

ManiVideo: Generating Hand-Object Manipulation Video with Dexterous and Generalizable Grasping

Supplementary Material

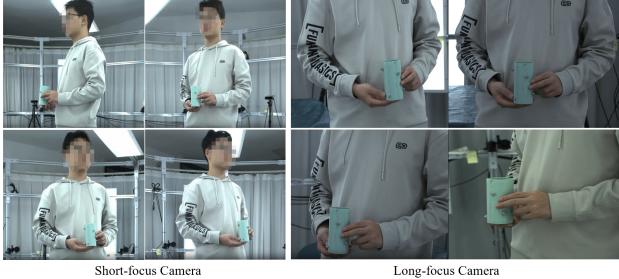


Figure 1. A sampling of our data.

Please refer to the demo video for more dynamic results.

A. Our Dataset

We train our ManiVideo on three types of datasets. For object data, we utilize Objaverse [2], and human data is sourced from Human4DiT [9]. For HOI video data, in addition to the public DexYCB dataset [1], We collect third-person view videos of participants standing and interacting with objects using bimanual hands. As shown in Tab. 1, our dataset contains 722 videos (376k frames) depicting daily tool-use behaviors covering 15 objects, 10 views and 8 participants. Compared to other datasets [3, 8], our data is human-centric and free from distractions caused by irrelevant objects (e.g., tables), making it particularly suitable for downstream applications like human-based HOI video generation in Sec.4.4. As illustrated in Fig. 1, the dataset comprises human-centered videos captured from multiple views using both long-focus and short-focus cameras.

Object model acquisition. We utilize the 3D object models from TACO [8] as our objects, with meshes consisting of up to 100K triangular faces to capture fine-grained geometric details. Specifically, we select commonly used objects from daily life, including spoons, shovels, bowls, cups, and boxes.

Data Capturing. Similar to TACO [8], hand motion is extracted from multi-view RGB videos, whereas object motion is captured using a motion capture system by tracking four markers affixed to the object’s surface. Moreover, our data capture system combines 12 synchronized industrial FLIR cameras with a NOKOV optical motion capture setup equipped with six Mars4H infrared cameras.

Data Annotating. We process hands and objects as separate entities. For hands, we employ RTMpose [6] to differentiate between the left and right hands and extract 2D keypoints. Subsequently, the MANO model is utilized to represent the

3D hand mesh, which is optimized using both 2D and 3D loss functions. For a detailed description of the process, please refer to TACO [8].

For objects, the 6D pose, comprising rotation and translation, is obtained using the motion capture system. Marker-to-surface correspondence is then optimized, and the refined object poses are computed by integrating the relative positions of markers on the object mesh with the captured marker motions.

B. Training Strategy

Due to the differences between different datasets, we propose a training strategy to integrate all dataset. As shown in Fig. 2, we apply distinct conditions to each of the three datasets. For the HOI training, all conditions outlined in the main paper are utilized. Deviations from HOI training in Objaverse training are highlighted in red, while differences in the human training are indicated in green. None conditions are filled to zero.

C. Comparisons on HO3Dv3 Dataset

In addition to the public dataset DexYCB, we also conduct comparison on dataset HO3Dv3 [4]. As shown in Fig. 3 and Fig. 4, our method achieves the best results when the fingers are stacked together. Specifically, directly learning the hand-object correspondence from 2D conditions poses an ill-posed problem, making it challenging for HOGAN to maintain consistency, particularly in scenarios where fingers are densely packed. In diffusion-based methods, the limited representation of fingers often causes details to be misinterpreted during the denoising process.

D. More Results on Our Dataset

In Fig. 5 and Fig. 6, we show more results on our dataset. Our method leverages the proposed multi-layer occlusion representation to effectively capture occlusion relationships based on comprehensive finger information. This approach tackles challenges such as self-occlusion of fingers, mutual occlusion between hands and objects, and the invisibility of bent fingers, leading to more accurate handling of these complex interactions.

E. Long Sequence Generation

Generating extended video sequences poses significant challenges for video diffusion models. To overcome this

Dataset	bimanual	functional manipulation	multi-view	mocap	sequence	frame
GRAB [10]	✓	✗	✗	✓	1.3K	1.6M
HO3D [4]	✗	✗	✓	✗	27	78K
DexYCB [1]	✗	✗	✓	✗	1.0K	582K
OakInk [11]	✗	✗	✓	✓	778	314K
HOI4D [7]	✗	✓	✗	✗	4.0K	1.2M
ARCTIC [3]	✓	✗	✓	✓	339	2.1M
AffordPose [5]	✗	✓	✗	✗	-	27K
Ours	✓	✓	✓	✓	722	376k

Table 1. Comparison of our dataset with existing 3D hand-object interaction datasets.

