DUETHOI: LANGUAGE-GUIDED BIMANUAL HAND–OBJECT MOTION GENERATION WITH ARTICULATION PLANNING AND CONTACT REFINEMENT

Anonymous authors
Paper under double-blind review

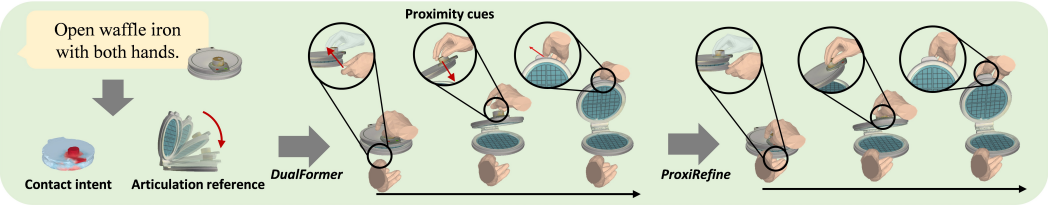


Figure 1: DuetHOI generation process. Contact intent and an articulation reference guide global interaction generation, after which ProxiRefine keeps the object trajectory fixed while correcting local bimanual geometry with proximity cues.

ABSTRACT

Bimanual articulated-object interaction generation requires a model to capture the evolution of object articulation, coordination between the two hands, and fine-grained hand–object contact. Existing methods typically encode object and hand motion as a unified high-dimensional sequence, making it difficult to explicitly accommodate the different scales of manipulation progress, relative bimanual motion, and local hand pose. We propose **DuetHOI**, a structured framework for bimanual articulated-object interaction generation. DuetHOI first uses Contact-VAE to predict contact intent on the object surface from language and object geometry. Conditioned on this spatial intent, ArtPlanner provides a trajectory-level articulation reference, and DualFormer generates the global interaction using object states, object-centered hand positions, and compact hand-pose tokens learned by ManiVAE. After global generation, ProxiRefine fixes the object trajectory and residually corrects the two hands using current hand–object proximity, improving local surface alignment. Experiments across articulated and rigid objects and single- and bimanual settings show that DuetHOI outperforms three adapted baselines on most contact and hand–object consistency metrics, with particularly strong results in bimanual articulated interaction generation.

1 INTRODUCTION

Opening a laptop, operating a microwave, or rotating a movable object part requires two hands to coordinate around a changing object state. Such language-driven motion generation is important for dexterous robotics Zhang et al. (2024b); Christen et al. (2022); Li et al. (2025a), embodied agents Li et al. (2025c), and virtual character animation Zhang et al. (2025). Language-driven bimanual articulated-object interaction generation is the focus of this work: given an object and a language instruction, the model infers action regions on the object surface and generates coherent object and two-hand motion.

Three coupled challenges arise in bimanual articulated manipulation: **articulation progress**, for which object-state errors alter the subsequent manipulation; **bimanual coordination**, which assigns complementary hand roles in the object-centered frame; and **local contact geometry**, where finger-pose errors cause contact failures. Recent data-driven HOI methods commonly concatenate object motion, hand trajectories, and hand poses into a unified sequence Jiang et al. (2021); Li et al. (2024);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Cha et al. (2024); Li et al. (2025b); Ye et al. (2023); Huang et al. (2025). This representation entangles large-scale state changes with fine-grained hand motion and offers limited control for correcting local contact after the global object trajectory is generated Grady et al. (2021); Jiang et al. (2021); Liu et al. (2023). In addition, an instruction can admit several valid contact locations, hand-role assignments, and trajectories Ye et al. (2023); Yang et al. (2024); Liu et al. (2023).

These challenges motivate decomposing the interaction at both the representation and generation levels. Object state, object-centered hand positions, and local hand poses describe manipulation progress, bimanual coordination, and contact configuration, respectively. The model first generates the global interaction and then refines local hand geometry with the generated object trajectory held fixed, preserving manipulation progress and improving hand–object alignment.

Accordingly, we propose **DuetHOI**. ContactVAE first infers object-side contact intent from language and canonical object geometry and represents it as a spatial contact map. Conditioned on this spatial prior, language, object geometry, and the initial articulation state, ArtPlanner predicts a trajectory-level articulation reference. ManiVAE encodes high-dimensional hand poses into compact latent tokens. DualFormer organizes object state, object-centered left/right hand positions, and left/right hand-pose tokens as a token–time grid, and generates the global interaction conditioned on the contact map and articulation reference. The generated object trajectory is then held fixed while ProxiRefine predicts hand residuals from current hand–object proximity to correct local surface alignment.

Experiments evaluate DuetHOI on articulated and rigid objects in single- and bimanual settings. The results show consistent improvements in contact quality, hand accuracy, and relative hand–object consistency, with particularly strong performance in bimanual articulated-object interactions.

Our main contributions are:

- We propose DuetHOI, a structured framework that represents object state, object-centered bimanual positions, and compact hand poses in separate streams, and stages global interaction generation with object-fixed local hand refinement.
- We introduce ContactVAE and ArtPlanner to provide complementary spatial and temporal conditions: a predicted object-side contact map and a contact-conditioned articulation reference.
- We develop ManiVAE, DualFormer, and ProxiRefine for structured generation and refinement: latent pose tokens and token–time attention model bimanual interaction, while proximity-guided residuals refine local hand–object alignment without altering object motion.

2 Related Work

2.1 Bimanual Hand Object Interaction

HOI generation methods are broadly physics-based or data-driven. Physics-based methods learn manipulation policies through reinforcement learning in simulation Zhang et al. (2024a;b); Xu et al. (2023); Christen et al. (2022), but depend on simulator fidelity and reward design, which can limit generalization. Data-driven methods learn HOI distributions from multimodal motion data Yang et al. (2022); Taheri et al. (2022); Liu et al. (2021); Corona et al. (2020); Turpin et al. (2022); Yang et al. (2024); Zheng et al. (2023), commonly using autoencoder-based Li et al. (2024); Huang et al. (2025); Jiang et al. (2021); Taheri et al. (2020); Wu et al. (2022) or diffusion-based models Li et al. (2025b); Cha et al. (2024). Most jointly predict object motion, hand trajectories, articulation, and hand poses in a single representation, obscuring the distinct roles of global manipulation progress, bimanual coordination, and local contact. DuetHOI separates articulation planning, latent bimanual generation, and hand-pose refinement.

2.2 Human Motion Priors and Pose Refinement

Learned motion priors reduce the complexity of high-dimensional pose generation through variational, discrete, or diffusion-based latent representations Ludwig et al. (2025); Rempe et al. (2021);

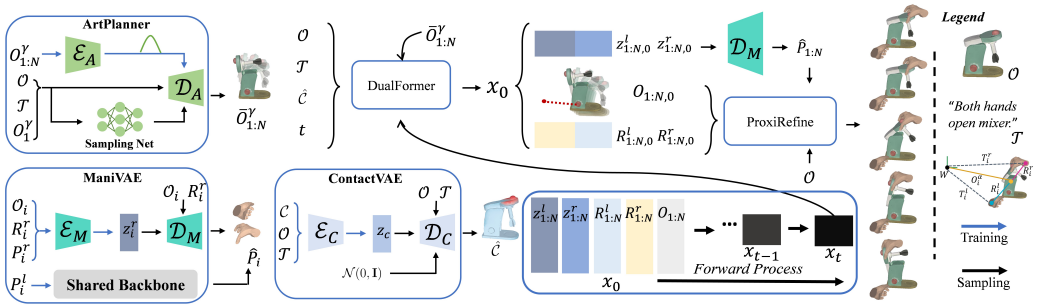


Figure 2: DuetHOI pipeline. ContactVAE predicts an object-side contact map that conditions both ArtPlanner and DualFormer. ArtPlanner provides a trajectory-level articulation reference for structured DualFormer generation, ManiVAE maps between hand poses and compact tokens, and ProxiRefine residually corrects the generated hands while preserving object motion. The lower-right inset visualizes the object-centered hand-position representation, with W denoting the world origin.

Petrovich et al. (2021; 2022); Guo et al. (2022); Tevet et al. (2022a); Zhou et al. (2024); Zhang et al. (2023); Guo et al. (2024). While these priors mainly regularize body motion, bimanual articulated HOI also demands precise contact and coordination with object articulation. Refinement methods correct local pose errors either generically Moon et al. (2019) or with contact prediction, test-time adaptation, and differentiable optimization for hand–object interaction Grady et al. (2021); Jiang et al. (2021); Liu et al. (2023). DuetHOI combines both ideas: ManiVAE encodes compact hand-pose tokens, and ProxiRefine corrects the generated hands in an object-fixed frame while preserving object motion.

3 Method

3.1 Overview of DuetHOI

Given object geometry $\mathcal{O}$ and a language instruction $\mathcal{T}$, DuetHOI first predicts an object-side contact map $\hat{\mathcal{C}} \in [0,1]^M$ and uses it to condition articulation planning and global interaction generation. The resulting output is an N-frame sequence of articulated object states and bimanual motion. At frame i, the hand state $H_i = \{T_i, P_i\}$ contains global translations $T_i = \{T_i^l, T_i^r\}$ and MANO poses Romero et al. (2022) $P_i = \{P_i^l, P_i^r\}$. The object state $O_i = \{O_i^\alpha, O_i^\beta, O_i^\gamma\}$ contains a 3D translation $O_i^\alpha \in \mathbb{R}^3$, a rot6d rotation Zhou et al. (2019) $O_i^\beta \in \mathbb{R}^6$, and a scalar articulation coordinate $O_i^\gamma \in \mathbb{R}$, which is set to zero for rigid objects. To remove the world translation shared by the object and hands, we use object-centered hand positions $R_i^h = T_i^h - O_i^\alpha$ for $h \in \{l, r\}$. Object geometry features are constructed in the same reference frame, allowing them and the relative hand positions to directly characterize hand–object spatial relations. This representation removes shared translation while preserving world-frame orientation; global hand translations are recovered by $\hat{T}_i^h = \hat{O}_i^\alpha + \hat{R}_i^h$. Text and geometry are encoded by frozen CLIP Radford et al. (2021) and PointNet++ Qi et al. (2017), respectively, except that ContactVAE uses a point-wise MLP.

Figure 2 shows the full pipeline. ContactVAE predicts an object-side contact map that captures contact intent. ArtPlanner uses this spatial prior to provide an articulation reference, while ManiVAE represents hand articulation with compact tokens. Conditioned on the contact map and articulation reference, DualFormer generates the global interaction, after which ProxiRefine adjusts the hands with the object trajectory fixed.

