled with subsequent hand motion (Johansson et al., 2001) and evolves jointly with head and body movements (Pelz et al., 2001). These relationships vary with task goals and motor demands (Land, 2006; Tatler et al., 2011), suggesting that eye–body coordination reflects both action execution and progression.

Computer-vision methods used gaze for future attention prediction (Zhang et al., 2017), task-dependent fixation modeling (Huang et al., 2018), and activity anticipation (Shen et al., 2018). For motion forecasting, GIMO (Zheng et al., 2022) treated gaze as an intent cue, GazeMotion (Hu et al., 2024) combined predicted gaze with motion history, and GazeMoDiff (Yan et al., 2023) conditioned stochastic generation on gaze.

However, existing work treats gaze as an auxiliary cue for motion forecasting, whereas we learn task-dependent eye–body coordination as predictive states, forecast states' evolution, use them to forecast future human motion and further use them to filter candidates that conflict with the predicted behavior.

3 METHODOLOGY

As illustrated in Fig. 2, CoordMotion consists of three stages. First, we construct eye–body coordination states from historical observations and atomic-action descriptions. Second, we predict the evolution of this state and use it to guide deterministic and stochastic motion prediction. Finally, we introduce a Coordination-Guided Verifier by evaluating the predicted motion candidates according to coordination consistency. The coordination states therefore connect historical observations, future motion forecasting, and candidate selection throughout the framework.

3.1 CONSTRUCTION OF EYE–BODY COORDINATION STATES

Given historical body motion $\mathbf{B}^h$, eye $\mathbf{E}^h$, visual features $\mathbf{V}^h$, device trajectory $\mathbf{T}^h$, and the current atomic action A_i, we construct the observed coordination states $\mathcal{C}_{\text{obs}}$.

Explicit and implicit modeling of eye–body coordination states. Atomic-action descriptions specify what a person does, but do not describe how eye and body motion jointly realize the action. We therefore use Qwen (Wang et al., 2024) offline to generate an evidence-based coordination description D_i for each atomic action. For two adjacent actions, we also generate a directed description $D_{i\to i+1}$ that characterizes the coordination change across their boundary. A frozen text encoder produces the coordination semantic representation:

$$\mathbf{e}_i = \text{norm}\left(E_{\text{text}}(D_i)\right), \qquad \mathbf{e}_{i\to i+1} = \text{norm}\left(E_{\text{text}}(D_{i\to i+1})\right). \tag{1}$$

Semantic text capture coordination meaning, whereas human motion reveals how each subject physically executes that coordination. For each gaze, head, root and body trajectory $\mathbf{x}_{j,1:T}$, we compute displacement, path length, and mean speed:

$$\Delta\mathbf{x}_j = \mathbf{x}_{j,T} - \mathbf{x}_{j,1}, \qquad L_j = \sum_{t=2}^{T} \left\|\mathbf{x}_{j,t} - \mathbf{x}_{j,t-1}\right\|_2, \qquad \bar{v}_j = \frac{L_j}{(T-1)\Delta t}. \tag{2}$$

After capturing response direction, path length, and execution rate. We project the resulting physical features into modality and joint tokens, encode them with a Transformer, and apply attention pooling:

$$\mathbf{H}_i^{\text{phys}} = \text{Tr}_{\text{phys}}(\mathbf{X}_i^{\text{phys}}), \qquad \mathbf{p}_i = \text{norm}\left(W_p \sum_n \alpha_{i,n}\mathbf{H}_{i,n}^{\text{phys}}\right), \quad \alpha_{i,n} = \text{softmax}_n(\mathbf{w}_a^\top \mathbf{H}_{i,n}^{\text{phys}}). \tag{3}$$

Observed eye–body coordination states. We fuse the action representation $\mathbf{a}_i$ and physical evidence $\mathbf{p}_i$, and align the fused representation with the semantic target $\mathbf{e}_i$:

$$\mathbf{z}_i = \text{norm}\left[\mathbf{a}_i + f_{\text{coord}}(\mathbf{a}_i, \mathbf{p}_i)\right], \qquad \mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{con}}(\mathbf{z}_i, \mathbf{e}_i). \tag{4}$$

We then project $\mathbf{z}_i$ into four complementary coordination representations:

$$\mathcal{C}_{\text{obs}} = \left\{\mathbf{c}_i^g, \mathbf{c}_i^r, \mathbf{c}_i^p, \mathbf{c}_i^t\right\} = \text{reshape}(W_f\mathbf{z}_i), \tag{5}$$

which encode the global behavioral context, eye–body relations, execution phase, and temporal evolution, respectively.

Since atomic-action description and generated descriptions are unavailable at deployment, we distill the pretrained representations into a causal encoder:

$$(\widehat{\mathbf{a}}_i, \widehat{\mathbf{p}}_i, \widehat{\mathbf{z}}_i) = E_{\text{causal}}(\mathbf{B}^h, \mathbf{G}^h, \mathbf{V}^h, \mathbf{T}^h), \qquad \widehat{\mathcal{C}}_{\text{obs}} = \text{reshape}(W_f\widehat{\mathbf{z}}_i). \tag{6}$$

3.2 COORDINATION-GUIDED MOTION PREDICTION

The observed coordination states describe the current execution, but future motion also depends on how this organization changes at action boundaries. We therefore predict future coordination states before human motion forecasting.

Cross-atomic state prediction. We fuse $(\widehat{\mathbf{a}}_i, \widehat{\mathbf{p}}_i, \widehat{\mathbf{z}}_i)$ to initialize a phase-conditioned GRU, which predicts $M = 16$ future events:

$$\mathbf{h}_m = \mathrm{GRU}(f_{\mathrm{in}}[\widehat{\mathbf{e}}_{m-1}, u_m], \mathbf{h}_{m-1}), \quad \widehat{\mathbf{e}}_m = \mathrm{norm}(W_e \mathbf{h}_m), \quad \widehat{\Delta t}_m = \mathrm{softplus}(W_t \mathbf{h}_m). \tag{7}$$

Event, continuation, time, and physical heads additionally produce continuation probabilities and six relational slots. Frame alignment maps these sparse predictions to $N_f = 60$ future positions and constructs

$$\widehat{\mathcal{C}}_{\mathrm{fut}} = \{\widehat{\mathbf{c}}^g, \widehat{\mathbf{C}}^r, \widehat{\mathbf{C}}^p, \widehat{\mathbf{C}}^t\}, \quad \begin{aligned} \widehat{\mathbf{c}}^g &= \mathrm{norm}\Big[\tfrac{1}{2}(\mathrm{Mean}\,\widehat{\mathbf{C}}^p + \mathrm{Mean}\,\widehat{\mathbf{C}}^r)\Big], \\ \widehat{\mathbf{C}}^t_n &= [\sum_m \omega_{n,m} W^e_f \widehat{\mathbf{e}}_m, \widehat{\mathbf{f}}^{\mathrm{imp}}_n]. \end{aligned} \tag{8}$$

