MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation

Bohan Zhou, Yi Zhan, Zhongbin Zhang, Zongqing Lu

03/12/2025

Motivation

北京大學 PEKING UNIVERSITY

MEgoHand: Generating hand-object motions from first-person views, instructions and initial states.

Wide Applications:

- Immersive virtual-real alignment in AR/VR.
- Robotic imitation learning from human demonstrations.

Key Challenges:

- Unstable and shifting viewpoints of egocentric views.
- Frequent self-occlusions of objects or hands.
- Strong perspective distortion and rapid scale changes because of close distance.
- Hard reasoning under partial observations and sparse visual cues.

Limitations of Existing approaches:

- Reliance on predefined 3D object priors (mass, geometry), limited generalization to novel objects.
- Multimodal approaches suffer from open-loop prediction errors, ambiguous generation or complex 3D hand-object correlation pipelines.

Related Work

Hand-Object Interaction (HOI) Prediction

Method Type	Examples	Strengths	Limitations
Object-Centric	GEARS, MACS	Explicit physical modeling	Relies on 3D object priors
Text-Conditioned	DiffH2O, Text2HOI	Text-guided motion	Needs object-specific info
Image-Based	SIGHT-Fusion	Occlusion resilience	Requires object detection
Multimodal	LatentAct	Vision-text-contact fusion	Complex contact map pipeline

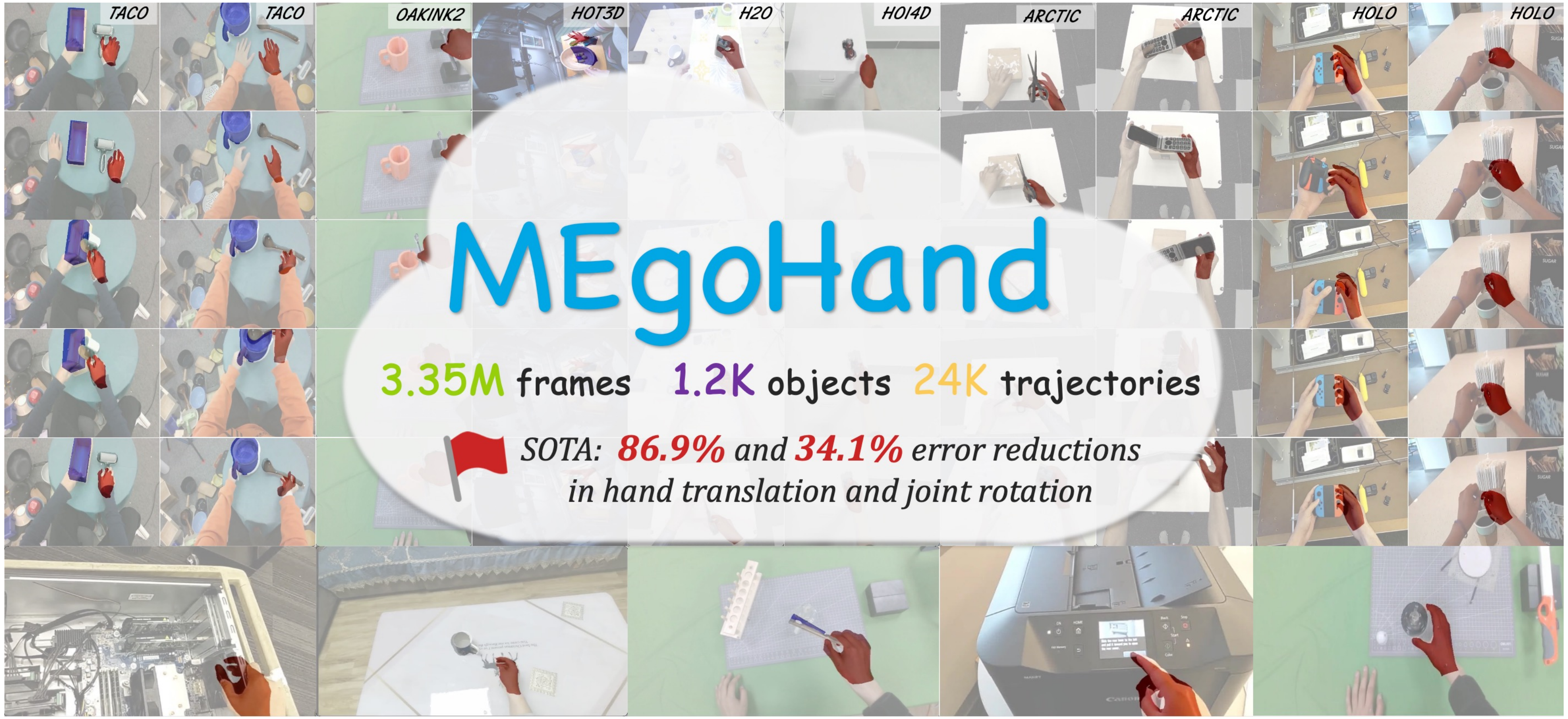
Introduction

Inputs:
- Task Description
- Egocentric RGB
- Initial MANO State

Outputs:
- Future MANO States

$$h = [\theta; \beta; r; t] \in \mathbb{R}^{109}$$

$$\theta \in \mathbb{R}^{15\times 6} \quad t \in \mathbb{R}^{3}$$
$$\beta \in \mathbb{R}^{10} \quad r \in \mathbb{R}^{1\times 6}$$

