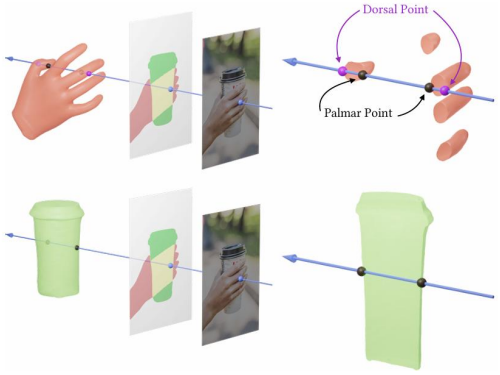
OpenHOI: Open-World Hand-Object Interaction Synthesis with Multimodal Large Language Model

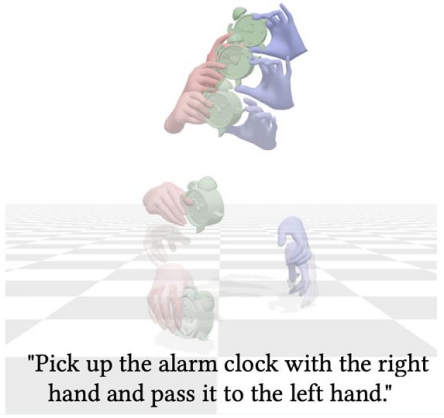
Zhenhao Zhang[1], Ye Shi[1*], Lingxiao Yang[1], Suting Ni[1], Qi Ye[2], Jingya Wang[1*]
[1]ShanghaiTech University [2]Zhejiang University
[*]Indicates Corresponding Author

Background: Hand-Object Interaction

GRAB[1]: Hand-Object Interaction Dataset



EasyHOI[2]: Hand-Object Interaction Reconstruct

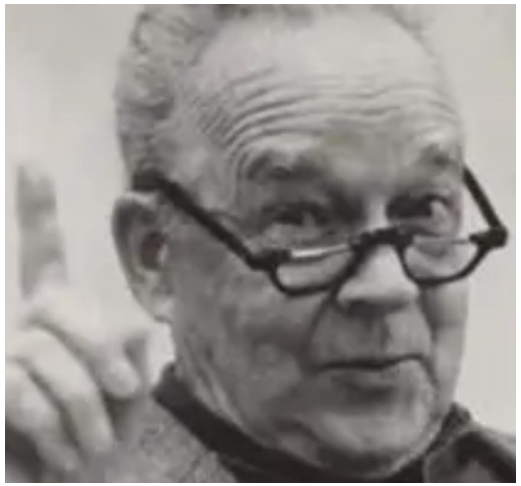


DiffH2O[3]: Hand-Object Interaction Generation

Hand-object Interaction (HOI) involves jointly modeling hand articulation and object dynamics to generate and interpret realistic manipulation sequences. This reflects one of the ***most pervasive*** human behaviors, deeply embedded in daily activities.

[1]Taheri, Omid, et al. "GRAB: A dataset of whole-body human grasping of objects." European conference on computer vision. Cham: Springer International Publishing, 2020.
[2]Liu, Yumeng, et al. "EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
[3]Christen, Sammy, et al. "Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions." SIGGRAPH Asia 2024 Conference Papers. 2024.

Background: Affordance

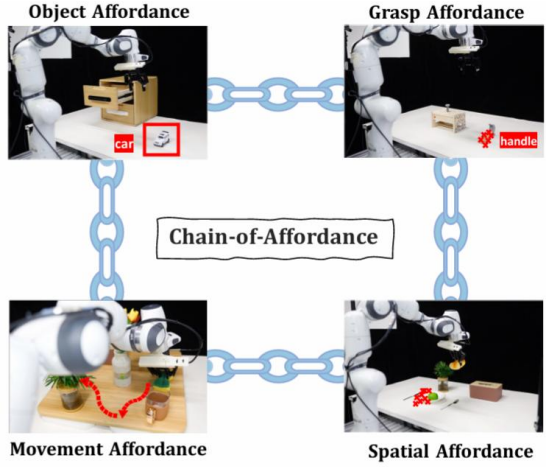


James J. Gibson

Donald Norman

Affordance was first proposed by cognitive psychologist ***James J. Gibson*** in 1979, and was later introduced to the field of human-computer interaction (HCI) by ***Donald Norman*** in 1988. Affordance refers to the property of an object that visually suggests its possible uses—that is, it provides cues about "***how it can be used.***". Nowadays, Affordance is widely used in ***Lots of Area***.

Background: Affordance Application in Embodied Interaction



CoA-VLA[1]: Affordance as bbox

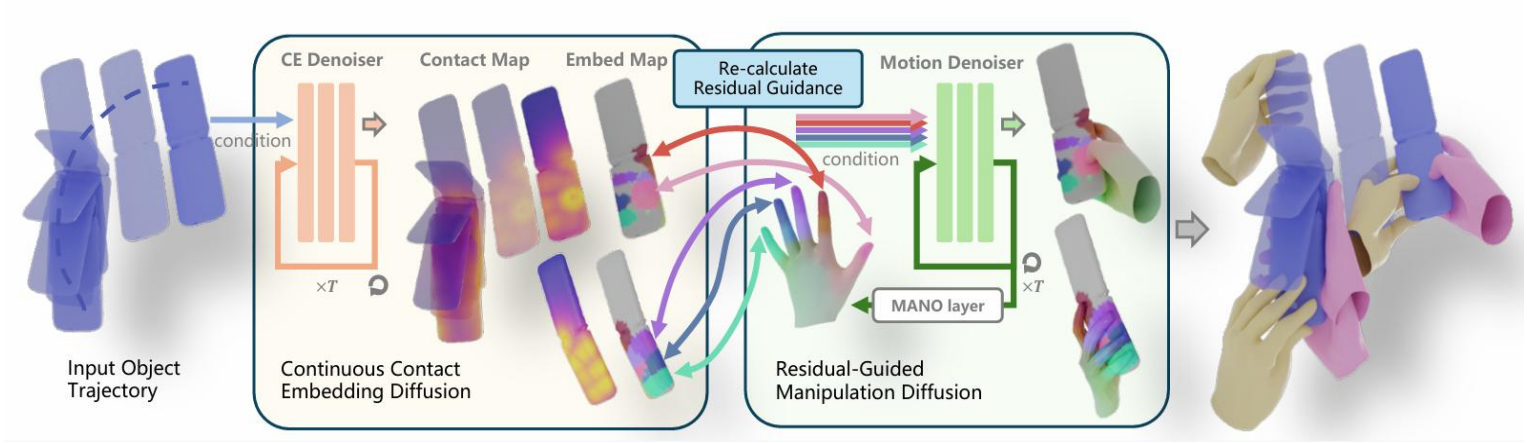
GLOVER++[2]: Affordance as a keypoint

SeqAfford[3]: Affordance as a 3D Mask

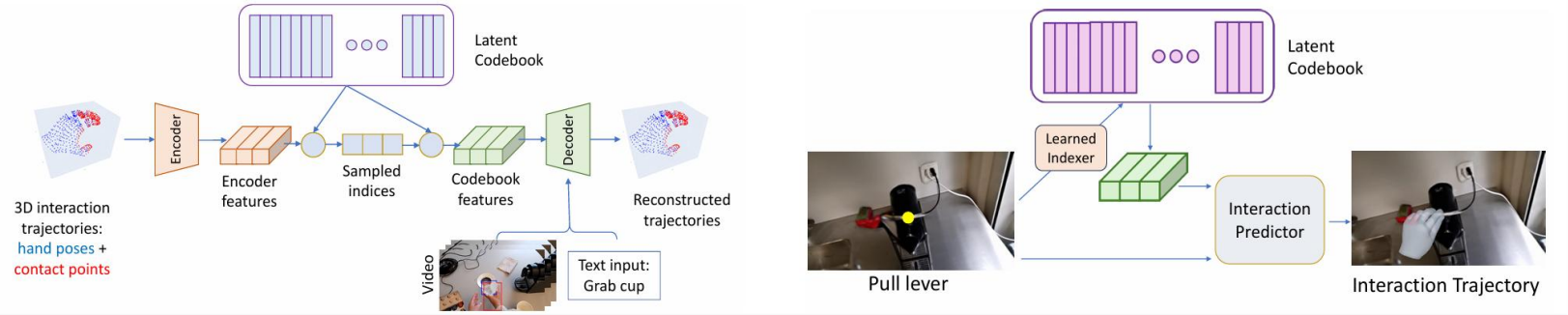
Affordance serve as a key in the area of ***Embodied AI***. Affordance can be expressed as bounding box, keypoints or 3D masks in the area of interaction. It is a ***powerful interaction prior*** for lots of manipulation tasks such as VLA[1], Learning from Video[2] and 3D Interaction[3]

[1]Li, Jinming, et al. "CoA-VLA: Improving Vision-Language-Action Models via Visual-Text Chain-of-Affordance." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.
[2]Ma, Teli, et al. "GLOVER++: Unleashing the Potential of Affordance Learning from Human Behaviors for Robotic Manipulation." arXiv preprint arXiv:2505.11865 (2025).
[3]Yu, Chunlin, et al. "Seqafford: Sequential 3d affordance reasoning via multimodal large language model." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Realted Works: Hand Motion Synthesis



