CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis Supplementary Material

A. Overview

In the supplementary paper, we first present the details of our method (Section B), our evaluation metrics (Section C), our experiments (Section D), and the data processing (Section E), respectively. We then compare our approach with an additional baseline (Section F) and conduct more ablation studies (Section G) to verify the effectiveness of different modules in our method.

Besides the paper, we strongly recommend watching our **video** to visualize our approach and its synthesis effects.

B. Method Details

Section B.1 shows the loss design of our **planner** (see Section 4.2 in the main paper) in detail, while Section B.2 provides the details of loss functions in our **synthesizer** (see Section 4.3 in the main paper). In the following subsections, we use hatted symbols (*e.g.* $\hat{\mathbf{C}}_{i,j}$) to denote network predictions, and vise versa (*e.g.* $\mathbf{C}_{i,j}$) for the ground truth values.

B.1. Definitions of Loss Functions in Planner

In this section, any notation with hat indicates a ground truth value. We first define our BCE loss $\mathcal{L}_{flag}$ as:

$$\begin{aligned}\mathcal{L}_{flag} = &\sum_{i=1}^{5}\sum_{j=1}^{M}\sum_{k=1}^{N} BCE(\hat{\mathbf{c}}_{ijk}, \mathbf{c}_{ijk}) \\ &+ \sum_{i=1}^{5}\sum_{j=1}^{M}\sum_{k=1}^{N}\sum_{t=1}^{T} BCE(\hat{\mathbf{f}}_{n,ijkt}, \mathbf{f}_{n,ijkt}) \\ &+ \sum_{i=1}^{5}\sum_{j=1}^{M}\sum_{k=1}^{N}\sum_{t=1}^{T} BCE(\hat{\mathbf{f}}_{c,ijkt}, \mathbf{f}_{c,ijkt}),\end{aligned} \tag{1}$$

in which the subscript $1 \le t \le T$ indicates a sampled frame index in time-continuous part.

Two L_2 loss functions $\mathcal{L}_{tip}$ and $\mathcal{L}_{vec}$ are defined as:

$$\begin{aligned}\mathcal{L}_{tip} = &\sum_{i=1}^{5}\sum_{j=1}^{M}\sum_{k=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{c}}_{ijk}(\|\hat{\mathbf{J}}^{tip}_{start,ijkt} - \mathbf{J}^{tip}_{start,ijkt}\|_2^2 \\ &+ \|\hat{\mathbf{J}}^{tip}_{end,ijk} - \mathbf{J}^{tip}_{end,ijk}\|_2^2), \\ \mathcal{L}_{vec} = &\sum_{joint\in\{dip,pip,mcp,root\}} \mathcal{L}_{vec}(joint), \\ \mathcal{L}_{vec}(joint) = &\sum_{i=1}^{5}\sum_{j=1}^{M}\sum_{k=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{c}}_{ijk}(\|\hat{\mathbf{D}}^{joint}_{start,ijkt} - \mathbf{D}^{joint}_{start,ijkt}\|_2^2 \\ &+ \|\hat{\mathbf{D}}^{joint}_{end,ijk} - \mathbf{D}^{joint}_{end,ijk}\|_2^2),\end{aligned} \tag{2}$$

where j denotes the index of the stage that contains frame t, and $k \in \{1, 2, 3, root\}$ indicates different joints on a finger.

The KL-Divergence loss $\mathcal{L}_{KLD}$ is defined as:

$$\mathcal{L}_{KLD} = KL(\mathbf{Q}(z|\mu, \theta^2)\|\mathcal{N}(0, I)). \tag{3}$$

