EggHand: A Multimodal Foundation Model for Egocentric Hand Pose Forecasting

Supplementary Material

A. More Studies for EggHand

This supplementary document provides additional details, qualitative results, and analyses to support the main paper. The contents are organized as follows:

- Forecasting Samples (Section A.1): Presents qualitative results demonstrating the model's robust prediction performance across diverse data conditions.
- Training Dynamics (Section A.2): Demonstrates that the proposed loss combination significantly improves training convergence speed and stability compared to the baseline.
- Implementation Details (Section B): Details the specific hyperparameter configurations for EggHand and the implementation protocols for the comparative baseline models.
- Backbone Model Structure (Section C): Elaborates on the detailed architecture of the VLM and Action modules, including the mechanism for hardware-specific control.
- Limitations (Section D): Discusses computational constraints due to model size and suggests potential avenues for future optimization via model compression.

A.1. Forecasting Samples

In this section, we present additional qualitative forecasting results of EggHand that were omitted from the main paper due to space constraints. Fig. S1 and Fig. S2 illustrate the model's consistent predictive performance across various interaction tasks from multiple perspectives. Specifically, Fig. S2 visualizes the predicted poses projected onto the 2D egocentric video frames to demonstrate visual alignment, while Fig. S1 directly visualizes the 3D joint points in spatial coordinates, highlighting the model's precise spatial understanding. EggHand consistently demonstrates robust performance across a diverse range of tasks and complex environments. Even in challenging scenarios involving rapid ego-motion or severe occlusions during the forecasting window, the model reliably generates stable and semantically coherent hand poses that fully preserve plausible grasping structures. Consequently, additional supplementary samples further confirm that EggHand maintains exceptional robustness across diverse, unconstrained real-world conditions.

A.2. Training Dynamics

We compare the training behavior of our full loss configuration against using only the Absolute Coordinate Loss ($\mathcal{L}_{\text{abs}}$). As shown in Fig. S3 and Fig. S4, Ours(Abs + Rel + Pair) exhibits substantially faster convergence and markedly smaller oscillation amplitude, demonstrating a much more stable training process overall. In contrast, using only $\mathcal{L}_{\text{abs}}$ leads to slower stabilization and larger early-stage fluctuations. These trends demonstrate that the additional geometric constraints introduced in our full loss reduce optimization instability and enable smoother, more reliable convergence than the absolute-only baseline.

Table S1. **Model configuration for EggHand.** Config values for EgoVideo and GR00T used in EggHand.

Component	Parameter	Value
EgoVideo	Frames	4
	Vision dim	768
	Projected dim	2048
	Tune vision	False
	Tune LLM	False
GR00T Action Head	Hidden size	1024
	Backbone Emb Dim	2048
	Max Act / State Dim	126 / 126
	Vision tokens	32
Diffusion	Layers	16
	Heads	32
	Head dim	48
	Cross-attn dim	2048
	Dropout	0.2
	Noise (α, β, s)	(1.5, 1.0, 0.999)

B. Implementation Details

B.1. Model Configuration

Tab. S1 summarizes the configuration of the GR00T-N1.5 [S1] action decoder and the EgoVideo [S3] backbone used in EggHand. We follow the GR00T-N1.5 defaults for hidden dimensions, attention, and diffusion depth, while adapting the input embeddings to our egocentric hand motion representation.

B.2. Adapter and Fusion Setup

We keep the EgoVideo backbone frozen and project the resulting video tokens into the GR00T [S1] latent space using a lightweight learnable linear adapter. This projection layer