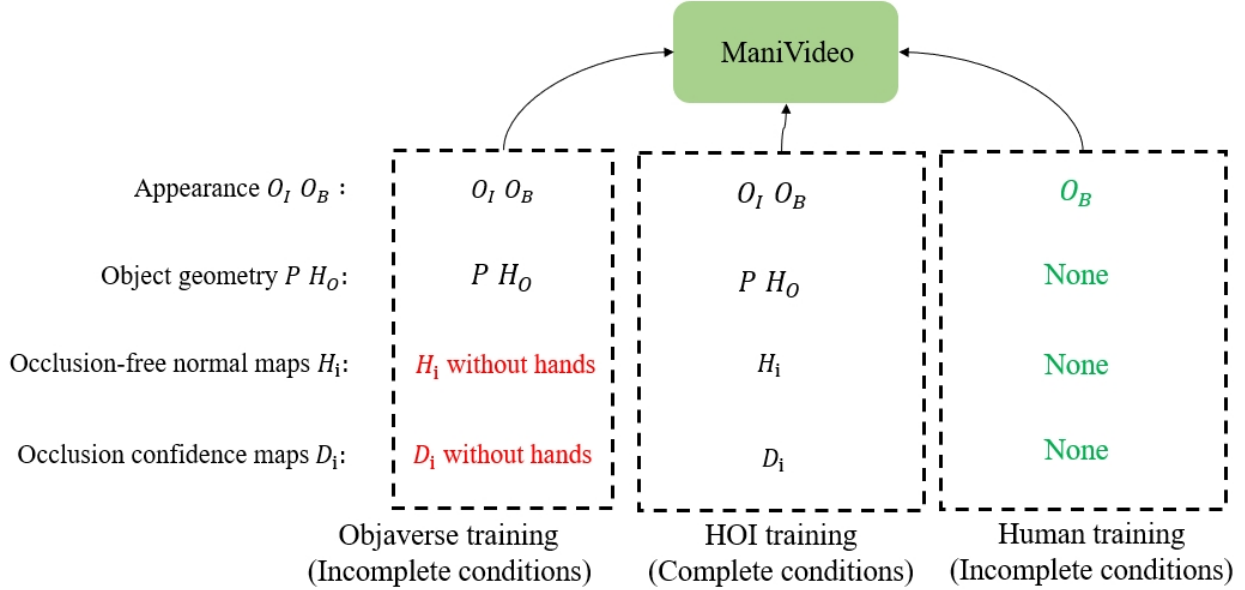


Figure 2. Training strategy. We apply distinct conditions to each of the three datasets. For the HOI training, all conditions outlined in the main paper are utilized. Deviations from HOI training in Objaverse training are highlighted in red, while differences in the human training are indicated in green. Missing conditions are filled to zero.

limitation, we adopt a temporal sliding window approach, which facilitates the generation of arbitrarily long videos while ensuring inter-frame consistency throughout. Let N denote the number of frames contained in the latent code z , and w represent the window size with a stride of s . Therefore, z will be divided into $\frac{N-w}{s} + 1$ parts. During DDIM sampling, each timestep employs a sliding window mechanism along the temporal dimension with a specified stride to sample all groups. Overlapping segments are averaged to maintain coherence. This process is repeated across successive timesteps, ensuring consistency in sequence generation. Formally, sampling the p -th window at each timestep is as follows:

$$\text{DDIM}(z_{[(p-1) \times s : p \times s + s]}) \quad (1)$$

where indexes are sliced in the temporal dimension. Temporal smoothing effectively addresses inconsistencies in over-

all brightness, hue, and style, which commonly arise due to noise variations and sampling discrepancies.

F. Ablation Studies

Here, we add further experiments about framework.

AppearanceNet R and Point cloud P : In all experiments, we use classifier-free guidance to control P and R , combining unconditional and conditional outputs via weighting coefficients. For results of the full model (Ours in all experiments), we set the coefficients to 3.5 to combine unconditional generation with the generation conditioned on P and R . Here, we report the unconditional generation of R and P by setting the coefficients to zero. The quantitative results are shown in the first two rows of Tab. 2. For w/o R , we use a reference image of the human interacting with objects as the first frame to provide the necessary appearance. R plays a