3.2 Language-Conditioned Contact Map Prediction

Language specifies manipulation intent but not the precise object regions acted on by the hands Ye et al. (2023); Yang et al. (2024); Liu et al. (2023). We sample M object-surface points and define a static spatial contact map $\mathcal{C} \in \{0,1\}^M$ that represents potential contact regions throughout the interaction. Here, $\mathcal{C}_m$ indicates whether the mth point belongs to such a region. ContactVAE predicts the corresponding probabilities $\hat{\mathcal{C}} \in [0,1]^M$ from the instruction and object point cloud. Its encoder infers $q_{\phi_C}(z_C \mid \mathcal{C}, \mathcal{O}, \mathcal{T})$ during training, and the decoder reconstructs $\mathcal{C}$. Because contact

is sparse, we combine binary cross-entropy with Dice reconstruction Milletari et al. (2016):

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{M}\sum_{m=1}^{M}\left[\mathcal{C}_m \log \hat{\mathcal{C}}_m + (1-\mathcal{C}_m)\log(1-\hat{\mathcal{C}}_m)\right], \qquad \mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2\sum_m \hat{\mathcal{C}}_m \mathcal{C}_m + \varepsilon_C}{\sum_m \hat{\mathcal{C}}_m + \sum_m \mathcal{C}_m + \varepsilon_C}. \tag{1}$$

The constant $\varepsilon_C > 0$ stabilizes the Dice ratio. The latent posterior is regularized for inference-time sampling by

$$\mathcal{L}_C^{\mathrm{KL}} = D_{\mathrm{KL}}(q_{\phi_C}(z_C \mid \mathcal{C}, \mathcal{O}, \mathcal{T}) \,\|\, \mathcal{N}(0, \mathbf{I})). \tag{2}$$

At inference, the contact map is generated as

$$z_C \sim \mathcal{N}(0, \mathbf{I}), \qquad \hat{\mathcal{C}} = D_C(z_C; \mathcal{O}, \mathcal{T}). \tag{3}$$

The principal ContactVAE objective is

$$\mathcal{L}_C = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{Dice}} + \beta_C \mathcal{L}_C^{\mathrm{KL}}. \tag{4}$$

The predicted contact probabilities condition both ArtPlanner and DualFormer.

3.3 Articulation Trajectory Planning

Object articulation provides a low-dimensional description of manipulation progress. ArtPlanner predicts its scalar trajectory $O^{\gamma}_{1:N}$ before full hand–object generation, conditioned on language $\mathcal{T}$, object geometry $\mathcal{O}$, the initial articulation O^{γ}_1, and the predicted contact probabilities. The scalar represents an angle for revolute joints and a displacement for prismatic joints; rigid objects use $O^{\gamma}_{1:N} = \mathbf{0}$.

Let $c_A = (\mathcal{O}, \mathcal{T}, O^{\gamma}_1, \hat{\mathcal{C}})$. During training, the encoder maps the ground-truth trajectory and condition to

$$q_{\phi_A}(z_A \mid O^{\gamma}_{1:N}, c_A) = \mathcal{N}(\mu^q_A, \sigma^q_A), \tag{5}$$

and reconstructs the training trajectory from a posterior sample:

$$z^q_A \sim q_{\phi_A}(z_A \mid O^{\gamma}_{1:N}, c_A), \qquad \bar{O}^{\gamma}_{1:N} = D_A(z^q_A; c_A). \tag{6}$$

The posterior-decoded reference is supervised by

$$\mathcal{L}_{\mathrm{traj}} = \frac{1}{N}\sum_{i=1}^{N}\left|\bar{O}^{\gamma}_i - O^{\gamma}_i\right|. \tag{7}$$

To generate an articulation reference from conditions alone at test time, we also learn a conditional prior

$$p_{\psi_A}(z_A \mid c_A) = \mathcal{N}(\mu^p_A, \sigma^p_A). \tag{8}$$

Here (μ^q_A, σ^q_A) and (μ^p_A, σ^p_A) parameterize the posterior and conditional-prior Gaussians, respectively. The conditional latent space is aligned by

$$\mathcal{L}_A^{\mathrm{KL}} = D_{\mathrm{KL}}(q_{\phi_A}(z_A \mid O^{\gamma}_{1:N}, c_A) \,\|\, p_{\psi_A}(z_A \mid c_A)). \tag{9}$$

The principal ArtPlanner objective is

$$\mathcal{L}_A = \mathcal{L}_{\mathrm{traj}} + \beta_A \mathcal{L}_A^{\mathrm{KL}}. \tag{10}$$

At inference, the deterministic main pipeline decodes the conditional-prior mean:

$$\bar{O}^{\gamma}_{1:N} = D_A(\mu^p_A; c_A). \tag{11}$$

For stochastic inference, we sample $z^p_A \sim p_{\psi_A}(z_A \mid c_A)$ and decode $\bar{O}^{\gamma}_{1:N} = D_A(z^p_A; c_A)$ as the articulation reference; rigid interactions use $\bar{O}^{\gamma}_{1:N} = \mathbf{0}$. DualFormer produces the final articulation prediction.

3.4 Local Hand Pose Prior Learning

Directly generating frame-wise MANO rotations substantially enlarges the structured motion space. ManiVAE encodes local hand articulation into a compact token Li et al. (2024; 2025b). We denote by $\mathcal{O}_i$ the object geometry $\mathcal{O}$ transformed by the rotation O^{β}_i and articulation O^{γ}_i at frame i, with the object translation O^{α}_i set to zero. Both the encoder and decoder are conditioned on the object-centered hand position R^h_i and the transformed object geometry $\mathcal{O}_i$. These conditions allow the latent token to focus on local hand articulation without redundantly encoding global placement, yielding a more compact pose representation and encouraging the decoded articulation to remain consistent with the current hand–object configuration.

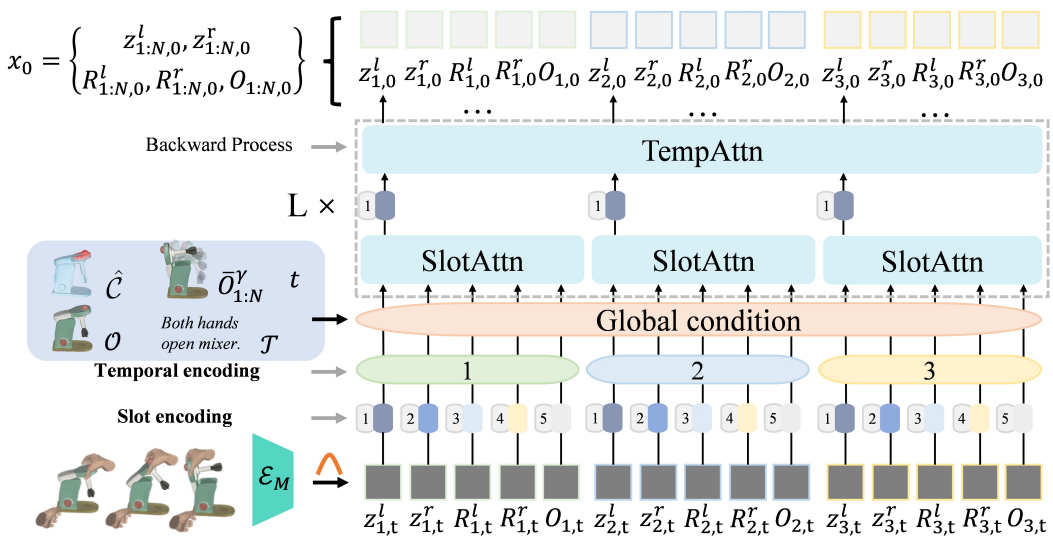


Figure 3: DualFormer. Slot attention models within-frame hand–object coordination across the five semantic streams, while temporal attention captures the evolution of each stream throughout the denoising process.

We train one ManiVAE for both hands in a canonical right-hand space. Right-hand samples remain unchanged; left-hand samples are mirrored before encoding and mapped back after decoding. Given the canonicalized pose P_i^h, ManiVAE encodes a hand-pose token and reconstructs the MANO articulation:

$$z_i^h \sim q_{\phi_M}(z_i^h \mid P_i^h, R_i^h, \mathcal{O}_i), \qquad \hat{P}_i^h = D_M(z_i^h, R_i^h, \mathcal{O}_i). \tag{12}$$

Global hand translation is recovered from the generated object translation and object-centered hand position.

We train ManiVAE with parameter- and geometry-space reconstruction. The former directly supervises the 6D MANO representation,

$$\mathcal{L}^M_{\text{param}} = \left\| P_i^h - \hat{P}_i^h \right\|_2^2. \tag{13}$$

Let $\mathcal{V}(\cdot)$ and $\mathcal{J}(\cdot)$ denote MANO vertices and joints. We supervise the decoded hand geometry by

$$\mathcal{L}^M_{\text{geom}} = \left\| \mathcal{V}(\hat{P}_i^h) - \mathcal{V}(P_i^h) \right\|_2^2 + \left\| \mathcal{J}(\hat{P}_i^h) - \mathcal{J}(P_i^h) \right\|_2^2. \tag{14}$$

The latent space is regularized by

$$\mathcal{L}^{\text{KL}}_M = D_{\text{KL}}\left(q_{\phi_M}(z_i^h \mid P_i^h, R_i^h, \mathcal{O}_i) \,\|\, \mathcal{N}(0, \mathbf{I}) \right). \tag{15}$$

The compact objective is

$$\mathcal{L}_M = \mathcal{L}^M_{\text{param}} + \mathcal{L}^M_{\text{geom}} + \beta_M \mathcal{L}^{\text{KL}}_M. \tag{16}$$

We freeze ManiVAE and use the posterior mean $z_i^h = \mu_{\phi_M}(P_i^h, R_i^h, \mathcal{O}_i)$ as the hand-token target for DualFormer.

3.5 Token-Temporal Interaction Generation

DualFormer generates the global interaction sequence on a token-temporal grid. After ArtPlanner provides an articulation reference and ManiVAE maps hand articulation into compact latent tokens, the clean diffusion target is represented with five semantic slots:

$$x_0 = \left[z^l_{1:N,0},\ z^r_{1:N,0},\ R^l_{1:N,0},\ R^r_{1:N,0},\ O_{1:N,0} \right]. \tag{17}$$

The five streams separate object motion, relative hand placement, and local hand articulation. We standardize each stream with training-set statistics and apply stream-specific input and output projections. The DDPM Ho et al. (2020) forward process gives

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{18}$$

where t is the diffusion timestep and $\bar{\alpha}_t$ is the cumulative noise schedule.

A naive flattened-attention design jointly models all NS tokens. DualFormer instead factorizes this operation into separate attention operations for within-frame coordination and temporal propagation. Each component $x^a_{i,t}$ at frame i and slot $a \in \{1, \ldots, S\}$ is mapped into a shared hidden space by the slot-specific projection $e^a_{i,t} = \mathrm{Proj}_a(x^a_{i,t}) + \mathrm{PE}_{\mathrm{time}}(i)$, which also encodes its stream identity. At a fixed diffusion timestep t, the projected streams form $e_t \in \mathbb{R}^{N\times S\times d}$, where $S = 5$ is the number of semantic streams and d is the hidden dimension. Inspired by factorized video modeling Arnab et al. (2021), at each frame i, slot attention jointly processes the S tokens $(e^1_{i,t}, \ldots, e^S_{i,t})$ to model hand–object coordination. For each semantic slot a, temporal attention then processes $(e^a_{1,t}, \ldots, e^a_{N,t})$ across the N frames. Each DualFormer block applies these two operations sequentially. Both operations are modulated by the DualFormer condition c_D. This factorization reduces the attention cost from $O(N^2S^2)$ for flattened attention to $O(NS^2 + SN^2)$.