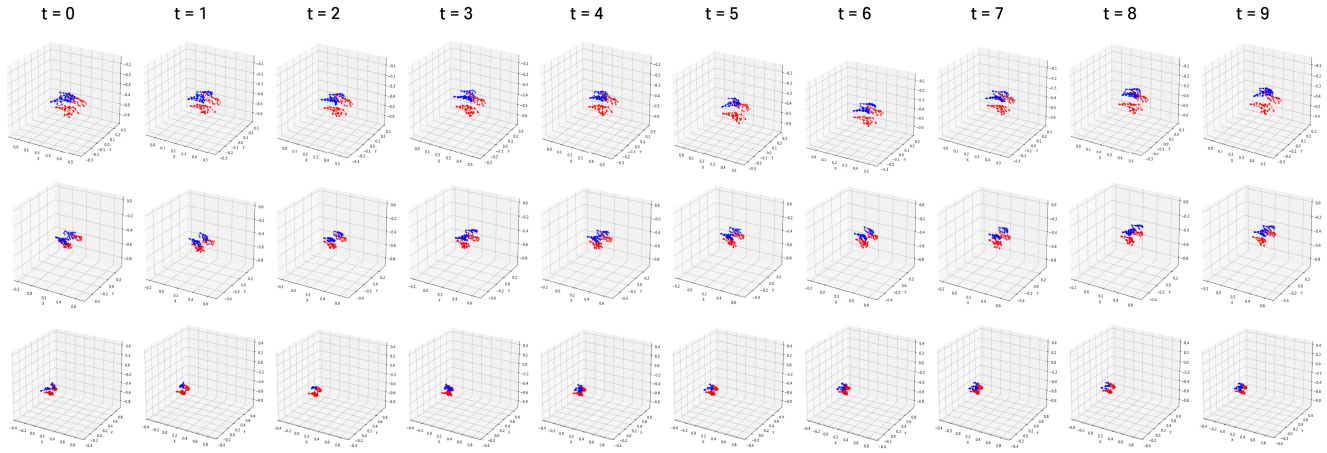

Figure S1. **3D qualitative results.** Predicted 3D hand joint positions across future time steps from t = 0 to t = 9. Blue dots represent the ground truth hand joints, while red dots indicate the predicted poses. Each row displays a distinct continuous motion sequence, demonstrating the model's ability to track and forecast hand movements over time.

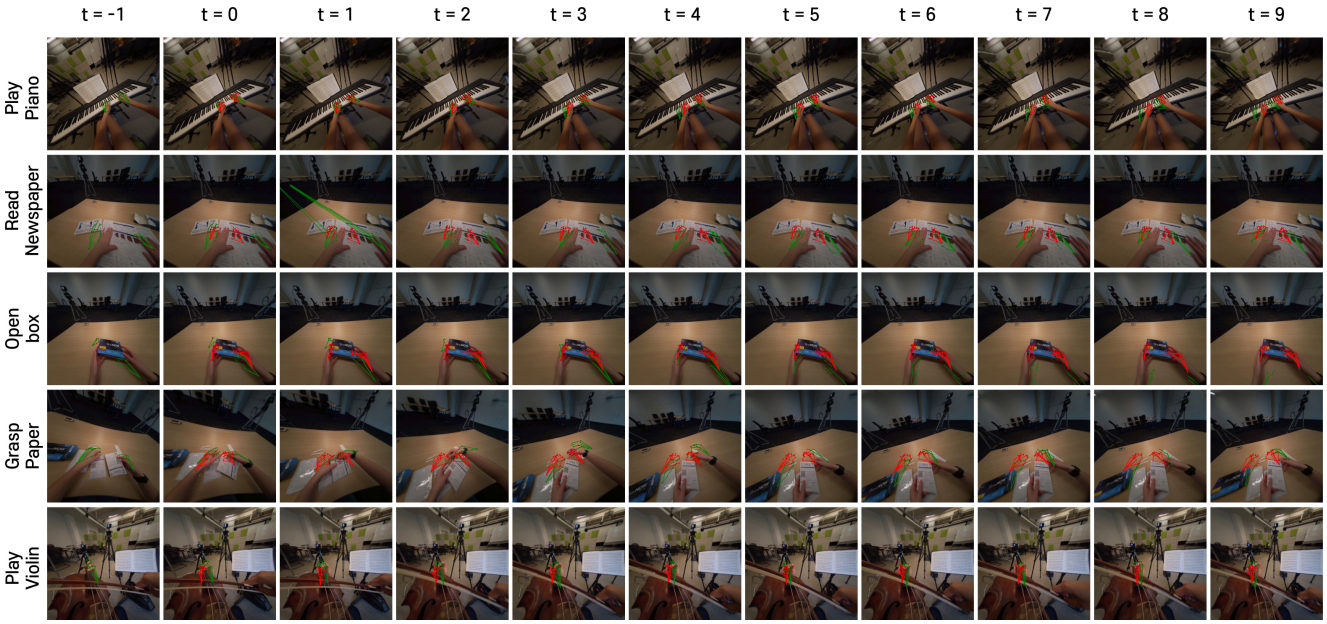

Figure S2. **Additional qualitative results.** Predicted hand poses overlaid on egocentric video frames from time step t = -1 to t = 9. The y-axis denotes the specific interaction tasks, while the x-axis represents the time steps. Green joints represent the ground truth hand poses, and red joints indicate the predicted poses.

is the only visual component updated during training, allowing efficient adaptation while preserving the pretrained visual features.

Hand joint states from the past 20 frames are embedded using a single linear layer and concatenated with the projected video tokens. The GR00T [S1] action decoder then fuses these representations through cross-attention, enabling the model to combine view-dependent context with fine-grained temporal motion cues.