Here $\widehat{\mathbf{C}}^p \in \mathbb{R}^{16\times d}$ contains event/phase tokens, $\widehat{\mathbf{C}}^r \in \mathbb{R}^{6\times d}$ contains physical relation slots, and $\widehat{\mathbf{C}}^t \in \mathbb{R}^{60\times d_t}$ is frame-aligned; predicted times and continuation probabilities determine $\omega_{n,m}$. The planner objective is

$$\mathcal{L}_{\mathrm{plan}} = \mathcal{L}_{\mathrm{event}} + 0.2\mathcal{L}_{\mathrm{continue}} + 0.5\mathcal{L}_{\mathrm{time}} + 0.5\mathcal{L}_{\mathrm{implicit}}. \tag{9}$$

We next forecast future coordination states and use them as the behavioral bridge from observed execution to future motion.

Deterministic prediction. A history encoder converts body states and velocity, visual features, device trajectory, and $\widehat{\mathcal{C}}_{\mathrm{obs}}$ into causal memory $\mathbf{M}$. Future queries attend to $\mathbf{M}$ while gated, frame-aligned explicit and implicit tokens from $\widehat{\mathcal{C}}_{\mathrm{fut}}$ condition each prediction time:

$$\mathbf{M} = E_{\mathrm{hist}}\left(\mathbf{B}^h, \mathbf{V}^h, \mathbf{T}^h, \widehat{\mathcal{C}}_{\mathrm{obs}}\right), \qquad \boldsymbol{\mu}_{0:6} = \mathrm{Dec}_{\mathrm{RL}}\left(D_{\mathrm{mot}}\left(\mathbf{Q} + \phi(t) + \mathrm{Gate}(\widehat{\mathcal{C}}_{\mathrm{fut}}), \mathbf{M}\right), \mathbf{B}^h_{-1}\right), \tag{10}$$

where $\mathrm{Dec}_{\mathrm{RL}}$ restores world coordinates from root-local corrections, yielding the full-horizon deterministic center $\boldsymbol{\mu}_{0:6}$. We train the center directly against all 60 future frames:

$$\mathcal{L}_{\mathrm{det}} = \mathcal{L}_{\mathrm{MPJPE}} + 0.25\mathcal{L}_{0:2} + 0.2\mathcal{L}_{\mathrm{vel}} + 0.1\mathcal{L}_{\mathrm{root}} + 0.1\mathcal{L}_{\mathrm{local}}, \tag{11}$$

where all terms use the future-validity mask, $\mathcal{L}_{0:2}$ emphasizes the deterministic prefix, and $\mathcal{L}_{\mathrm{local}}$ is computed in the root-relative body frame.

Stochastic prediction. Uncertainty grows with the forecasting horizon, but sampling the entire future can corrupt the reliable near-term prediction. We therefore retain $\boldsymbol{\mu}_{0:2}$ and diffuse only the root-local residual over 2–6 s:

$$\widetilde{\mathbf{R}}_0 = \mathbf{g} \odot \mathrm{RootLocal}\left(\mathbf{Y}_{2:6} \ominus \boldsymbol{\mu}_{2:6}\right), \qquad \widetilde{\mathbf{R}}_t = \sqrt{\bar{\alpha}_t}\widetilde{\mathbf{R}}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{12}$$

where the short cosine ramp $\mathbf{g}$ preserves continuity at the deterministic–stochastic boundary. A v-prediction Transformer conditions on the center, causal memory, and predicted future coordination states:

$$\widehat{\mathbf{v}}_t = D_\theta\left(\widetilde{\mathbf{R}}_t, t, \boldsymbol{\mu}_{2:6}, \mathbf{M}, \widehat{\mathcal{C}}_{\mathrm{fut}}\right), \qquad \mathcal{L}_{\mathrm{diff}} = \mathbb{E}\,\|\widehat{\mathbf{v}}_t - \mathbf{v}_t\|_2^2. \tag{13}$$

Reverse DDIM sampling generates $K = 16$ motion candidates:

$$\widehat{\mathbf{Y}}^{(k)}_{0:2} = \boldsymbol{\mu}_{0:2}, \qquad \widehat{\mathbf{Y}}^{(k)}_{2:6} = \boldsymbol{\mu}_{2:6} \oplus \widehat{\widetilde{\mathbf{R}}}^{(k)}, \quad k = 1, \ldots, 16. \tag{14}$$

Thus, the candidates share a stable near future but express different coordination-consistent long-term outcomes.

3.3 Coordination-Guided Verification

Stochastic sampling broadens coverage but can produce futures inconsistent with the predicted coordination evolution. We therefore use the same four future states to causally re-rank the candidates. For candidate k, a history-conditioned temporal decoder encodes root-local motion and its first difference over the 2–6 s suffix:

$$\mathbf{H}_k = E_{\mathrm{cand}}(W_m[\mathbf{x}_{k,1:40}, \Delta\mathbf{x}_{k,1:40}] + \boldsymbol{\pi}, \mathbf{M}). \tag{15}$$

Table 1: **Deterministic motion forecasting results on EE4D-Motion and Nymeria.** EE4D-Motion uses 2 s of history to predict 2 s, whereas Nymeria uses a 5 s-to-5 s protocol; both datasets are evaluated over the first 2 s of the future. Lower is better. † denotes results reproduced by us. Best results are shown in **bold**, and second-best results are underlined.