ManiDext[1]: Generate Hand Motion with ***Object Trajectoty*** as Condition

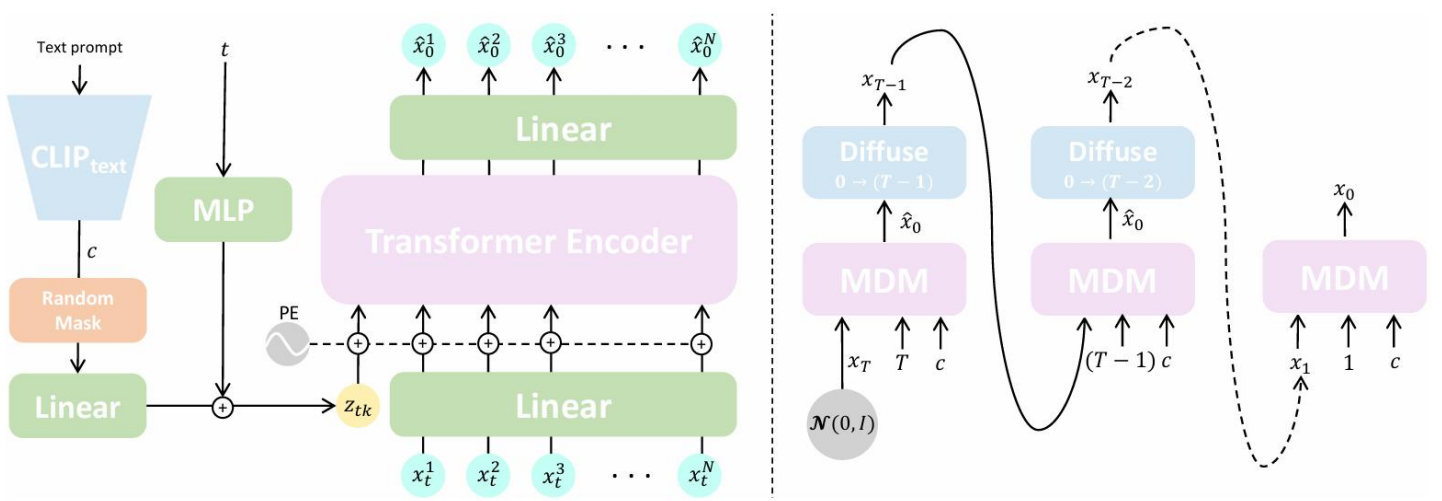


LatentAct[2]: Synthesizing 3D Hand Motion and Contacts from ***Video***

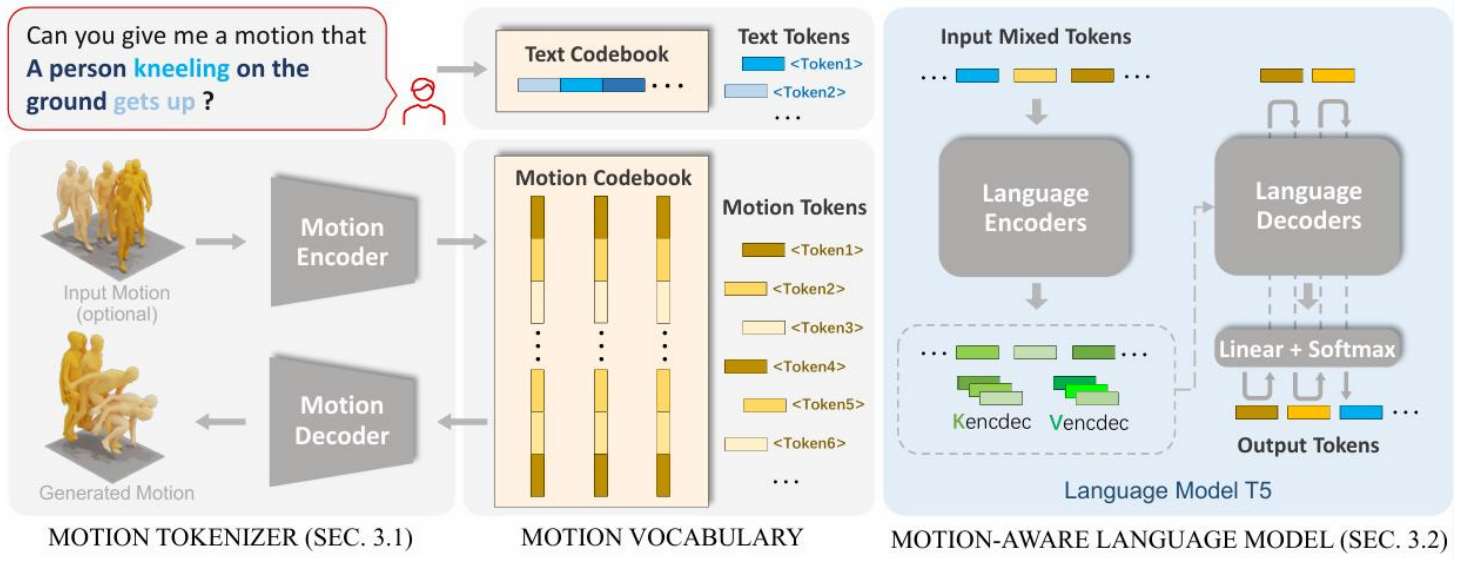
Can We Synthesis Hand Motion with language guided?

[1]Zhang, Jiajun, et al. "Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion." IEEE Transactions on Pattern Analysis and Machine Intelligence.
[2]Prakash, Aditya, et al. "How do i do that? synthesizing 3d hand motion and contacts for everyday interactions." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Realted Works: Text to Human Motion Generation



MDM[1]: Generate Human Motion with ***Text*** Prompt

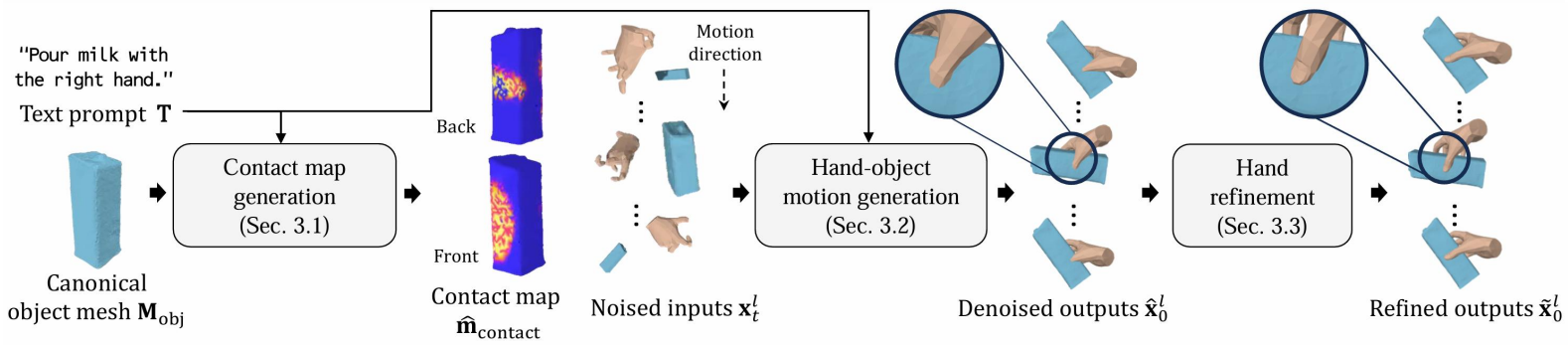


MotionGPT[2]: Generate Human Motion with Large Language Model from Free-form ***Language***

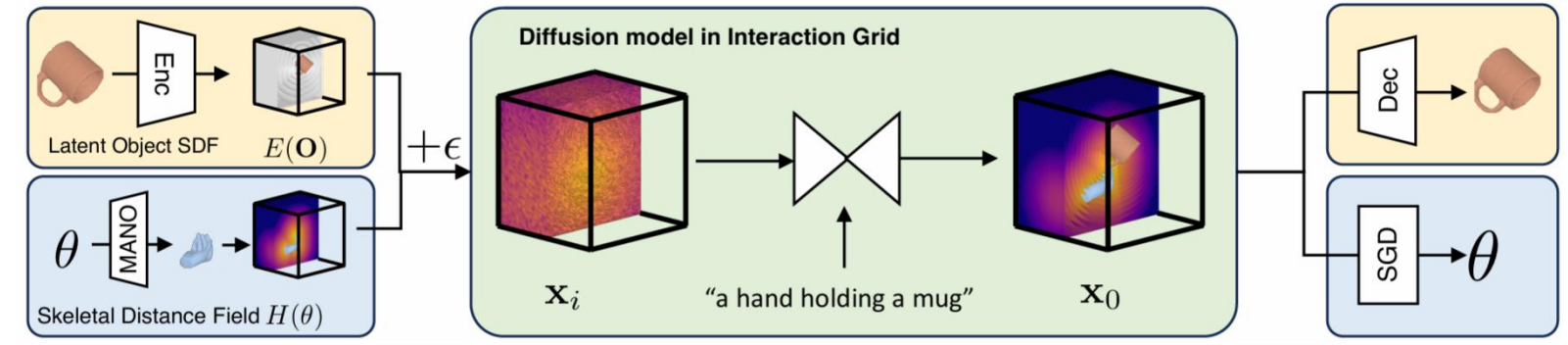
Can We used them into Hand Motion Generation?

