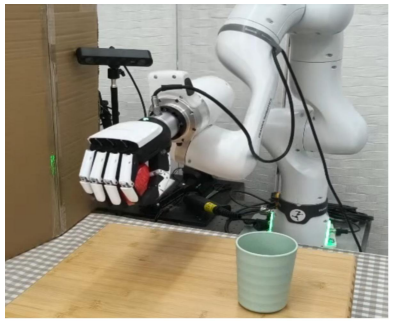
上海科技大学
ShanghaiTech University

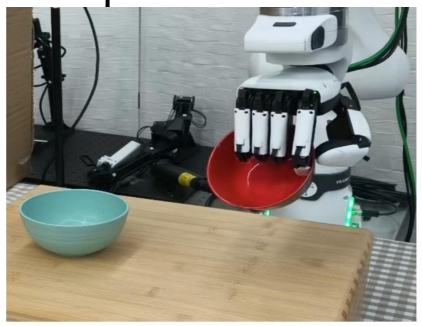
ICLR

UniHM: Unified Dexterous Hand Manipulation with Vision Language Model

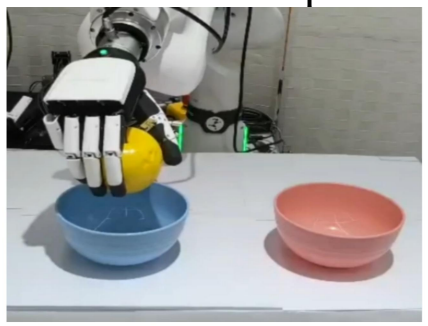
Zhenhao Zhang[1,2*], Jiaxin Liu[1*], Ye Shi[1,2†], Jingya Wang[1,2†]
[1]ShanghaiTech University [2]InstAdapt
[*]Indicates Equal Contribution [†]Indicates Corresponding Author



Grab the apple!

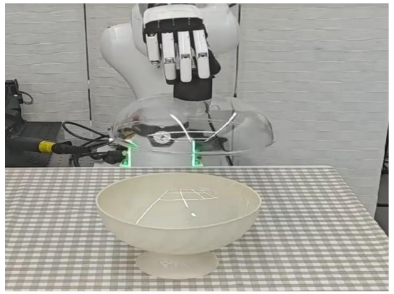


Grab the bowl!



Pick&Place the fruit!

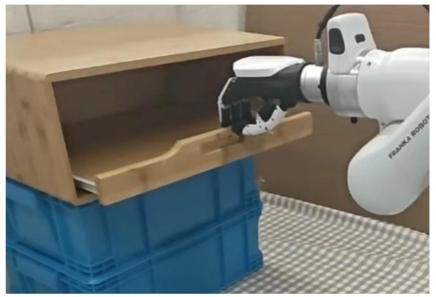
Pick&Place the bottle!



Open the lid!

Close the cabinet!

Pull the drawer!



Push the drawer!

Overview

+

"Grab the power drill."

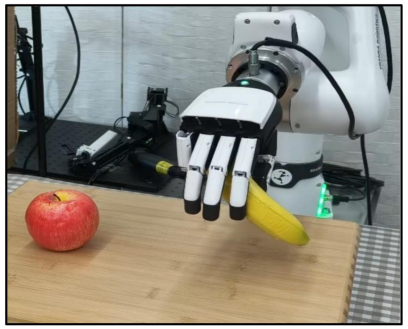
"Pick up the cracker box."

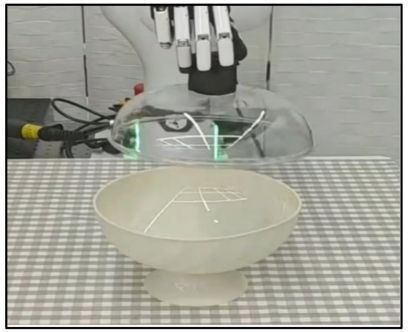
"Use the mustard bottle."

Training on **Closed-set** HOI Datasets.

UniHM

Generalize to **Open-world** Tasks!