B.3. Baseline Implementation Details

The EgoH4 [S2] baseline is a diffusion-based forecasting model that leverages full-body motion cues to estimate future 3D hand trajectories and hand poses, including cases where hands are partially or fully out-of-view during the observation period. Instead of relying solely on visible hand locations, EgoH4 jointly denoises *body joints and hand joints*, allowing the model to exploit structural constraints from the body—e.g., shoulders or elbows—to infer plausible hand positions when direct visual evidence is ab-

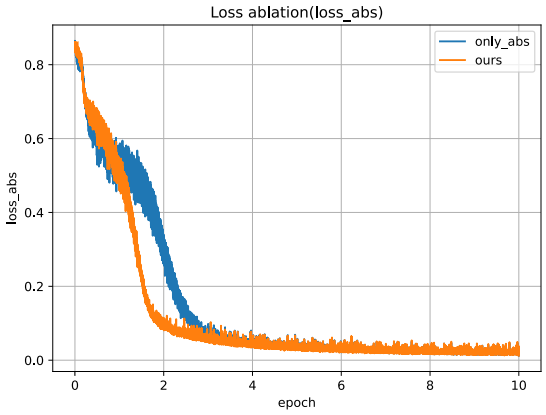

Figure S3. **Absolute Coordination Loss.** Training curve comparing Ours(Abs + Rel + Pair) against the absolute-only baseline. Our configuration converges faster and exhibits significantly smaller oscillation amplitude.

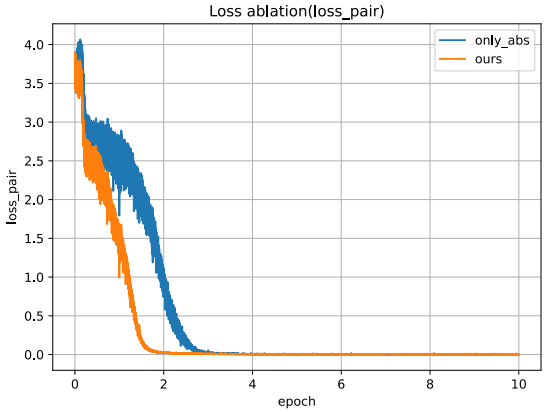

Figure S4. **Pairwise Distance Loss.** Pairwise structural consistency loss during training. Ours stabilizes rapidly compared to the absolute-only setting, confirming that geometric constraints improve optimization stability.

sent. EgoH4 incorporates three main components that enhance stability and visibility awareness. First, a visibility classifier predicts whether each hand is in- or out-of-view at every observation timestep, which improves forecasting for occluded hands by explicitly modeling visibility transitions. Second, a 3D-to-2D reprojection loss enforces geometric consistency between the predicted 3D wrist positions and their observed 2D image coordinates, providing an additional regularization signal that anchors the 3D predictions to image evidence when available. Third, conditioning on image features, 2D hand detections, and gravity-aligned head poses stabilizes the denoising process and improves forecasting accuracy across both in-view and out-of-view sequences. We retrained the EgoH4 baseline following its original coordinate convention and training configuration.

B.4. Metrics

We evaluate the predictive performance using standard trajectory and pose forecasting metrics. Since our model generates multimodal predictions via stochastic sampling, we report the Best-of-K metrics, where the sample closest to the ground truth is selected for evaluation. Let $P^i_{t,j} \in \mathbb{R}^3$ denote the ground truth position of the j-th joint for subject i at time step t in world coordinates. Similarly, $\hat{P}^{i,(k)}_{t,j}$ represents the k-th predicted sample among K generated hypotheses. To evaluate the local pose articulation independent of global trajectory, we also define the root-relative coordinates as $\tilde{P}_{t,j} = P_{t,j} - P_{t,\text{root}}$, where the root joint (e.g., wrist) is aligned to the origin. The total number of joints is J, and the prediction horizon is T.

- **Average Displacement Error (ADE)**: The mean Euclidean distance between the predicted root trajectory and the ground truth over all time steps. We report the minimum ADE across K samples, averaged over all N sequences:

$$\text{ADE} = \frac{1}{N}\sum_{i=1}^{N}\min_{k=1}^{K}\left(\frac{1}{T}\sum_{t=1}^{T}\|\hat{P}^{i,(k)}_{t,\text{root}} - P^i_{t,\text{root}}\|_2\right) \tag{1}$$

- **Final Displacement Error (FDE):** The Euclidean distance between the predicted root position and the ground truth at the final time step T. Similarly, we report the minimum FDE:

$$\text{FDE} = \frac{1}{N}\sum_{i=1}^{N}\min_{k=1}^{K}\left(\|\hat{P}^{i,(k)}_{T,\text{root}} - P^i_{T,\text{root}}\|_2\right) \tag{2}$$