Let $c_D = (\mathcal{O}, \mathcal{T}, \hat{\mathcal{C}}, \bar{O}^{\gamma}_{1:N}, t)$ denote the DualFormer condition. It combines object geometry, language, predicted contact probabilities, the ArtPlanner trajectory, and the diffusion timestep. Each value in $\hat{\mathcal{C}}$ is appended to its corresponding object point before encoding. Adaptive normalization injects c_D into each block Peebles & Xie (2023). After the final block, the five slot features within each frame are concatenated and passed to stream-specific output heads to predict the clean interaction target $\hat{x}_0 = D_\theta(x_t, c_D)$. The frozen ManiVAE decoder maps each generated hand token back to a MANO pose, conditioned on the corresponding generated object-centered hand position and transformed object geometry.

We train DualFormer with clean-target prediction and component-wise supervision over the five semantic streams:

$$\mathcal{L}_{\mathrm{stream}} = \left\|\hat{O}_{1:N} - O_{1:N}\right\|_2^2 + \sum_{h\in\{l,r\}} \left(\left\|\hat{R}^h_{1:N} - R^h_{1:N}\right\|_2^2 + \left\|\hat{z}^h_{1:N} - z^h_{1:N}\right\|_2^2 \right). \tag{19}$$

Beyond rot6d reconstruction in the object stream, we supervise object orientation in $SO(3)$. Let $Q_i = \mathrm{Rot}(O^\beta_i)$ and $\hat{Q}_i = \mathrm{Rot}(\hat{O}^\beta_i)$; the geodesic loss is

$$\mathcal{L}_{\mathrm{rot}} = \frac{1}{N}\sum_{i=1}^{N} \arccos\left(\frac{\mathrm{tr}(\hat{Q}_i^\top Q_i) - 1}{2}\right). \tag{20}$$

The principal DualFormer objective is

$$\mathcal{L}_D = \mathcal{L}_{\mathrm{stream}} + \lambda_{\mathrm{rot}}\mathcal{L}_{\mathrm{rot}}. \tag{21}$$

At inference, standard reverse sampling produces the global interaction passed to ProxiRefine.

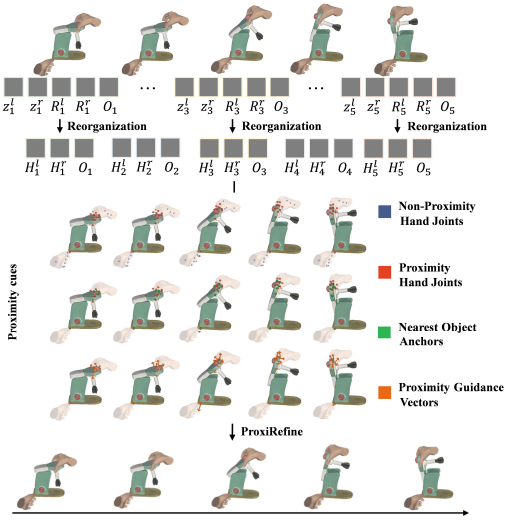


Figure 4: Object-fixed ProxiRefine. ProxiRefine derives joint-level proximity cues from the generated interaction and residually corrects bimanual position and pose while keeping the object trajectory fixed.

3.6 Object-Fixed Hand Refinement

Decoded ManiVAE poses may still contain local contact errors Grady et al. (2021); Jiang et al. (2021); Liu et al. (2023). ProxiRefine fixes the DualFormer object trajectory $\hat{O}_{1:N}$ and predicts bimanual position and pose residuals from current hand–object proximity.

ProxiRefine operates in direct MANO pose space. Let $\hat{\mathcal{O}}_i$ denote the translation-free object geometry transformed according to the generated rotation and articulation at frame i. For each $h \in \{l, r\}$, we decode the generated hand-pose token as $\hat{P}^h_i = D_M(\hat{z}^h_i, \hat{R}^h_i, \hat{\mathcal{O}}_i)$, recover its global translation as $\hat{T}^h_i = \hat{O}^\alpha_i + \hat{R}^h_i$, and form $\hat{H}^h_i = \{\hat{T}^h_i, \hat{P}^h_i\}$. ProxiRefine constructs joint-level proximity cues from the DualFormer result. Let $\mathcal{J}(\hat{H}^h_i)$ denote the MANO joints, and let $\Pi_{\hat{O}_i}(\cdot)$ return their nearest points on the generated object surface at frame i. We construct the proximity cue for hand h as

$$C^h_i = \left[\mathcal{J}(\hat{H}^h_i) - \Pi_{\hat{O}_i}(\mathcal{J}(\hat{H}^h_i)),\ d^h_i,\ \exp(-\eta d^h_i)\right]. \tag{22}$$

Dataset	Setting	Method	Phy↑	C-F1↑	MPJPE$_A$↓	OTE↓	ORE↓	SD↑	OD↑	C-Rate↑	C-Stab↑	NN-Rel↓
ARCTIC	Bim.-Art.	Text2HOI	0.830	0.698	13.901	0.405	12.670	4.540	8.338	0.619	0.955	0.227
		LatentHOI	0.955	0.767	14.183	0.425	14.439	7.205	11.638	0.799	0.957	0.162
		MDM	0.867	0.743	14.032	0.380	13.909	**8.007**	**11.746**	0.713	0.960	0.209
		DuetHOI (Ours)	**0.997**	**0.834**	**12.043**	**0.332**	**12.186**	4.489	10.271	**0.873**	**0.970**	**0.095**
	Bim.-Rig.	Text2HOI	0.875	0.631	14.002	0.551	**6.027**	3.770	6.522	0.608	**0.959**	0.188
		LatentHOI	0.845	0.694	14.329	0.545	9.438	**6.859**	10.243	0.642	0.954	0.178
		MDM	0.810	0.643	12.879	0.422	6.993	6.370	10.562	0.572	0.957	0.193
		DuetHOI (Ours)	**1.000**	**0.746**	**9.614**	**0.357**	7.301	4.047	**10.722**	**0.705**	0.956	**0.118**
GRAB	Bim.-Rig.	Text2HOI	0.789	0.725	29.240	0.765	**31.062**	6.866	10.191	0.617	0.949	0.156
		LatentHOI	0.752	0.644	30.966	1.247	87.927	5.008	6.375	0.527	0.938	0.162
		MDM	0.658	0.580	31.029	0.786	31.238	**18.048**	**21.801**	0.444	0.981	0.186
		DuetHOI (Ours)	**0.975**	**0.775**	**28.806**	**0.748**	48.124	9.318	15.858	**0.657**	**0.986**	**0.088**

Table 1: Main bimanual results on ARCTIC and GRAB. Arrows indicate the preferred metric direction.

where d_i^h is the joint-to-surface distance and η controls the distance decay. The cue combines surface offsets, distances, and proximity weights, while a distance threshold defines the proximity mask. As visualized in Figure 4, the resulting proximity mask marks the selected joints in red and the remaining joints in dark blue. For each selected joint, the green anchor is its nearest point on the generated object surface, and the orange arrow represents the corresponding surface-to-joint offset vector. The refiner takes the generated hands, object sequence, and proximity cues as input:

$$H_{1:N}^{l,\mathrm{ref}}, H_{1:N}^{r,\mathrm{ref}} = F_P(\hat{H}_{1:N}^{l}, \hat{H}_{1:N}^{r}, \hat{O}_{1:N}, C_{1:N}^{l}, C_{1:N}^{r}). \tag{23}$$

The refiner predicts residuals relative to its input hands, $H_i^{h,\mathrm{ref}} = \hat{H}_i^h + \Delta_i^h$ for $h \in \{l, r\}$. Translation residuals are added directly, whereas pose residuals are applied to the 6D representation of each MANO joint rotation and then converted back to rotation matrices. Temporal position encodings and hand-role embeddings couple the two hand streams. The residual heads are zero-initialized to preserve the input motion at the start of training.

The hand reconstruction term supervises relative placement and 6D MANO pose:

$$\mathcal{L}_{\mathrm{hand}}^P = \lambda_R^P \left(\left\| R_{1:N}^{l,\mathrm{ref}} - R_{1:N}^{l} \right\|_2^2 + \left\| R_{1:N}^{r,\mathrm{ref}} - R_{1:N}^{r} \right\|_2^2 \right) + \lambda_P^P \left(\left\| P_{1:N}^{l,\mathrm{ref}} - P_{1:N}^{l} \right\|_2^2 + \left\| P_{1:N}^{r,\mathrm{ref}} - P_{1:N}^{r} \right\|_2^2 \right). \tag{24}$$

Let $\mathcal{K}_i^h$ contain joints selected by the distance threshold, and let $|\mathcal{K}| = \sum_{h,i} |\mathcal{K}_i^h|$. We define the contact loss as

$$\mathcal{L}_{\mathrm{contact}} = \frac{1}{\max(|\mathcal{K}|, 1)} \sum_{h \in \{l,r\}} \sum_{i} \sum_{j \in \mathcal{K}_i^h} \left\| \mathcal{J}_j(H_i^{h,\mathrm{ref}}) - \Pi_{\hat{O}_i}\left(\mathcal{J}_j(\hat{H}_i^h) \right) \right\|_1. \tag{25}$$

We set the loss to zero when $|\mathcal{K}| = 0$. The surface targets are computed from the DualFormer output, preserving its contact assignment during refinement. The principal ProxiRefine objective is

$$\mathcal{L}_P = \mathcal{L}_{\mathrm{hand}}^P + \lambda_{\mathrm{contact}} \mathcal{L}_{\mathrm{contact}}. \tag{26}$$

4 Experiments

4.1 Experimental Setup

We evaluate on ARCTIC Fan et al. (2023) and GRAB Taheri et al. (2020) using four modes: single-hand articulated (Single-Art.), single-hand rigid (Single-Rig.), bimanual articulated (Bim.-Art.), and bimanual rigid (Bim.-Rig.). ARCTIC covers all four modes, whereas GRAB is evaluated on rigid objects. Each model is jointly trained on all available modes and evaluated per mode; absent-hand streams are masked. Bim.-Art. is our primary setting, while the others test the unified representation across hand usage and articulation. We adapt Text2HOI Cha et al. (2024), LatentHOI Li et al. (2025b), and MDM Tevet et al. (2022b) to the same language-conditioned HOI task, providing the same raw task inputs, data splits, and evaluation protocol while preserving their core modeling principles. We report Phy, C-F1, MPJPE$_A$, OTE, ORE, SD, OD, C-Rate, C-Stab, and NN-Rel for physical plausibility, contact, hand and object accuracy, diversity, stability, and relative hand–object consistency. Baseline adaptation and metric definitions are provided in the supplementary material.