B.2. Definitions of Loss Functions in Synthesizer

The L2 loss $\mathcal{L}_{tip}$ for measuring fingertip positions is defined as:

$$\begin{aligned}\mathcal{L}_{\text{tip}}(\theta_t) = &\sum_{i=1}^{5}\sum_{k=1}^{N}(\mathbf{f}_n)_{i,t,k} \\ &\|\text{MANO}(\mathbf{J}^{tip}_i; \beta, \theta_t) - \mathbf{J}^{tip}_{i,t,k}\|^2,\end{aligned} \tag{4}$$

where $\text{MANO}(\mathbf{J}^{tip}_i; \beta, \theta_t)$ [3] computes the tip position of finger i given the MANO parameter $\{\beta, \theta_t\}$, and $\mathbf{J}^{tip}_{i,t,k}$ indicates the target fingertip position in the t-th frame with respect to the information from the k-th object part, which is the linear interpolation from the two directly predicted fingertip positions $\{(\mathbf{J}_1)^{tip}_{i,t,k}, (\mathbf{J}_2)^{tip}_{i,t,k}\}$ in the finger embeddings $(\mathbf{F}_1)_{i,t,k}, (\mathbf{F}_2)_{i,t,k}$. The target fingertip position is then transformed into the world coordinate frame.

The L2 loss $\mathcal{L}_{joint}$ for measuring finger joint direction vectors is defined as:

$$\begin{aligned}\mathcal{L}_{\text{joint}}(\theta_t) = &\sum_{i=1}^{5}\sum_{k=1}^{N}\sum_{joint}(\mathbf{f}_n)_{i,t,k} \\ &\|\text{MANO}(\mathbf{J}_i^{joint};\beta,\theta_t) \\ &-\text{MANO}(\mathbf{J}_i^{tip};\beta,\theta_t) - \mathbf{D}_{i,t,k}^{joint}\|^2,\end{aligned} \tag{5}$$

where $\text{MANO}(\mathbf{J}_i^{joint};\beta,\theta_t)$ is each of the joint positions of finger i given the MANO parameter $\{\beta,\theta_t\}$, and $\mathbf{D}_{i,t,k}^{joint}$ is the target joint orientation, similarly interpolated and transformed from the predicted $\mathbf{D}$ as in Equation (4).

The temporal smoothness loss $\mathcal{L}_{\text{smooth}}$ is defined as:

$$\mathcal{L}_{\text{smooth}}(\theta) = \sum_{t=2}^{T}\|\theta_t - \theta_{t-1}\|^2. \tag{6}$$

We use $j(t)$ to denote the stage index containing the time t. To define the contact loss $\mathcal{L}_{\text{contact}}$, we first find all frame-finger index tuples (i,t,k) satisfying $(\mathbf{f}_c)_{i,t,k}=1$, indicating where contact should take place. For each of such tuple (i,t,k), we search for a local surface section $(\mathbf{M}'_k)_{t,i}$ of the contacting object part $\mathbf{M}_k$. $(\mathbf{M}'_k)_{t,i}$ is set to be the closest section to the predicted reference frame $\mathbf{V}_{i,j(t),k}$ where all the vertex normals in such a local section are within a fixed included angle range $\alpha = 45^\circ$ compared to the predicted contact normal $\mathbf{N}_{i,j(t),k}$. We calculate the signed distances $\{\mathbf{d}_{i,t,k,l}\}_{l=1}^{n(i)}$ from each of the finger vertices to $(\mathbf{M}'_k)_{t,i}$, as well as the corresponding nearest points $\{\mathbf{p}_{i,t,k,l}\}_{l=1}^{n(i)}$ on $(\mathbf{M}'_k)_{t,i}$, where $n(i)$ is the number of vertices on the i-th finger. The contact loss $\mathcal{L}_{\text{contact}}$ is designed to attract the nearby finger vertices to the local surface section:

$$\begin{aligned}\mathcal{L}_{\text{contact}}(\theta') = &\sum_{(\mathbf{f}_c)_{i,t,k}=1}\sum_{l=1}^{n(i)} c_{i,t,k,l} \\ &\cdot\|\text{MANO}(\mathbf{vertex}_{i,l};\beta,\theta'_t) - \mathbf{p}_{i,t,k,l}\|^2,\end{aligned} \tag{7}$$

where $c_{i,t,k,l} = e^{-(\|\mathbf{d}_{i,t,k,l}\|^2 - \min_{l^*}\|\mathbf{d}_{i,t,k,l^*}\|^2)}$ provides centralized coefficients for such finger vertices.

