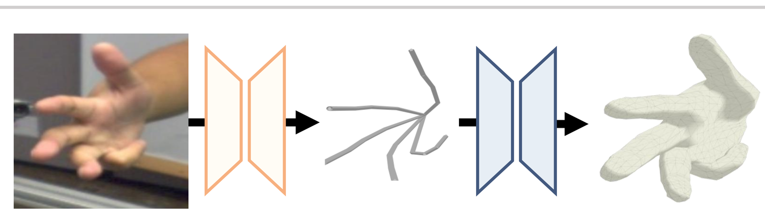
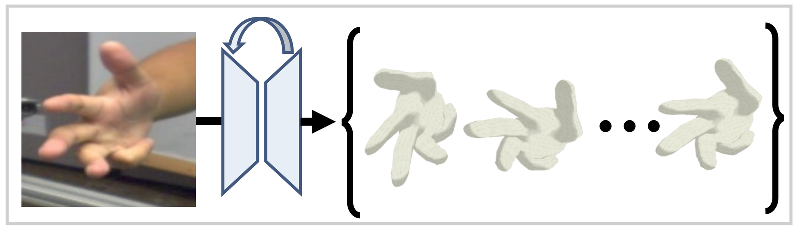
Cascaded Diffusion Framework for Probabilistic Coarse-to-Fine Hand Pose Estimation

