Learning Predictive Visuomotor Coordination

Supplementary Material

6. Additional Video Demo

Due to space limitations, the main paper only presents static visualizations. In this supplementary material, we provide a richer set of qualitative demos showcasing our model's predictions across diverse scenarios. These include extended visualizations of predicted head pose, gaze, and upper-body motion, highlighting both successful cases and failure modes. We also analyze common prediction errors, such as subtle signal leads to wrong predictions, and inherent challenge from unforseeable human motion, to better illustrate the model's strengths and limitations.

We encourage readers to view the full set of qualitative results in the provided demo videos for a more comprehensive understanding of our model's performance.

7. EgoExo4D Data Cleaning Pipeline

EgoExo4D provides detailed pose annotations by running off-the-shelf human pose estimation models on multiple exocentric camera views. However, occlusions frequently lead to missing body parts, resulting in occasional inaccuracies in the automatically generated annotations. Even with manually annotated corrections, these issues remain common due to unavoidable viewpoint limitations.

While Fiction [3] improves annotation quality by re-annotating filtered sequences using narration-based semantic cues, our goal is to model general visuomotor coordination without restricting the dataset based on activity type. Instead of manually filtering data based on semantics, we apply a 5-second sliding window to correct annotation errors using temporally adjacent frames. If a missing or incorrect joint annotation cannot be recovered within a reasonable range, we discard that frame.

After this cleaning process, our training and testing samples are drawn from the valid index list using a 20 steps sliding window, with a stride of 10 steps. This ensures that our model learns from reliable annotations while maintaining a broad range of natural visuomotor behaviors, aligning with our goal of capturing general coordination patterns rather than task-specific motions.

8. Per-Class Performance Analysis

Table 3 presents the per-step prediction performance of our model across different skilled activities. The results reveal a strong correlation between motion variability and prediction difficulty, with activities exhibiting larger motion change amplitudes (Δ_{Avg}) leading to higher errors. Structured tasks like Cooking and Health yield lower errors, while dynamic or fine-grained activities such as Basketball and fixing Bike introduce more uncertainty due to abrupt gaze shifts and complex hand-eye coordination. These findings highlight the challenge of modeling high-motion scenarios, suggesting that improving robustness in such conditions is crucial for advancing visuomotor prediction models.

Task / Δ_{Avg}	Basketball	Cooking	Bike	Health
PA-MPJPE	78 / 116	47 / 59	52 / 61	38 / 48
Head Pos.	16 / 35	12 / 18	11 / 19	11 / 17
Gaze Pos.	195 / 546	89 / 164	85 / 137	64 / 94
Hand Pos.	304 / 733	128 / 208	124 / 166	87 / 117
Head Rot.	16 / 35	12 / 18	11 / 19	11 / 17
Count	2,037	887	1,136	1,066

Table 3. Per-class Average Prediction Error vs. Motion Change Amplitude across different activity categories.

Time Step	t+1	t+3	t+5	t+7	t+10	Mean
PA-MPJPE	29	48	60	68	78	59
Head Pos.	16	54	94	136	200	106
Gaze Pos.	26	69	112	156	226	124
Hand Pos.	61	130	181	228	294	188
Head Rot.	2.5	7.6	12.3	16.7	23.2	13.2

Table 4. Per-Step Performance vs. Mean Performance.

Dataset	Head Pos.	Gaze Pos.	Hand Pos.	Head Rot.
Nymeria	63	89	110	12.8
EgoExo4D	106	124	188	13.2

Table 5. Results on a Nymeria and EgoExo4D Cooking Subset.

9. Per-Step Performance Analysis.

Table 4 presents the per-step prediction performance of our model across different future time steps. While the diffusion model generates the entire trajectory at once, errors increase over longer horizons due to growing uncertainty in future motion. At $t + 1$, predictions are highly accurate, with head position error at 16 and head rotation at 2.5 degrees. However, as the time step extends, errors grow significantly, reaching 200 for head position and 226 for gaze position at $t+10$, reflecting the increasing difficulty of modeling long-range dependencies where future states become less constrained by recent observations.

Hand motion exhibits the largest variation, with error rising from 61 at $t + 1$ to 294 at $t + 10$, suggesting that fine-grained hand movements are harder to predict due to their higher variability and dependence on external factors. In contrast, head rotation remains more stable, increasing

gradually to 23.2 at $t + 10$, indicating that head orientation follows smoother, more predictable patterns. These results suggest that incorporating trajectory-level constraints or enhancing long-range temporal dependencies could improve long-horizon stability, particularly for hand motion.

10. Experiment on a Nymeria Subset: Cooking Scene

We also include a small-scale experiment on the recently released Nymeria dataset[41], as it uses the same data collection device as EgoExo4D and provides precise upper-body joint annotations. We evaluated our model on 5,966/1,464 training/testing samples from the Cooking Scene, with results reported in Table 5, offering a direct comparison with EgoExo4D.

While Nymeria contains high-quality annotations, key differences from EgoExo4D make it less suitable for our visuomotor coordination task. Unlike EgoExo4D, which focuses on skilled activities, Nymeria captures a broader range of behaviors due to its larger, unconstrained recording environment. The absence of exocentric cameras, which in EgoExo4D helps provide multiple viewpoints, allows Nymeria to include simpler navigation behaviors. As a result, its prediction errors are numerically lower (Table 5), not necessarily due to better performance but rather because it contains more low-movement activities, reducing overall motion complexity.

Another key limitation is the presence of an observer role in Nymeria, which significantly alters the natural behavior of the camera wearer. Instead of fully engaging in skilled activities, the wearer frequently exhibits behaviors such as glancing, waiting, and social interactions. In many cases, their trajectory is modified to avoid the observer, introducing motion patterns that do not align with our visuomotor coordination focus. While filtering out these behaviors could make Nymeria more applicable, it would require additional annotations or heuristics to distinguish observer-influenced motions from task-driven actions, making it non-trivial.

Despite these limitations, Nymeria remains a valuable dataset due to its clean annotations. Future work could explore methods for selectively filtering out non-task-relevant behaviors, potentially making it more suitable for visuomotor research. However, such efforts fall outside the scope of this paper.

11. Limitations and Future Work

While our model effectively predicts future visuomotor states, it assumes consistent egocentric input quality and does not explicitly account for external environmental factors such as object affordances or scene constraints. This limits its applicability in tasks requiring fine-grained interaction reasoning. Additionally, our current framework focuses on short-term predictions, and extending it to long-horizon forecasting or real-time inference remains an open challenge. Future work can explore integrating external scene context, improving long-term prediction robustness, and adapting the model for interactive task modeling, further bridging the gap between human cognition and computational visuomotor learning.