Contributions

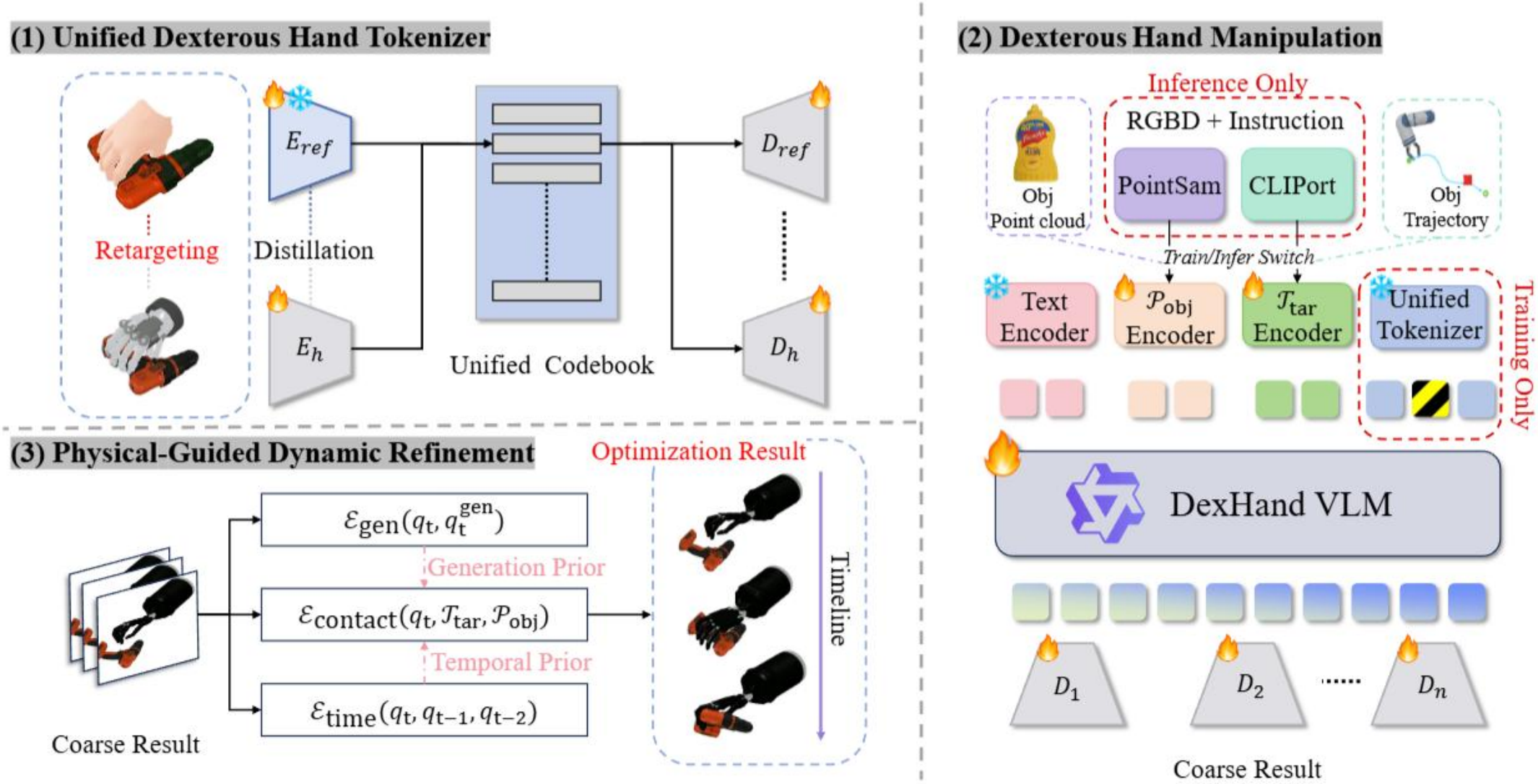
- We introduce **UniHM**, the first unified, language-conditioned framework for dynamic dexterous hand manipulation beyond static grasps directly from images and open-vocabulary instructions.

- We introduce a **unified codebook** with cross dexterous hand consistency that maps heterogeneous hand kinematics into one discrete action lattice and decodes tokens into hand-specific joint trajectories, which enables direct token reuse and transfer across robotic and anthropomorphic hands.

- We employ a tailored physical-guided refinement stack that fuses a **generative hoi prior** shaping feasible pose manifolds, a **temporal prior** enforcing smooth velocity–acceleration profiles and retime-aware consistency, and contact-aware dynamic trajectory optimization to optimize the generation result.

- Our framework eliminates the dependency on expensive teleoperation data by learning dexterous manipulation skills from **human videos**. This paradigm achieves robust generalization to unseen scenes and instructions, significantly lowering the barrier to developing dexterous manipulation systems.