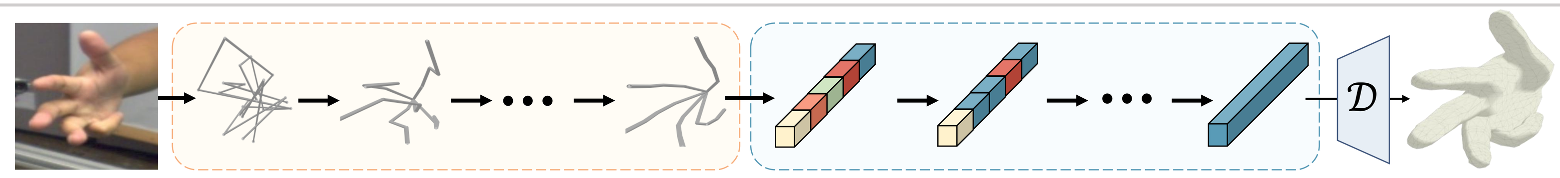
ECCV 2026 | MALMÖ | SEPT 8-12

Taeyun Woo, Jinah Park, and Tae-Kyun Kim
KAIST

Motivation

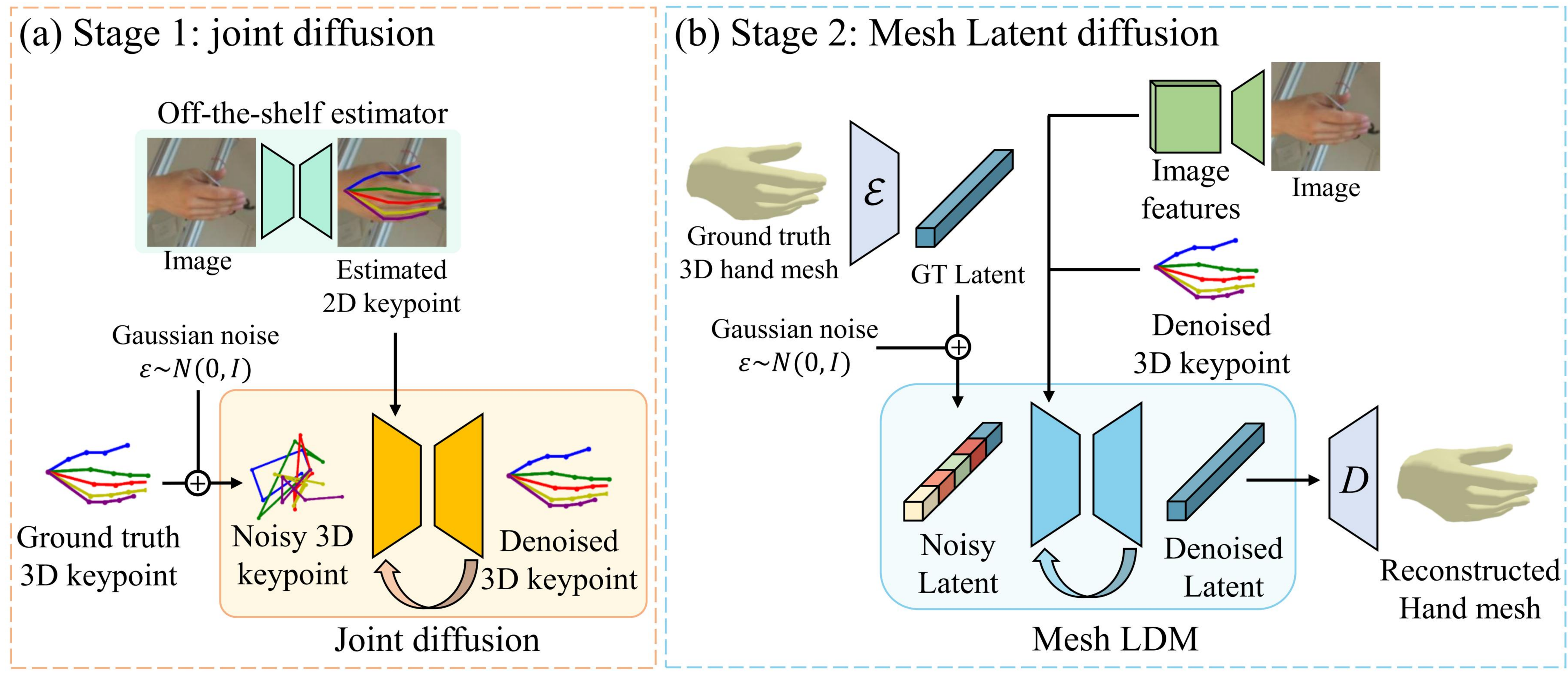

Deterministic cascaded model


Probabilistic single-stage model

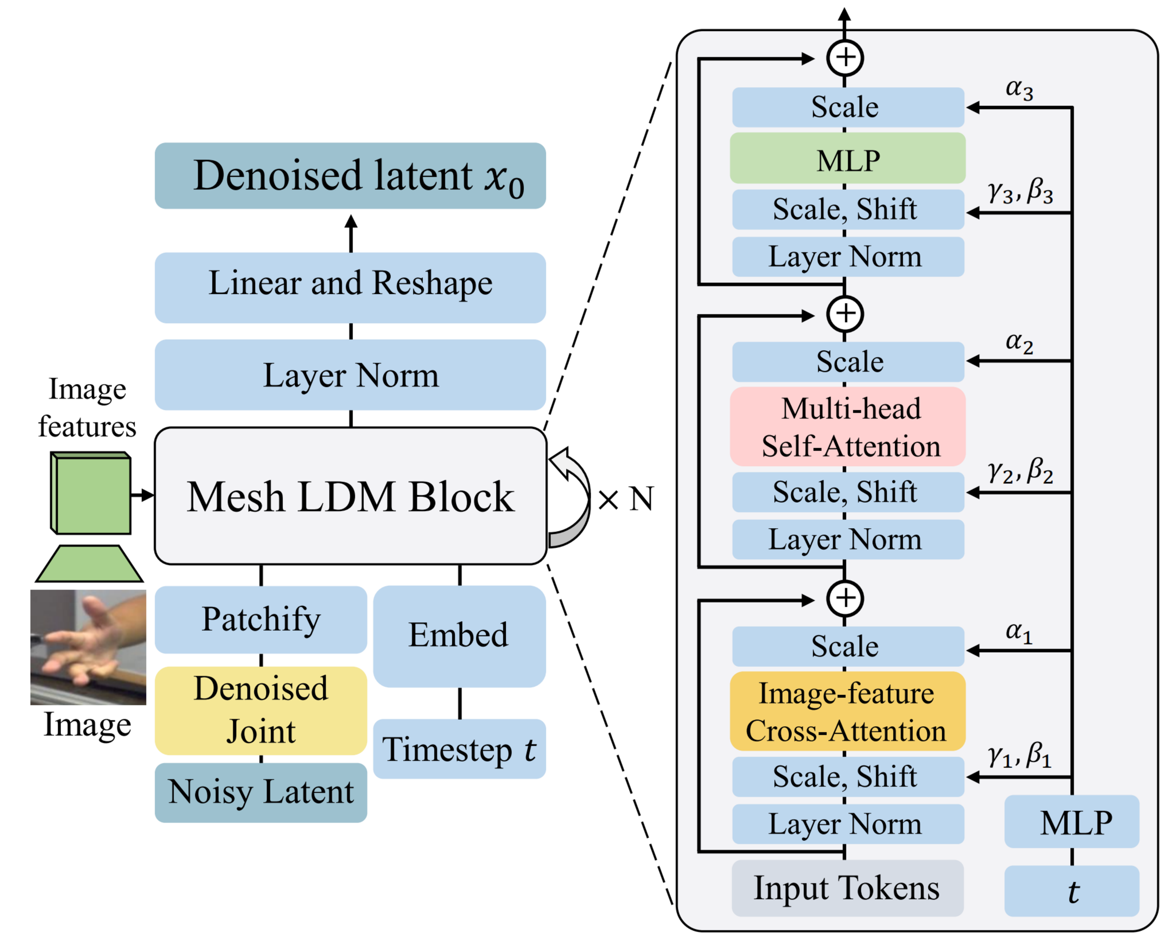

(Ours) Probabilistic cascaded model

Method	Deterministic cascaded	Probabilistic single-stage	(Ours) Cascaded diffusion
Pipeline	Coarse-to-fine	Single-step	Coarse-to-fine
Diverse hypothesis	✗	✓	✓
Refinement	Step-by-step	Lacks fine details	Step-by-step (+Latent)
Joint condition	Single (fixed)	None (image only)	Diverse
Under occlusion	Error propagates	No joint guidance	Remains stable