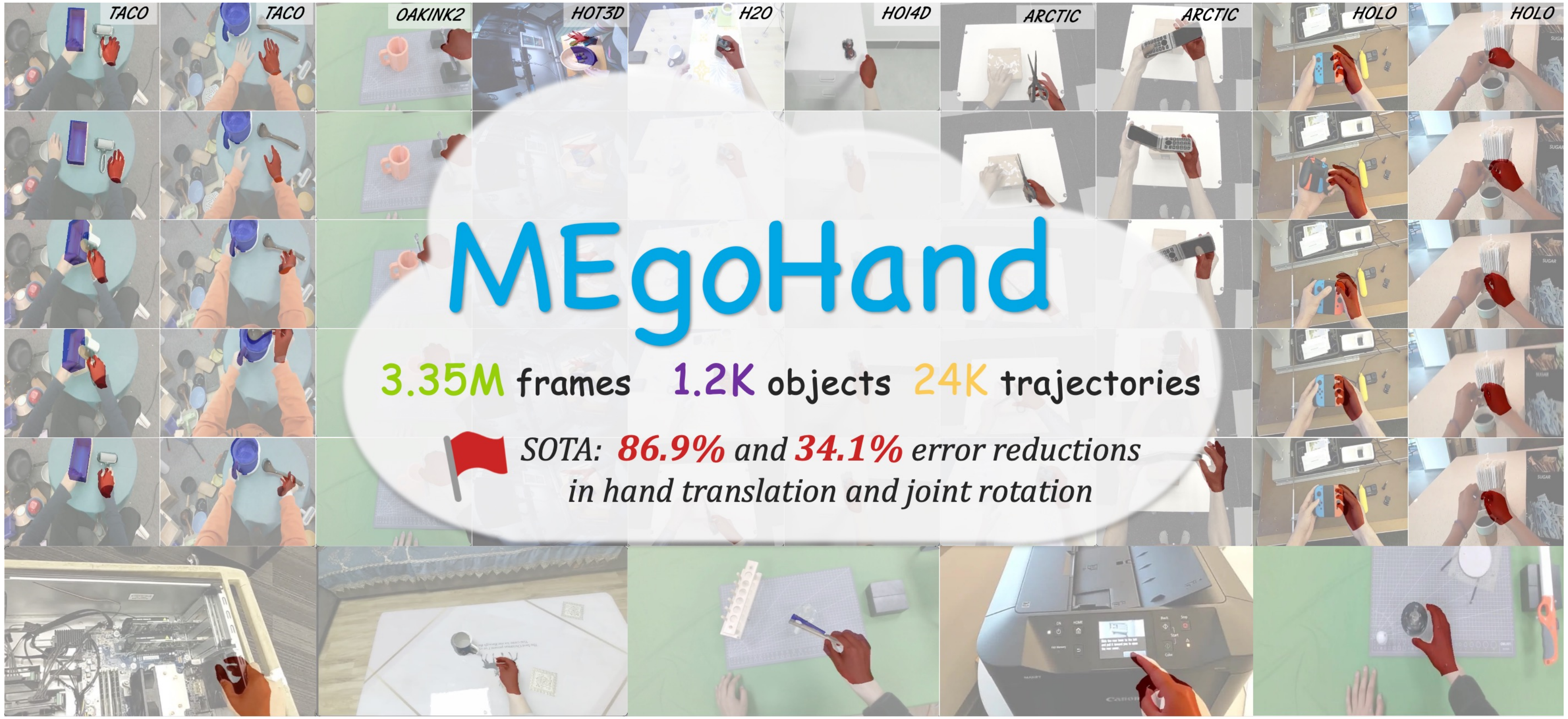


$$\mathcal{H}_k = \{h_{k+1}, h_{k+2}, \ldots, h_{k+l}\} = \text{MEgoHand}(\mathcal{T}, \mathcal{V}_k, h_k)$$

Introduction

Core Contributions

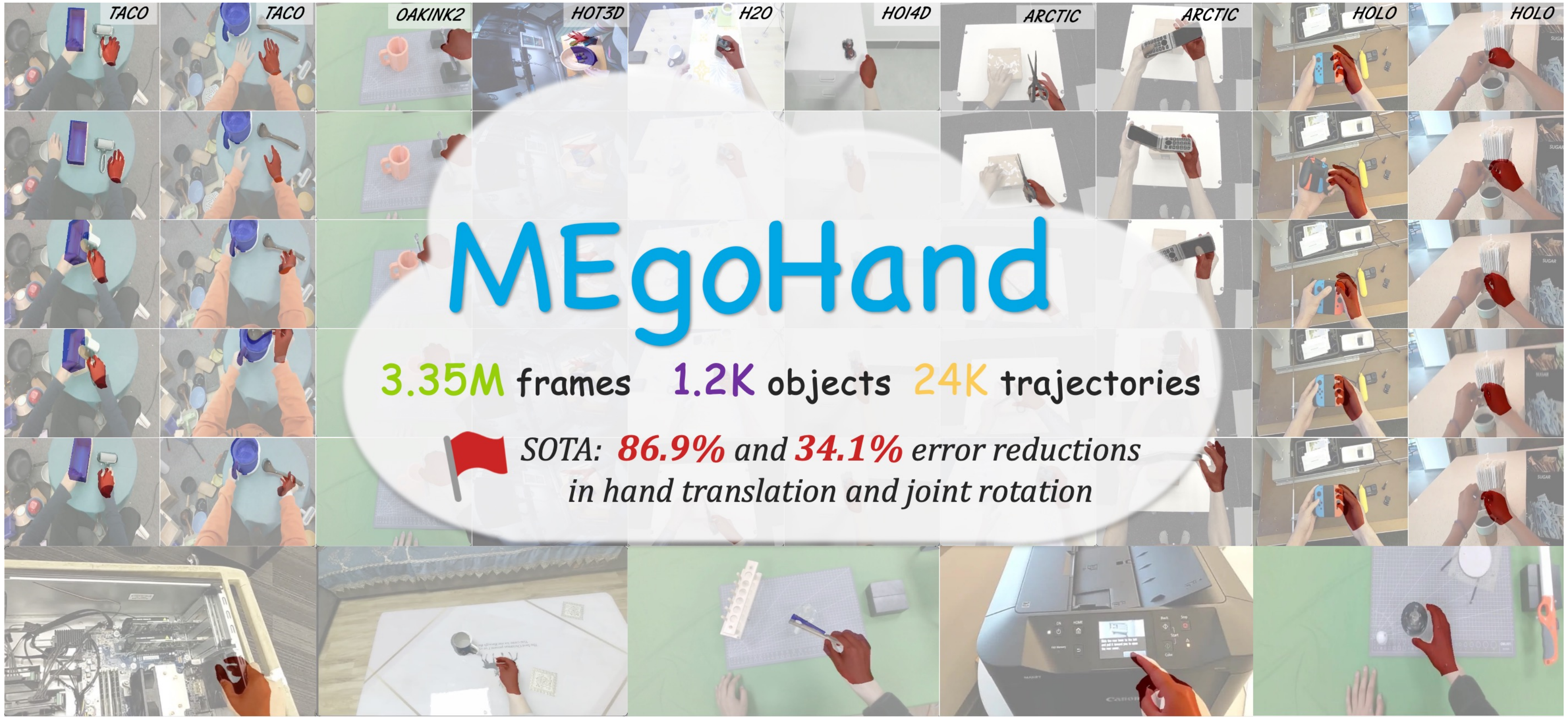
- Large-Scale Dataset
- Architecture Design
- Decoding Strategy
- Benchmarks



Methodology

Core Contributions:

- Large-Scale Dataset
- Architecture Design
- Decoding Strategy
- Benchmarks



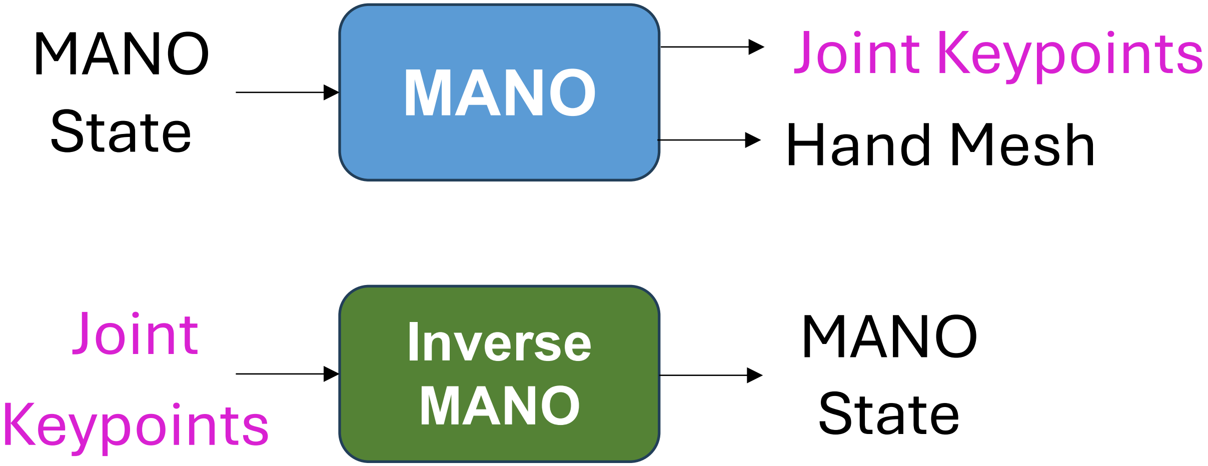
Dataset Integration

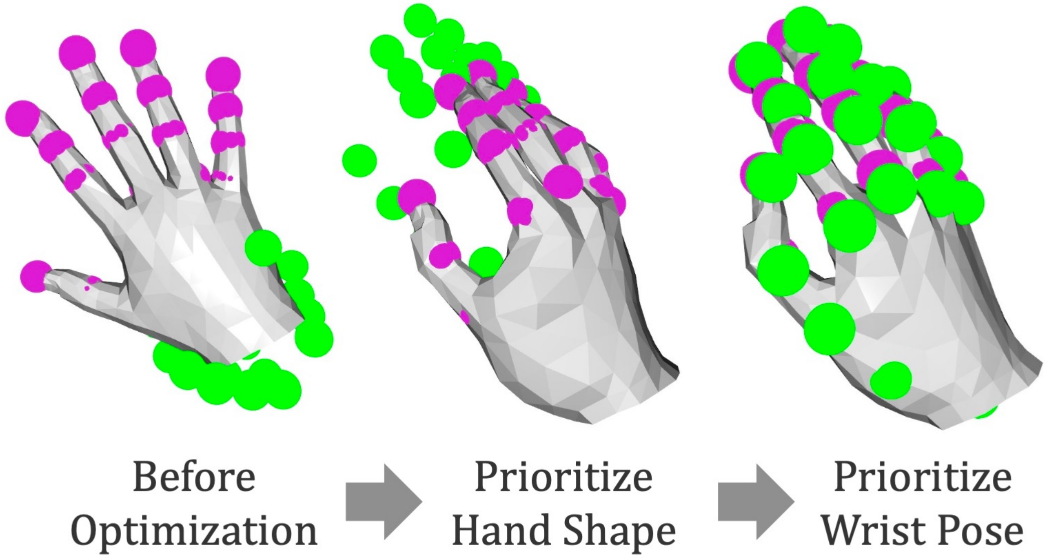
Dataset	Frame	Trajectory	Object	Mesh	RGBD
OakInk2 [42]	600K	2.5K	75	✓	✗
HOT3D [2]	400K	3K	33	✓	✗
HOI4D [26]	400K	3K	800	✓	✓
TACO [25]	300K	2.2K	218	✓	✓
H2O [19]	100K	1K	8	✓	✓
FPHA [15]	100K	1.3K	26	✓	✓
ARCTIC [12]	250K	1K	12	✓	✗
HOLO [37]	1200K	10K	40	✗	✓
Total	**3.35M**	**24K**	**1.2K**		

We integrate and preprocess large-scale public datasets into a unified and standardized training corpus by filling the missing modalities: (1) **MANO State Labels** (2) **Depth Input**.

Dataset Integration

Inverse MANO Retargeting: recover hand meshes from joint keypoints





$$\mathcal{L}_1 = w_1 \mathcal{L}_{\text{shape}}\left(\phi(j), \theta, \beta\right) + \mathcal{L}_{\text{recon}}\left(\text{MANO}(\phi(j)), j\right),$$

$$\mathcal{L}_2 = w_2 \mathcal{L}_{\text{pose}}\left(\phi(j), r, t\right) + \mathcal{L}_{\text{recon}}\left(\text{MANO}(\phi(j)), j\right),$$