Pipeline


(1) Unified Dexterous Hand Tokenizer
Retargeting
Distillation
E_{ref}
E_h
Unified Codebook
D_{ref}
D_h
(2) Dexterous Hand Manipulation
Inference Only
RGBD + Instruction
PointSam
CLIPort
Obj Point cloud
Obj Trajectory
Train/Infer Switch
Text Encoder
$\mathcal{P}_{obj}$ Encoder
$\mathcal{T}_{tar}$ Encoder
Unified Tokenizer
Training Only
DexHand VLM
D_1
D_2
D_n
Coarse Result
(3) Physical-Guided Dynamic Refinement
Optimization Result
$\mathcal{E}_{gen}(q_t, q_t^{gen})$
Generation Prior
$\mathcal{E}_{contact}(q_t, \mathcal{T}_{tar}, \mathcal{P}_{obj})$
Temporal Prior
$\mathcal{E}_{time}(q_t, q_{t-1}, q_{t-2})$
Timeline
Coarse Result

Experiments(Main Result)

Table 1: Main Result on DexYCB

	Method	MPJPE↓	FOL↓	FPL↓	FID ↓	Diversity →
	GT	-	-	-	-	125.53
Seen	TM2T(Guo et al., 2022)	$85.33^{\pm 3.41}$	$36.57^{\pm 1.46}$	$24.10^{\pm 0.96}$	$54.83^{\pm 2.19}$	$37.12^{\pm 1.48}$
Seen	MDM(Tevet et al., 2023)	$88.06^{\pm 3.52}$	$33.40^{\pm 1.34}$	$23.06^{\pm 0.92}$	$52.33^{\pm 2.09}$	$33.95^{\pm 1.36}$
Seen	FlowMDM(Barquero et al., 2024)	$82.75^{\pm 3.31}$	$31.24^{\pm 1.25}$	$21.55^{\pm 0.86}$	$48.05^{\pm 1.92}$	$61.25^{\pm 2.45}$
Seen	MotionGPT3(Zhu et al., 2025)	$74.80^{\pm 2.99}$	$28.76^{\pm 1.15}$	$19.32^{\pm 0.77}$	$43.35^{\pm 1.73}$	$\mathbf{72.51}^{\pm 2.90}$
Seen	**Ours**	$\mathbf{61.40}^{\pm 1.93}$	$\mathbf{23.14}^{\pm 0.65}$	$\mathbf{12.15}^{\pm 0.24}$	$\mathbf{31.24}^{\pm 1.02}$	$39.62^{\pm 0.66}$
Unseen	TM2T(Guo et al., 2022)	$94.22^{\pm 3.77}$	$37.25^{\pm 1.49}$	$27.03^{\pm 1.08}$	$55.94^{\pm 2.24}$	$31.25^{\pm 1.25}$
Unseen	MDM(Tevet et al., 2023)	$93.05^{\pm 3.72}$	$39.04^{\pm 1.56}$	$25.89^{\pm 1.04}$	$55.13^{\pm 2.21}$	$29.0^{\pm 1.16}$
Unseen	FlowMDM(Barquero et al., 2024)	$86.13^{\pm 3.45}$	$32.67^{\pm 1.31}$	$24.09^{\pm 0.96}$	$51.33^{\pm 2.05}$	$58.21^{\pm 2.33}$
Unseen	MotionGPT3(Zhu et al., 2025)	$77.93^{\pm 3.12}$	$30.55^{\pm 1.22}$	$21.48^{\pm 0.86}$	$46.14^{\pm 1.85}$	$\mathbf{75.84}^{\pm 3.03}$
Unseen	**Ours**	$\mathbf{63.56}^{\pm 2.08}$	$\mathbf{27.29}^{\pm 0.43}$	$\mathbf{13.06}^{\pm 0.43}$	$\mathbf{41.03}^{\pm 1.65}$	$42.70^{\pm 1.19}$