Method



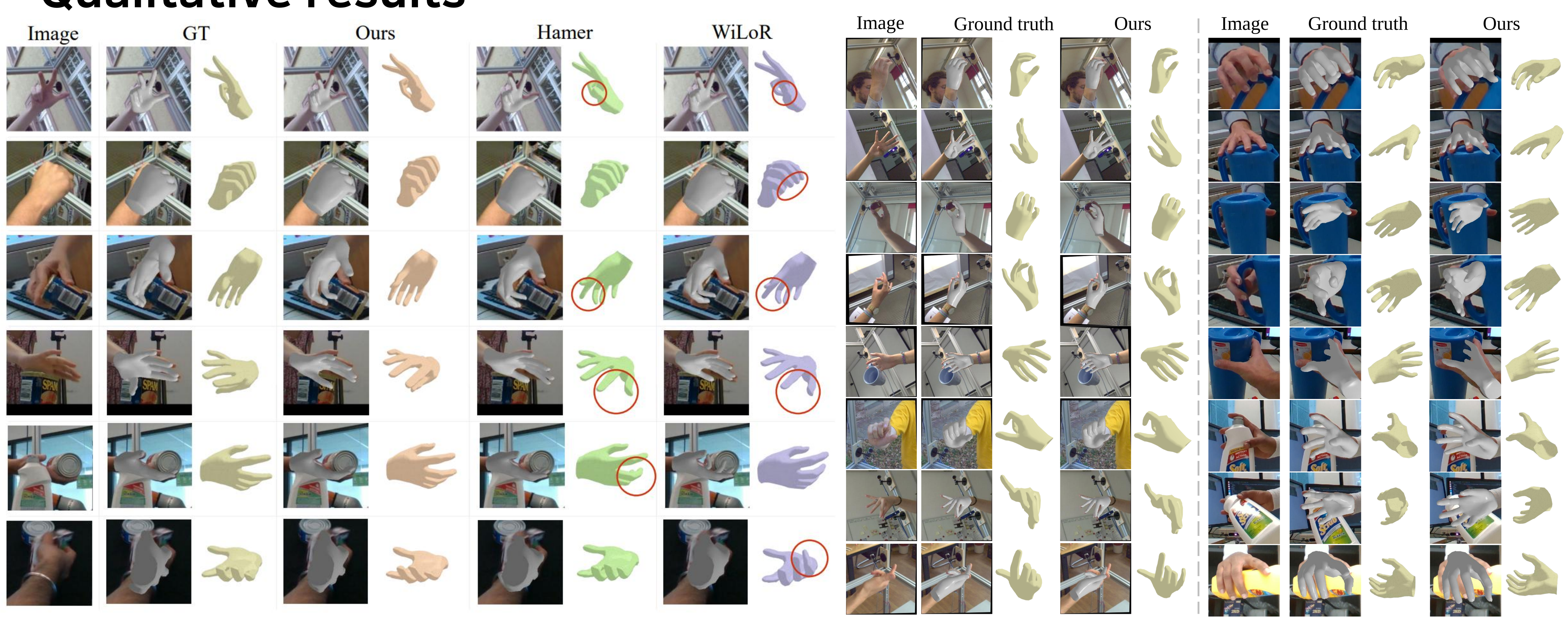
Probabilistic coarse-to-fine framework



DiT blocks

- Stage 1 (Joint Diffusion) samples diverse, plausible 3D joint hypotheses from noise, conditioned on off-the-shelf 2D keypoints.
- Stage 2 (Mesh LDM) denoises in the learned latent space using a DiT-based framework, conditioned on sampled joints and image features.
- Our key idea, distribution-aware learning, exposes the mesh stage to diverse joint hypotheses as a training condition, so reconstruction remains stable under coarse joint errors and occlusion.