[1]Tevet, Guy, et al. "Human Motion Diffusion Model." The Eleventh International Conference on Learning Representations..
[2]Jiang, Biao, et al. "Motiongpt: Human motion as a foreign language." Advances in Neural Information Processing Systems 36 (2023): 20067-20079.

Motivation: Closed-Set HOI



Text2HOI[1]: Generate Hand-Object Interaction Sequence with ***Text*** and prediction contact map

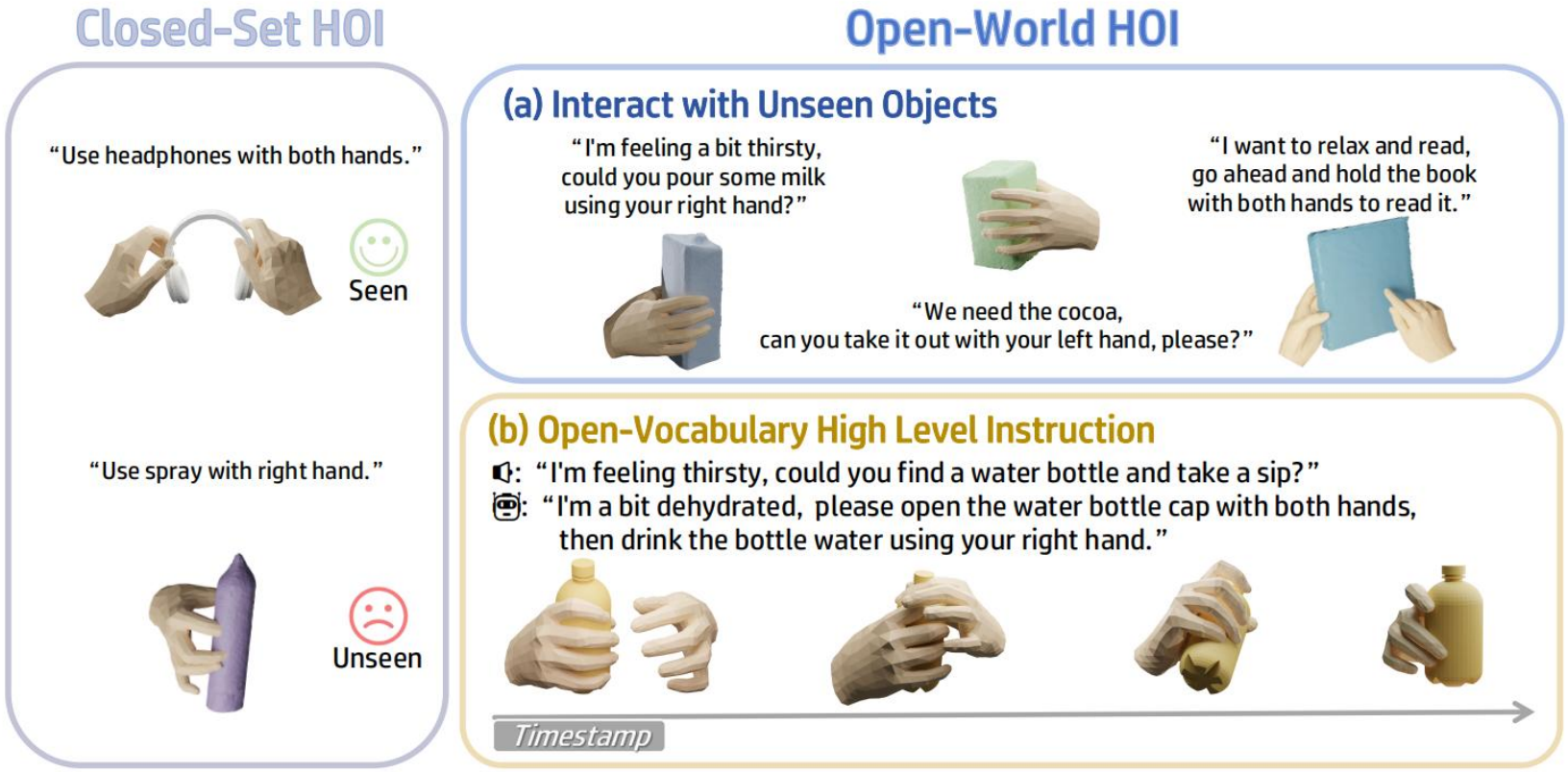


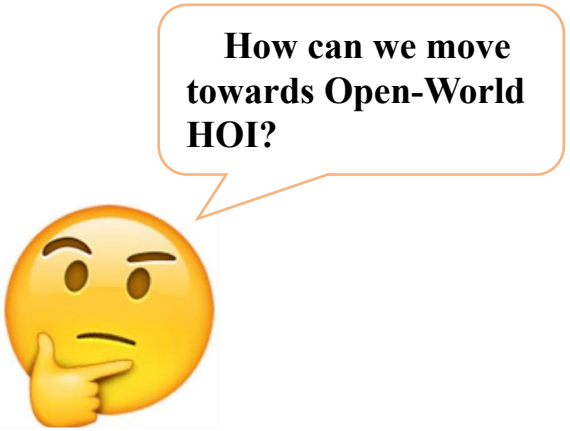
G-HOP[2]: Modeling HOI Synthesis with Diffusion Model in Interaction Grid with ***Language***

How can we get a better HOI Prior?

[1]Cha, Junuk, et al. "Text2hoi: Text-guided 3d motion generation for hand-object interaction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[2]Ye Y, et al. "G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Motivation: Closed-Set HOI vs. Open-World HOI





Contributions

How could we process ***unseen*** objects?

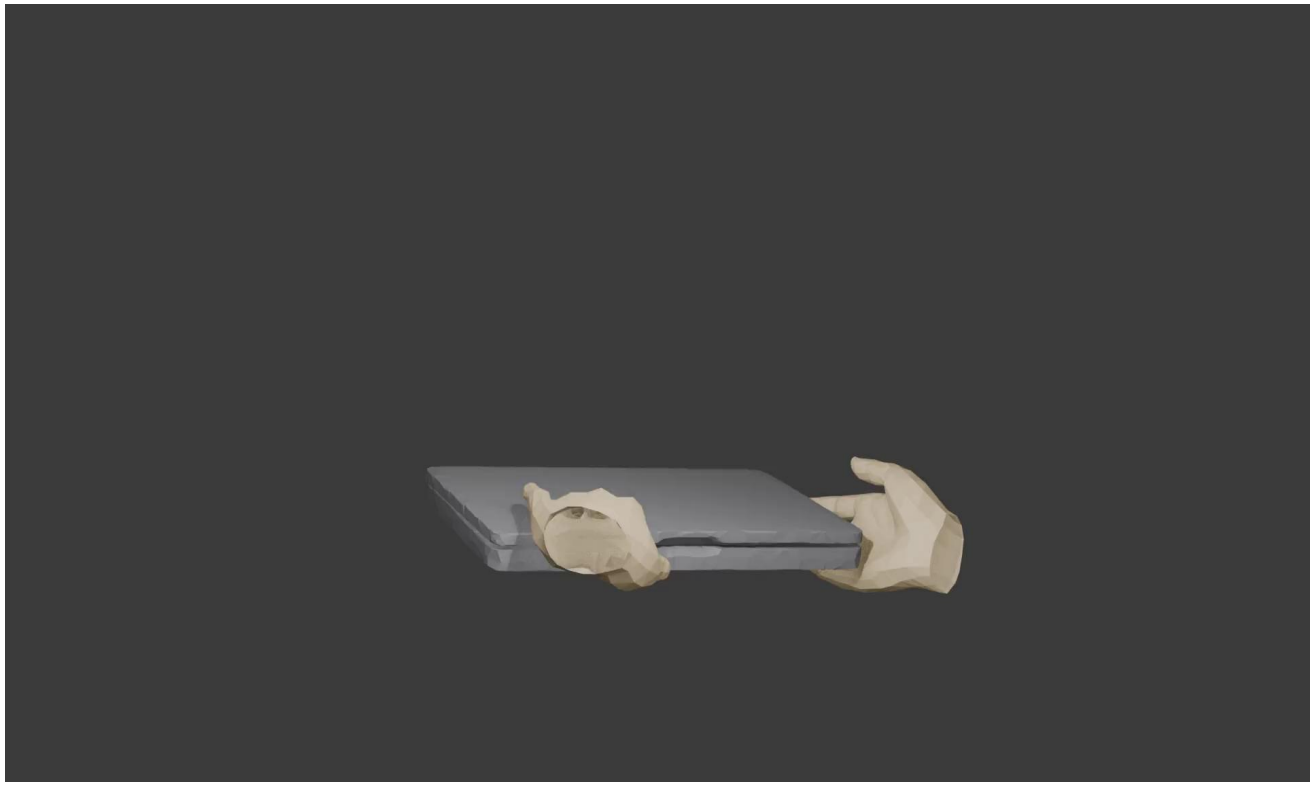
OpenHOI involves fine-tuning a 3D MLLM to learn ***Open-World affordance*** as powerful HOI Generation ***Priors***

How could we process the ***open-vocabulary*** instruction and generate ***long-horizon*** sequences?