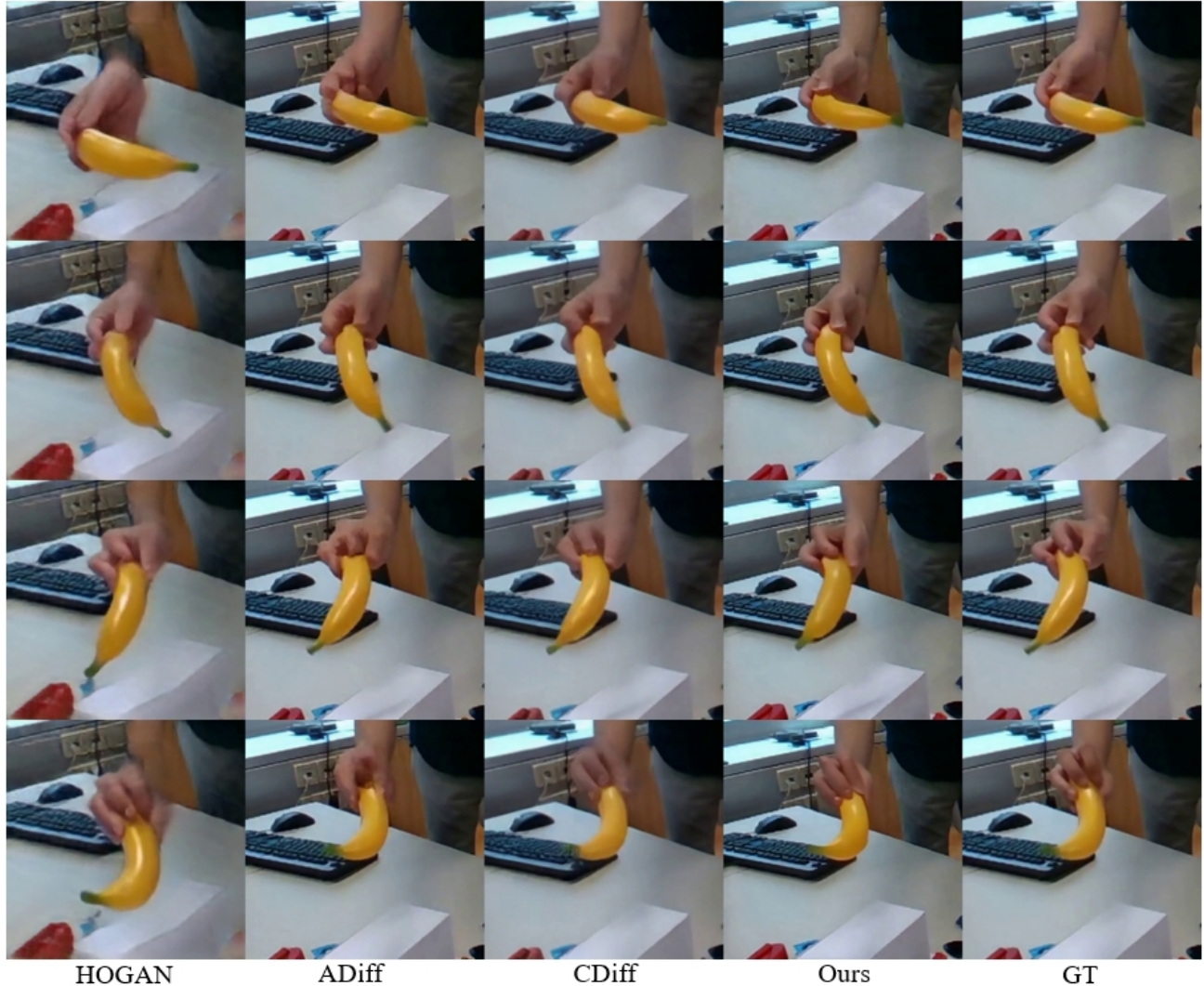


Figure 3. Qualitative comparison of different methods on HO3Dv3. Our approach achieves the best results.

crucial role in integrating appearance details, the absence of R leads to inconsistent identities. P offers geometric information about objects within 3D space, and the omission of P can compromise the structural integrity. As shown in rows three to five of Tab. 2, the model utilizes the guidance of D to address issues such as finger dislocation in the absence of D (w/o D).

G. Future Work

In this work, driving motion sequences are extracted from our dataset. However, our method can be integrated with motion generation methods to achieve end-to-end hand-object manipulation video generation. For example, given the object model and motion trajectory, we first use ManiDext [12] to generate motion sequences of HOI. Then, our method

leverages appearance and generated motion as inputs to produce temporally coherent and visually plausible hand-object manipulation videos, which is consistent with the main paper.