$$\mathcal{L}_{\text{inv}} = \sigma \mathcal{L}_1 + (1-\sigma)\mathcal{L}_2$$

$$w_1 = 4.0 \;\; \sigma = 1 \;\; \text{and} \; w_2 = 5.0 \;\; \sigma = 0$$

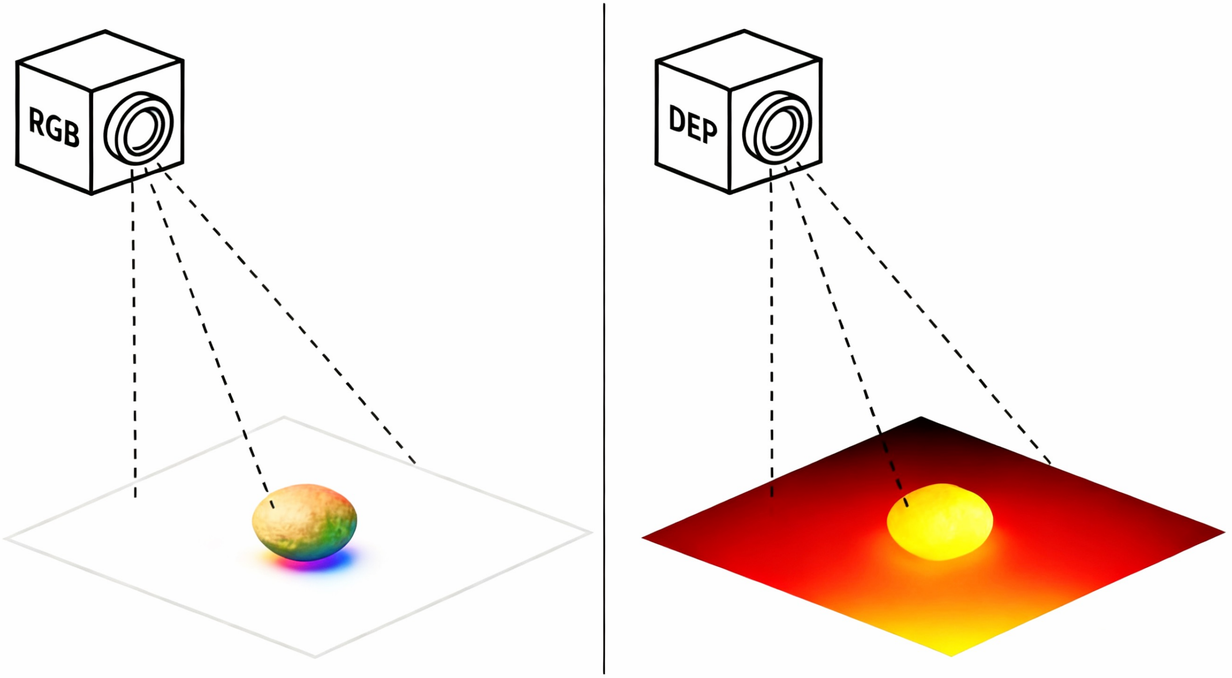
Dataset Integration

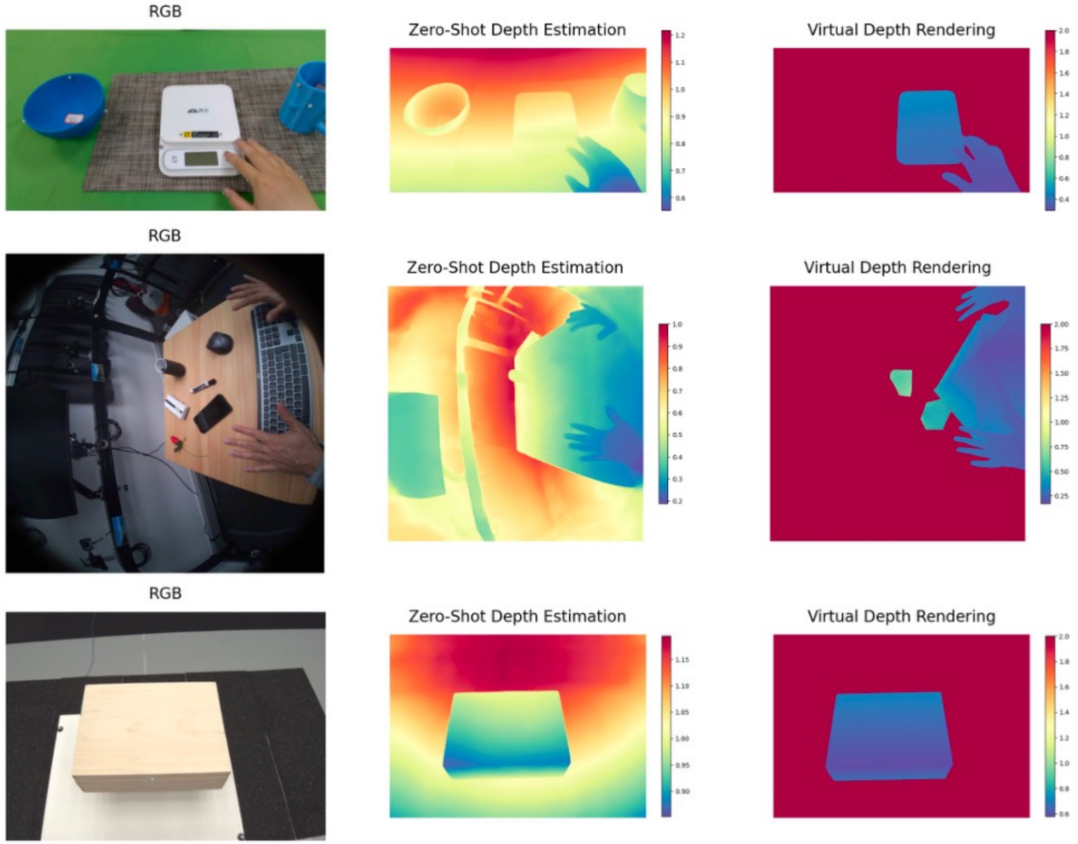
PEKING UNIVERSITY

Inverse MANO Retargeting: recover hand meshes from joint keypoints

Dataset Integration

Virtual RGB-D Rendering: synthesize depth images aligned with the RGB frames





Dataset Integration

北京大學 PEKING UNIVERSITY

Virtual RGB-D Rendering: synthesize depth images aligned with the RGB frames

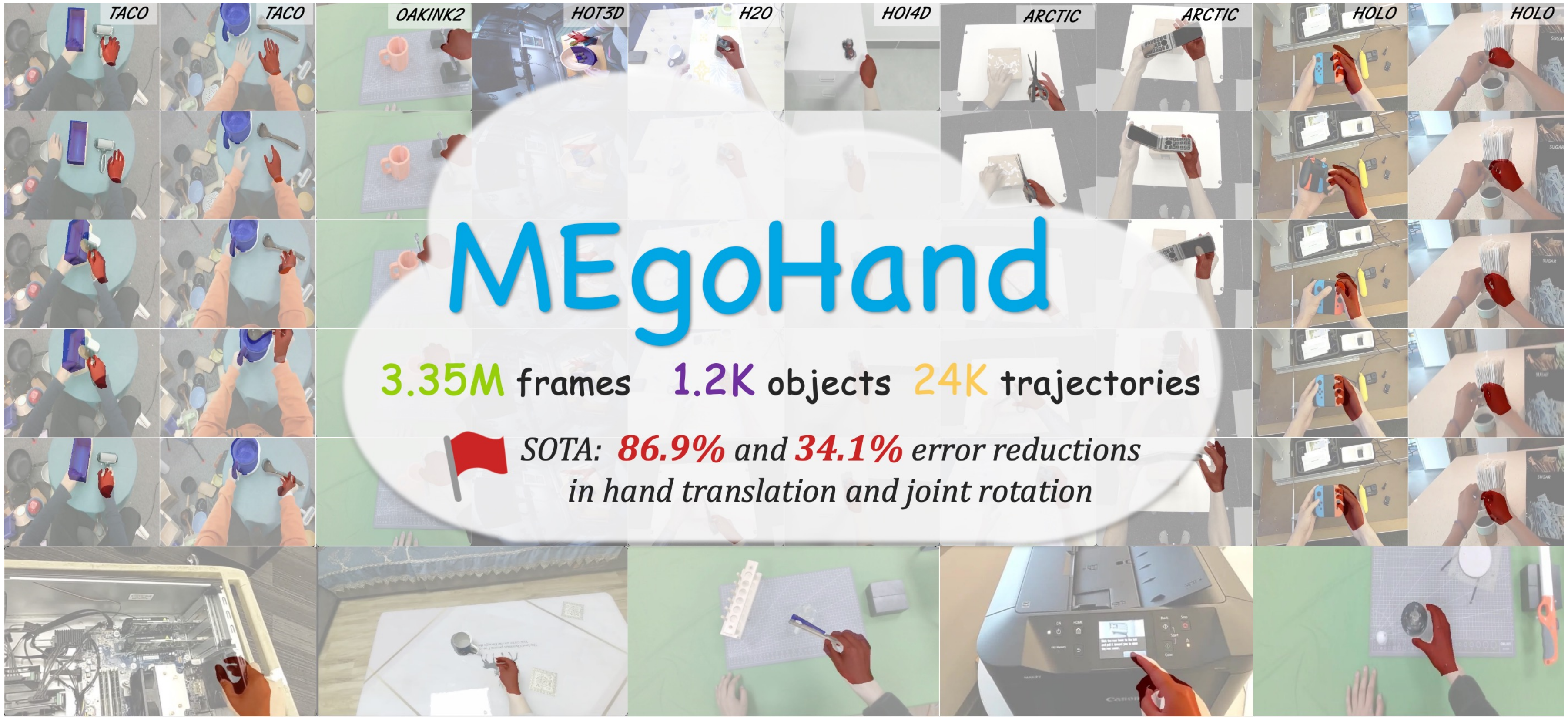
1. Transform to camera frame: $P_c = T_{cw} \cdot P_w^T$
2. Extract camera-space depth: $Z_c = (P_c)_z$
3. Project to homogeneous pixels: $\tilde{p}_{uv} = K \cdot (P_c \oslash Z_c)$
4. Convert to integer pixel indices: $p_{uv} = \pi\left(\tilde{p}_{uv}\right) = \operatorname{round}\left(\left[(\tilde{p}_{uv})_u, (\tilde{p}_{uv})_v\right]^T\right)$
5. Update depth map (for visible points): $D[v, u] = \min\left(D[v, u], Z_c^{(i)}\right)$