OpenHOI involves MLLM to learn ***semantic task decomposition***, and use the Affordance-Driven HOI Diffusion to generate ***long-horizon*** HOI Sequences

We introduce **OpenHOI**, the first ***open-world*** hand-object interaction synthesis framework capable of generating ***long-horizon*** manipulation sequences for ***unseen*** objects guided by ***open-vocabulary*** instructions.

Result

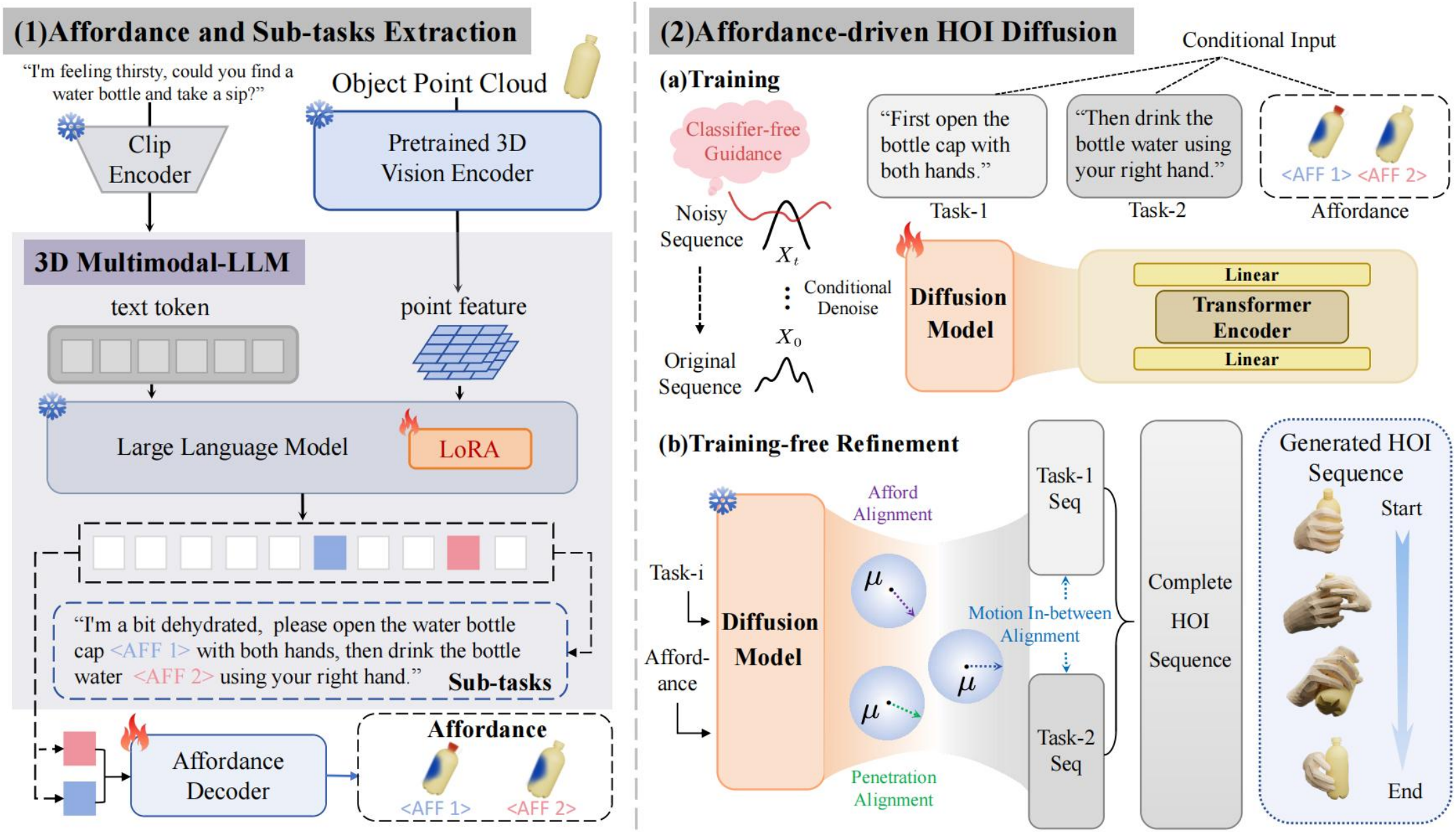


I want to write NeurIPS paper with my laptop， what should I do?

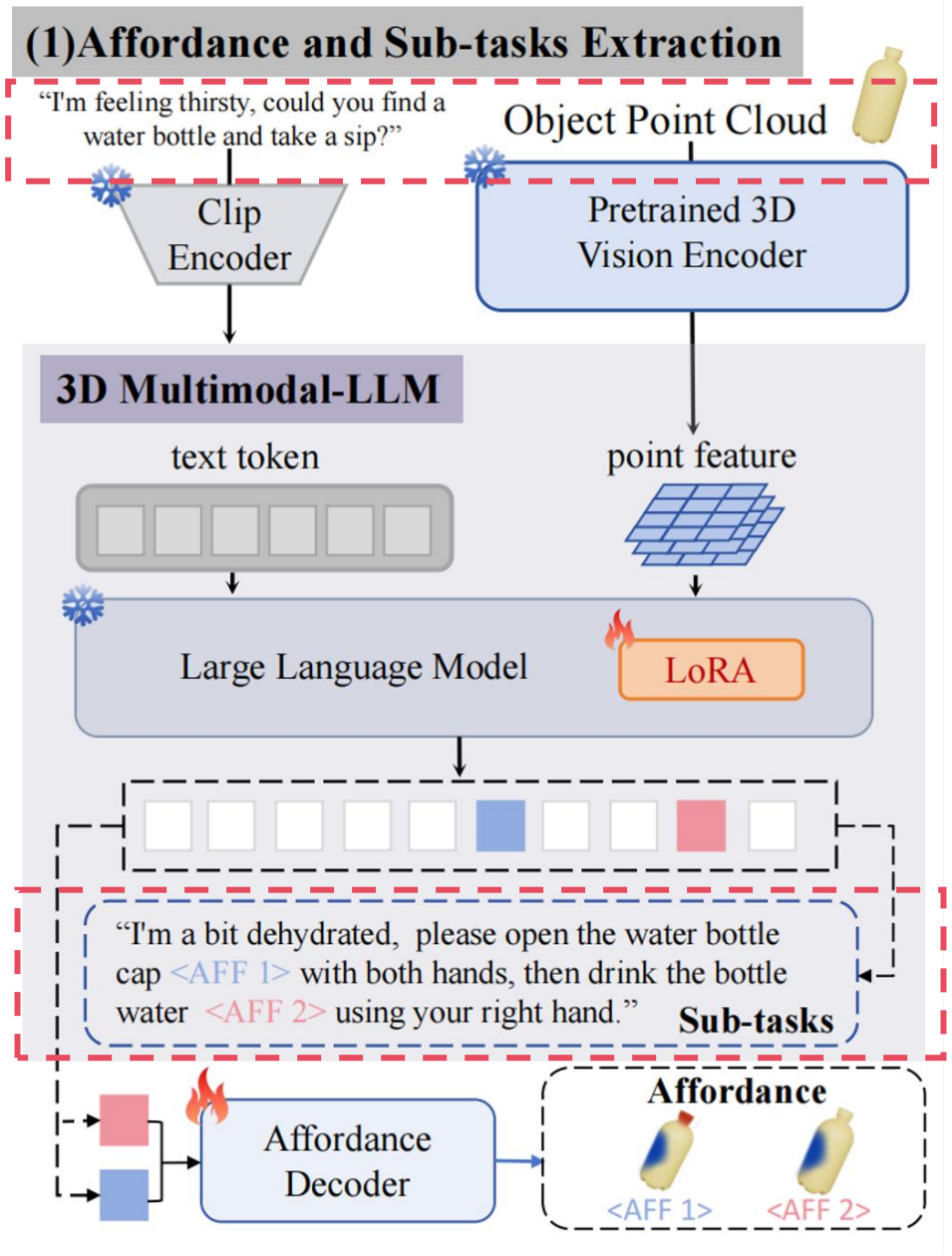
First, ***open*** your laptop with both hands
Then, ***type*** the laptop with both your hands to write NeurIPS!
At last, ***close*** your laptop and have a break

How could we get this ***fancy*** result?