	EE4D-Motion						Nymeria					
Method	$J_{0\text{-}2}\downarrow$	$J_{0.2}\downarrow$	$J_{0.5}\downarrow$	$J_{1.0}\downarrow$	$J_{1.5}\downarrow$	$J_{2.0}\downarrow$	$J_{0\text{-}2}\downarrow$	$J_{0.2}\downarrow$	$J_{0.5}\downarrow$	$J_{1.0}\downarrow$	$J_{1.5}\downarrow$	$J_{2.0}\downarrow$
Zero velocity	14.767	4.291	9.488	16.146	21.281	24.608	16.886	4.174	9.611	16.982	23.285	28.676
HisRep (Mao et al., 2020)†	11.983	3.408	7.432	12.833	17.426	20.795	12.806	3.426	7.443	12.602	17.460	22.234
PGBIG (Ma et al., 2022)†	12.701	4.757	8.334	13.885	17.713	20.863	13.743	8.053	9.481	13.473	16.700	20.601
Transformer†	12.757	3.491	7.294	12.081	16.172	19.396	19.562	12.321	14.358	18.862	23.495	28.527
PhysMoP (Zhang et al., 2024)†	12.168	3.382	7.019	12.334	17.602	21.697	13.423	**2.833**	7.680	13.499	18.416	23.304
EgoCast (Escobar et al., 2025)†	16.266	9.857	11.952	16.195	20.124	23.217	13.702	6.522	8.845	13.057	17.429	22.451
UniEgoMotion (Patel et al., 2025)	21.303	12.547	15.825	21.468	26.399	30.120	–	–	–	–	–	–
UniEgoMotion (Patel et al., 2025)†	21.359	11.389	14.257	20.013	26.870	34.404	15.133	5.076	8.551	14.654	20.576	25.831
Ours	**10.820**	**3.272**	**6.748**	**11.429**	**15.648**	**18.987**	**11.583**	4.034	**6.638**	**10.941**	**15.515**	**20.488**

To preserve their distinct structure, Global, Relational, Phase, and Temporal states enter four independent cross-attention paths:

$$\mathbf{f}_k^v = \mathrm{Pool}\Big[\mathrm{CA}_v(\mathrm{Pool}_{L_v}(\mathbf{H}_k), \widehat{\mathbf{C}}^v_{\mathrm{fut}})\Big], qquadv \in \{g, r, p, t\}, \quad L_v \in \{1, 6, 16, 40\}. \tag{16}$$

We fuse these aligned features with the candidate motion feature, then use a two-layer candidate-set encoder to compare all proposals jointly:

$$\mathbf{f}_k = F_{\mathrm{fuse}}[\mathrm{Pool}(\mathbf{H}_k), \mathbf{f}_k^g, \mathbf{f}_k^r, \mathbf{f}_k^p, \mathbf{f}_k^t], \qquad \{\bar{\mathbf{f}}_k\}_{k=1}^{16} = E_{\mathrm{set}}(\{\mathbf{f}_k\}_{k=1}^{16}). \tag{17}$$

Quality, physical-plausibility, and coordination-consistency heads give

$$q_k = F_Q(\bar{\mathbf{f}}_k), \quad p_k = F_P(\bar{\mathbf{f}}_k), \quad c_k = F_C(\bar{\mathbf{f}}_k), \qquad s_k = q_k + \log\sigma(p_k) + \log\sigma(c_k). \tag{18}$$

Ground-truth futures provide training targets only. With candidate error ε_k, physical label y_k^P, coordination label y_k^C, and $\pi_k^\star = \mathrm{softmax}(-\varepsilon_k/\tau)$, we optimize

$$\mathcal{L}_{\mathrm{ver}} = -\sum_k \pi_k^\star \log \mathrm{softmax}(\mathbf{s}/\tau)_k + 5\sum_k \mathrm{softmax}(\mathbf{s}/\tau)_k \varepsilon_k + 0.25\,\mathrm{BCE}(\mathbf{p}, \mathbf{y}^P) + 0.25\,\mathrm{BCE}(\mathbf{c}, \mathbf{y}^C). \tag{19}$$

At inference, only generated motions, history memory, and predicted states are available; the verifier retains the eight highest-scoring candidates:

$$\mathcal{S}_{\mathrm{retained}} = \mathrm{TopK}(\{s_k\}_{k=1}^{16}, 8). \tag{20}$$

Hence, verification improves retained-set quality without observing the future ground truth.

4 Experiments

We describe the experimental setup in Sec. 4.1, report the main forecasting results in Sec. 4.2, and analyze the contributions of the coordination states in Sec. 4.3.

4.1 Experimental Setup

Datasets. We evaluate CoordMotion on two large-scale human-motion datasets: **EE4D-Motion** (Patel et al., 2025) is derived from Ego-Exo4D and provides synchronized egocentric video, device trajectories, and body motion across diverse procedural activities. **Nymeria** (Ma et al., 2024) records long-horizon daily activities with multimodal egocentric sensing and synchronized body motion, offering diverse subjects, environments, and skill executions. Together, they support evaluation across both structured task-oriented activities and unconstrained everyday human behavior.

Table 2: Stochastic long-term motion forecasting results on EE4D-Motion (Patel et al., 2025) over future 2–6 s. † denotes our input-matched implementation or adaptation on EE4D-Motion. All errors are reported in centimeters; lower is better except APD. Best results are shown in **bold**, and second-best results are underlined.

Method	$\mathrm{ADE}_{2\text{–}4} \downarrow$	$\mathrm{ADE}_{4\text{–}6} \downarrow$	$\mathrm{ADE}_{2\text{–}6} \downarrow$	$\min \mathrm{ADE}_{4s} \downarrow$	$\min \mathrm{ADE}_{6s} \downarrow$	FDE ↓	APD ↑
Residual diffusion†	33.914	39.649	36.707	23.419	26.704	24.449	30.450
SkeletonDiffusion (Curreli et al., 2025)†	35.118	42.558	38.661	22.806	26.337	24.403	**32.330**
UniEgoMotion (Patel et al., 2025)	34.175	40.460	37.318	22.736	26.219	27.389	27.391
UniEgoMotion (Patel et al., 2025)†	34.131	45.639	39.885	21.923	27.925	28.569	31.044
Ours w/o verifier	28.842	34.915	31.680	**21.200**	**24.239**	**22.544**	20.912
Ours	**27.060**	**32.119**	**29.560**	21.423	24.776	23.045	15.464

Table 3: Motion forecasting results on Nymeria under the 5-second-history to 5-second-future setting. All positional errors are reported in millimeters; lower is better except Diversity.