Method	FID↓	LPIPS↓	PSNR↑	SSIM↑	MPJPE↓
w/o R	53.12	0.151	24.87	0.784	40.55
w/o P	40.33	0.115	28.88	0.901	33.76
w/o D	44.07	0.117	28.03	0.868	38.89
D^1	40.09	0.113	28.72	0.896	36.72
D^2	42.99	0.114	28.46	0.893	37.20
Ours	37.70	0.113	29.59	0.905	32.89

Table 2. Quantitative comparison for ablation studies.

References

- [1] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for

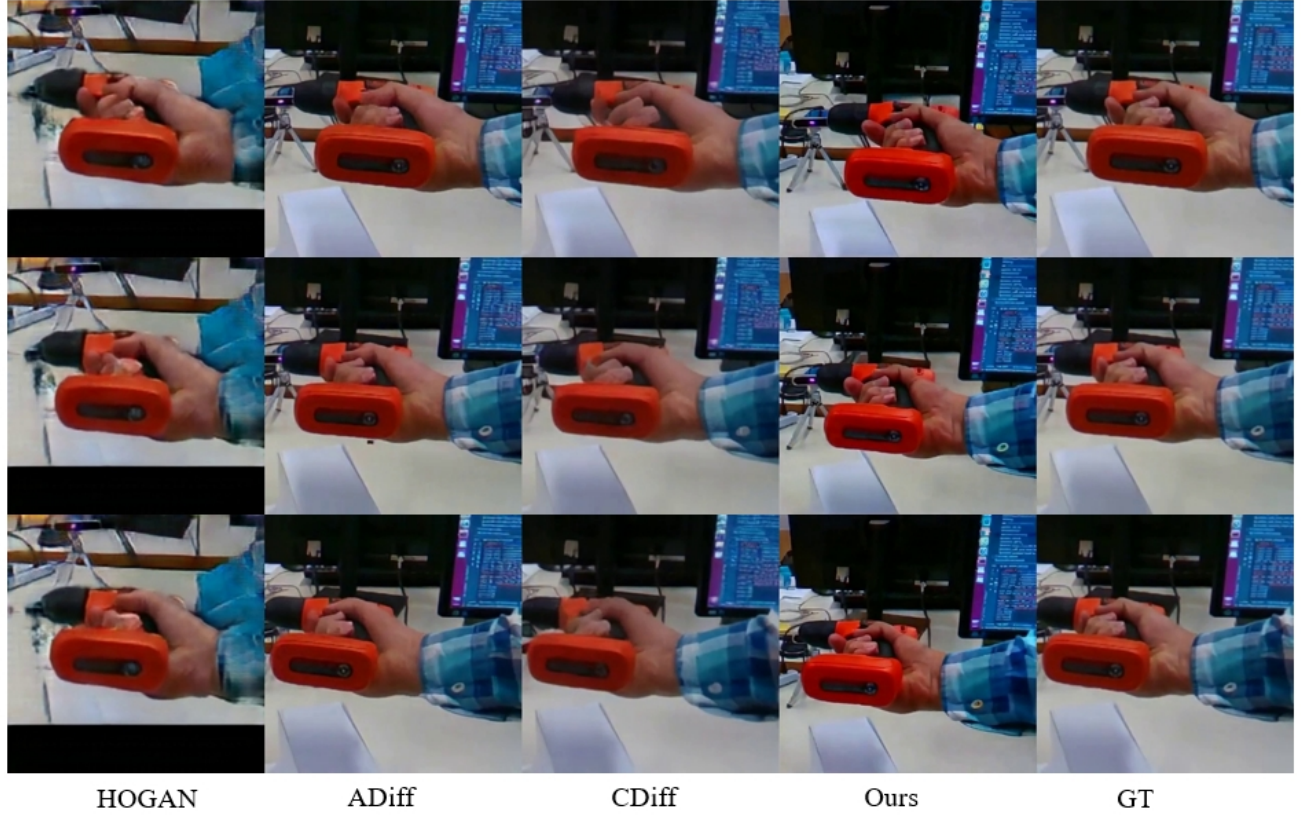


Figure 4. Qualitative comparison of different methods on HO3Dv3. Our approach achieves the best results.

capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021. [1](#), [2](#), [7](#)

- [2] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. [1](#)
- [3] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. [1](#), [2](#)
- [4] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. [1](#), [2](#)
- [5] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*, pages 14713–14724, 2023. [2](#)
- [6] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RtmPose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. [1](#)
- [7] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, pages 21013–21022, 2022. [2](#)
- [8] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *CVPR*, pages 21740–21751, 2024. [1](#)
- [9] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *ACM TOG*, 43(6), 2024. [1](#)
- [10] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. [2](#)
- [11] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, pages 20953–20962, 2022. [2](#)
- [12] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *arXiv preprint arXiv:2409.09300*, 2024. [3](#)