   where $(u, v) = p_{uv}^{(i)} \in$ image bounds and $Z_c^{(i)} > 0$

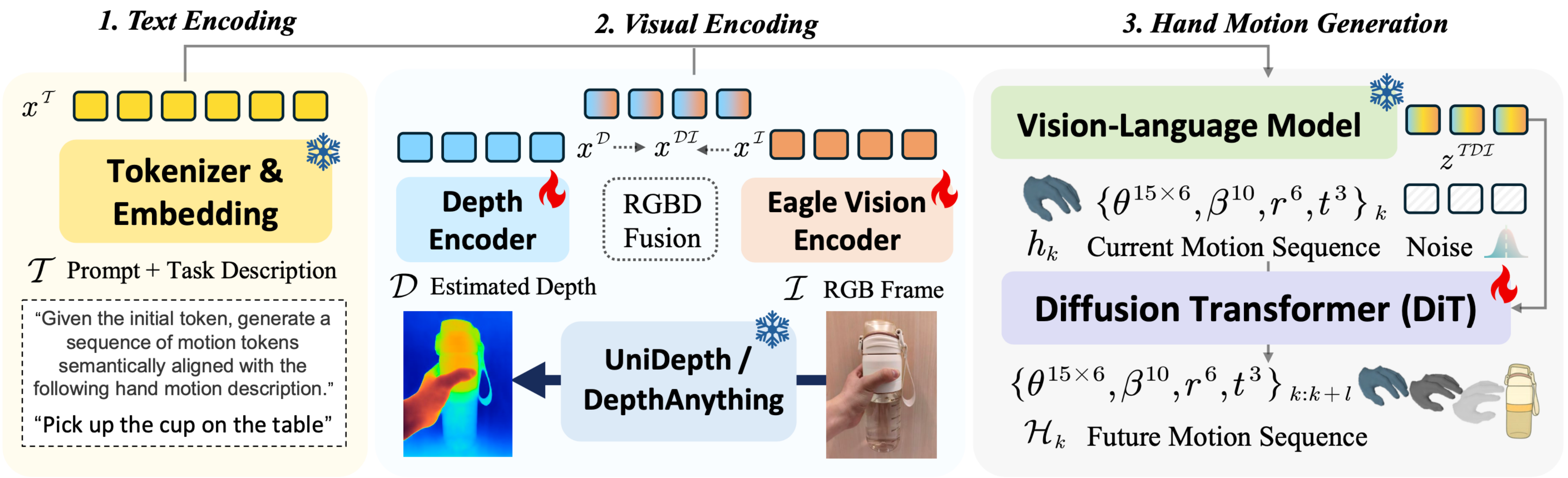
Methodology

Core Contributions:

- Large-Scale Dataset
- Architecture Design
- Decoding Strategy
- Benchmarks



Architecture Design

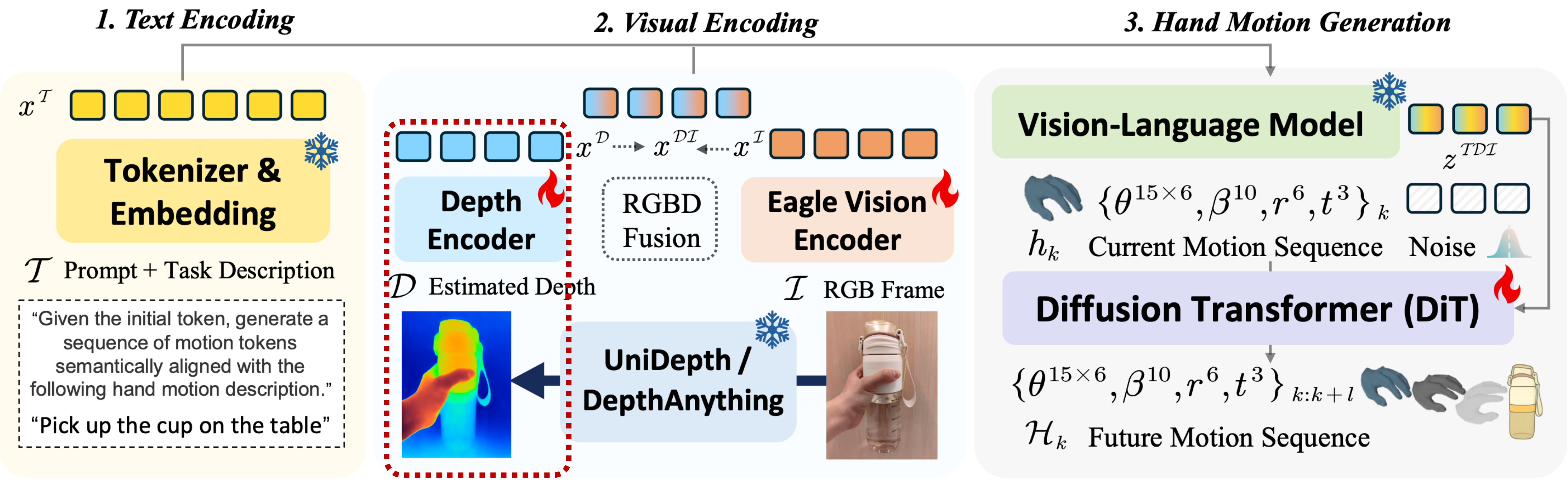


- Backbone: pretrained **Eagle-2** [1], (**SmolLM2** language + **SigLIP-2** vision encoder)
- Decomposition: high-level ("**cerebrum**") reasoning and low-level ("**cerebellum**") DiT head

[1] Li, Zhiqi, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. ARXIV 2025.