Method	APE↓	JPE↓	$\mathrm{ADE}_{0\text{–}5} \downarrow$	$\mathrm{ADE}_{2s} \downarrow$	FDE↓	$\mathrm{FDE}_{2s} \downarrow$	Diversity↑
FIction (Ashutosh et al., 2025)	206.2	564.7	558.3	416.7	904.3	494.9	0.7432
EgoLM (Hong et al., 2025) (GT motion)	168.8	583.9	540.0	299.0	1059.5	494.8	0.0055
EgoLM (Hong et al., 2025) (instruction tuning)	184.9	579.4	552.6	329.9	983.6	519.2	0.6142
UniEgoMotion (Patel et al., 2025)†	192.7	386.5	366.0	188.4	660.4	295.1	**1.3048**
Ego3DLM (Bae et al., 2026)	**147.9**	364.5	343.9	205.9	648.1	312.6	1.0624
Ours w/o verifier	164.3	333.5	**320.8**	**180.8**	**560.7**	**261.7**	1.0340
Ours	163.6	**333.2**	321.9	181.5	561.0	263.2	1.0379

Baselines. We compare with representative deterministic and stochastic forecasting paradigms. Deterministic baselines include zero-velocity extrapolation, HisRep (Mao et al., 2020), PGBIG (Ma et al., 2022), Transformer, PhysMoP (Zhang et al., 2024), EgoCast (Escobar et al., 2025), and UniEgoMotion (Patel et al., 2025). For stochastic forecasting on EE4D-Motion, we compare with residual diffusion, SkeletonDiffusion (Curreli et al., 2025), and UniEgoMotion (Patel et al., 2025). On Nymeria, we additionally compare with FIction (Ashutosh et al., 2025), EgoLM (Hong et al., 2025), and Ego3DLM (Bae et al., 2026) under its official evaluation protocol. Published results are reported when protocols match; † denotes our implementation using the corresponding dataset split, forecasting horizon, and joint representation.

Evaluation metrics. Following UniEgoMotion (Patel et al., 2025) on EE4D-Motion (Patel et al., 2025) and Ego3DLM (Bae et al., 2026) on Nymeria (Ma et al., 2024), we adopt dataset-specific metrics to enable direct comparison with prior work. For EE4D-Motion, we report joint position error (JPE), average displacement error (ADE), final displacement error (FDE), average pairwise and distance (APD). For Nymeria, we report average positional error (APE), JPE, ADE, FDE, and Diversity. To analyze horizon-dependent performance, we additionally report MPJPE over disjoint future intervals, denoted by $\mathrm{J}_{a\text{–}b}$. For stochastic forecasting, minimum ADE measures oracle coverage. Errors are reported in centimeters on EE4D-Motion and millimeters on Nymeria; lower is better except APD and Diversity.

Implementation details. For a controlled comparison, on EE4D-Motion, we follow UniEgoMotion (Patel et al., 2025) and use 2 s of body motion, egocentric video, and device trajectory to forecast the subsequent 6 s. On Nymeria (Ma et al., 2024), we follow the 5 s-history-to-5 s-future protocol of Ego3DLM (Bae et al., 2026). We use a 300-step diffusion schedule with 30-step DDIM sampling. Additional architectural and optimization details are provided in the appendix.

4.2 Main Results

Results on deterministic human motion prediction. Table 1 follows the UniEgoMotion protocol on EE4D-Motion and the Ego3DLM protocol on Nymeria. On EE4D-Motion (Patel et al., 2025), CoordMotion reduces $J_{0\text{–}2}$ from 11.983 to 10.820 cm, a 9.7% improvement over the strongest reproduced baseline, and achieves the lowest error at every reported horizon. On Nymeria, it reduces $J_{0\text{–}2}$ from 12.806 to 11.583 cm and achieves the best results from 0.5 s onward. CoodMotion explicitly models eye–body coordination and obtains the best $J_{0\text{–}2}$ on both datasets and the lowest error at

Table 4: **Component ablations on EE4D-Motion.** (a) Deterministic prediction using explicit semantic (Exp.) and implicit coordination (Imp.). (b) Stochastic forecasting using Exp., implicit coordination (Imp.), and the Coordination-Guided Verifier.

(a) Deterministic predictor

Exp.	Imp.	$J_{0.2}\downarrow$	$J_{1.0}\downarrow$	$J_{2.0}\downarrow$	$\mathrm{ADE}_{0\text{–}2}\downarrow$
–	–	3.437	12.008	19.162	11.757
✓	–	3.406	11.912	19.226	11.599
–	✓	3.412	11.826	19.080	11.891
✓	✓	**3.272**	**11.429**	**18.987**	**10.820**

(b) Stochastic predictor and verifier

Exp.	Imp.	Verifier	$\mathrm{ADE}_{2\text{–}4}\downarrow$	$\mathrm{ADE}_{4\text{–}6}\downarrow$	$\mathrm{ADE}_{2\text{–}6}\downarrow$	APD↑
–	–	–	30.937	36.851	33.894	**26.605**
–	–	✓	28.511	33.282	31.897	17.587
✓	✓	–	28.842	34.915	31.680	20.912
✓	✓	✓	**27.060**	**32.119**	**29.560**	15.464

reported horizons. These gains show that explicitly forecasting coordination states improves motion accuracy beyond direct multimodal conditioning.

Results on stochastic human motion prediction. On EE4D-Motion, the coordination-guided predictor achieves the strongest oracle coverage over both the 2–4 s and 4–6 s horizons. The verifier further reduces expected $\mathrm{ADE}_{2\text{–}6}$ from 31.680 to 29.560 cm by removing candidates that conflict with the predicted coordination states. Its slightly higher oracle error and lower APD reflect the intended quality–coverage trade-off: the retained set is less diverse but more coordination-consistent, without using future ground truth. Compared with Ego3DLM on Nymeria, our predictor reduces ADE from 343.9 to 320.8 mm, while verification further improves JPE from 333.5 to 333.2 mm. These results show that coordination modeling improves long-term future coverage, while the verifier turns diverse proposals into a more reliable plausible future set.

Table 5: **Analysis of learned coordination states.** (a) Linear probing evaluates coordination relations. (b) Retrieval and hard-negative evaluation measure semantic alignment. Relation-HN Acc. evaluates hard negatives sharing coordination relations.

(a) Relation probing

Representation	mAUC	mF1	BAcc
Action	0.878	0.678	0.831
Coord. states	**0.897**	**0.707**	**0.837**
Action pair	0.728	0.636	0.676
Coord. transition	**0.773**	**0.667**	**0.704**

(b) Semantic alignment

Metric	Froz. MPNet	Action encoder	Ours
Recall@1	0.226	0.246	**0.341**
Recall@5	0.526	0.604	**0.712**
MRR	0.376	0.408	**0.507**
Relation-HN Acc.	0.338	0.367	**0.490**

4.3 Ablation Study

Effect of the main components. Table 4 evaluates the main components of CoordMotion. Explicit and implicit coordination individually provide modest deterministic gains, while their combination reduces $J_{0\text{–}2}$ from 11.757 to 10.820,cm. For stochastic forecasting, coordination reduces $\mathrm{ADE}_{2\text{–}6}$ from 33.894 to 31.680 cm, and the verifier further lowers it to 29.560 cm by filtering coordination-inconsistent candidates. The accompanying APD reduction reflects the expected quality–diversity trade-off.