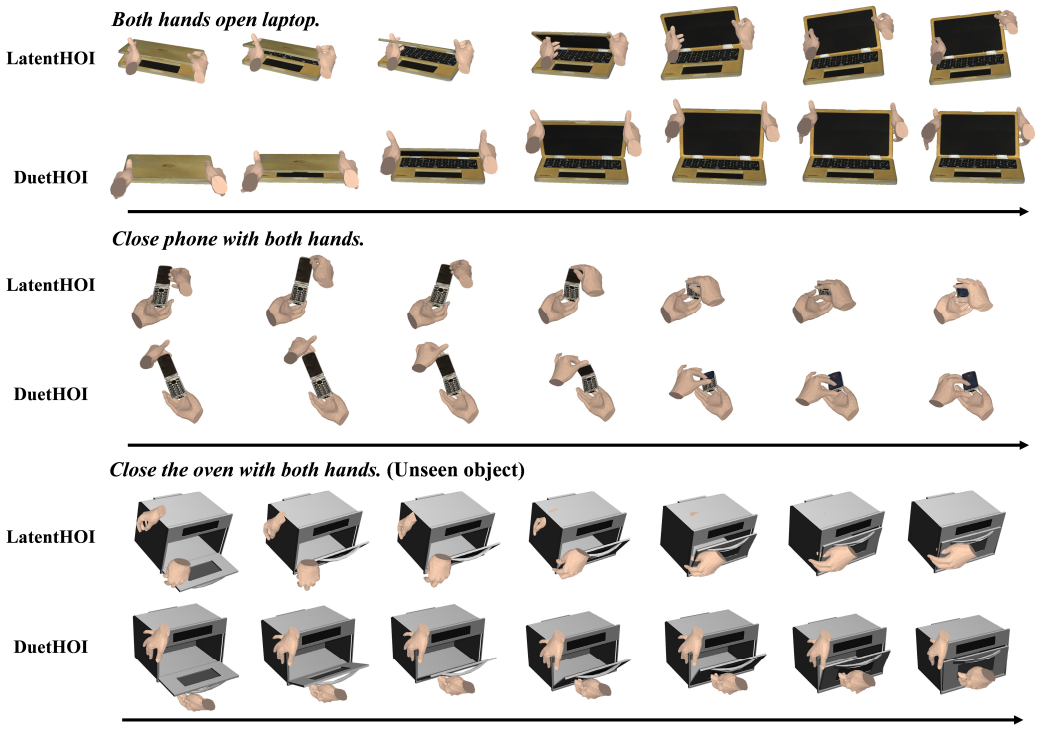


Figure 5: Qualitative results. DuetHOI preserves bimanual coordination during articulation, including for unseen ovens.

Method	Phy↑	C-F1↑	MPJPE$_A$ ↓	OTE↓	ORE↓
ARCTIC, Single-Art.					
Text2HOI	0.664	0.518	12.535	0.504	11.587
LatentHOI	0.757	0.569	13.249	0.488	13.550
MDM	0.702	0.571	11.647	0.425	11.663
DuetHOI (Ours)	**0.992**	**0.660**	**9.573**	**0.361**	**11.411**
ARCTIC, Single-Rig.					
Text2HOI	0.750	0.471	14.403	0.497	10.834
LatentHOI	0.689	0.559	15.901	0.602	12.106
MDM	0.700	0.515	14.211	0.520	11.461
DuetHOI (Ours)	**0.947**	**0.633**	**11.875**	**0.481**	**10.454**
GRAB, Single-Rig.					
Text2HOI	0.646	0.580	14.594	1.178	31.749
LatentHOI	0.678	0.608	15.253	1.393	58.235
MDM	0.699	0.599	14.699	1.078	**31.560**
DuetHOI (Ours)	**0.981**	**0.783**	**14.286**	**1.043**	46.262

(a) Single-hand results of the jointly trained model.

Variant	Phy↑	C-F1↑	MPJPE$_A$ ↓	OTE↓	ORE↓
Bim.-Art.					
w/o ManiVAE	0.951	0.775	12.458	0.341	12.191
w/o ContactVAE	0.836	0.679	13.690	0.346	12.212
w/o ArtPlanner	0.962	0.779	13.272	0.342	12.202
w/ slot context	0.952	0.776	12.628	0.333	**11.872**
Flattened atten.	0.968	0.784	12.429	0.335	12.995
Full	**0.969**	**0.786**	**12.243**	**0.332**	12.186
Bim.-Rig.					
w/o ManiVAE	0.833	0.700	10.512	0.367	7.373
w/o ContactVAE	0.765	0.683	9.700	0.376	7.447
w/ slot context	0.828	0.670	11.060	0.360	8.369
Flattened atten.	0.839	**0.713**	11.745	0.375	7.356
Full	**0.842**	**0.713**	**9.636**	**0.357**	**7.301**

(b) DualFormer ablation on ARCTIC before ProxiRefine; rigid objects omit ArtPlanner.

Table 2: Single-hand generalization and DualFormer ablations. Full single-hand metrics are provided in the supplement.

4.2 Main Results on Bimanual Interaction

On ARCTIC Bim.-Art. in Table 1, DuetHOI leads eight of ten metrics, with **0.997** Phy, **0.834** C-F1, and **0.095** NN-Rel. Against the strongest baselines, C-F1 improves by **0.067**, while MPJPE$_A$ and NN-Rel decrease by **13.4%** and **41.4%**. C-Rate and C-Stab confirm persistent contact. These gains support separating articulation progress, object-centered bimanual placement, and local hand pose before object-fixed contact refinement. Rigid-interaction gains persist: on ARCTIC Bim.-Rig., DuetHOI reaches **0.746** C-F1 and reduces MPJPE$_A$ from 12.879 to **9.614** by **25.4%**; on GRAB, it reaches **0.775** C-F1 and lowers NN-Rel from 0.156 to **0.088** by **43.6%**. Baselines retain advantages in rotation accuracy or diversity. Figure 5 shows coherent bimanual manipulation as articulation evolves, including the unseen oven category.

4.3 Performance on Single-Hand Interaction

In Table 2(a), with one hand stream masked, DuetHOI achieves **0.992** Phy, **0.660** C-F1, and **9.573** MPJPE_{A} on ARCTIC Single-Art., outperforming all baselines. It leads all five metrics on ARCTIC Single-Rig., including **0.947** Phy and **0.633** C-F1. On GRAB, it raises C-F1 from 0.608 to **0.783**, although MDM obtains a lower ORE. This robustness follows from the semantic-stream design: masking the absent hand preserves object and active-hand streams, while ManiVAE provides a common compact pose prior for either hand.

4.4 Articulation-Aware Interaction Generation

Before ProxiRefine, full DualFormer reaches **0.969** Phy and **0.786** C-F1 on Bim.-Art. in Table 2(b). Without ContactVAE, Phy/C-F1 fall to 0.836/0.679 on Bim.-Art. and 0.765/0.683 on Bim.-Rig., confirming its surface prior anchors placement across articulation types. Without ArtPlanner, articulated MPJPE_{A} increases from **12.243** to 13.272 because its trajectory reference aligns hand motion with evolving geometry. Restricting output heads to their own slots lowers C-F1 to 0.776/0.670 and raises MPJPE_{A} to 12.628/11.060, exposing weaker cross-stream coordination. The larger rigid MPJPE increase without ManiVAE highlights compact pose tokens, especially without an articulation reference. Flattened attention raises rigid MPJPE_{A} to 11.745, supporting factorization. Overall, results support cross-stream context, compact pose tokens, and factorized attention.

Variant	C-F1↑	MPJPE_{A} ↓	C-Rate↑	C-Stab↑	NN-Rel↓
Bim.-Art.					
w/o $\mathcal{L}_{\text{contact}}$	0.797	12.119	0.829	0.962	0.113
w/o C_i^h	0.827	12.065	0.859	0.968	0.103
Full	**0.834**	**12.043**	**0.873**	**0.970**	**0.095**
Bim.-Rig.					
w/o $\mathcal{L}_{\text{contact}}$	0.689	9.622	0.632	0.950	**0.112**
w/o C_i^h	0.730	9.652	0.688	0.954	0.119
Full	**0.746**	**9.614**	**0.705**	**0.956**	0.118

Table 3: Ablation of the contact loss and joint-level proximity cue in object-fixed ProxiRefine on ARCTIC.

4.5 Object-Fixed Proximity Refinement

Table 3 shows that ProxiRefine raises C-F1 from 0.786/0.713 to **0.834/0.746** on Bim.-Art./Bim.-Rig., while MPJPE_{A} changes only from 12.243/9.636 to **12.043/9.614**. Thus, refinement improves contact without rewriting global motion. Moreover, C-Stab reaches **0.970/0.956**, while the fixed object trajectory attributes these gains to hand-geometry correction rather than altered manipulation progress. Removing $\mathcal{L}_{\text{contact}}$ lowers C-F1 to 0.797/0.689 and C-Rate to 0.829/0.632, making direct contact supervision the primary refinement signal.

Removing C_i^h lowers C-F1 to 0.827/0.730 and C-Rate to 0.859/0.688; this supports the cue's localization role. Its slight rigid NN-Rel advantage (0.112 vs. **0.118**) comes at substantially lower C-F1 and C-Rate.

5 Conclusion

We presented DuetHOI for language-conditioned bimanual articulated-object interaction generation. It decomposes motion into object state, object-centered hand placement, and compact local hand articulation. ContactVAE and ArtPlanner provide spatial contact intent and a trajectory-level articulation reference, while ManiVAE and DualFormer generate coordinated global interaction. ProxiRefine corrects local hand geometry while fixing the object trajectory. Experiments on ARCTIC and GRAB show consistent gains in physical plausibility, contact quality, hand accuracy, and relative hand–object consistency, especially for bimanual articulated interactions. These results support structured generation with object-fixed local refinement.

Limitations. DuetHOI represents articulation with one scalar and generates one instruction-conditioned interaction at a time. Multi-joint objects require vector-valued planning and constraints among movable parts, whereas multi-task sequences require consistent contacts, hand roles, and object states across stages. The current model captures neither coupled joints nor inter-task transitions. Future work will study hierarchical multi-joint planning, coupled-part constraints, and compositional manipulation over longer task horizons.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and LaTeX formatting. All AI-assisted changes were reviewed and revised by the authors, who take full responsibility for the final content, claims, and presentation of this work.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20577–20586, 2022.

Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5031–5041, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12943–12954, 2023.

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1471–1481, 2021.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7136–7146, 2025.

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*, pp. 11107–11116, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. In *European Conference on Computer Vision*, pp. 109–127. Springer, 2024.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6991–7003, 2025a.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17416–17425, 2025b.

Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, et al. Object-centric prompt-driven vision-language-action model for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27638–27648, 2025c.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20609–20620, 2023.

Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7(1):470–477, 2021.

Katja Ludwig, Yuliia Oksymets, Robin Schön, Daniel Kienzle, and Rainer Lienhart. Efficient 2d to full 3d human pose uplifting including joint rotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5852–5861, 2025.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7773–7781, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10985–10995, 2021.

Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European conference on computer vision*, pp. 480–497. Springer, 2022.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499, 2021.

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

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020.

Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13263–13273, 2022.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pp. 358–374. Springer, 2022a.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022b.

Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, pp. 201–221. Springer, 2022.

Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, pp. 257–274. Springer, 2022.

Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4737–4746, 2023.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20953–20962, 2022.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Junming Zhang, Jiefeng Li, and Cewu Lu. Learning a contact potential field for modeling the hand-object interaction. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5645–5662, 2024.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22479–22489, 2023.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024a.

Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pp. 235–246. IEEE, 2024b.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

Jiaxu Zhang, Xianfang Zeng, Xin Chen, Wei Zuo, Gang Yu, and Zhigang Tu. Mikudance: Animating character art with mixed motion dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19689–19699, 2025.

Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 585–594, 2023.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pp. 18–38. Springer, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.

A Overview

This supplementary material first provides notation, implementation and evaluation details, baseline adaptations, full single-hand results, and metric definitions. It then directly analyzes ContactVAE contact intent, ArtPlanner articulation references, and ManiVAE hand-pose tokens in terms of prediction accuracy, sample variation, and reconstruction quality, before presenting qualitative results across all four interaction modes and a representative failure case. Additional media resources are available for download from the **Media Supplement**.

B Notation

Following the definitions in the main paper, given object geometry $\mathcal{O}$ and a language instruction $\mathcal{T}$, the model generates an object sequence $\hat{O}_{1:N}$ and a bimanual motion sequence $\hat{H}_{1:N}$ of length N. The hand state at frame i is

$$H_i = \{T_i, P_i\}, \qquad T_i = \{T_i^l, T_i^r\}, \quad P_i = \{P_i^l, P_i^r\},$$

where T_i^h and P_i^h denote the global translation and MANO pose Romero et al. (2022), respectively, of hand $h \in \{l, r\}$. The object state is

$$O_i = \{O_i^\alpha, O_i^\beta, O_i^\gamma\},$$

where $O_i^\alpha \in \mathbb{R}^3$ is the object translation, $O_i^\beta \in \mathbb{R}^6$ is its rot6d rotation representation Zhou et al. (2019), and $O_i^\gamma \in \mathbb{R}$ is its scalar articulation state. Rigid objects satisfy $O_{1:N}^\gamma = \mathbf{0}$. The object-centered hand position used in the main paper is

$$R_i^h = T_i^h - O_i^\alpha,$$

and the global hand translation is recovered by $T_i^h = O_i^\alpha + R_i^h$. We use $\mathcal{O}_i$ to denote the object geometry transformed by O_i^β and O_i^γ but without the translation O_i^α. A hat above a symbol denotes a model prediction.

The binary supervision target of ContactVAE is $\mathcal{C} \in \{0,1\}^M$, and its predicted probabilities are $\hat{\mathcal{C}} \in [0,1]^M$, where M is the number of sampled object-surface points. The target is defined over the entire sequence: an object point is labeled positive if it is contacted by either valid hand in any frame. Each predicted probability is concatenated with the corresponding object-point coordinate before object encoding. The same predicted map $\hat{\mathcal{C}}$ conditions both ArtPlanner and DualFormer. We denote the articulation reference generated by ArtPlanner by $\bar{O}_{1:N}^\gamma$ and the left- and right-hand pose latents of ManiVAE by z_i^l and z_i^r, respectively. These symbols are consistent with those in the main paper.

We consider four interaction settings: single-hand articulated-object interaction (Single-Art.), single-hand rigid-object interaction (Single-Rig.), bimanual articulated-object interaction (Bim.-Art.), and bimanual rigid-object interaction (Bim.-Rig.). Bim.-Art. is the primary setting in the following analysis.

C Implementation Details

Computing infrastructure. All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU with 48 GB of GPU memory and an AMD Ryzen Threadripper PRO 3995WX CPU, using Ubuntu 24.04, Python 3.10, PyTorch 2.10, and CUDA 12.8.

Table 4 summarizes the architecture of each component. The VAE components follow variational auto-encoding Kingma & Welling (2013), and the attention-based components build on the Transformer architecture **?**. All components are optimized with AdamW **?**. The same component designs are used for single- and bimanual generation, with validity masks applied to absent hand streams.

Hyperparameter	ContactVAE	ArtPlanner	ManiVAE	DualFormer	ProxiRefine
Model Design					
Core architecture	MLP CVAE	Residual MLP CVAE	Residual MLP VAE	Factorized Transformer	Transformer
Geometry encoder	Point-wise MLP	PointNet++	PointNet++	PointNet++	–
Core depth	2/3-layer MLP	2 ResBlocks	4 ResBlocks	8 blocks	4 blocks
Attention heads	–	–	–	4	8
Hidden dimension	256	512	768	512	256
VAE latent dimension	64	64	32	–	–
Training and Inference					
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	2×10^{-4}	1×10^{-4}	1×10^{-4}	5×10^{-5}	1×10^{-4}
Weight decay	1×10^{-4}	1×10^{-2}	0	1×10^{-2}	1×10^{-4}
LR schedule	Constant	Warmup + cosine	Constant	Warmup + cosine	Constant
DDPM / DDIM steps	–	–	–	1,000 / 100	–

Table 4: Model design and optimization settings for the DuetHOI components.

Shared encoders and preprocessing. We use a frozen CLIP ViT-B/32 encoder Radford et al. (2021) to obtain a 512-D text feature and sample 1,024 points from each object surface. ContactVAE preserves point correspondence through a point-wise MLP. ArtPlanner, ManiVAE, and DualFormer use PointNet++ Qi et al. (2017) to produce 1,024-D geometry features.

Staged optimization and inference. ContactVAE, ArtPlanner, and ManiVAE are optimized separately. CLIP remains frozen, and ManiVAE is also frozen when DualFormer is optimized; its posterior mean serves as the target pose token. DualFormer follows a clean-target DDPM formulation Ho et al. (2020) with a cosine noise schedule **?** and deterministic DDIM inference **?**. ProxiRefine is subsequently optimized using DualFormer outputs, preserves the generated object stream, and predicts only residual hand corrections.

Canonical hand space and proximity cues. ManiVAE maps left-hand samples to a canonical right-hand space by reflecting object points and relative translations across the YZ plane and transforming rotations as $R' = SRS$, where $S = \mathrm{diag}(-1, 1, 1)$. Its decoded left-hand poses are reflected back at inference. For ProxiRefine, each of the 21 MANO joints contributes a nearest-surface offset, its distance, and an exponential proximity weight, yielding a 105-D cue per hand; joints whose nearest surface point lies within 2 cm are treated as local proximity evidence.

D Datasets and Evaluation Protocol

ARCTIC. ARCTIC Fan et al. (2023) was originally collected for reconstructing hands and articulated objects from RGB images. We use 4,597 processed motions annotated with 11 action labels. Each motion provides MANO hand parameters, an object mesh, object rotation and translation, and an articulation angle with a predefined axis. Since the original motion data do not contain free-form language instructions, we use the annotations prepared by Text2HOI Cha et al. (2024): 644 manually written sentences describing the action, object category, and interacting hand or hands.

GRAB. GRAB Taheri et al. (2020) contains 1,335 motions in which two captured hands interact with one of 51 everyday rigid objects of varying shapes and sizes. It covers 29 action labels and provides MANO hand parameters, object meshes, and object rotation and translation. We use the language annotations prepared by Text2HOI Cha et al. (2024), which automatically combine the action, object, and active-hand labels into 1,104 distinct sentences.

Interaction-mode partition. We partition the evaluation set along two binary attributes: object type and active-hand count. An interaction is articulated if its object articulation changes over the valid sequence and rigid otherwise; it is bimanual if both hands are active and single-hand if exactly one hand is active. This yields Single-Art., Single-Rig., Bim.-Art., and Bim.-Rig. ARCTIC contains all four modes, while GRAB contributes the two rigid modes. Every metric is computed separately on the sequences assigned to a mode and uses only their valid object and active-hand frames, with aggregation following the definitions below. Left- and right-hand single-hand sequences share the same mode score.

Dataset	Setting	Method	Phy↑	C-F1↑	MPJPE$_A$ ↓	OTE↓	ORE↓	SD↑	OD↑	C-Rate↑	C-Stab↑	NN-Rel↓
ARCTIC	Single-Art.	Text2HOI	0.664	0.518	12.535	0.504	11.587	3.610	11.382	0.653	0.967	0.079
		LatentHOI	0.757	0.569	13.249	0.488	13.550	**6.190**	**14.180**	0.733	0.965	0.101
		MDM	0.702	0.571	11.647	0.425	11.663	5.804	12.939	0.713	0.969	0.069
		DuetHOI (Ours)	**0.992**	**0.660**	**9.573**	**0.361**	**11.411**	3.596	13.397	**0.836**	**0.974**	**0.043**
	Single-Rig.	Text2HOI	0.750	0.471	14.403	0.497	10.834	3.409	8.644	0.665	**0.970**	**0.074**
		LatentHOI	0.689	0.559	15.901	0.602	12.106	**7.320**	13.871	0.617	0.961	0.122
		MDM	0.700	0.515	14.211	0.520	11.461	6.619	12.812	0.651	0.967	0.107
		DuetHOI (Ours)	**0.947**	**0.633**	**11.875**	**0.481**	**10.454**	4.885	**14.628**	**0.697**	0.962	0.102
GRAB	Single-Rig.	Text2HOI	0.646	0.580	14.594	1.178	31.749	4.190	6.596	0.700	0.959	0.273
		LatentHOI	0.678	0.608	15.253	1.393	58.235	2.176	4.299	0.713	0.936	0.735
		MDM	0.699	0.599	14.699	1.078	**31.560**	**12.343**	**15.263**	0.778	0.986	0.582
		DuetHOI (Ours)	**0.981**	**0.783**	**14.286**	**1.043**	46.262	5.099	6.934	**0.984**	**0.997**	**0.011**

Table 5: Complete single-hand results from the jointly trained model. Arrows indicate the preferred metric direction.

First-frame alignment. MPJPE$_A$, OTE, and ORE compare motion after expressing each predicted or ground-truth sequence in its own first-frame object coordinate system; the first frame itself is excluded. This removes discrepancies in the initial global object reference frame and focuses the scores on subsequent hand–object and object-motion evolution. Such alignment is important for a one-to-many generation task, in which multiple initial global placements can be plausible for the same instruction, and prevents this ambiguity from dominating comparisons between methods. It does not remove errors in the hand placement relative to the object.

E Baseline Adaptation

Text2HOI Cha et al. (2024), LatentHOI Li et al. (2025b), and MDM Tevet et al. (2022b) were originally designed with different task definitions and motion representations. We preserve the core generation mechanism of each method while extending its input and output interfaces to the unified language-conditioned hand–object interaction task. All three baselines predict the object state and the available left- and right-hand motion streams, and use the same validity masks as DuetHOI for an absent hand.

Text2HOI retains its direct diffusion formulation for hand–object motion, with the output interface extended to the object state and the left- and right-hand states. LatentHOI retains latent-space diffusion and uses a U-Net denoiser to generate the complete sequence in the unified hand and object representation. MDM retains its Transformer-based text-conditioned motion diffusion architecture, with each frame represented by the concatenation of the left-hand state, right-hand state, and object state.

All methods use the same training, validation, and test splits, sequence length, valid-frame masks, and object point clouds. All metrics are computed using the same evaluation code.

F Full Single-Hand Results

Table 5 reports the complete single-hand evaluation corresponding to the selected metrics in the main paper.