Architecture Design

北京大學 PEKING UNIVERSITY



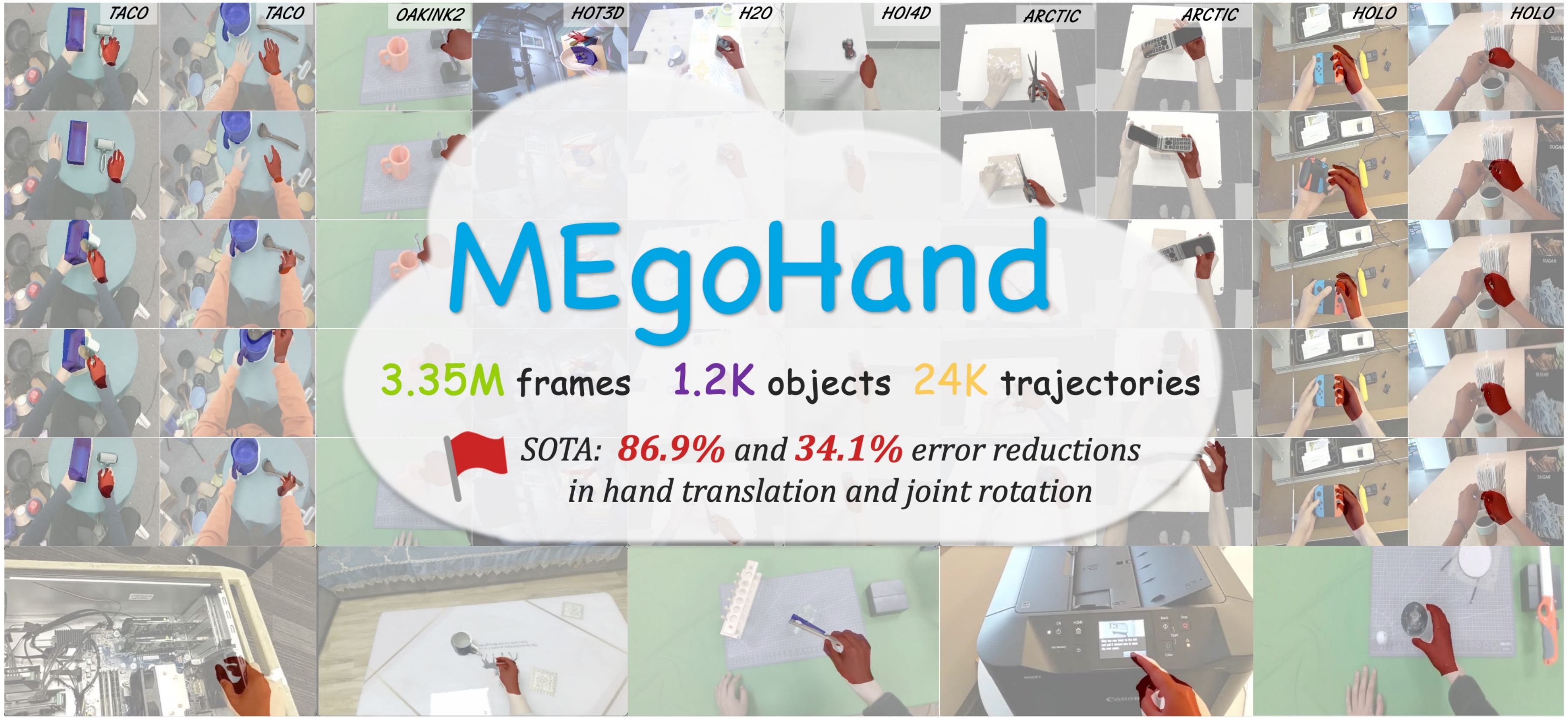
- Depth Encoder: pretrained **ResNet-50** (stack 3-channel depth image as inputs)
- Depth Input: **real / rendered** depth during training and **estimated** depth during inference
- Training Objective: conditional **flow matching** loss

$$\mathcal{L}^{\mathcal{T}}(\theta) = \mathbb{E}_{p(\mathcal{H}_k|h_k, z_k^{TDI}), q(\mathcal{H}_k^{\tau}|\mathcal{H}_k)} \left[\|\nu_\theta(\mathcal{H}_k^{\tau}, h_k, z_k^{TDI}) - \mathbf{u}(\mathcal{H}_k^{\tau}|\mathcal{H}_k)\|^2 \right] \quad \mathbf{u}(\mathcal{H}_k^{\tau}|\mathcal{H}_k) = \epsilon - \mathcal{H}_k$$

Methodology

Core Contributions:

- Large-Scale Dataset
- Architecture Design
- Decoding Strategy
- Benchmarks

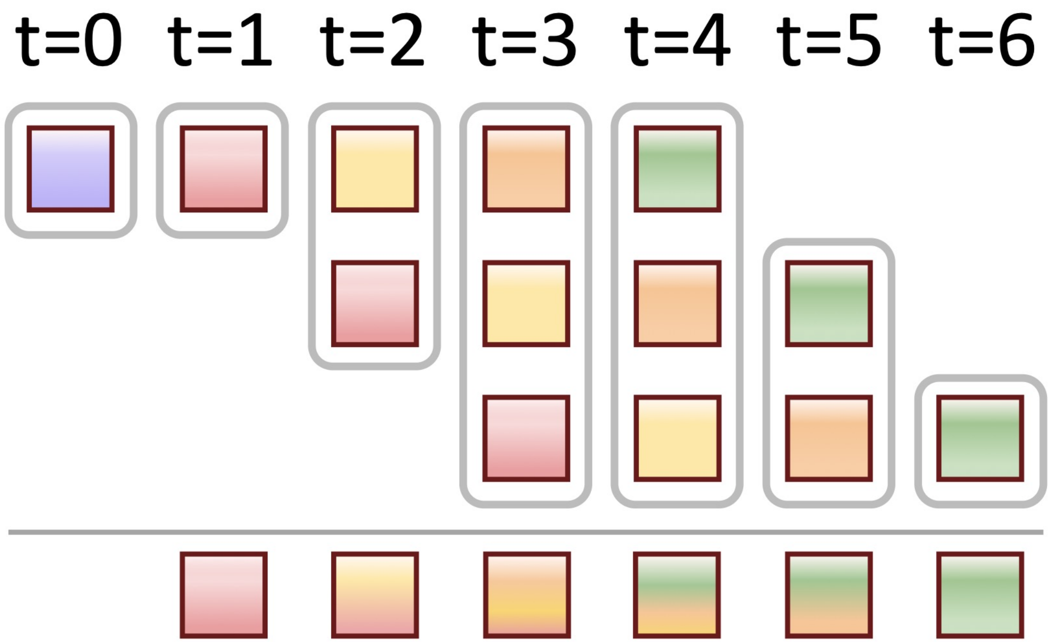


Decoding Strategy

Temporal Orthogonal Filtering (TOF)

A training-free decoding strategy to denoise predicted rotation sequences.

- **Temporal convolution** aggregates all rotation and translation estimates
- resulting convolved rotation is then projected onto the closest valid SO(3) manifold via **SVD**.



$$\tilde{R}_k = \underset{R \in \mathrm{SO}(3)}{\arg\min} \left\| R - \bar{R}_k \right\|_F = UV^\top, \quad \text{where } USV^\top = \mathrm{SVD}(\bar{R}_k), \bar{R}_k = \frac{1}{l} \sum_{t=1}^{l} \hat{R}_k^{k-t}$$