Table 6: **Ablation of structured coordination states.** Each row uses only the checked states except the complete model. $J_{a\text{–}b}$ denotes mean MPJPE over the indicated future interval.

Global	Rel.	Phased	Temp.	$J_{0\text{–}2}\downarrow$	$J_{2\text{–}4}\downarrow$	$J_{4\text{–}6}\downarrow$
✓	–	–	–	11.208	28.738	34.253
–	✓	–	–	11.207	29.881	33.666
–	–	✓	–	11.176	29.648	32.246
–	–	–	✓	11.181	28.981	31.315
✓	✓	✓	✓	**10.820**	**27.060**	**32.119**

Table 6 shows that the global, relational, phased and temporal states provide complementary cues. Their combination achieves the best performance across all horizons, demonstrating that predicting structured coordination states provide effective guidance for human motion forecasting.

Discussion on eye–body coordination representation. Table 5 examines whether the learned coordination states capture behavioral structure beyond action identity. For controlled comparison, Table 5(a) applies identical linear probes to action versus coordination-state features, and to action-pair versus coordination-transition features. The coordination states outperform the corresponding action representations, showing that they capture task-dependent eye–body relations and their evolution across actions. Table 5(b) compares retrieval using frozen MPNet embeddings, action-trained

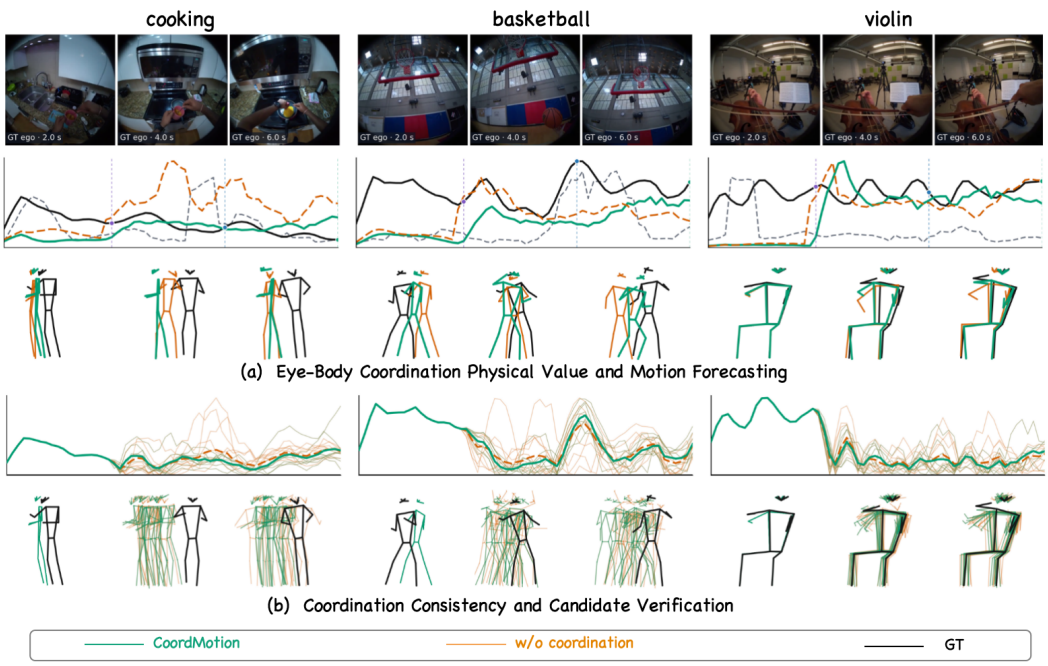


Figure 3: **Activity-wise qualitative results on EE4D-Motion.** **(a)** The upper curves show normalized eye and body-response dynamics, with the vertical axis denoting relative temporal change magnitude. **(b)** The lower curves show candidates' frame-wise absolute body-response errors to GT, where lower indicates stronger coordination consistency.

features, and our learned coordination states. Our representation improves Recall@1 from 0.226 to 0.341 and relation-aware hard-negative accuracy from 0.338 to 0.490. Together, these results show that the learned coordination states capture task-dependent eye–body relations, their temporal evolution, and execution specific coordination beyond action identity.

Visual discussion on coordination for human motion forecasting. Figure 4.3 covers distinct activity dynamics: target-centered coordination in cooking, eye–whole-body coordination in basketball, and frequent coordination changes in violin playing. The upper panels show that CoordMotion better follows the normalized ground-truth body-response dynamics and future poses than the variant without coordination. The lower panels further show that the verifier removes candidates with large frame-wise coordination-response errors and retains a more consistent future set. These results demonstrate the effectiveness of predictive eye–body coordination states for motion forecasting and candidate verification.

5 Conclusion

In this work we presented CoordMotion, which models eye–body coordination as structured predictive behavioral states for human motion forecasting. We learned coordination states from complementary evidence: explicit coordination semantics from atomic-action descriptions and multimodal observations, and implicit fine-grained physical evidence from eye and body dynamics. We predicted global, relational, phased, and temporal coordination states and used them to guide short-term deterministic and long-term stochastic forecasting. Finally, we further demonstrated that the predicted coordination states can effectively guide candidate ranking through a causal Coordination-Guided Verifier. This work demonstrates that predictive eye–body coordination states provide effective behavioral representations to improve human motion forecasting.

AI USE STATEMENT

We used a Qwen multimodal large language model accessed through an API as part of CoordMotion to generate explicit eye–body coordination descriptions from atomic-action annotations and synchronized visual, gaze, and body observations. We also used generative AI tools for Chinese–English translation, language editing, figure prototyping, and feedback on the presentation of the methodology and experiments. These tools did not generate experimental measurements or determine the final scientific claims. The authors manually inspected the generated coordination descriptions, checked AI-assisted text and citations against the original sources, and verified all reported results against the experimental outputs. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work used the existing EE4D-Motion and Nymeria datasets and did not recruit new human participants or collect additional personal data. We followed the access conditions and research-use licenses specified by the dataset providers. Because gaze and body motion may reveal sensitive behavioral information, we used only the released data and annotations for the research purposes described in this paper. The learned coordination states may inherit biases from the activities, participants, annotations, and sensing conditions represented in the training data. Moreover, errors in gaze estimation, atomic-action annotations, or MLLM-generated descriptions may affect the predicted motion. The method is intended for research on human motion understanding and human-aware systems; deployment in surveillance or safety-critical applications would require additional privacy safeguards, bias evaluation, and domain-specific validation.

