DISCONTACT: GENERATING HAND MOTION AND DISCRETE CONTACT MAPS WITH GEOMETRY-COUPLED HYBRID DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating hand–object interactions requires modeling two tightly coupled but structurally different quantities: continuous articulated hand motion and discrete contact over the hand surface. Existing approaches place both in a shared continuous or latent space, which ignores this difference. We introduce **DisContact**, a hybrid diffusion framework that generates each modality in its native state space and couples them through the geometry they share. Hand motion follows a Gaussian DDPM, while per-vertex binary contact follows a Bernoulli D3PM whose stationary distribution is the empirical contact frequency of each vertex. At every denoising step, the current motion estimate is decoded through differentiable MANO kinematics into vertex-aligned geometry that conditions contact denoising, and pooled contact features refine the motion in return. A structured denoiser further exploits mesh connectivity and anatomical hand regions. On HoloAssist and ARCTIC, DisContact improves contact F1 and motion accuracy over prior work in nearly all settings, raising contact F1 on ARCTIC from 0.41 to 0.59. Controlled comparisons show that most of the benefit of coupling comes from explicit geometry rather than generic feature exchange, and new object-free diagnostics show that the generated motion and contact are more consistent with each other. These results indicate that keeping each modality in its own generative process, while linking them through the geometry they describe, is an effective way to generate structured hand–object interactions.

1 INTRODUCTION

Digital twins, mixed-reality assistance, and embodied AI need realistic demonstrations of *how* people manipulate objects, in tasks such as assembly, maintenance, or machine operation (Wang et al., 2023). Such a demonstration involves two things: how the hand moves, and which parts of it touch the object over time. The same trajectory can correspond to quite different interactions depending on where contact occurs, so both are needed for realistic and interpretable generation (Fig. 1).

Most generative approaches do not treat contact as something to generate. It is either precomputed on the object surface and used as a condition (Cha et al., 2024; Zhang et al., 2025; Adossehoun & Wannous, 2026), enforced through guidance at inference time (Christen et al., 2024), or folded into a latent grasp representation (Li et al., 2025a). These methods assume access to object geometry, and none of them generates contact together with the evolving hand articulation.

LatentAct (Prakash et al., 2025) removes the need for object geometry. From a single RGB image, an action description, and a sparse 3D contact point, it predicts MANO hand motion, using the 6D rotation representation of Zhou et al. (2019), together with binary per-vertex contact on the hand surface. The price is a more underconstrained problem, since both articulated motion and dense contact must be inferred from sparse cues.

The two outputs also differ in nature. Motion is continuous and evolves through articulated rotations and translations. Contact is a sparse binary field on the MANO surface: neighboring vertices tend to switch together, and grasping or releasing can change it abruptly. LatentAct, like later tokenization-based methods (Huang et al., 2025), encodes both in a shared VQ-VAE latent space (Van Den Oord et al., 2017). This is convenient, but it forces two very different state structures into a single repre-

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.

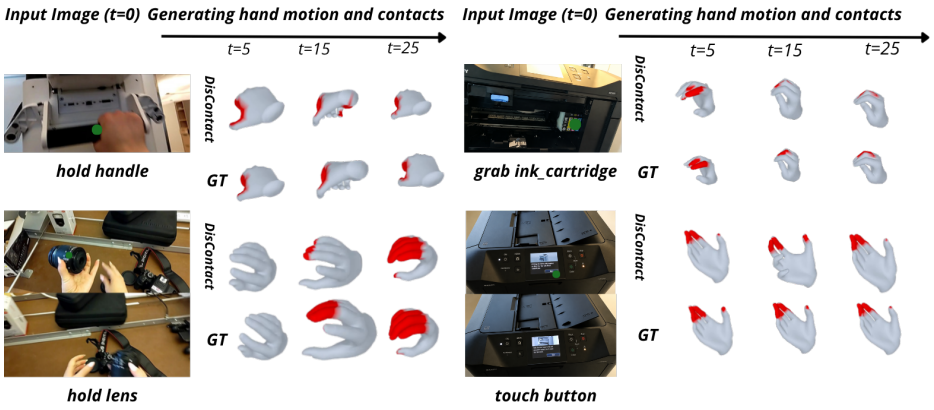

Figure 1: **Two modalities, two probabilistic spaces.** DisContact generates continuous hand articulation and a discrete per-vertex contact field (red) as two coupled diffusion processes, each in its own state space. Inputs are a single frame, an action description, and a 3D contact point (green). HoloAssist examples at frames $\tau = 5, 15, 25$, with ground truth below each prediction.

sentation and blurs the discrete, highly non-uniform nature of contact. Modeling the two separately is not enough either, because a contact variable only acquires a spatial meaning once the motion has produced a hand geometry. This raises a central modeling question: *how can motion and contact be generated in their own native state spaces while preserving their geometric dependency throughout generation?*

Joint continuous–discrete diffusion is a natural starting point. JointDiff (Capellera et al., 2026) generates multi-agent trajectories and possession events together by diffusing each modality in its own state space with a shared denoiser. This works because both modalities are indexed the same way: at each timestep, a single categorical event is concatenated with the agents' trajectories. Hand interaction breaks this assumption. Contact consists of $T_{\text{seq}} \times V$ binary variables indexed by mesh vertices, while motion lives in articulation parameters with no vertex-level counterpart. The two processes share a timeline, but relating them also requires the hand geometry that the motion produces, and this geometry has to be made explicit.