Qualitative results



Quantitative results

Method	P-MPJPE ↓	P-MPVPE ↓	F@5 ↑	F@15 ↑
I2L-MeshNet [46]	7.4	7.6	0.681	0.973
Pose2Mesh [12]	7.7	7.8	0.674	0.969
I2UV-HandNet [6]	6.7	6.9	0.707	0.977
METRO [40]	6.5	6.3	0.731	0.984
Tang *et al.* [63]	6.7	6.7	0.724	0.981
Lin *et al.* [41]	5.9	6.0	0.764	0.986
MobRecon [7]	5.7	5.8	0.784	0.986
AMVUR [29]	6.2	6.1	0.767	0.987
HaMeR [48]	6.0	5.7	0.785	0.990
Hamba† [13]	5.8	5.5	0.798	0.991
WiLoR† [51]	5.5	**5.1**	**0.825**	**0.993**
HandOS† [8]	**5.0**	5.3	0.812	0.991
HHMR [37]	5.8	5.8	-	-
HHMR (best) [37]	5.3	5.4	-	-
Proposed	**5.0**	5.2	0.816	0.992
Proposed (best)	4.2	4.5	0.866	0.995

FreiHAND

Method	P-MPJPE ↓	P-MPVPE ↓	F@5 ↑	F@15 ↑
Hasson *et al.* [23]	11.0	11.2	0.464	0.939
Hampali *et al.* [20]	10.7	10.6	0.506	0.942
I2L-MeshNet [47]	11.2	13.0	0.409	0.932
Pose2Mesh [12]	12.5	12.7	0.441	0.909
Liu *et al.* [43]	9.9	9.5	0.528	0.956
I2UV-HandNet [6]	9.9	10.1	0.500	0.943
METRO [41]	10.4	11.1	0.484	0.946
ArtiBoost [70]	11.4	10.9	0.488	0.944
MobRecon [7]	9.2	9.4	0.538	0.957
Keypoint Trans. [21]	10.8	-	-	-
HandOccNet [48]	9.1	8.8	0.564	0.963
AMVUR [30]	8.3	8.2	0.608	0.965
HaMeR [49]	7.7	7.9	0.635	0.980
Hamba† [13]	**7.5**	7.7	**0.648**	0.982
WiLoR† [52]	**7.5**	7.7	0.646	**0.983**
Proposed	**7.5**	**7.5**	0.633	0.982
Proposed (best)	7.4	7.4	0.639	0.982

HO3Dv2

Method	P-MPJPE ↓	P-MPVPE ↓	AUC_J ↑
Spurr *et al.* [64]	6.8	–	0.864
MobRecon [7]	6.4	5.6	–
HandOccNet [48]	5.8	5.5	–
H2ONet [68]	5.7	5.5	–
Zhou *et al.* [74]	5.5	5.5	–
WiLoR [52]	4.9	4.8	0.902
HandOS [8]	5.2	5.0	0.896
Proposed	**4.5**	**4.5**	**0.911**
Proposed (best)	4.2	4.2	0.917

DexYCB

- Competitive across FreiHAND, HO3Dv2; strongest on DexYCB (4.5 P-MPJPE / 0.911 AUC)
- "Proposed (best)" rows are oracle upper bounds (GT-selected among N samples), shown for analysis..
- Averaged prediction matches or approaches SOTA without ground-truth selection.

Ablation studies

Joint condition	P-MPJPE ↓	P-MPVPE ↓	F@5 ↑	F@15 ↑
off-the-shelf [52]	5.1	5.4	0.800	0.991
Averaged joint	5.1	5.4	0.805	0.992
Proposed	5.0	5.2	0.816	0.992
Proposed (best)	4.2	4.5	0.866	0.995

Diffusion space	FreiHAND		HO3Dv2	
	P-MPJPE ↓	P-MPVPE ↓	P-MPJPE ↓	P-MPVPE ↓
MANO	5.70	5.82	7.81	7.81
Ours (latent)	5.00	5.23	7.50	7.52

- Joint condition: Diverse joints improve over single condition → diversity as training signal works
- Diffusion space: Latent-space diffusion outperforms direct MANO-space → latent is more effective

Conclusions

- Coarse-to-fine cascaded diffusion: joint diffusion (Stage 1) + Mesh LDM (Stage 2)
- Joint diffusion (Stage 1) generates diverse 3D joint hypotheses; Mesh LDM (Stage 2) reconstructs mesh conditioned on a joint
- Diverse hypotheses used as a training signal (not final outputs) → distribution-aware joint-mesh learning.
- Competitive on FreiHAND & HO3Dv2, best on DexYCB; remains stable under occlusion & pose ambiguity.

Acknowledgements: This work was supported by NST grant (CRC 21015, CRC23023-000, MSIT), IITP grant (RS-2023-00228996, RS-2024- 00459749, RS-2025-25443318, RS-2025- 25441313, RS-2026-25526850, RS-2026-25522885, MSIT) , KOCCA grant (RS- 2024-00442308, MCST) and InnoCORE program (N10260110, MSIT)

Paper

Homepage