Similar to $\{\mathbf{d}_{i,t,k,l}\}_{l=1}^{n(i)}$ and $\{\mathbf{p}_{i,t,k,l}\}_{l=1}^{n(i)}$, we compute $\{\mathbf{d}'_{i,t,k,l}\}_{l=1}^{n(i)}$ and $\{\mathbf{p}'_{i,t,k,l}\}_{l=1}^{n(i)}$ that are towards the whole object part $\mathbf{M}_k$ but not the local $(\mathbf{M}'_k)_{t,i}$, and thus define the penetration loss $\mathcal{L}_{\text{penetr}}$ as:

$$\begin{aligned}\mathcal{L}_{\text{penetr}}(\theta') = &\sum_{i,t,k}\sum_{l=1}^{n(i)} p_{i,t,k,l}\cdot c'_{i,t,k,l} \\ &\cdot\|\text{MANO}(\mathbf{vertex}_{i,l};\beta,\theta'_t) - \mathbf{p}'_{i,t,k,l}\|^2,\end{aligned} \tag{8}$$

where $c'_{i,t,k,l} = e^{-(\|\mathbf{d}'_{i,t,k,l}\|^2 - \min_{l^*}\|\mathbf{d}'_{i,t,k,l^*}\|^2)}$, and $p_{i,t,k,l} = 1$ only if the hand vertex penetrates the object part in frame t.

C. Metric Details

In this section, we introduce our evaluation metrics (see Section 5.4 in the main paper) in detail.

C.1. Contact-Movement Consistency

We evaluate whether the object's movement can align with the contact forces produced by hand-object contacts, using the same physics model as in ManipNet [4].

Following [4], for each part of the object in a single video frame, we calculate the change of linear momentum $\dot{P}$ and angular momentum $\dot{L}$.

$$\begin{aligned}&P(t) = Mv(t),\\ &\dot{P}(t) = M\dot{v}(t),\\ &L(t) = I(t)\omega(t),\\ &\dot{L}(t) = \dot{I}(t)\omega(t) + I(t)\dot{\omega}(t),\\ &I(t) = R(t)I_0R(t)^T,\\ &\dot{I}(t) = \dot{R}(t)I_0R(t)^T + R(t)I_0\dot{R}(t)^T,\\ &\dot{R}(t)^T = [\omega(t)]R(t).\end{aligned} \tag{9}$$

The linear and angular velocities v,ω and accelerations $\dot{v},\dot{\omega}$ of a part can be calculated from its trajectory, and $R(t)$ is the rotation matrix of the part in frame t. The mass M is set to 1 since we do not set constraints on the attitude of forces on the fingers. We stack $\dot{P}$ and $\dot{L}$ into $\mathbf{b}$.

For each frame, we also calculate the force the fingers can apply on the object. First, we find all the contact points on the object mesh, which are less than 2mm from the fingers. To model frictional contact, we use four bases to approximate Coulomb's friction cone. We use 0.35 as the friction coefficient. We additionally add opposite force bases with the same coefficient variables on both parts of an articulated object, perpendicular to the spin axis of the object, modeling the articulation. We also add the supporting force basis of the part if the part is considered on the world's ground, which is to have an altitude above the lowest within 5mm.

Let K denote the number of force bases. We compute the force and torque for each part in frame t:

$$\begin{aligned}F(t) &= \sum_{i=1}^{K} V_i(t)x_i(t) + Mg,\\ \tau(t) &= \sum_{i=1}^{K}[c_i(t) - o(t)]V_i(t)x_i(t),\end{aligned} \tag{10}$$

where V is the force basis at each contact point, c is the corresponding contact point, and o is the part center of mass. $\mathbf{x} = [x_1(t), x_2(t), \cdots, x_K(t)]^T$ are the non-negative

coefficients that we can apply along each force basis. Since both $F(t)$ and $\tau(t)$ in Equation (10) are linear w.r.t. $\mathbf{x}$, we denote the right part of Equation (10) as a linear transformation $\mathbf{Ax} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{(3+3)\times K}$ and $\mathbf{b} \in \mathbb{R}^{3+3}$.