- **MPJPE (Mean Per Joint Position Error):** To evaluate the accuracy of the local hand pose independent of global trajectory errors, we compute the MPJPE after aligning the root joint (wrist) of the prediction to the ground truth. Let $\tilde{P}_{t,j} = P_{t,j} - P_{t,\text{root}}$ be the wrist-relative coordinates:

$$\text{MPJPE} = \frac{1}{N}\sum_{i=1}^{N}\min_{k=1}^{K}\left(\frac{1}{T\cdot J}\sum_{t=1}^{T}\sum_{j=1}^{J}\|\hat{\tilde{P}}^{i,(k)}_{t,j} - \tilde{P}^i_{t,j}\|_2\right) \tag{3}$$

- **MPJPE-F (Mean Per Joint Position Error Final):** The wrist-relative pose error calculated specifically at the final time step T:

$$\text{MPJPE-F} = \frac{1}{N}\sum_{i=1}^{N}\min_{k=1}^{K}\left(\frac{1}{J}\sum_{j=1}^{J}\|\hat{\tilde{P}}^{i,(k)}_{T,j} - \tilde{P}^i_{T,j}\|_2\right) \tag{4}$$

C. Backbone Model Structure

EggHand comprises a VLM module responsible for high-dimensional vision–language feature extraction and an Action module dedicated to robotic control. As shown in Fig. S5, the Vision Encoder in the VLM module adopts a

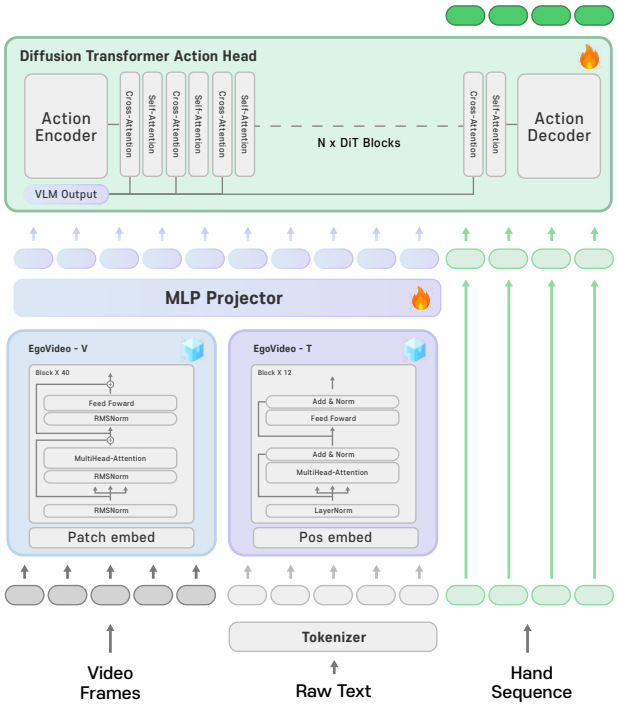

Figure S5. EggHand uses the Transformer block of the EgoVideo VLM and the DiT architecture from the GR00T action head. Since the Action Encoder applies distinct weights to each robot embodiment ID, we finetuned the model for the newly introduced embodiment ID.

large-scale InternVideo-based ViT architecture. To maintain training stability and computational efficiency in its deep 40-layer design, it integrates Pre-Norm, LayerScale, and Flash Attention. In contrast, the Text Encoder employs a standard BERT-based Post-Norm architecture to embed the context of user commands. The Action Module adopts a Diffusion Transformer backbone, generating action trajectories by leveraging the output of the Perception Module as the conditioning input for cross-attention. To smoothly leverage the pretrained weights of the action decoder, it is necessary to align the input state and action dimensions with the decoder's latent space. To achieve this, we introduce a single projection layer to transform the state and action features. Specifically, this projection layer is implemented as a single layer MLP, which minimizes additional computational overhead while effectively ensuring compatibility with the pretrained model.

D. Limitations

EggHand integrates two pretrained foundation components—an egocentric video–text encoder and a VLA action decoder—which makes the overall model somewhat larger than conventional task-specific hand forecasting architectures. This increases the computational footprint and may limit direct deployment on lightweight edge devices. However, this limitation is not fundamental: the modular structure of EggHand allows each component to be independently compressed through knowledge distillation, pruning, or substitution with lightweight backbones. Since the proposed framework does not rely on full-body pose or external tracking, compact distilled variants can retain most of the forecasting capability. Thus, while the current model size is modestly higher than specialized baselines, it remains amenable to future on-device optimization.

References

[S1] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 1, 2

[S2] Masashi Hatano, Zhifan Zhu, Hideo Saito, and Dima Damen. The invisible egohand: 3d hand forecasting through egobody pose estimation. *arXiv preprint arXiv:2504.08654*, 2025. 2

[S3] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 1