OpenHOI Pipeline



3D Affordance MLLM



Open-vocabulary instructions Decomposition:

$$\tilde{\mathbf{T}}_{\text{sub\_tasks}} = \text{MLLM}\left(\mathbf{F}_{\text{obj}},\ \mathbf{T}_{\text{ins}}\right)$$

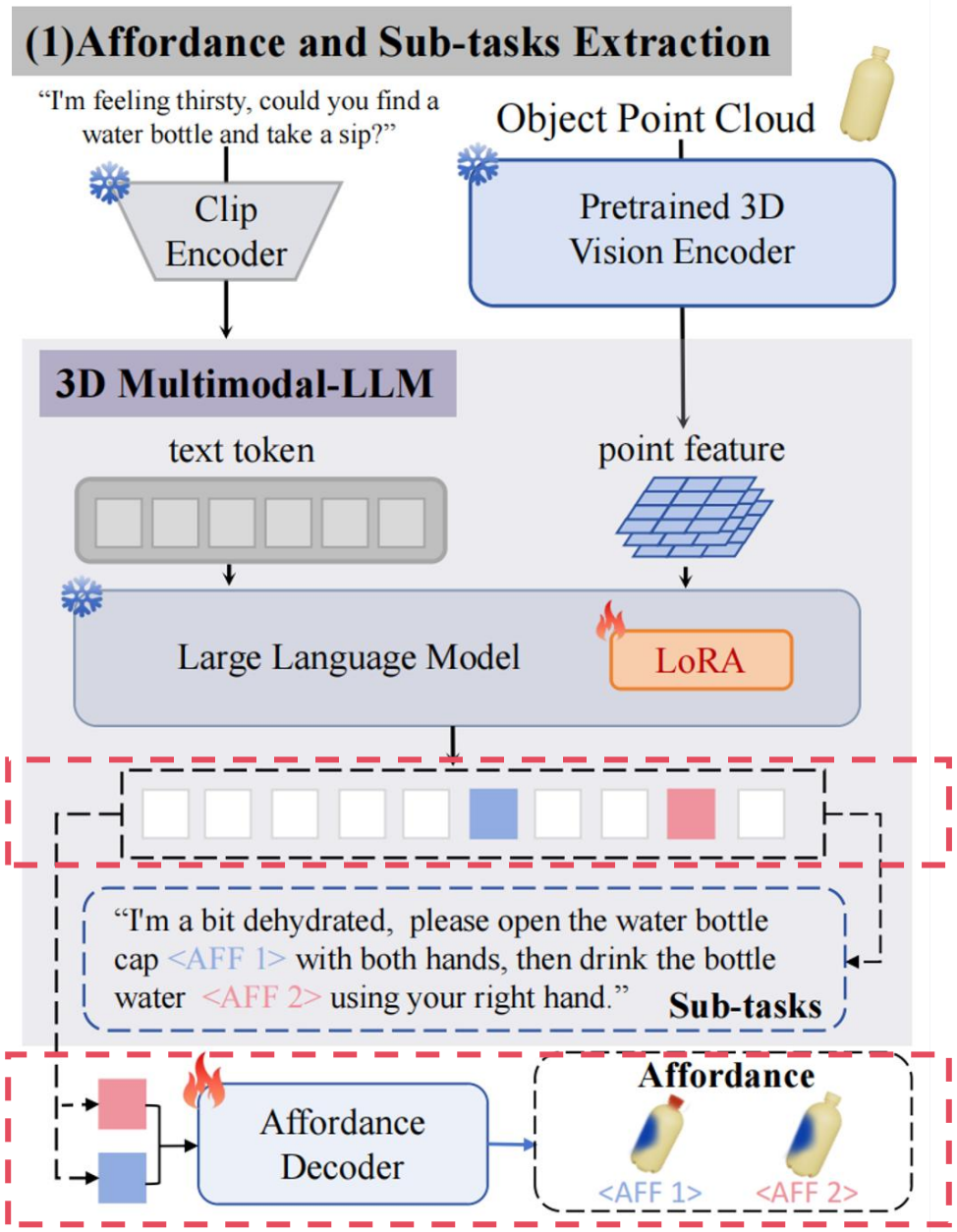
Affordance Map Decoder:

$$\tilde{\mathbf{A}}_{\text{obj}}^{(i)} = \text{Decoder}_{\text{aff}}\left(\mathbf{F}_{\text{obj}}, \mathbf{h}_{\text{aff}}^{(i)}\right), \quad i = 0, \ldots, S-1.$$

Coarse-to-Fine Affordance Tuning:

$$\text{L} = \lambda_{\text{task}}\text{L}_{\text{task}}\left(\mathbf{Y}_{\text{sub\_tasks}}, \tilde{\mathbf{Y}}_{\text{sub\_tasks}}\right) + \lambda_{\text{aff}}\text{L}_{\text{aff}}\left(\mathbf{A}_{\text{obj}}, \tilde{\mathbf{A}}_{\text{obj}}\right)$$

3D Affordance MLLM



Open-vocabulary instructions Decomposition:

$$\tilde{\mathbf{T}}_{\text{sub\_tasks}} = \text{MLLM}\left(\mathbf{F}_{\text{obj}},\ \mathbf{T}_{\text{ins}}\right)$$

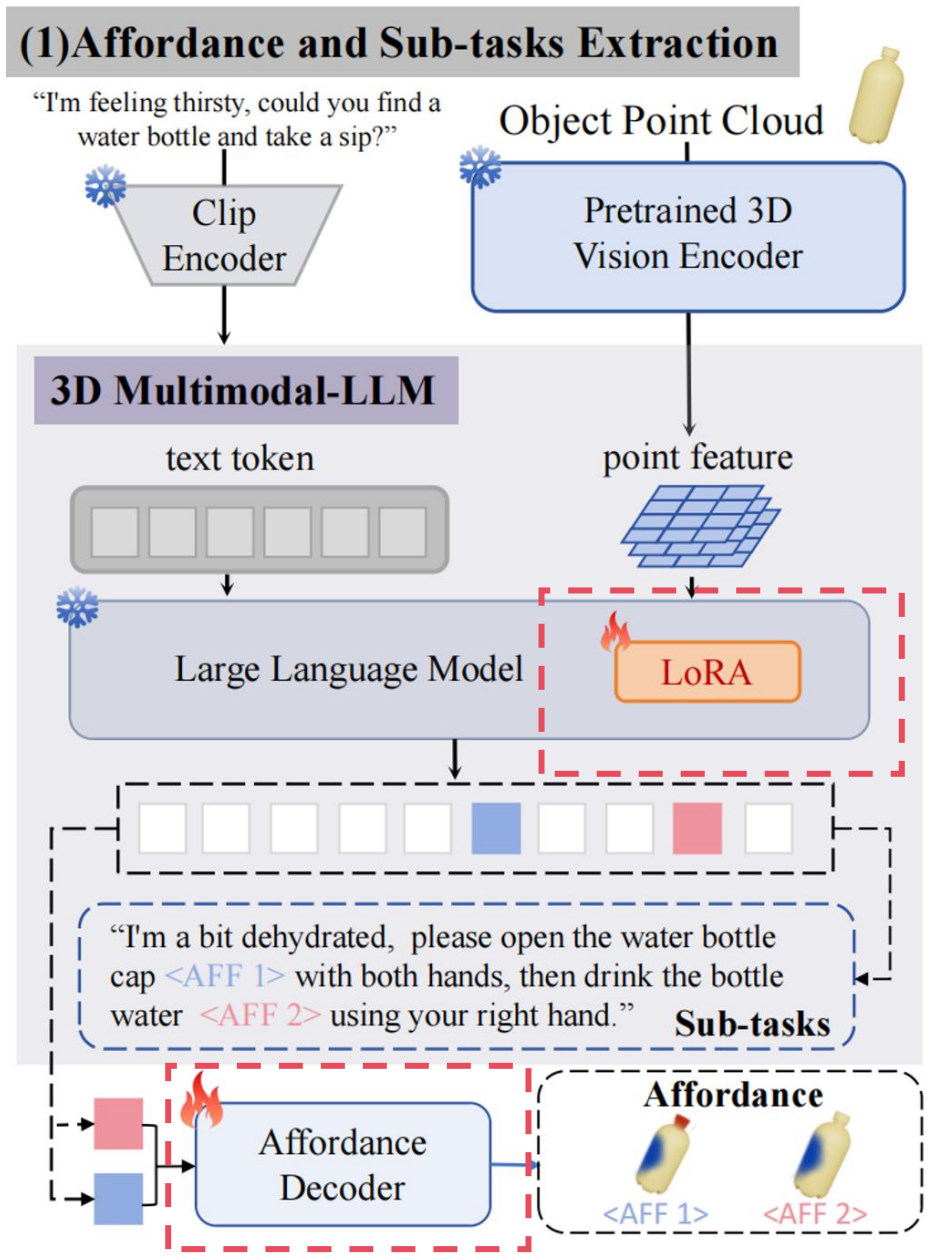
Affordance Map Decoder:

$$\tilde{\mathbf{A}}_{\text{obj}}^{(i)} = \text{Decoder}_{\text{aff}}\left(\mathbf{F}_{\text{obj}}, \mathbf{h}_{\text{aff}}^{(i)}\right), \quad i = 0, \ldots, S-1.$$