Given a pose trajectory for an object part, we can approximate the velocity and acceleration of the object part, hence computing the expected values of $F(t)$ and $\tau(t)$ that could force the object to move along such a trajectory. Let $\mathbf{c}$ denotes the 6D concatenation $[F(t), \tau(t)]^T$. We thus find an optimal $\mathbf{x}$ that minimize $\|\mathbf{Ax} + \mathbf{b} - \mathbf{c}\|_2$ with the constraint $\mathbf{x} \geq 0$. Only if $\min_{\mathbf{x}} \|\mathbf{Ax} + \mathbf{b} - \mathbf{c}\|_2 < 0.01$ do we consider a frame is consistent between contact and object movement. As one of our evaluation metrics, we compute the proportion of such consistent frames among all video frames.

C.2. Articulation Consistency

We evaluate whether the hand pose can control the object state in a human-like manner for an articulated object with a single revolute joint. The key insight is that, for each part of such an articulated object, the torque imposed by a human hand should be along the direction of the revolute joint. We consider all the contact points on the object surface that are less than 2mm from the fingers. To simulate the forces on the object, we first add the gravity as well as the supporting force of the world's ground into our calculation similar to Section C.1. We then suppose that a unified force is applied along the normal direction of each contact point, and compute the torque of each force w.r.t. the revolute joint of the articulated object. We normalize these torques with the inertia of the object.

Let $\vec{d}$ denote the direction of the object revolute joint. For each object part, we compute $E_{Art} = \max_{\tau} \tau \cdot \vec{d}$, where τ is the torque applied at a contact point. We consider a video frame achieves articulation consistency, only if each object part in this frame satisfies $E_{Art} > 0.3$. We calculate the proportion of such qualified video frames as one of our evaluation metrics.

C.3. Perceptual Score

We collect human perceptual scores to judge the naturalness of the motion sequences.

We invite 23 people who are not familiar with motion synthesis and have no information on our method or any of the baselines to rate the generated animation videos from both our approach and baselines. These people are given 10 different result videos per method per object category, where all videos are shuffled, and the corresponding method names are blind to the people. After that, they are required to rate each video from 1 to 5. The scoring rules are:

- 1 point: This video is very different from human behaviors, with lots of physical unrealities such as penetration and hand-object separation;
- 3 points: This video is human-like and physically plausible to some extent, but the tester can still detect the difference with human behavior or obvious physical defects;
- 5 points: This video is basically consistent with human behavior and is physically realistic.

For each method, we first compute the mean rating score of each object category and then report the average score among all categories in the main paper.

D. Experimental Details

Given triangular part meshes $\{\mathbf{M}_i\}_{i=1}^N$ of a manipulated object, we sample 1000 points from each part mesh. We use a batch size of 64 for training, and the training procedure contains 500 epochs for Laptops and 1000 epochs for other categories. To balance the effects of different loss functions, we empirically set $\lambda_{flag} = 0.1$, $\lambda_{pos} = 500$, $\lambda_{dir} = 100$, $\lambda_{tip} = 100$, $\lambda_{vec} = 1$, and $\lambda_{kld} = 5$ during training.

In our synthesizer, we first optimize MANO hand pose parameter θ in 2000 epochs for fitting finger embedding. We simply set $\lambda_{\text{tip}} = 50$, $\lambda_{\text{joint}} = 1$, and $\lambda_{\text{smooth}} = 0.05$ and 1000 respectively for the 45D MANO pose parameters and 3D wrist position parameters in θ. To further optimize contact and penetration, we then iteratively use 6 steps to progressively improve the θ, while in each step, the optimization process contains 500 epochs. We empirically set $\lambda_{\text{contact}} = 80$, $\lambda_{\text{trans}} = 1$, $\lambda_v = 5$ and $\lambda_a = 20$. To progressively improve the smoothness of our synthesis results, the parameter λ_{smooth} is set to 1 in the first two steps, 10 in the following two steps and 500 in the last two steps.

E. Data Processing for HOI4D

To better leverage the HOI4D [2] dataset for synthesis purposes, we performed several augmentation and modification steps to the raw data. We first split the object instances into training and testing sets in a proportion of $7:3$. Due to the limitation of the data collection method, the original HOI4D has a non-negligible problem of noise and penetration. To eliminate the noise, we extract several keyframes for each segment in training data and perform smooth interpolations between the keyframes. To solve the penetration problem, we use a contact optimization technique that is almost the same as our synthesizer module, in which we manually specify the contact points in each keyframe.