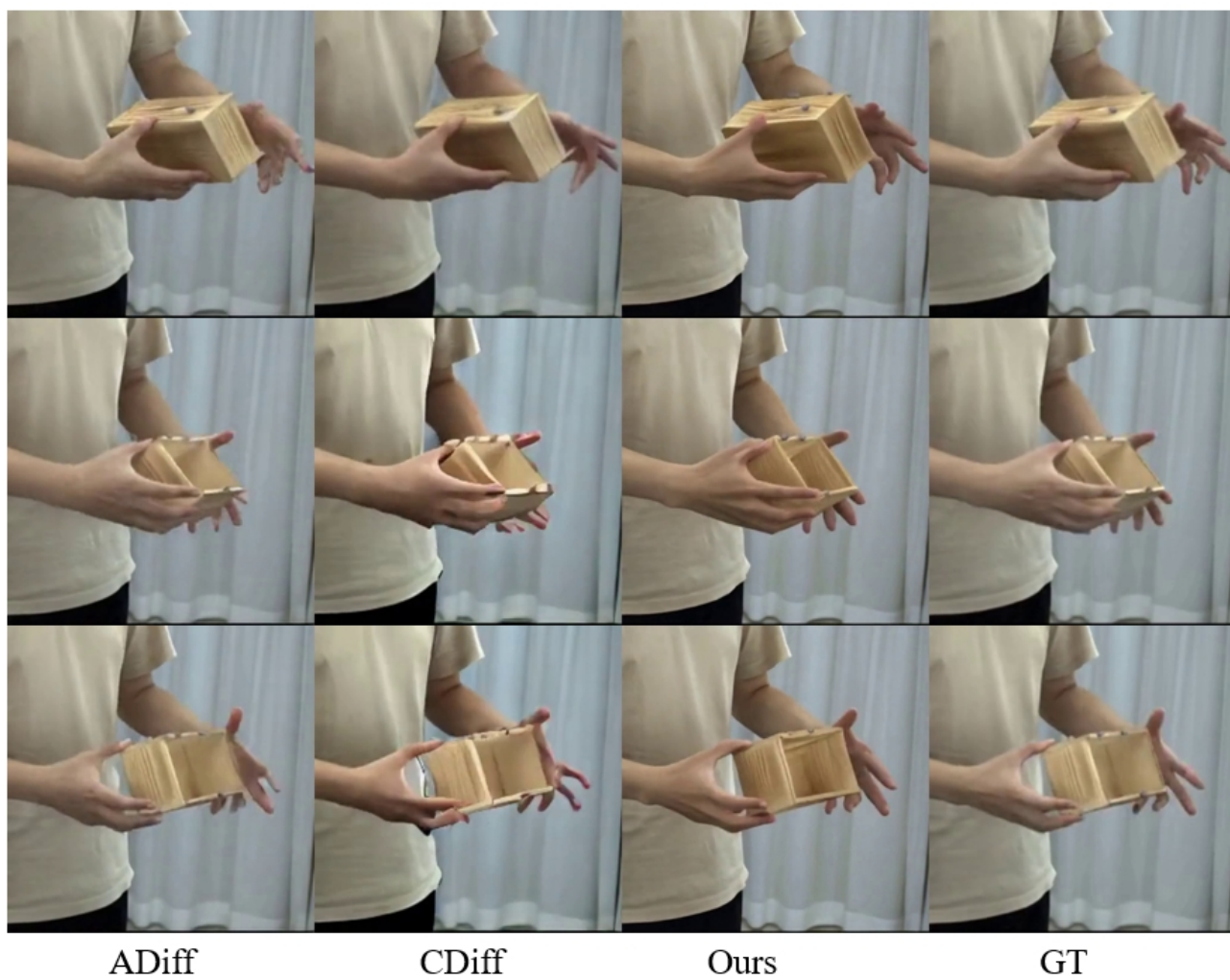


Figure 5. Qualitative comparison of different methods on our dataset. Our approach achieves the best results.



ADiff

CDiff

Ours

GT

Figure 6. Qualitative comparison of different methods on our dataset. Our approach achieves the best results.



Figure 7. Qualitative comparison of different methods on DexYCB dataset [1]. Our results perform best in cases of hand-object mutual occlusion and finger self-occlusion.



Figure 8. Qualitative comparison of different methods on videos we collect. Our approach achieves the best results.