REPRODUCIBILITY STATEMENT

The main paper specifies the architecture, coordination-states construction, forecasting objectives, verifier, datasets, evaluation protocols, and metrics. The appendix provides the preprocessing procedure, motion representations, sampling settings, baseline adaptations, and additional implementation details. To make the MLLM-based supervision reproducible despite possible API changes, we record the exact Qwen model version, prompt template, decoding settings, and generated coordination descriptions. We also provide the data splits, configuration files and random seeds in appendix.

REFERENCES

Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. Fiction: 4d future interaction prediction from video. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17613–17625. IEEE, 2025.

Yujin Bae, Jaewoo Jeong, Hyeonseong Kim, and Kuk-Jin Yoon. Ego-human motion prediction with 3d-aware llm. *arXiv preprint arXiv:2607.07001*, 2026.

Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. MotionMixer: MLP-based 3d human body pose forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 791–798, 2022. doi: 10.24963/IJCAI.2022/111.

Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, and Daniel Cremers. Nonisotropic gaussian diffusion for realistic 3d human motion prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1871–1882. IEEE, 2025.

Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbelaez. Egocast: Forecasting egocentric human pose in the wild. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5831–5841. IEEE, 2025.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, 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*, 2024.

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5344–5354. IEEE, 2025.

Zhiming Hu, Syn Schmitt, Daniel F. B. Häufle, and Andreas Bulling. GazeMotion: Gaze-guided human motion forecasting. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 13017–13022, 2024. doi: 10.1109/IROS58592.2024.10802548.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision*, pp. 754–769, 2018.

Roland S. Johansson, Göran Westling, Anders Bäckström, and J. Randall Flanagan. Eye–hand coordination in object manipulation. *The Journal of Neuroscience*, 21(17):6917–6932, 2001. doi: 10.1523/JNEUROSCI.21-17-06917.2001.

Michael F. Land. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296–324, 2006. doi: 10.1016/j.preteyeres.2006.01.002.

Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25–26):3559–3565, 2001a. doi: 10.1016/S0042-6989(01)00102-X.

Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25–26):3559–3565, 2001b. doi: 10.1016/S0042-6989(01)00102-X.

Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed forecasting of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2144–2154, 2024.

Ye Lu, Jie Wang, Tianyi Liu, Jianjun Gao, and Kim-Hui Yap. Kinemadiff: Towards diffusion for coherent and physically plausible human motion prediction. In *International Conference on Learning Representations*, 2026.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024.

Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6427–6436. IEEE, 2022.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pp. 474–489. Springer, 2020.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Multi-level motion attention for human motion prediction. *International Journal of Computer Vision*, 129(9):2513–2535, 2021. doi: 10.1007/s11263-021-01483-7.

Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2900, 2017.

Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10318–10329, 2025.

Jeff B. Pelz, Mary Hayhoe, and Russ Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139(3):266–277, 2001. doi: 10.1007/s002210100745.

Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *Proceedings of the European Conference on Computer Vision*, pp. 197–212, 2018.

Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11209–11218, 2021.

Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):5, 2011. doi: 10.1167/11.5.5.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Haodong Yan, Zhiming Hu, Syn Schmitt, and Andreas Bulling. Gazemodiff: Gaze-guided diffusion model for stochastic human motion prediction. *arXiv preprint arXiv:2312.12090*, 2023.

Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10081–10091, 2019.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4372–4381, 2017.

Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6152–6162. IEEE, 2024.

Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J. Guibas. GIMO: Gaze-informed human motion prediction in context. In *Proceedings of the European Conference on Computer Vision*, 2022.

A Appendix

This supplementary material provides additional details on data preprocessing, coordination-supervision construction, model architecture, optimization, evaluation protocols, baseline adaptation, and reproducibility. All forecasting modules follow a strictly causal protocol: future motion, future atomic annotations, and future coordination descriptions are used only as training targets or evaluation annotations and are never provided as model inputs at inference time.

A Dataset and Preprocessing Details

A.1 EE4D-Motion

We use EE4D-Motion (Patel et al., 2025), which is derived from Ego-Exo4D (Grauman et al., 2024) and provides synchronized egocentric video, exocentric full-body motion, Aria-device trajectories, and procedural activity annotations. We resample all modalities to 10 Hz and construct each forecasting sample from 20 observed frames and 60 future frames, corresponding to 2 s of history and 6 s of future motion.

The observed input is

$$\mathcal{O}^h = \left(\mathbf{B}^h, \mathbf{G}^h, \mathbf{V}^h, \mathbf{T}^h\right), \tag{21}$$

where $\mathbf{B}^h \in \mathbb{R}^{20\times17\times3}$ is the observed Body17 motion, $\mathbf{G}^h$ denotes synchronized gaze observations, $\mathbf{V}^h \in \mathbb{R}^{20\times1024}$ contains frozen DINOv2 visual features, and $\mathbf{T}^h$ contains the historical Aria-device trajectory. The prediction target is $\mathbf{Y} \in \mathbb{R}^{60\times17\times3}$.

We use all valid training windows for optimization. For comparisons in the main paper, all locally reproduced methods are evaluated on the same 4,405-window validation manifest. This fixed manifest ensures that differences between methods cannot be attributed to modality validity filtering, temporal stride, or inconsistent sample subsets.

A.2 NYMERIA

Nymeria contains long-horizon egocentric recordings, head trajectories, eye, and full-body motion for diverse indoor activities. Following the Ego3DLM-aligned setting, we use 5 s of observations to forecast 5 s of future motion. We retain the dataset's Body23 representation for evaluation and report local pose error, global joint error, and head-trajectory error. We use a frozen scene-disjoint split to prevent temporally adjacent recordings from the same scene from appearing in both training and evaluation.

Because the published Ego3DLM scene list is unavailable, our locally evaluated results use a reproducible scene-disjoint split. We report this distinction explicitly and do not use target-domain future motion or coordination descriptions to construct causal inputs.

A.3 TEMPORAL SYNCHRONIZATION

We synchronize gaze, device pose, body motion, visual features, and atomic annotations using their timestamps. For a target timestamp t, we associate a modality observation only when its nearest timestamp lies within the modality-specific synchronization tolerance. Invalid observations are masked rather than replaced with future information. All normalization statistics are estimated from the training split and then frozen for validation and test data.