Coarse-to-Fine Affordance Tuning:

$$\text{L} = \lambda_{\text{task}}\text{L}_{\text{task}}\left(\mathbf{Y}_{\text{sub\_tasks}}, \tilde{\mathbf{Y}}_{\text{sub\_tasks}}\right) + \lambda_{\text{aff}}\text{L}_{\text{aff}}\left(\mathbf{A}_{\text{obj}}, \tilde{\mathbf{A}}_{\text{obj}}\right)$$

3D Affordance MLLM



Open-vocabulary instructions Decomposition:

$$\tilde{\mathbf{T}}_{\text{sub\_tasks}} = \text{MLLM}\left(\mathbf{F}_{\text{obj}},\ \mathbf{T}_{\text{ins}}\right)$$

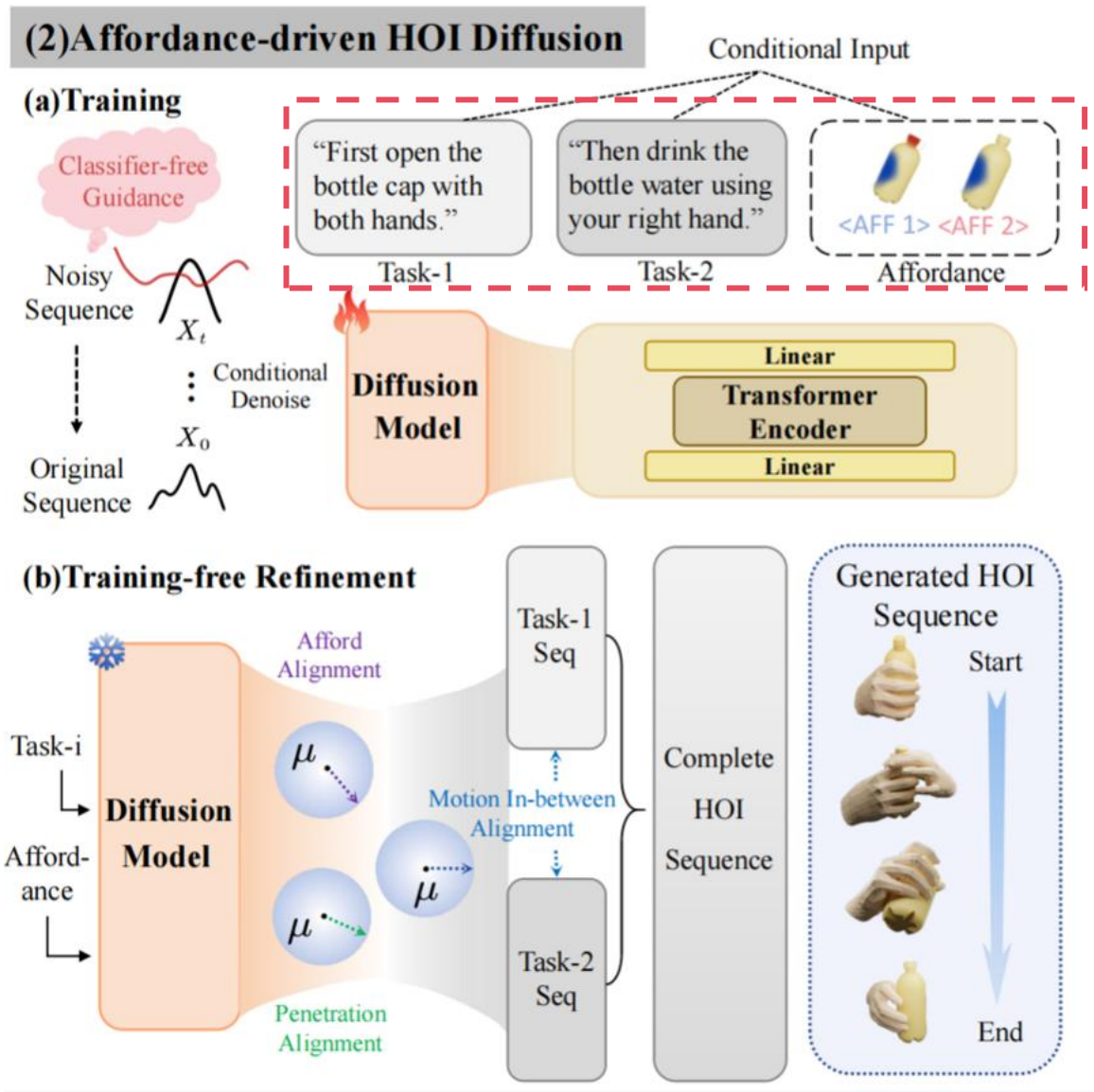
Affordance Map Decoder:

$$\tilde{\mathbf{A}}_{\text{obj}}^{(i)} = \text{Decoder}_{\text{aff}}\left(\mathbf{F}_{\text{obj}}, \mathbf{h}_{\text{aff}}^{(i)}\right), \quad i = 0, \ldots, S-1.$$

Coarse-to-Fine Affordance Tuning:

$$\text{L} = \lambda_{\text{task}}\text{L}_{\text{task}}\left(\mathbf{Y}_{\text{sub\_tasks}}, \tilde{\mathbf{Y}}_{\text{sub\_tasks}}\right) + \lambda_{\text{aff}}\text{L}_{\text{aff}}\left(\mathbf{A}_{\text{obj}}, \tilde{\mathbf{A}}_{\text{obj}}\right)$$

Affordance-driven HOI Diffusion



Condition: Interaction Prior from MLLM:

$$\mathbf{C} = [\tilde{\mathbf{A}}_{\text{obj}}, f^{\text{clip}}(\tilde{\mathbf{T}}_{\text{sub\_tasks}}), \mathbf{F}_{\text{obj}}]$$

Training Loss for HOI Diffusion:

$$L_{\text{hoi\_train}}(\hat{X}_\theta, \mathbf{C}) = L_{\text{hoi\_diff}}(\hat{X}_\theta, \mathbf{C}) + L_{\text{hoi\_distance}}(\hat{X}_\theta) + L_{\text{hoi\_orient}}(\hat{X}_\theta)$$

Classifier-free Guidance for Better Alignment:

$$X^s_\theta(X_t, t, \mathbf{C}) = X_\theta(X_t, t, \emptyset) + s \cdot (X_\theta(X_t, t, \mathbf{C}) - X_\theta(X_t, t, \emptyset))$$

Affordance Refinement:

$$l_{\text{aff}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{J}_{\text{lhand}}, \tilde{P}^{\text{ljoint}}_{\text{obj}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{J}_{\text{rhand}}, \tilde{P}^{\text{rjoint}}_{\text{obj}})\|^2$$

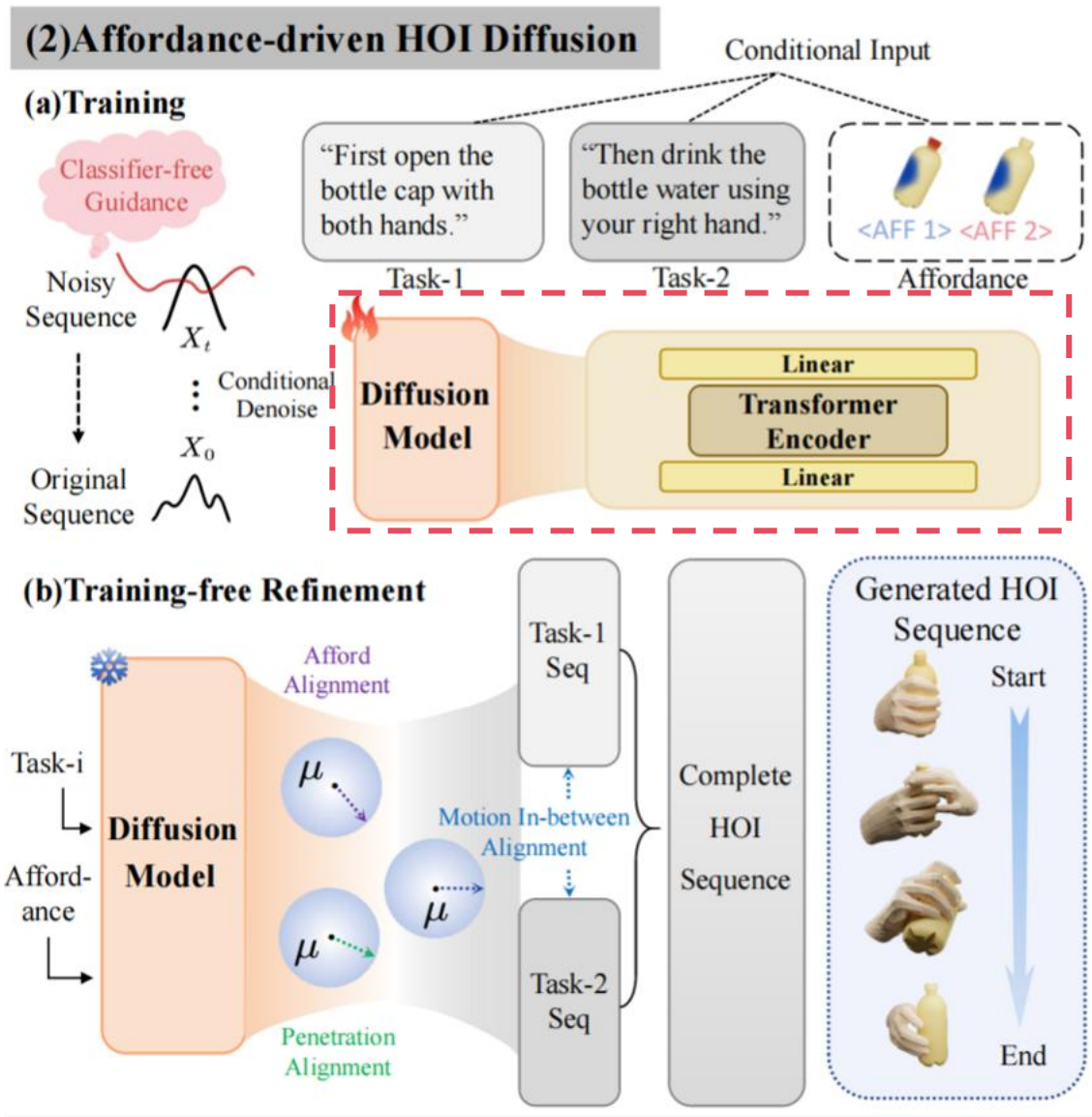
Penetration Refinement:

$$l_{\text{penetration}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{V}_{\text{lhand}}, \tilde{P}^{\text{lvert}}_{\text{obj}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{V}_{\text{rhand}}, \tilde{P}^{\text{rvert}}_{\text{obj}})\|^2$$

Motion Inbetween Refinement:

$$l_{\text{transition}} = ||\hat{V}^0_{\text{trans}} - V^T_{\text{pre}}||^2_2 + ||\hat{V}^T_{\text{trans}} - V^0_{\text{after}}||^2_2$$

Affordance-driven HOI Diffusion



Condition: Interaction Prior from MLLM:

$$\mathbf{C} = [\tilde{\mathbf{A}}_{\text{obj}}, f^{\text{clip}}(\tilde{\mathbf{T}}_{\text{sub\_tasks}}), \mathbf{F}_{\text{obj}}]$$

Training Loss for HOI Diffusion:

$$L_{\text{hoi\_train}}(\hat{X}_\theta, \mathbf{C}) = L_{\text{hoi\_diff}}(\hat{X}_\theta, \mathbf{C}) + L_{\text{hoi\_distance}}(\hat{X}_\theta) + L_{\text{hoi\_orient}}(\hat{X}_\theta)$$

Classifier-free Guidance for Better Alignment:

$$X_\theta^s(X_t, t, \mathbf{C}) = X_\theta(X_t, t, \emptyset) + s \cdot (X_\theta(X_t, t, \mathbf{C}) - X_\theta(X_t, t, \emptyset))$$

Affordance Refinement:

$$l_{\text{aff}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{J}_{\text{lhand}}, \tilde{P}_{\text{obj}}^{\text{ljoint}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{J}_{\text{rhand}}, \tilde{P}_{\text{obj}}^{\text{rjoint}})\|^2$$

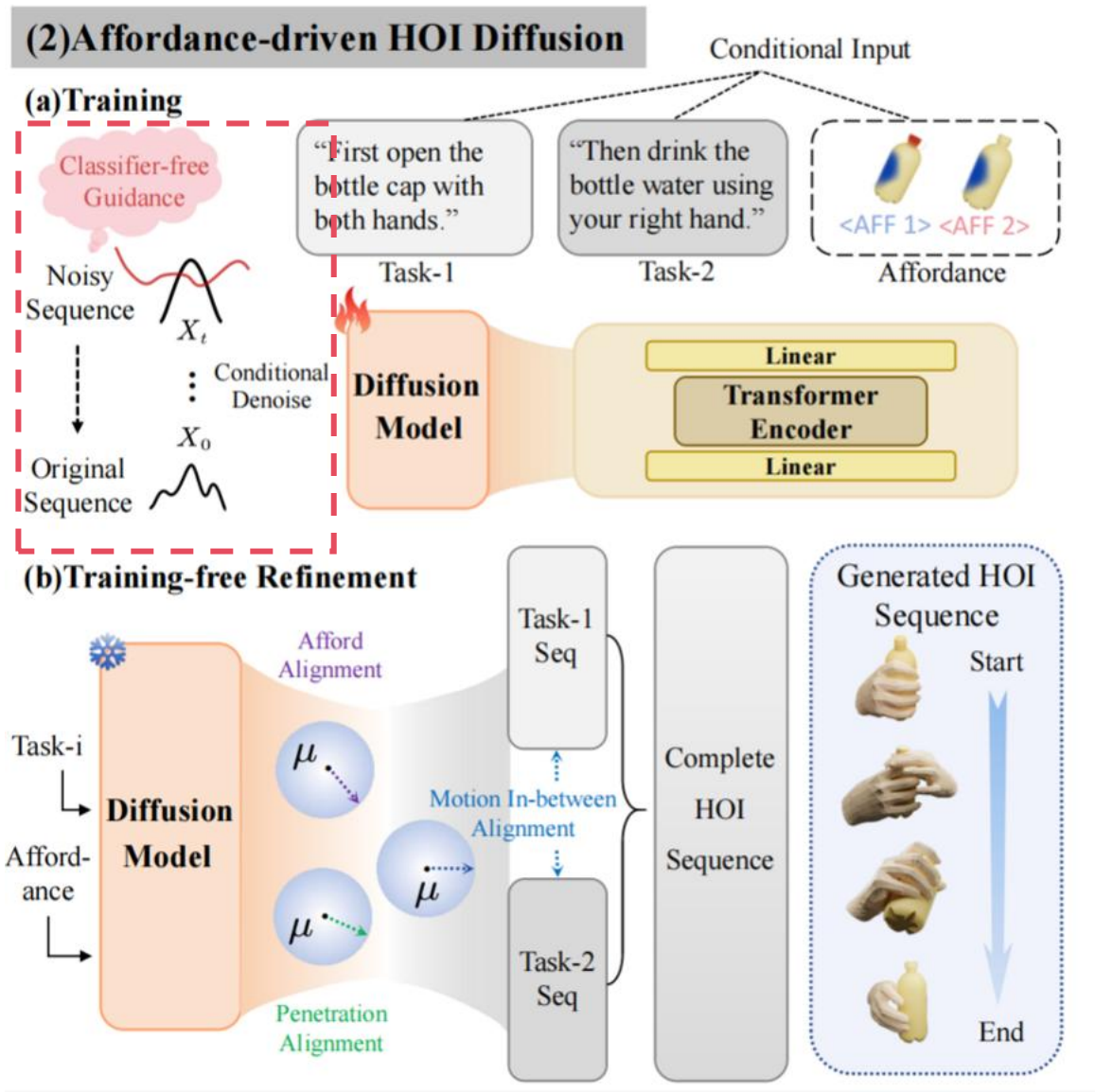
Penetration Refinement:

$$l_{\text{penetration}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{V}_{\text{lhand}}, \tilde{P}_{\text{obj}}^{\text{lvert}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{V}_{\text{rhand}}, \tilde{P}_{\text{obj}}^{\text{rvert}})\|^2$$

Motion Inbetween Refinement:

$$l_{\text{transition}} = ||\hat{V}_{\text{trans}}^0 - V_{\text{pre}}^T||_2^2 + ||\hat{V}_{\text{trans}}^T - V_{\text{after}}^0||_2^2$$

Affordance-driven HOI Diffusion



Condition: Interaction Prior from MLLM:

$$\mathbf{C} = [\tilde{\mathbf{A}}_{\text{obj}}, f^{\text{clip}}(\tilde{\mathbf{T}}_{\text{sub\_tasks}}), \mathbf{F}_{\text{obj}}]$$

Training Loss for HOI Diffusion:

$$L_{\text{hoi\_train}}(\hat{X}_\theta, \mathbf{C}) = L_{\text{hoi\_diff}}(\hat{X}_\theta, \mathbf{C}) + L_{\text{hoi\_distance}}(\hat{X}_\theta) + L_{\text{hoi\_orient}}(\hat{X}_\theta)$$

Classifier-free Guidance for Better Alignment:

$$X^s_\theta(X_t, t, \mathbf{C}) = X_\theta(X_t, t, \emptyset) + s \cdot (X_\theta(X_t, t, \mathbf{C}) - X_\theta(X_t, t, \emptyset))$$

Affordance Refinement:

$$l_{\text{aff}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{J}_{\text{lhand}}, \tilde{P}^{\text{ljoint}}_{\text{obj}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{J}_{\text{rhand}}, \tilde{P}^{\text{rjoint}}_{\text{obj}})\|^2$$