G Evaluation Metrics

Let $\mathcal{S}$ denote the set of sequences for the current dataset and interaction mode, with $s \in \mathcal{S}$ indexing a sequence. We use $\mathcal{I}_s$ and $\mathcal{I}_s^h$ to denote the valid object frames and the valid frames of hand $h \in \{l, r\}$ in sequence s, respectively. The operators $\mathcal{V}(\cdot)$ and $\mathcal{J}(\cdot)$ denote mesh vertices and MANO joints, respectively; thus, $\mathcal{V}(H^h_{s,i})$, $\mathcal{J}(H^h_{s,i})$, and $\mathcal{V}(O_{s,i})$ correspond to the hand vertices, hand joints, and object surface points at frame i. All contact metrics below use a distance threshold of $\tau_c = 5$ mm, and $\mathbb{1}[\cdot]$ denotes the indicator function.

G.1 Ground-Truth-Aligned Motion and Contact Metrics

Physical validity (Phy, ↑). For frame i of sequence s, we define the object-lift indicator as

$$\hat{g}_{s,i} = \mathbb{1}\left[\min_{v \in \mathcal{V}(\hat{O}_{s,i})} [v]_z > \varepsilon_z\right], \qquad \varepsilon_z = 10^{-4}\ \mathrm{m},$$

and the ratio of hand vertices in contact as

$$\hat{r}^h_{s,i} = \frac{1}{|\mathcal{V}(\hat{H}^h_{s,i})|} \sum_{v \in \mathcal{V}(\hat{H}^h_{s,i})} \mathbb{1}\left[\min_{o \in \mathcal{V}(\hat{O}_{s,i})} \|v - o\|_2 < \tau_c\right].$$

A lifted frame maintains valid contact when at least one valid hand satisfies $\hat{r}^h_{s,i} \geq \rho_c$, where $\rho_c = 0.01$. Phy is first computed over the lifted frames of each sequence and then averaged across sequences:

$$\mathrm{Phy} = \frac{1}{|\mathcal{S}^+|} \sum_{s \in \mathcal{S}^+} \frac{\sum_{i \in \mathcal{I}_s} \hat{g}_{s,i} \mathbb{1}[\max_h \hat{r}^h_{s,i} \geq \rho_c]}{\sum_{i \in \mathcal{I}_s} \hat{g}_{s,i}},$$

where $\mathcal{S}^+$ contains generated sequences with at least one lifted frame.

Contact F1 (C-F1, ↑). Let $\mathcal{H}_s \subseteq \{l, r\}$ be the set of active hands in sequence s, and let $\mathcal{I}^C_s = \mathcal{I}_s \cap \bigcap_{h \in \mathcal{H}_s} \mathcal{I}^h_s$ be the contact-evaluation frames. The ground-truth and predicted frame-level contact states are

$$c_{s,i} = \bigvee_{h \in \mathcal{H}_s} \mathbb{1}\left[\min_{j,m} \left\| \mathcal{J}_j(H^h_{s,i}) - \mathcal{V}_m(O_{s,i}) \right\|_2 < \tau_c\right],$$
$$\hat{c}_{s,i} = \bigvee_{h \in \mathcal{H}_s} \mathbb{1}\left[\min_{j,m} \left\| \mathcal{J}_j(\hat{H}^h_{s,i}) - \mathcal{V}_m(\hat{O}_{s,i}) \right\|_2 < \tau_c\right].$$

Let TP, FP, and FN be the micro-averaged counts obtained from $\hat{c}_{s,i}$ and $c_{s,i}$ over all $i \in \mathcal{I}^C_s$. Then

$$\text{C-F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

Anchor-aligned mean per-joint position error ($\mathrm{MPJPE_A}$, ↓). Let $a_s = \min \mathcal{I}_s$ be the first valid object frame of sequence s, and let $Q_{s,i} = \mathrm{Rot}(O^\beta_{s,i})$ and $\hat{Q}_{s,i} = \mathrm{Rot}(\hat{O}^\beta_{s,i})$ denote the ground-truth and predicted object rotation matrices at frame i, respectively. The predicted and ground-truth joints are transformed into their respective first-frame object coordinate systems:

$$J^{h,c}_{s,i,j} = \left(\mathcal{J}_j(H^h_{s,i}) - O^\alpha_{s,a_s}\right) Q_{s,a_s},$$
$$\hat{J}^{h,c}_{s,i,j} = \left(\mathcal{J}_j(\hat{H}^h_{s,i}) - \hat{O}^\alpha_{s,a_s}\right) \hat{Q}_{s,a_s}.$$

Let Ω_H contain all valid hand-joint indices (s, h, i, j) with $i > a_s$. Then

$$\mathrm{MPJPE_A} = \frac{100}{|\Omega_H|} \sum_{(s,h,i,j) \in \Omega_H} \left\| \hat{J}^{h,c}_{s,i,j} - J^{h,c}_{s,i,j} \right\|_2.$$

The factor 100 converts meters to centimeters. No separate rigid alignment is applied to the hand.

Object translation and rotation errors (OTE/ORE, ↓). Object motion is likewise measured relative to the first valid object frame. We define

$$
\begin{aligned}
t^c_{s,i} &= (O^\alpha_{s,i} - O^\alpha_{s,a_s}) Q_{s,a_s}, \\
\hat{t}^c_{s,i} &= (\hat{O}^\alpha_{s,i} - \hat{O}^\alpha_{s,a_s}) \hat{Q}_{s,a_s}, \\
Q^c_{s,i} &= Q^\top_{s,a_s} Q_{s,i}, \\
\hat{Q}^c_{s,i} &= \hat{Q}^\top_{s,a_s} \hat{Q}_{s,i}.
\end{aligned}
$$

Let $\Omega_O = \{(s,i) : i \in \mathcal{I}_s,\ i > a_s\}$. OTE and ORE are

$$
\begin{aligned}
\mathrm{OTE} &= \frac{10}{|\Omega_O|} \sum_{(s,i)\in\Omega_O} \left\| \hat{t}^c_{s,i} - t^c_{s,i} \right\|_2, \\
\mathrm{ORE} &= \frac{1}{|\Omega_O|} \sum_{(s,i)\in\Omega_O} \frac{180}{\pi} \\
&\qquad \arccos\left(\mathrm{clip}\left(\frac{\mathrm{tr}((\hat{Q}^c_{s,i})^\top Q^c_{s,i}) - 1}{2}, -1, 1 \right) \right).
\end{aligned}
$$

OTE follows the display scale used in the main tables by multiplying the error in meters by 10; ORE is reported in degrees.

G.2 Diversity Metrics

Following Text2HOI Cha et al. (2024), we generate $K = 5$ samples for each object–instruction condition and flatten all valid hand vertices over time into $f^{(k)}_s \in \mathbb{R}^{D_s}$. With $d(f,g) = \|f - g\|_2 / \sqrt{D_s}$, sample diversity is

$$
\mathrm{SD} = \frac{100}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \frac{2}{K(K-1)} \sum_{1\le k<q\le K} d\Big(f^{(k)}_s, f^{(q)}_s \Big).
$$

For overall diversity, let $\mathcal{F}$ contain all samples in the current interaction mode and let $L = |\mathcal{F}|$. Then

$$
\mathrm{OD} = 100 \cdot \frac{2}{L(L-1)} \sum_{1\le a<b\le L} d(\mathcal{F}_a, \mathcal{F}_b).
$$

Both metrics are computed separately for single- and bimanual settings.

G.3 Relative Interaction Metrics

For each valid hand–frame pair (h,i), we define the ground-truth and predicted minimum hand–object distances and contact states as

$$
\begin{aligned}
\delta^h_{s,i} &= \min_{v\in\mathcal{V}(H^h_{s,i}),\, o\in\mathcal{V}(O_{s,i})} \|v - o\|_2, \quad b^h_{s,i} = \mathbb{1}[\delta^h_{s,i} < \tau_c], \\
\hat{\delta}^h_{s,i} &= \min_{v\in\mathcal{V}(\hat{H}^h_{s,i}),\, o\in\mathcal{V}(\hat{O}_{s,i})} \|v - o\|_2, \quad \hat{b}^h_{s,i} = \mathbb{1}[\hat{\delta}^h_{s,i} < \tau_c].
\end{aligned}
$$

Let $\mathcal{P}_s$ contain all valid hand–frame pairs, $\mathcal{E}_s \subseteq \mathcal{P}_s$ contain pairs with a valid preceding frame, and $\mathcal{I}^{lr}_s$ contain frames in which both hands are valid. We further define the object-relative hand centroid as $q^h_{s,i} = \mathrm{cent}(\mathcal{V}(H^h_{s,i})) - \mathrm{cent}(\mathcal{V}(O_{s,i}))$. The five ground-truth statistics are

$$
\begin{aligned}
d_s^{\mathrm{HO}} &= \frac{100}{|\mathcal{P}_s|} \sum_{(h,i)\in\mathcal{P}_s} \delta_{s,i}^h, \\
\text{C-Rate}_s &= \frac{1}{|\mathcal{P}_s|} \sum_{(h,i)\in\mathcal{P}_s} b_{s,i}^h, \\
\text{C-Stab}_s &= 1 - \frac{1}{L_s - 1} \sum_{n=1}^{L_s-1} \mathbb{1}[b_{s,n+1} \neq b_{s,n}], \\
u_s^{\mathrm{rel}} &= \frac{100}{|\mathcal{E}_s|} \sum_{(h,i)\in\mathcal{E}_s} \|q_{s,i}^h - q_{s,i-1}^h\|_2, \\
d_s^{\mathrm{LR}} &= \frac{100}{|\mathcal{I}_s^{lr}|} \sum_{i\in\mathcal{I}_s^{lr}} \left\| \mathrm{cent}(\mathcal{V}(H_{s,i}^l)) - \mathrm{cent}(\mathcal{V}(H_{s,i}^r)) \right\|_2 .
\end{aligned}
$$

Here, $(b_{s,1}, \ldots, b_{s,L_s})$ lists the valid contact states in temporal order; C-Stab is set to 1 when $L_s = 1$. Each sequence is summarized by

$$
\begin{aligned}
\phi_s &= \left[d_s^{\mathrm{HO}},\ \text{C-Rate}_s,\ \text{C-Stab}_s,\ u_s^{\mathrm{rel}},\ d_s^{\mathrm{LR}} \right], \\
\hat{\phi}_s &= \left[\hat{d}_s^{\mathrm{HO}},\ \widehat{\text{C-Rate}}_s,\ \widehat{\text{C-Stab}}_s,\ \hat{u}_s^{\mathrm{rel}},\ \hat{d}_s^{\mathrm{LR}} \right].
\end{aligned}
$$

Hatted quantities are computed identically from the predictions. Distances are measured in centimeters and u_s^{rel} in centimeters per frame; d_s^{LR} and $\hat{d}_s^{\mathrm{LR}}$ are omitted for single-hand sequences. The tables report the predicted C-Rate and C-Stab, with hats omitted from the headers.

For feature dimension j, we normalize using the 5th and 95th percentiles, r_j^- and r_j^+, computed from the real sequences in the same evaluation split:

$$
\tilde{\phi}_{s,j} = \frac{\phi_{s,j} - r_j^-}{\max(r_j^+ - r_j^-, 1)}, \quad \widetilde{\hat{\phi}}_{s,j} = \frac{\hat{\phi}_{s,j} - r_j^-}{\max(r_j^+ - r_j^-, 1)}.
$$

Nearest-neighbor relative interaction distance (NN-Rel, ↓). Let $\widetilde{\Phi}$ and $\widetilde{\hat{\Phi}}$ be the normalized real and generated feature sets, respectively, for the current interaction mode. We define

$$
\text{NN-Rel} = \frac{1}{|\widetilde{\hat{\Phi}}|} \sum_{\widetilde{\hat{\phi}}\in\widetilde{\hat{\Phi}}} \min_{\tilde{\phi}\in\widetilde{\Phi}} \left\| \widetilde{\hat{\phi}} - \tilde{\phi} \right\|_2 .
$$

A lower NN-Rel indicates that the contact, distance, and relative-velocity statistics of the generated motion are closer to those of real interactions.