Decoding Strategy

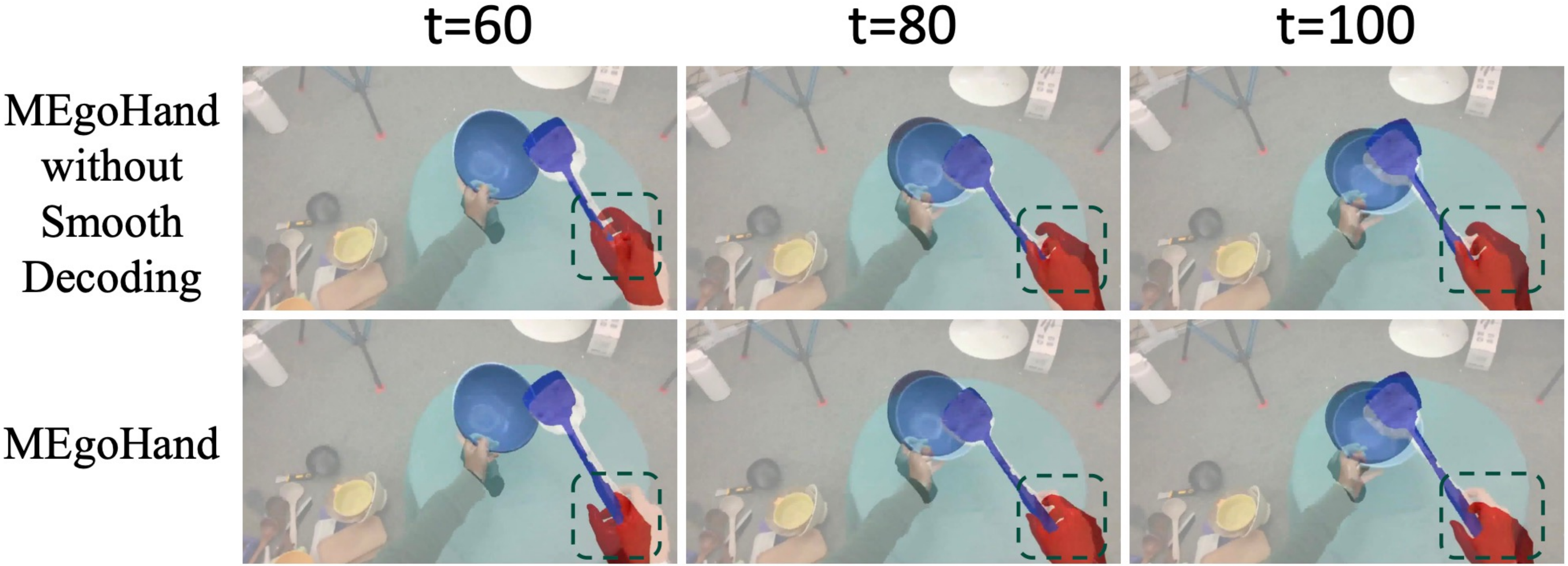


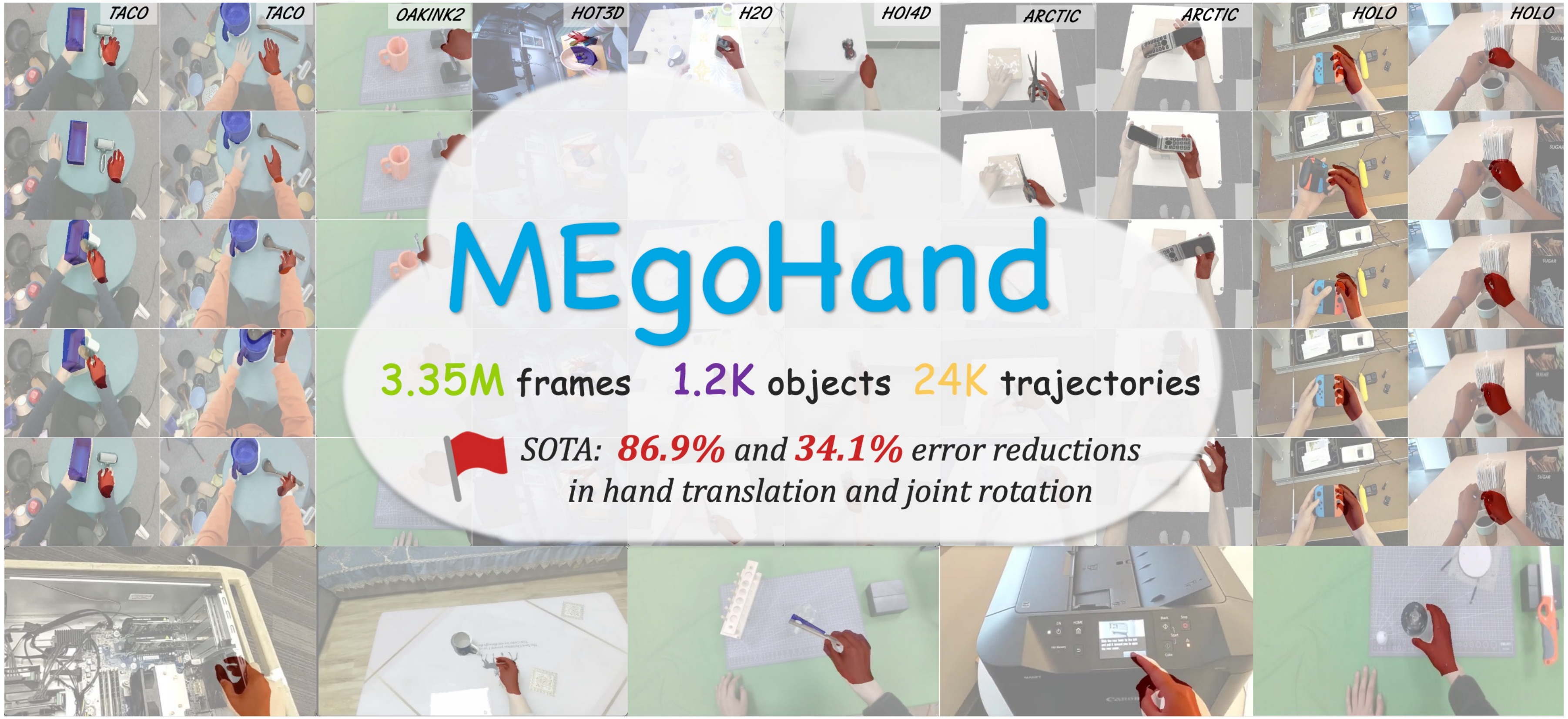
Figure 3: Frames randomly sampled from task *"Stir the bowl with spatula"* of TACO. Without decoding strategy, the predicted trajectory exhibits more fluctuations.

Methodology

北京大學
PEKING UNIVERSITY

Core Contributions:

- Large-Scale Dataset
- Architecture Design
- Decoding Strategy
- Benchmarks



In-Domain Evaluation

北京大學 PEKING UNIVERSITY

Table 1: Average metrics of in-domain evaluation across 5 datasets: TACO, HOI4D, H2O, HOT3D, and OakInk2. The unit for MRE is radians, and the remaining metrics are measured in centimeters.

Method	MPJPE↓	MPJPE-PA↓	MPVE↓	MPVE-PA↓	MWTE↓	MRE↓
LatentAct	7.726	1.478	7.696	1.453	7.221	0.937
– no concat map	8.523	1.481	8.476	1.464	7.813	0.947
LatentAct-Diff	7.819	1.498	7.787	1.483	7.322	0.941
– no concat map	8.802	1.582	8.752	1.564	8.051	0.950
MEgoHand-T	8.328	0.477	8.282	0.460	7.637	0.145
MEgoHand-I	6.269	0.480	6.120	0.457	5.521	0.143
MEgoHand-ID	5.969	0.470	5.920	0.453	5.213	0.137
MEgoHand-TI	5.683	0.476	5.632	0.459	4.889	0.136
MEgoHand (ours)	**5.425**	**0.425**	**5.381**	**0.409**	**4.756**	**0.123**