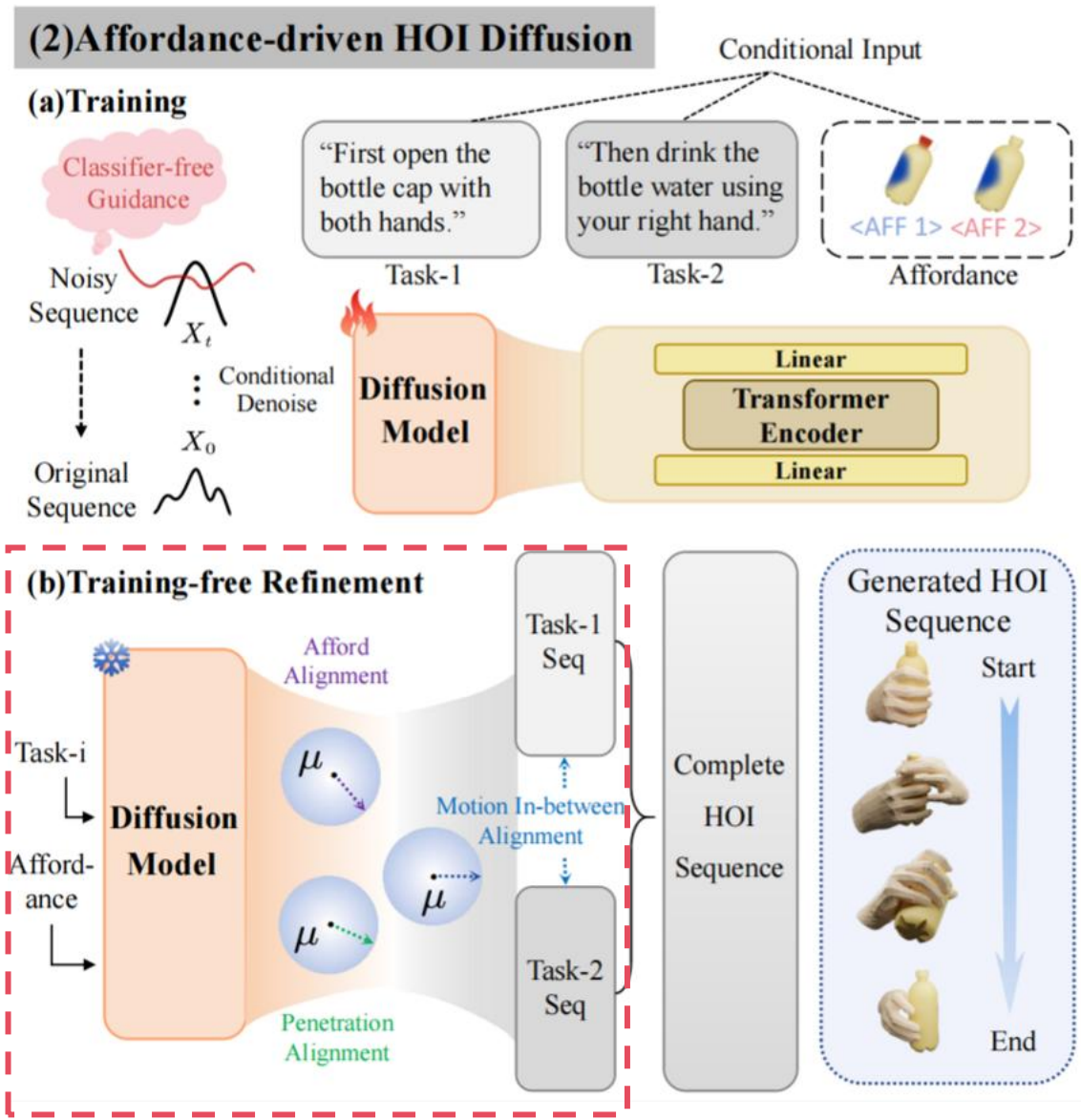
Penetration Refinement:

$$l_{\text{penetration}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{V}_{\text{lhand}}, \tilde{P}^{\text{lvert}}_{\text{obj}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{V}_{\text{rhand}}, \tilde{P}^{\text{rvert}}_{\text{obj}})\|^2$$

Motion Inbetween Refinement:

$$l_{\text{transition}} = ||\hat{V}^0_{\text{trans}} - V^T_{\text{pre}}||^2_2 + ||\hat{V}^T_{\text{trans}} - V^0_{\text{after}}||^2_2$$

Affordance-driven HOI Diffusion



Condition: Interaction Prior from MLLM:

$$\mathbf{C} = [\tilde{\mathbf{A}}_{\text{obj}}, f^{\text{clip}}(\tilde{\mathbf{T}}_{\text{sub\_tasks}}), \mathbf{F}_{\text{obj}}]$$

Training Loss for HOI Diffusion:

$$L_{\text{hoi\_train}}(\hat{X}_\theta, \mathbf{C}) = L_{\text{hoi\_diff}}(\hat{X}_\theta, \mathbf{C}) + L_{\text{hoi\_distance}}(\hat{X}_\theta) + L_{\text{hoi\_orient}}(\hat{X}_\theta)$$

Classifier-free Guidance for Better Alignment:

$$X_\theta^s(X_t, t, \mathbf{C}) = X_\theta(X_t, t, \emptyset) + s \cdot (X_\theta(X_t, t, \mathbf{C}) - X_\theta(X_t, t, \emptyset))$$

Affordance Refinement:

$$l_{\text{aff}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{J}_{\text{lhand}}, \tilde{P}_{\text{obj}}^{\text{ljoint}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{J}_{\text{rhand}}, \tilde{P}_{\text{obj}}^{\text{rjoint}})\|^2$$

Penetration Refinement:

$$l_{\text{penetration}} = \mathbb{1}_{\text{left}} \cdot \|d(\hat{V}_{\text{lhand}}, \tilde{P}_{\text{obj}}^{\text{lvert}})\|^2 + \mathbb{1}_{\text{right}} \cdot \|d(\hat{V}_{\text{rhand}}, \tilde{P}_{\text{obj}}^{\text{rvert}})\|^2$$

Motion Inbetween Refinement:

$$l_{\text{transition}} = ||\hat{V}_{\text{trans}}^0 - V_{\text{pre}}^T||_2^2 + ||\hat{V}_{\text{trans}}^T - V_{\text{after}}^0||_2^2$$

Experiments

Table 2: **Main Results on GRAB.**

	Method	MPJPE↓	FOL↓	FID ↓	Diversity →	MModality ↑
	GT	-	-	-	4.66	-
Seen	MDM[34]	74.92±2.25	0.62±0.02	62.37±1.56	3.28±0.10	12.77±0.45
	TM2T[8]	59.27±1.19	0.46±0.06	57.41±2.30	3.60±0.07	21.28±0.82
	MotionGPT[12]	63.94±2.56	0.43±0.01	52.03±1.82	3.61±0.08	20.26±0.51
	Text2HOI[1]	56.29±2.13	0.44±0.03	33.72±1.27	3.41±0.16	17.71±0.87
	Ours	**47.64±1.03**	**0.26±0.02**	**26.43±0.77**	**3.69±0.27**	**24.59±2.01**
Unseen	MDM[34]	92.97±1.86	0.69±0.03	75.59±1.89	3.07±0.11	11.15±0.85
	TM2T[8]	61.07±1.34	0.55±0.02	66.43±1.66	3.37±0.07	14.03±0.67
	MotionGPT[12]	66.26±1.99	0.51±0.01	56.49±1.98	2.852±0.07	16.36±0.53
	Text2HOI[1]	60.67±1.80	0.41±0.02	36.96±0.77	1.80±0.05	10.98±0.44
	Ours	**51.34±0.85**	**0.27±0.01**	**28.29±0.62**	**3.61±0.09**	**19.91±0.63**

Table 3: **Main Results on ARCTIC.**

	Method	MPJPE↓	FOL↓	FID ↓	Diversity →	MModality ↑
	GT	-	-	-	3.39	-
Seen	MDM[34]	72.67±0.63	0.60±0.05	33.66±0.19	2.35±0.05	8.20±0.20
	TM2T[8]	54.39±0.64	0.41±0.04	34.12±0.49	1.67±0.02	13.60±0.17
	MotionGPT[12]	60.17±0.72	0.41±0.03	31.58±0.46	1.89±0.02	13.23±0.09
	Text2HOI[1]	52.16±0.41	0.33±0.01	23.35±0.33	2.43±0.02	11.21±0.20
	Ours	**45.15±0.94**	**0.25±0.04**	**19.74±0.16**	**2.65±0.03**	**15.25±1.44**
Unseen	MDM[34]	86.75±1.35	0.64±0.01	41.53±1.37	1.58±0.04	7.13±0.63
	TM2T[8]	55.57±1.26	0.53±0.03	37.22±0.75	1.54±0.12	11.23±0.44
	MotionGPT[12]	64.41±0.73	0.43±0.04	33.99±2.43	1.50±0.09	11.08±0.79
	Text2HOI[1]	57.83±1.61	0.39±0.01	25.22±0.59	1.61±0.06	7.11±0.25
	Ours	**47.25±0.39**	**0.28±0.03**	**20.05±0.80**	**2.49±0.08**	**12.66±0.71**

Thank you for listening!

OpenHOI

Open-World Generalization with *Affordance*
Open-Vocabulary Instructions with 3D *MLLM*
***Long-horizon* and *Complex* Manipulation**