Unless otherwise specified, metrics are accumulated over all valid frames and then aggregated either per sequence or through global counts. Invalid frames of either hand are excluded from hand metrics.

G.4 Module-Level Metrics

ContactVAE. Let $\mathcal{U}$ be the set of sequence–object-point pairs. For $u \in \mathcal{U}$, let $y_u \in \{0, 1\}$ and $\hat{p}_u \in [0, 1]$ denote the label and predicted probability, and let $\hat{y}_u = \mathbb{1}[\hat{p}_u \geq \tau_C]$. The threshold τ_C is selected on the validation set by micro-F1. Let TP, FP, and FN be the point-level counts on the test set; after sorting points by $\hat{p}_u$ in descending order, let $y_{(r)}$ be the label at rank r and $N_{\mathrm{pos}} = \sum_u y_u$. The accuracy metrics are

$$
\begin{aligned}
\mathrm{AP} &= \frac{1}{N_{\mathrm{pos}}} \sum_{r=1}^{|\mathcal{U}|} y_{(r)} \frac{\sum_{q=1}^{r} y_{(q)}}{r}, \\
\mathrm{F1}_{\mathrm{pt}} &= \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \\
\mathrm{IoU}_{\mathrm{pt}} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \\
\mathrm{BCE} &= -\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left[y_u \log \hat{p}_u + (1 - y_u) \log(1 - \hat{p}_u) \right].
\end{aligned}
$$

For each condition s, sample variation is the mean pairwise Jaccard distance d_J among K thresholded contact-intent maps:

$$
\begin{aligned}
\mathrm{Div}_C = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} & \frac{2}{K(K-1)} \\
& \cdot \sum_{1 \le k < q \le K} d_J\left(\hat{\mathcal{B}}_s^{(k)}, \hat{\mathcal{B}}_s^{(q)} \right),
\end{aligned}
$$

where $d_J(A, B) = 1 - |A \cap B| / |A \cup B|$, with an empty union assigned zero distance.

ArtPlanner. Let $\mathcal{S}_A$ be the articulated-object sequences and $(i_1, \ldots, i_{n_s})$ their ordered valid frames. Define $\Delta \bar{O}^{\gamma}_{s,i_r} = \bar{O}^{\gamma}_{s,i_r} - \bar{O}^{\gamma}_{s,i_{r-1}}$ and define $\Delta O^{\gamma}_{s,i_r}$ analogously. The four metrics are

$$
\begin{aligned}
\text{Art-MAE} &= \frac{1}{|\mathcal{S}_A|} \sum_{s \in \mathcal{S}_A} \frac{1}{n_s} \sum_{r=1}^{n_s} \left| \bar{O}^{\gamma}_{s,i_r} - O^{\gamma}_{s,i_r} \right|, \\
\text{Art-FDE} &= \frac{1}{|\mathcal{S}_A|} \sum_{s \in \mathcal{S}_A} \left| \bar{O}^{\gamma}_{s,i_{n_s}} - O^{\gamma}_{s,i_{n_s}} \right|, \\
\text{Art-VE} &= \frac{1}{|\mathcal{S}_A|} \sum_{s \in \mathcal{S}_A} \frac{1}{n_s - 1} \sum_{r=2}^{n_s} \left| \Delta \bar{O}^{\gamma}_{s,i_r} - \Delta O^{\gamma}_{s,i_r} \right|, \\
\text{OutRate} &= \frac{1}{|\mathcal{S}_A|} \sum_{s \in \mathcal{S}_A} \frac{1}{n_s} \sum_{r=1}^{n_s} \mathbb{1}\left[\bar{O}^{\gamma}_{s,i_r} \notin [L_{c_s}, U_{c_s}] \right].
\end{aligned}
$$

Angular errors are converted to degrees, and Art-VE is reported in degrees per frame.

ManiVAE. Let Ω_M contain all valid hand frames, and let J_H, V_H, and $J_R = 16$ be the numbers of hand joints, vertices, and local MANO rotations. The reconstruction metrics are

$$
\begin{aligned}
\mathrm{MPJPE}_M &= \frac{1000}{|\Omega_M| J_H} \sum_{(s,h,i) \in \Omega_M} \sum_{j=1}^{J_H} \left\| \mathcal{J}_j(\hat{H}^h_{s,i}) - \mathcal{J}_j(H^h_{s,i}) \right\|_2, \\
\mathrm{MPVPE}_M &= \frac{1000}{|\Omega_M| V_H} \sum_{(s,h,i) \in \Omega_M} \sum_{v=1}^{V_H} \left\| \mathcal{V}_v(\hat{H}^h_{s,i}) - \mathcal{V}_v(H^h_{s,i}) \right\|_2, \\
\mathrm{GeoErr}_M &= \frac{1}{|\Omega_M| J_R} \sum_{(s,h,i) \in \Omega_M} \sum_{j=1}^{J_R} \frac{180}{\pi} \\
&\qquad \arccos\left(\mathrm{clip}\left(\frac{\mathrm{tr}((\hat{R}^h_{s,i,j})^{\top} R^h_{s,i,j}) - 1}{2}, -1, 1 \right) \right).
\end{aligned}
$$

MPJPE_M and MPVPE_M are reported in millimeters, and GeoErr_M in degrees.

Setting	AP↑	F1↑	IoU↑	BCE↓	Div.↑
Prior mean	**0.562**	**0.545**	**0.374**	**0.432**	–
Prior sample	0.495	0.493	0.327	0.519	0.633

Table 6: Contact-intent accuracy and sample variation of ContactVAE on the ARCTIC test set.

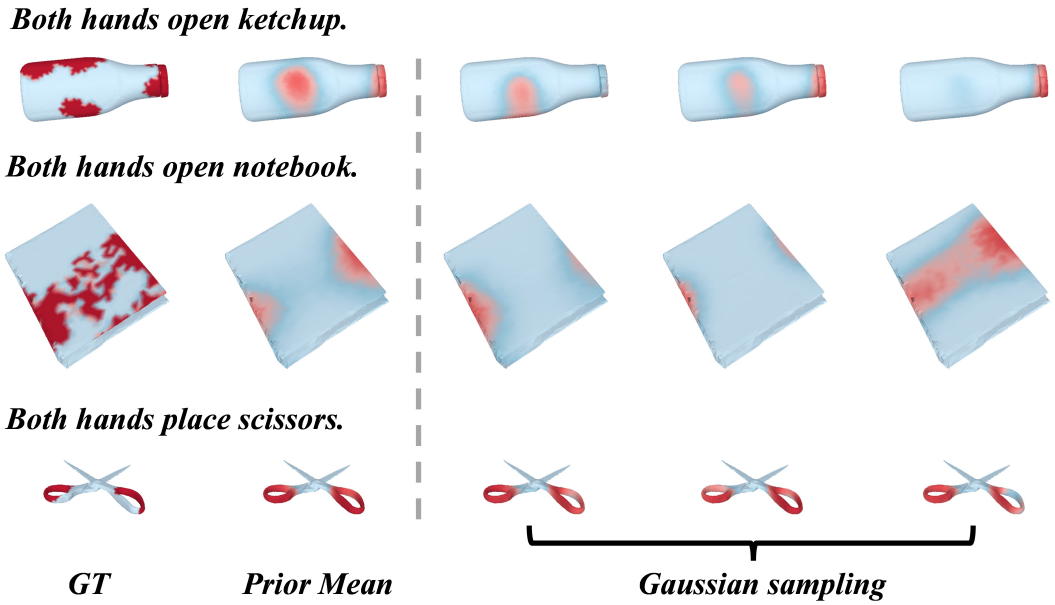


Figure 6: Qualitative ContactVAE predictions. Each row shows one test condition. From left to right, the columns show the ground-truth sequence-level contact intent, the standard-Gaussian prior-mean prediction, and three predictions obtained from independently sampled standard-normal latent codes. Red denotes the predicted or ground-truth object-side contact intent.

H Module-Level Evaluation

The main-paper ablations measure the contribution of each module to final interaction generation. The following experiments evaluate the corresponding intermediate representations directly. ContactVAE is evaluated by contact-intent accuracy and sample variation, ArtPlanner by articulation-trajectory accuracy under predicted contact intent, and ManiVAE by hand-pose reconstruction accuracy. These evaluations respectively examine contact intent, manipulation progress, and local hand configuration.

H.1 ContactVAE: Contact-Intent Accuracy and Sample Variation

ContactVAE infers sequence-level object-side contact intent from language and object geometry. We evaluate agreement with the annotated contact-intent target using point-wise AP, F1, IoU, and BCE, and quantify variation among random samples using pairwise $1 - \mathrm{IoU}$. The binarization threshold is selected exclusively on the validation set using micro-F1 and then fixed to $\tau_C = 0.52$ for testing. The test set contains 920 conditions, and five random samples are generated for each condition. The reported prior-sample accuracy metrics are averaged over the five test-set results.

Prior-mean decoding achieves the highest contact-intent accuracy, with an AP of 0.562 and an F1 score of 0.545. Random sampling reduces AP and F1 to 0.495 and 0.493, respectively, while producing a pairwise $1 - \mathrm{IoU}$ of 0.633. These results indicate a trade-off between point-wise accuracy and spatial diversity: prior-mean decoding offers higher point-wise accuracy against the ground-truth contact intent, whereas random sampling produces contact-intent maps with greater spatial diversity but lower point-wise accuracy. Figure 6 further shows spatially distinct contact intent across independent samples. Together with the main-paper ablation, these results support using predicted contact intent as a variable spatial condition for downstream generation.

ContactVAE latent	MAE↓ (°)	FDE↓ (°)	VE↓ (°/f)	OutRate↓ (%)
ARCTIC, Single-Art.				
Mean latent	6.140	5.490	**1.126**	**9.969**
Sampled latent	**5.886**	**5.326**	**1.126**	10.901
ARCTIC, Bim.-Art.				
Mean latent	12.377	**8.591**	2.540	11.169
Sampled latent	**12.045**	8.597	**2.505**	**10.070**

Table 7: ArtPlanner trajectory accuracy with mean and sampled ContactVAE latents.

Subset	Frames	MPJPE↓ (mm)	MPVPE↓ (mm)	GeoErr↓ (°)
All	37,013	2.053	2.006	2.531
Left	17,541	2.040	1.995	2.434
Right	19,472	2.065	2.015	2.618

Table 8: Posterior-mean reconstruction accuracy of the 32-dimensional ManiVAE representation.