Ego-Exo4D atomic annotations generally provide action timestamps rather than dense phase boundaries. When an action interval is required, we define its boundaries using the midpoints between adjacent atomic timestamps. Future atomic timestamps are used only for transition-centered evaluation and never enter the forecasting model.

B EYE–BODY COORDINATION SUPERVISION

B.1 EVIDENCE PREPARATION

For each valid atomic action, we collect the following synchronized evidence:

$$\mathcal{E}_i = (I_i^{\mathrm{ego}}, I_i^{\mathrm{gaze}}, I_i^{\mathrm{exo}}, A_i, A_{\mathrm{ctx}}, \mathbf{G}_i, \mathbf{H}_i, \mathbf{B}_i)\,, \tag{22}$$

where I_i^{ego} is the raw egocentric image, I_i^{gaze} is the corresponding gaze-overlay image, I_i^{exo} is the synchronized exocentric image, A_i is the atomic description, and A_{ctx} denotes the enclosing activity or skill context. The synchronized gaze, head, and body observations provide physical evidence for the generated description.

We reject examples without valid causal synchronization rather than generating coordination descriptions from the atomic label alone. This filtering prevents the MLLM from fabricating gaze–body relations that are not supported by the observations.

B.2 MLLM-GENERATED COORDINATION DESCRIPTIONS

We use Qwen3.7-Plus as the multimodal large language model. The MLLM receives only synchronized evidence associated with the current atomic action.

We disable hidden reasoning and request a JSON object. We do not explicitly override temperature, top-p, or maximum output length; the provider defaults are therefore used. Each sample is identified by a unique sample ID and written once after local schema validation.

The user-message template is:

```
Generate coordination semantics from the following
synchronized evidence.  Describe only relations supported
```

```
by the images, gaze observation, body motion, and atomic
context.  Return one JSON object that follows the provided
schema.
```

The system prompt applies the following constraints:

1. describe eye–head–root–body relations rather than restating the action label;
2. distinguish observed evidence from uncertain interpretation;
3. specify the attended target only when supported by gaze evidence;
4. describe relative onset, response, monitoring, and transition relations when observable;
5. avoid unsupported object identities, intentions, and future actions;
6. separate body-part coordination from whole-body strategy;
7. return only the required structured JSON fields.

B.3 Cross-atomic transition descriptions

For two adjacent atomic actions (A_i, A_{i+1}), we additionally construct a directed transition description $D_{i\to i+1}$. Unlike a concatenation of two action labels, the transition description characterizes how the coordination pattern changes:

$$D_{i\to i+1} = \mathrm{MLLM}\left(\mathcal{E}_i, \mathcal{E}_{i+1}, A_i, A_{i+1}\right). \tag{23}$$

The direction is important: $D_{i\to i+1}$ and $D_{i+1\to i}$ are treated as different targets. This supervision teaches the representation to distinguish continuing the current coordination pattern from reorganizing gaze and body motion for the next action.

C Coordination Representations

C.1 Explicit semantic representation

We concatenate the body-part and whole-body coordination fields into the final description D_i and encode it with a frozen `all-mpnet-base-v2` sentence encoder:

$$\mathbf{e}_i = \mathrm{norm}\left(E_{\mathrm{text}}(D_i)\right), \qquad \mathbf{e}_i \in \mathbb{R}^{768}. \tag{24}$$

The normalized vector $\mathbf{e}_i$ serves as semantic supervision rather than a future input. The same encoder maps the directed cross-atomic description to $\mathbf{e}_{i\to i+1}$.

C.2 Implicit physical representation

We characterize physical coordination through the synchronized trajectories of gaze, head, root, and Body17 joints. For each trajectory $\mathbf{x}_{1:T}$, we compute

$$\Delta\mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1}, \qquad v_t = \frac{\|\Delta\mathbf{x}_t\|_2}{\Delta t}, \tag{25}$$

together with net displacement and path length:

$$d = \|\mathbf{x}_T - \mathbf{x}_1\|_2, \qquad \ell = \sum_{t=2}^{T} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2. \tag{26}$$

These quantities summarize how far a signal moves, how quickly it changes, and how much cumulative motion it exhibits.

We project gaze, head, root, and joint-wise body descriptors into physical tokens and add learned identities to distinguish the signal sources:

$$\mathbf{X}_i^{\mathrm{phys}} = \left\{W_q\boldsymbol{\phi}_{i,q} + \mathbf{u}_q\right\}_{q\in\{G,H,R,B_1,\ldots,B_{17}\}}, \tag{27}$$

where $\boldsymbol{\phi}_{i,q}$ contains the displacement, mean-speed, and path-length descriptors for source q. A temporal Transformer followed by attention pooling produces

$$\mathbf{p}_i = \mathrm{norm}\left[W_p \sum_n \alpha_{i,n}\, \mathrm{Tr}_{\mathrm{phys}}(\mathbf{X}_i^{\mathrm{phys}})_n\right]. \tag{28}$$

C.3 Structured coordination states

We fuse the action representation $\mathbf{a}_i$, explicit semantic target $\mathbf{e}_i$, and implicit physical evidence $\mathbf{p}_i$ into a coordination representation:

$$\mathbf{z}_i = \mathrm{norm}\left[\mathbf{a}_i + f_{\mathrm{coord}}(\mathbf{a}_i, \mathbf{p}_i)\right]. \tag{29}$$

We align $\mathbf{z}_i$ with the corresponding semantic target using an in-batch contrastive objective:

$$\mathcal{L}_{\mathrm{sem}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{e}_i/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}_i^\top \mathbf{e}_j/\tau)}. \tag{30}$$

Rather than using $\mathbf{z}_i$ as one undifferentiated condition, we derive four coordination states:

$$\mathcal{C}_i = \left\{\mathbf{c}_i^g, \mathbf{c}_i^r, \mathbf{c}_i^p, \mathbf{c}_i^t\right\}, \tag{31}$$

where:

- $\mathbf{c}_i^g$ represents the global behavioral context;
- $\mathbf{c}_i^r$ represents eye–head–root–body relations;
- $\mathbf{c}_i^p$ represents the execution phase;
- $\mathbf{c}_i^t$ represents temporal evolution and transition tendency.

Each state uses an independent projection:

$$\mathbf{c}_i^q = f_q(\mathbf{z}_i), \qquad q \in \{g, r, p, t\}. \tag{32}$$

This state-specific parameterization prevents the dominant global or action semantics from overwhelming the relational, phase, and temporal information.