Table 2: Main Result on OakInk

	Method	MPJPE↓	FOL↓	FPL↓	FID ↓	Diversity →
	GT	-	-	-	-	147.40
Seen	TM2T(Guo et al., 2022)	$71.08^{\pm 2.84}$	$91.25^{\pm 3.65}$	$34.51^{\pm 1.38}$	$311.90^{\pm 12.48}$	$277.38^{\pm 11.10}$
Seen	MDM(Tevet et al., 2023)	$67.55^{\pm 2.70}$	$93.8^{\pm 3.75}$	$30.06^{\pm 1.20}$	$285.22^{\pm 11.41}$	$275.42^{\pm 11.02}$
Seen	FlowMDM(Barquero et al., 2024)	$60.74^{\pm 2.43}$	$85.43^{\pm 3.42}$	$26.47^{\pm 1.06}$	$249.08^{\pm 9.96}$	$189.54^{\pm 7.58}$
Seen	MotionGPT3(Zhu et al., 2025)	$56.29^{\pm 2.25}$	$79.24^{\pm 3.17}$	$23.98^{\pm 0.96}$	$221.10^{\pm 8.84}$	$247.10^{\pm 9.88}$
Seen	**Ours**	$\mathbf{52.73}^{\pm 2.08}$	$\mathbf{72.32}^{\pm 0.55}$	$\mathbf{19.86}^{\pm 1.38}$	$\mathbf{204.91}^{\pm 7.64}$	$\mathbf{165.47}^{\pm 6.30}$
Unseen	TM2T(Guo et al., 2022)	$75.34^{\pm 3.01}$	$125.33^{\pm 5.01}$	$45.51^{\pm 1.82}$	$337.08^{\pm 13.48}$	$362.08^{\pm 14.48}$
Unseen	MDM(Tevet et al., 2023)	$72.90^{\pm 2.92}$	$112.94^{\pm 4.52}$	$42.93^{\pm 1.72}$	$325.58^{\pm 13.02}$	$354.93^{\pm 14.20}$
Unseen	FlowMDM(Barquero et al., 2024)	$65.39^{\pm 2.62}$	$101.25^{\pm 4.05}$	$36.14^{\pm 1.45}$	$298.04^{\pm 11.92}$	$224.67^{\pm 8.99}$
Unseen	MotionGPT3(Zhu et al., 2025)	$61.95^{\pm 2.48}$	$93.65^{\pm 3.75}$	$28.25^{\pm 1.13}$	$272.69^{\pm 10.91}$	$316.58^{\pm 12.66}$
Unseen	**Ours**	$\mathbf{58.62}^{\pm 2.35}$	$\mathbf{83.27}^{\pm 1.17}$	$\mathbf{22.87}^{\pm 0.52}$	$\mathbf{253.41}^{\pm 13.05}$	$\mathbf{153.28}^{\pm 9.48}$

Experiments(Real-World Experiments and Ablation Study)

Table 3: Real-World Experiments

Split	Method	Success Rate			
		Grab	Pick&Place	Pull&Push	Open&Close
Seen	MDM+Dex-Retargeting	20%	10%	0%	5%
	MotionGPT3+Dex-Retargeting	30%	15%	25%	25%
	Ours	**65%**	**50%**	**60%**	**55%**
Unseen	MDM+Dex-Retargeting	5%	0%	0%	5%
	MotionGPT3+Dex-Retargeting	45%	25%	15%	20%
	Ours	**60%**	**35%**	**55%**	**45%**

Table 4: Ablation Result on DexYCB

	Method	MPJPE↓	FOL↓	FPL↓	FID ↓	Diversity →
	GT	-	-	-	-	125.53
Seen	w/o Depth Input	$85.47^{\pm 3.42}$	$33.41^{\pm 1.34}$	$20.97^{\pm 0.84}$	$56.36^{\pm 2.25}$	$66.40^{\pm 2.66}$
	w/o Masked Training	$73.41^{\pm 2.94}$	$28.22^{\pm 1.13}$	$14.42^{\pm 0.58}$	$44.87^{\pm 1.79}$	$\mathbf{73.09}^{\pm 2.92}$
	w/o Physical Refinement	$65.78^{\pm 2.63}$	$25.06^{\pm 1.00}$	$15.35^{\pm 0.61}$	$33.57^{\pm 1.34}$	$38.06^{\pm 1.52}$
	Ours	$\mathbf{61.40}^{\pm 1.93}$	$\mathbf{23.14}^{\pm 0.65}$	$\mathbf{12.15}^{\pm 0.24}$	$\mathbf{31.24}^{\pm 1.02}$	$39.62^{\pm 0.66}$
Unseen	w/o Depth Input	$90.12^{\pm 3.60}$	$39.77^{\pm 1.59}$	$21.70^{\pm 0.87}$	$77.38^{\pm 3.10}$	$67.53^{\pm 2.70}$
	w/o Masked Training	$74.63^{\pm 2.99}$	$28.08^{\pm 1.12}$	$17.25^{\pm 0.69}$	$43.09^{\pm 1.72}$	$\mathbf{74.88}^{\pm 3.00}$
	w/o Physical Refinement	$65.39^{\pm 2.62}$	$28.55^{\pm 1.14}$	$16.05^{\pm 0.64}$	$45.06^{\pm 1.80}$	$41.03^{\pm 1.64}$
	Ours	$\mathbf{63.56}^{\pm 2.08}$	$\mathbf{27.29}^{\pm 0.43}$	$\mathbf{13.06}^{\pm 0.43}$	$\mathbf{41.03}^{\pm 1.65}$	$42.70^{\pm 1.19}$