F. Additional Baseline

GraspTTA [1]+TOCH [5]: TOCH was developed to refine the hand poses given a coarse hand-object manipulation. Benefiting from the dense field representation, TOCH

	Pliers			Scissors			Laptop		
	Pen (%) ↓	Mov ↑	Art ↑	Pen (%) ↓	Mov ↑	Art ↑	Pen (%) ↓	Mov ↑	Art ↑
Ground Truth	0.000	1.000	1.000	0.046	1.000	0.970	0.316	1.000	1.000
GraspTTA	0.555	0.779	0.420	0.454	0.993	0.849	5.211	**1.000**	0.997
GraspTTA+TOCH	0.124	0.918	0.511	0.124	0.947	0.298	1.596	**1.000**	0.978
CAMS (Ours)	**0.004**	**1.000**	**1.000**	**0.080**	**0.999**	**0.989**	**0.906**	**1.000**	**1.000**

	Kettle			Overall		
	Pen (%) ↓	Mov ↑	Art ↑	Pen (%) ↓	Mov ↑	Art ↑
Ground Truth	0.602	1.000	N/A	0.241	1.000	0.990
GraspTTA	4.852	0.586	N/A	2.768	0.839	0.755
GraspTTA+TOCH	2.002	0.897	N/A	0.961	0.941	0.596
CAMS (Ours)	**0.098**	**0.915**	N/A	**0.272**	**0.978**	**0.996**

Table 1. **Quantitative results compared with GraspTTA [1] and GraspTTA+TOCH [5].** "Pen" denotes the average percentage of hand vertices penetrated in the object. "Mov" denotes the average proportion of frames that are contact-movement consistent. "Art" denotes the average proportion of frames that are articulation consistent.

	Pen (%) ↓	Mov ↑	Art ↑
Ground Truth	0.000	1.000	1.000
CAMS (w/o Obj Can)	0.147	**1.000**	**1.000**
CAMS (w/o Contact Can)	0.021	**1.000**	0.879
CAMS (Abs Finger)	0.009	0.998	**1.000**
CAMS	**0.004**	**1.000**	**1.000**

Table 2. **Additional ablation studies on the "Pliers" category.**

could precisely sense the local geometry of contact. Thus we additionally implement a baseline that further refines the result of GraspTTA using TOCH. To better adapt the imperfect generation results from ManipNet, we train TOCH networks with ground truth manipulation trajectories added with random noise, and test in unseen manipulation animations generated from GraspTTA.

Quantitative result: As shown in Table 1, GraspTTA+TOCH could generate physically more realistic results than simply using GraspTTA, whereas our approach still significantly outperforms the baselines.

G. Additional Ablation Studies

Object-centric Canonicalization We ablate the root-level canonicalization and thus generate contact reference frames directly in the original object frame rather than the scale-normalized one. As shown in Table 2 (2nd line), our CAMS performs worse without object-centric canonicalization, and the generalizability of our framework also decreases.

Contact-centric Canonicalization We also design experiments to demonstrate the necessity of canonicalizing finger embedding into the contact reference frames. Table 2 (3rd line) shows that both penetration rate and articulation rate significantly decrease after removing the contact-centric canonicalization, indicating that our contact-centric canonicalization could help improve the synthesis quality.

Absolute Finger Embedding Instead of representing the canonicalized finger embedding as $\mathbf{F}_i = (\mathbf{J}_i^{tip}, \mathbf{D}_i^{dip}, \mathbf{D}_i^{pip}, \mathbf{D}_i^{mcp}, \mathbf{D}_i^{root})$, we use similar representations to $\mathbf{J}_i^{tip}$ for other joints, denoted as $\mathbf{F}_i^{abs} = (\mathbf{J}_i^{tip}, \mathbf{J}_i^{dip}, \mathbf{J}_i^{pip}, \mathbf{J}_i^{mcp}, \mathbf{J}_i^{root})$, which are absolute positions of the i-th finger's joints respectively. Table 2 (4th line) shows that using the absolute position for finger joints could harm the whole framework.

References

[1] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 3, 4

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 3

[3] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 1

[4] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2

[5] Keyang Zhou, Bharat Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. *arXiv preprint arXiv:2205.07982*, 2022. 3, 4