H.2 ArtPlanner: Articulation Planning under Predicted Contact Intent

ArtPlanner converts language, object geometry, the initial articulation state, and the predicted contact-intent probabilities $\hat{\mathcal{C}}$ into a trajectory-level articulation reference for DualFormer. We directly compare this reference with the ground-truth trajectory on the articulated-object subset of the ARCTIC test set. We report Art-MAE, Art-FDE, Art-VE, and OutRate as defined in Section G.4. Evaluation uses the continuous probabilities $\hat{\mathcal{C}}$ predicted by ContactVAE. All articulated objects in ARCTIC use revolute joints, so errors are converted to degrees.

We consider two ContactVAE latent settings. Mean latent sets ContactVAE to the standard-Gaussian prior mean $z_C = \mathbf{0}$ and decodes the conditional-prior mean of ArtPlanner. In Sampled latent, ContactVAE draws five latent samples from $\mathcal{N}(0, \mathbf{I})$ for each condition, while ArtPlanner continues to decode its conditional-prior mean, matching the main inference pipeline. The table reports their average.

ContactVAE inference using the mean or sampled latent produces comparable trajectory errors in both interaction modes. The Single-Art. and Bim.-Art. subsets differ in object categories, language, and articulation ranges, so their absolute errors are reported separately. Across the two settings, approximately 89%–90% of the predicted reference values lie within the training articulation range of the corresponding object category. Figure 7 further shows that the reference follows the range and temporal progression of the ground-truth trajectory.

H.3 ManiVAE: Reconstruction with a Compact Hand Representation

ManiVAE aims to preserve the geometry required for local contact configurations while substantially reducing the dimensionality of hand pose. We use a 32-dimensional latent variable and evaluate its capacity through posterior-mean reconstruction on all valid hand frames in the test set. The left hand is mirrored into the canonical right-hand space for encoding and decoding and then mapped back to its original coordinate system; the right hand is evaluated directly in the canonical right-hand space.

The 32-dimensional latent variable reconstructs hand joints and vertices with errors of approximately 2 mm, and the similar left- and right-hand errors indicate that canonicalization does not introduce a marked handedness imbalance. This experiment isolates the reconstruction capacity of ManiVAE, while the direct-pose ablation in the main paper evaluates its contribution to full-sequence generation. Together, the reconstruction results and system-level ablation support using low-dimensional pose tokens in place of direct regression of high-dimensional MANO poses.

I Additional Visualizations and a Failure Case

This section complements the quantitative evaluation with qualitative examples of complete generated interactions. We first examine the primary Bim.-Art. setting, then demonstrate the remaining three interaction modes, and finally discuss a representative failure case involving local interpenetration.

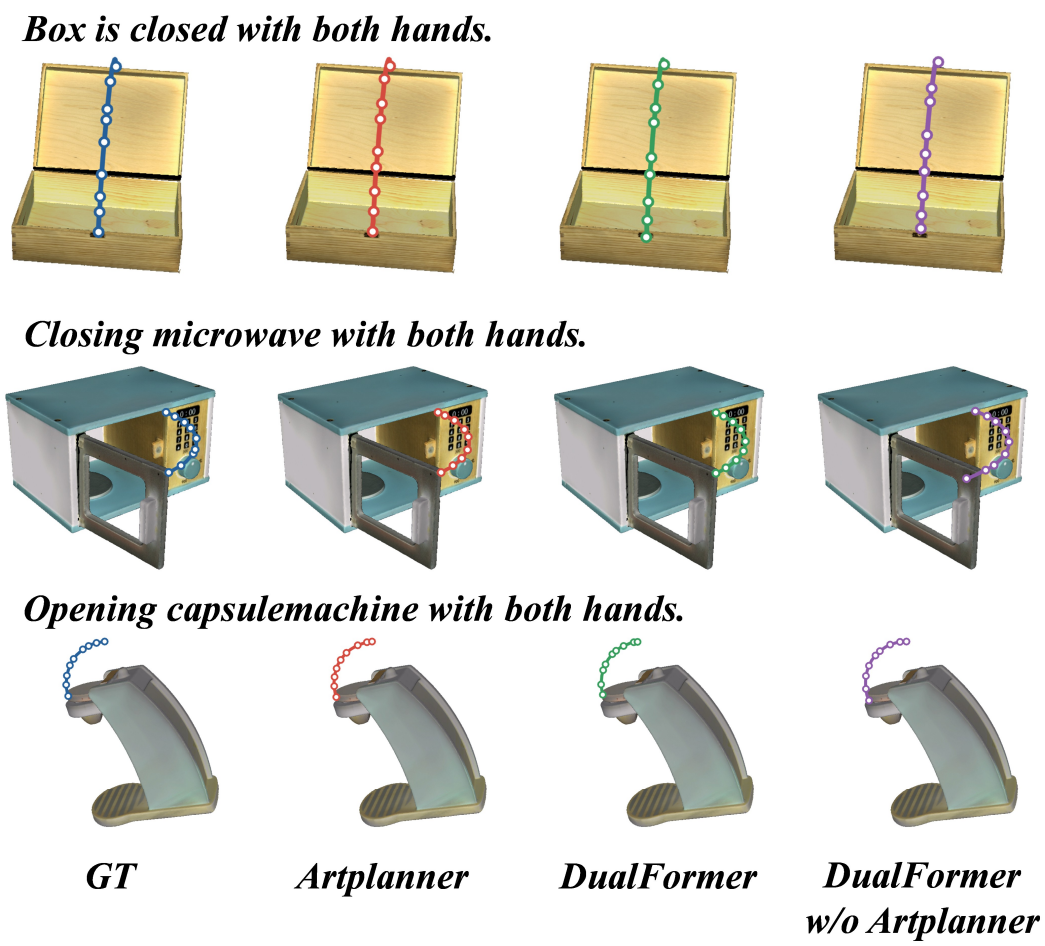


Figure 7: Qualitative comparison of articulation trajectories. Each row shows one articulated object. From left to right, the columns show the ground-truth trajectory, the ArtPlanner reference under prior-mean ContactVAE contact intent, the DualFormer output conditioned on this reference, and the DualFormer output without ArtPlanner.

I.1 Bimanual Articulated-Object Interactions

Figure 8 presents generated Bim.-Art. sequences for three objects with distinct articulation patterns. Across the sequences, the two hands assume complementary manipulation and support roles while adapting to the evolving object state.

I.2 Additional Interaction Modes

Beyond the primary Bim.-Art. setting, DuetHOI also supports interactions with different numbers of active hands and object types. Figure 9 shows generated sequences for the remaining three interaction modes.

I.3 Representative Failure Case

Figure 10 shows that the overall notebook-opening motion and bimanual interaction pattern remain coherent, but local interpenetration appears around the moving panel in later frames. This example indicates that object-fixed refinement does not completely eliminate collisions near thin articulated parts.

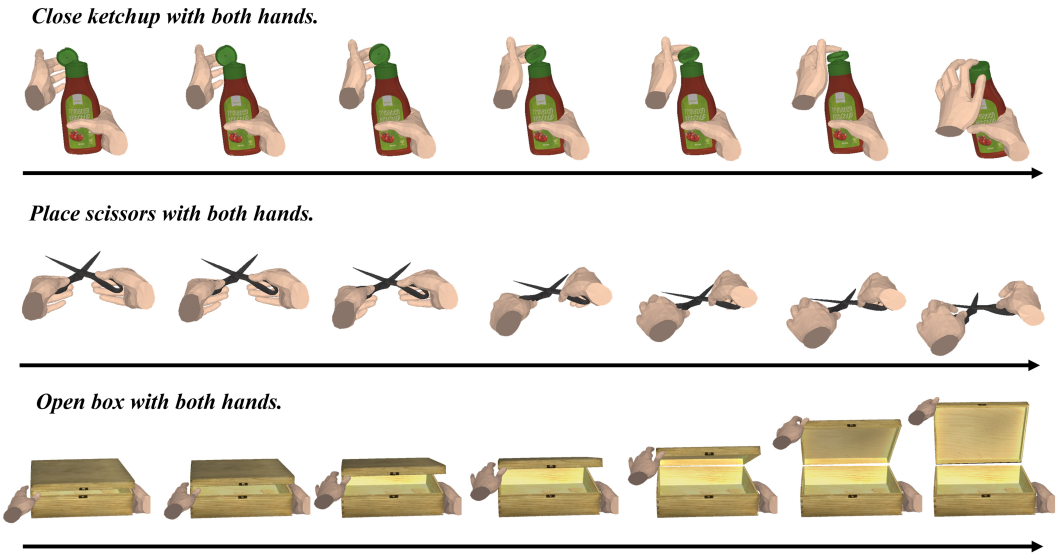


Figure 8: Qualitative results on Bim.-Art. interactions. From top to bottom, the rows show generated interactions with a ketchup bottle, scissors, and a box. Columns show successive frames from left to right.

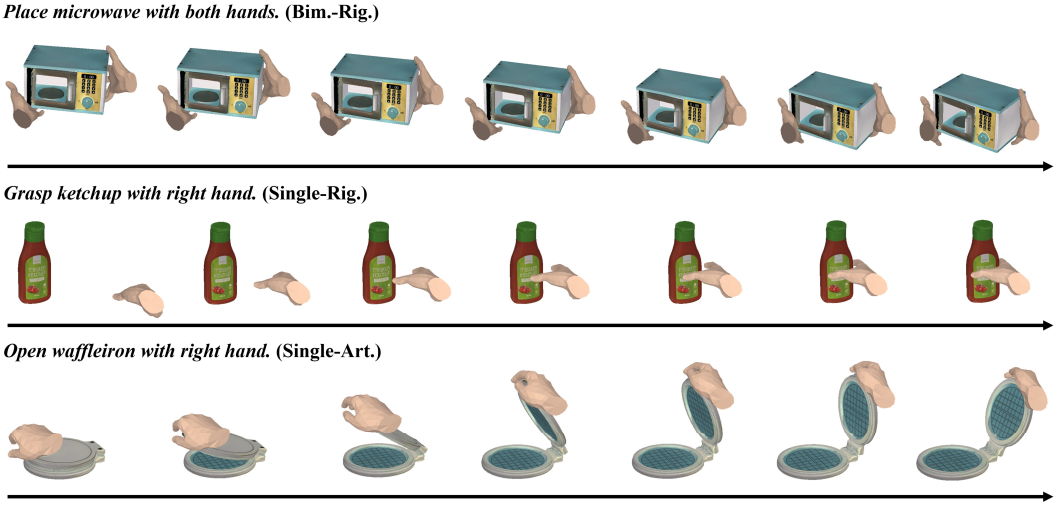


Figure 9: Qualitative results on additional interaction modes. From top to bottom, the rows show Bim.-Rig., Single-Rig., and Single-Art. interactions with a microwave, ketchup bottle, and waffle iron, respectively. Columns show successive frames from left to right.

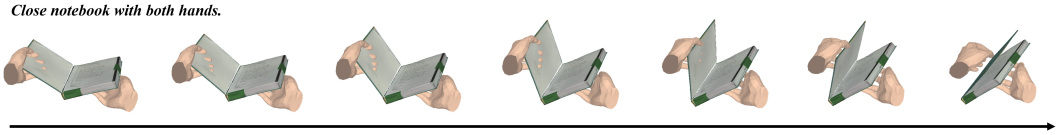


Figure 10: A notebook-interaction failure case. Later frames show local hand–object interpenetration near the moving panel.