Cross-Domain Evaluation

北京大學 PEKING UNIVERSITY

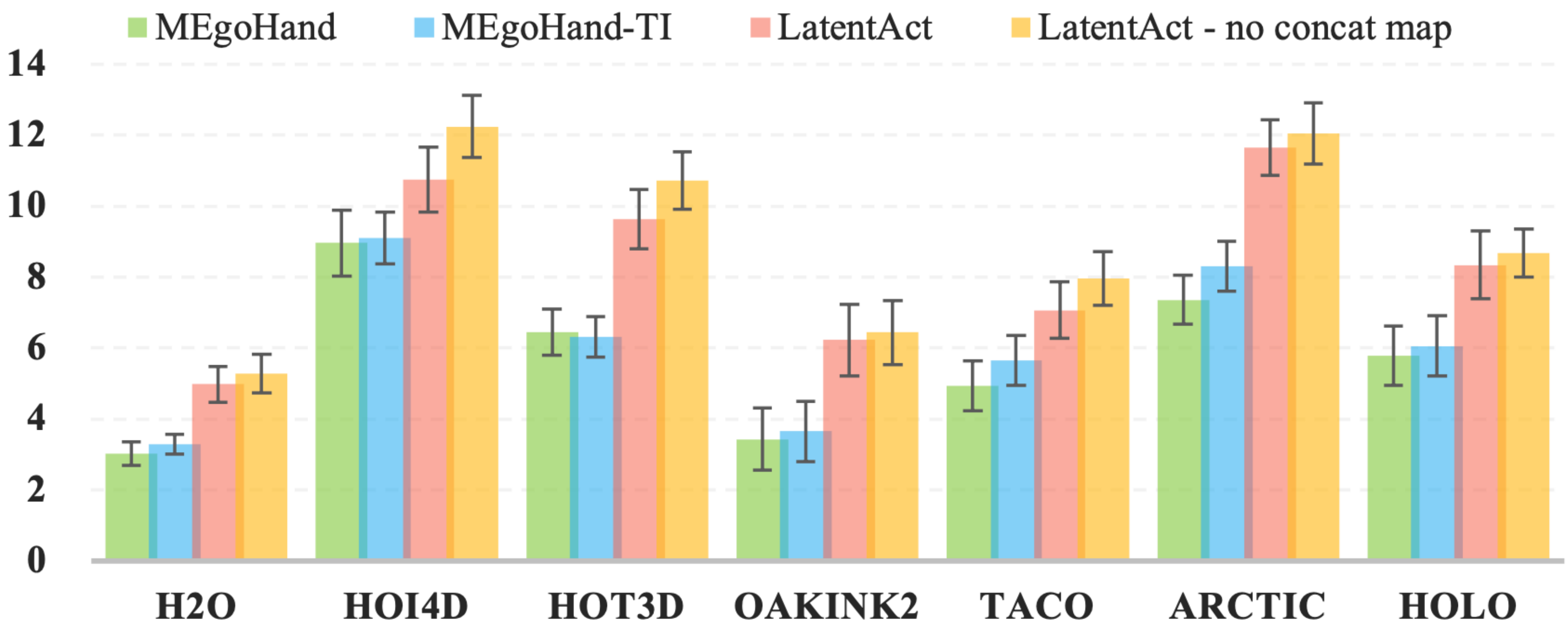


Figure 4: The evaluation of our two methods and two baseline variants on five in-domain (H2O, HOI4D, HOT3D, OAKINK2, TACO) and two cross-domain datasets (ARCTIC, HOLO), using MPJPE as metric (unit: cm, lower is better).

Ablations on Depth

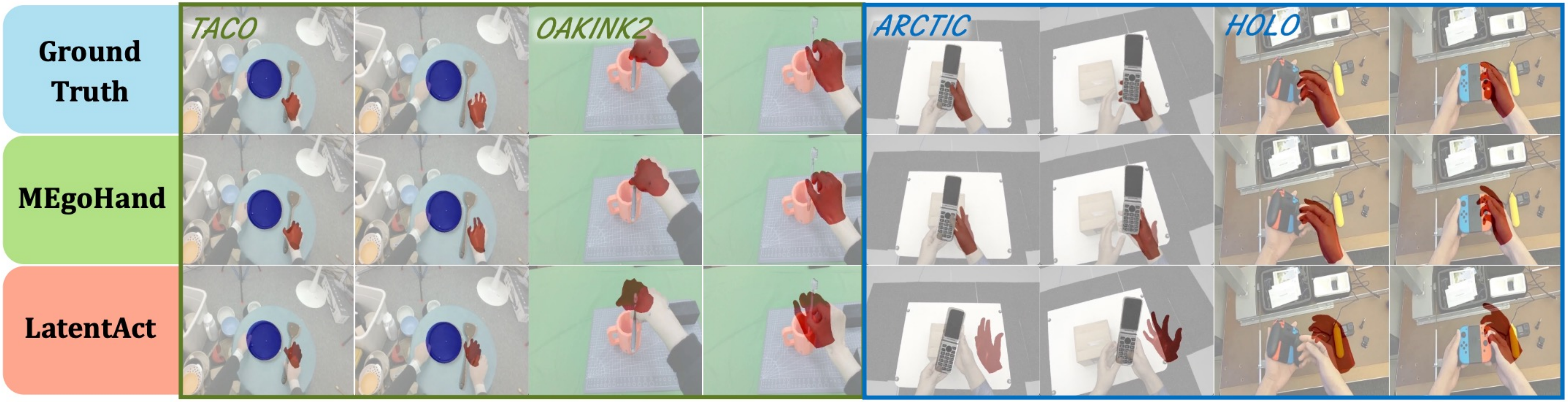
北京大學 PEKING UNIVERSITY

Dataset	Method	MPJPE↓	MPJPE-PA↓	MPVE↓	MPVE-PA↓	MWTE↓	MRE↓
Evaluation Datasets	**MEgoHand**	**5.425**	**0.425**	**5.381**	**0.409**	**4.756**	**0.123**
	– depthanythingv2	5.671	0.475	5.621	0.457	4.895	0.137
	– no depth supervision	5.725	0.492	5.671	0.473	4.900	0.142
	– relative depth	5.610	0.444	5.564	0.427	4.895	0.128
ARCTIC	**MEgoHand**	**7.358**	1.161	**7.268**	1.106	**5.958**	**0.398**
	– depthanythingv2	8.240	1.220	8.141	1.203	6.287	0.544
	– no depth supervision	8.174	1.140	8.092	1.092	6.608	0.436
	– relative depth	7.564	**1.121**	7.485	**1.091**	6.082	0.473
HOLO	**MEgoHand**	**5.775**	0.697	**5.747**	0.673	5.437	**0.271**
	– depthanythingv2	6.094	0.895	6.055	0.873	5.512	0.331
	– no depth supervision	6.434	0.835	6.397	0.837	5.889	0.473
	– relative depth	5.879	**0.663**	5.841	**0.643**	**5.418**	0.280

- MEgoHand is compatible with various depth estimators (UniDepth / DepthAnythingV2)
- Auxiliary depth supervision is imperative
- Metric depth is more sensitive to scenarios with drastic camera shifting.

Visualizations

北京大學 PEKING UNIVERSITY



MEgoHand consistently outperforms LatentAct with more accurate hand poses and finer geometric alignment, particularly in **wrist pose and finger joint rotations**. We analyze that **metric depth** inputs play an important role in the generation of higher precision.

Conclusion

北京大学
PEKING UNIVERSITY

Contributions

- Standardized dataset pipeline (3.35M frames, 1.2K objects, 24K tasks) solving annotation/representation inconsistencies.
- First framework combining VLMs and depth for egocentric HOI motion generation.
- SOTA performance on 7 datasets, with robust generalization to novel domains.

Future Extensions

- Annotate more HOI datasets with pretrained Inverse MANO network to scale up.
- Use modern hand pose detectors (e.g., HaMeR, HaWor) to label in-the-wild videos.

北京大学
PEKING UNIVERSITY

Thanks

Bohan Zhou

2025.12.03