C.4 Directed transition representation

We represent the transition from action A_i to A_{i+1} as

$$\boldsymbol{\delta}_{i\to i+1} = E_{\mathrm{trans}}\left(\mathcal{C}_i, \mathbf{a}_i, \mathbf{a}_{i+1}\right). \tag{33}$$

We train it to retrieve $\mathbf{e}_{i\to i+1}$ and distinguish the correct directed transition from action-pair and reversed-order negatives. The transition loss is

$$\mathcal{L}_{\mathrm{trans}} = \mathcal{L}_{\mathrm{retrieve}} + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{direction}}. \tag{34}$$

D Optimization and Implementation Details

We train the components in the following order:

1. construct and validate offline coordination descriptions;
2. pretrain explicit semantic, physical, and transition encoders;
3. freeze the teacher encoders and train the causal coordination-state predictor;
4. train the cross-atomic coordination planner;
5. train the deterministic motion center;
6. freeze the center and train the stochastic residual generator;
7. generate candidate sets and train the causal verifier.

E Evaluation Metrics

E.1 Deterministic forecasting

For a prediction $\widehat{\mathbf{Y}}$ and ground truth $\mathbf{Y}$, MPJPE at frame t is

$$J_t = \frac{1}{J}\sum_{j=1}^{J} \left\|\widehat{\mathbf{y}}_{t,j} - \mathbf{y}_{t,j}\right\|_2. \tag{35}$$

The interval error is

$$J_{a\text{–}b} = \frac{1}{|\mathcal{T}_{a:b}|} \sum_{t \in \mathcal{T}_{a:b}} J_t. \tag{36}$$

Thus, the ADE values in our Body17 tables are temporal averages of per-frame MPJPE rather than a separate root-only trajectory metric:

$$\mathrm{ADE} = \frac{1}{T} \sum_{t=1}^{T} J_t, \qquad \mathrm{FDE} = J_T. \tag{37}$$

PA-MPJPE removes global translation, rotation, and scale using Procrustes alignment before computing joint error. Root-relative MPJPE subtracts the root position but does not rotate or rescale the skeleton.

E.2 Stochastic forecasting

Given K candidates, the expected interval error is

$$\mathbb{E}J_{a\text{–}b} = \frac{1}{K} \sum_{k=1}^{K} J_{a\text{–}b}^{(k)}. \tag{38}$$

Oracle coverage is

$$\min J_{a\text{–}b}@K = \min_k J_{a\text{–}b}^{(k)}. \tag{39}$$

These minimum metrics are non-causal because they select the best candidate using future ground truth; we use them only to measure whether the generated set covers the true future.

Average pairwise diversity is

$$\mathrm{APD}\,@K = \frac{2}{K(K-1)} \sum_{k<l} \frac{1}{TJ} \sum_{t,j} \left\| \widehat{\mathbf{y}}_{t,j}^{(k)} - \widehat{\mathbf{y}}_{t,j}^{(l)} \right\|_2. \tag{40}$$

Higher APD denotes greater candidate diversity, but it does not by itself imply greater physical plausibility or forecasting accuracy.

E.3 Verifier metrics

For a retained candidate set $\mathcal{S}$, we report:

- expected error over candidates in $\mathcal{S}$;
- minimum error within $\mathcal{S}$;
- retained-set APD;
- oracle retention, indicating whether the best raw candidate remains in $\mathcal{S}$.

The verifier generally reduces expected error while also reducing APD. This is a quality–coverage trade-off rather than a contradiction: the verifier removes diverse candidates that conflict with predicted coordination or physical dynamics.

F Additional Analysis of Coordination States

F.1 Relation probing

We freeze the representation encoder and train only a class-balanced linear classifier:

$$\mathbf{z}_i = E_{\mathrm{frozen}}(\mathcal{E}_i), \qquad \widehat{\mathbf{y}}_i = W\mathbf{z}_i + \mathbf{b}. \tag{41}$$

The input is a frozen action, action-pair, coordination-state, or coordination-transition representation. The targets are the LLM-derived coordination-relation labels contained in the validated structured descriptions. We report macro AUC, macro F1, and balanced accuracy. Because the probe is linear, improved performance indicates that coordination relations are already organized in the frozen representation rather than learned by a powerful downstream classifier.

F.2 SEMANTIC ALIGNMENT

We evaluate held-out coordination-description retrieval. The query is a frozen representation and the candidates are unseen coordination descriptions embedded by the frozen text encoder. Recall@K and mean reciprocal rank measure retrieval quality. Relation-aware hard-negative accuracy compares descriptions with similar action content but different coordination relations, preventing performance from being explained by action identity alone.

F.3 FOUR-STATE ABLATION

To isolate the contribution of the structured states, we train controlled predictors using only one of the global, relational, phased, or temporal states. The complete model uses all four state-specific projections and FiLM branches. All variants use identical observations, model capacity outside the state branches, optimizer, training schedule, and validation manifest.

The ablation should be reported as:

G TRANSITION-CENTERED EVALUATION

To test whether coordination modeling is especially useful when behavior changes, we partition validation windows using the next atomic timestamp. A window is classified as *stable* if no transition proxy occurs within the first two future seconds and as *transition* otherwise. For transition windows, we additionally evaluate the 0.5 s interval immediately following the transition proxy.

Future timestamps are used only to form evaluation subsets. They are not provided to any method. This analysis tests whether the learned transition states help the model forecast motion near action boundaries rather than merely improving static or slowly varying sequences.

H QUALITATIVE VISUALIZATION PROTOCOL

For each activity, we visualize synchronized egocentric frames, deterministic or stochastic Body17 predictions, and physical coordination curves. The gaze-dynamics curve is computed from the angular change between consecutive world-coordinate gaze directions:

$$\omega_t^g = \frac{\arccos\left[\mathrm{clip}\left(\mathbf{g}_t^\top \mathbf{g}_{t-1}, -1, 1\right)\right]}{\Delta t}. \tag{42}$$

Head dynamics use the geodesic rotation difference of consecutive Aria poses. Body response uses the mean velocity of root-relative Body17 joints.

Because gaze angular speed, head angular speed, and body velocity have different units, we apply robust per-window normalization:

$$\widehat{z}_t = \mathrm{clip}\left(\frac{z_t - P_5(z)}{P_{95}(z) - P_5(z)}, 0, 1\right). \tag{43}$$

The resulting vertical axis therefore represents normalized physical change, not absolute pose or position.

For stochastic visualization, thin curves denote individual candidates, thick curves denote candidate-set summaries, and the black reference denotes ground truth. The verifier visualization compares the raw proposal set with the causally retained set and illustrates whether verification removes candidates whose body responses conflict with the predicted coordination states.