DisContact makes this geometry explicit. Motion follows a Gaussian DDPM (Ho et al., 2020), and contact a Bernoulli D3PM (Austin et al., 2021) whose stationary distribution is the empirical contact frequency of each vertex rather than a uniform prior. At every reverse step, the current motion estimate is decoded through differentiable MANO kinematics into hand vertices, so that each contact variable sees the geometry of its own vertex. In return, pooled contact features refine the motion estimate. Our contributions are as follows:

(1) Geometry-coupled hybrid diffusion. We extend joint continuous–discrete diffusion to discrete variables indexed by the vertices of a surface that the continuous process itself generates, and couple the two reverse processes through differentiable MANO kinematics.

(2) Structure-aware contact modeling. We model sparse, non-uniform contact with vertex-specific stationary marginals and a denoiser that reasons over the MANO surface through sparse mesh attention, anatomical regions derived from skinning weights, and region-density supervision.

(3) Evaluating joint generation. DisContact improves motion and contact accuracy over prior work on HoloAssist and ARCTIC. We further introduce object-free diagnostics of motion–contact consistency, and show through controlled comparisons that both native discrete modeling and geometry-mediated coupling are needed to achieve it.

2 Related Work

Hand–object interaction generation. Generative models of hand–object interactions have moved from static grasp synthesis to full sequences conditioned on language, body motion, or visual in-

put. Text-conditioned methods generate hand–object motion from action descriptions (Cha et al., 2024; Christen et al., 2024), and a related line of work synthesizes full-body human–object interactions (Ghosh et al., 2023; Xu et al., 2023; Diller & Dai, 2024; Peng et al., 2025). In most of these methods, contact is defined on the object and enters as a condition, a guidance term, or a latent grasp code (Sec. 1). Grasp and affordance models predict contact regions from object geometry (Liu et al., 2023; Jiang et al., 2021; Karunratanakul et al., 2020; Hasson et al., 2019), but they target static grasps rather than contact that evolves over time.

LatentAct (Prakash et al., 2025) is the closest to our setting: it generates hand motion and per-vertex contact from an image, an action description, and a 3D contact point, without an object mesh. HOIGPT (Huang et al., 2025) also represents interactions with discrete tokens, while Adossehoun & Wannous (2026) introduce geometry-constrained, contact-aware diffusion for text-driven synthesis. We adopt the setting of LatentAct, but treat motion and contact as continuous and discrete generative variables, respectively.

Continuous, discrete, and hybrid diffusion. DDPMs (Ho et al., 2020) are now standard for continuous data, including human motion (Tevet et al., 2023). Discrete diffusion (Sohl-Dickstein et al., 2015; Austin et al., 2021; Hoogeboom et al., 2021) brings the same idea to categorical and binary variables through Markov transition matrices, with uniform, absorbing, or data-dependent corruption. DiGress (Vignac et al., 2022) showed that transitions preserving the empirical marginals work well for graph generation; our contact process applies this idea to binary variables, with one marginal per MANO vertex. Diffusion over mixed continuous and discrete variables has mostly targeted static data, such as layouts (Levi et al., 2023), molecules (Vignac et al., 2023), CAD sketches (Chereddy & Femiani, 2025), and joint image–text generation (Li et al., 2025b). JointDiff (Capellera et al., 2026) extends it to dynamic data, but assumes that both modalities share the same index space, which does not hold for hand interaction (Sec. 1). DisContact instead couples the two processes through the hand geometry that the motion generates.

3 DisContact

3.1 Problem Formulation

Given a single RGB image I, an action description s, and a 3D contact point $\mathbf{p} \in \mathbb{R}^3$, jointly denoted by $\mathbf{y} = (I, s, \mathbf{p})$, we generate an interaction trajectory

$$\mathbf{X}_0 = (\mathbf{M}_0, \mathbf{C}_0), \qquad \mathbf{M}_0 \in \mathbb{R}^{T_{\text{seq}} \times 99}, \qquad \mathbf{C}_0 \in \{0,1\}^{T_{\text{seq}} \times V}, \tag{1}$$

where $V = 778$ is the number of MANO vertices. At frame τ, $\mathbf{m}^\tau$ contains the global orientation, 15 articulated joint rotations in continuous 6D representation, and global translation, while c_v^τ denotes the binary contact state of vertex v. Following Prakash et al. (2025), $\mathbf{p}$ is encoded as a Gaussian heatmap on a fixed $16 \times 16 \times 16$ metric grid. We model $p(\mathbf{M}_0, \mathbf{C}_0 \mid \mathbf{y})$.

Motion and contact live in different state spaces but are tied by geometry. Let $\mathcal{K}$ be the differentiable MANO forward model, with shape fixed to the mean ($\boldsymbol{\beta} = \mathbf{0}$). At frame τ, vertex v lies at $\mathbf{x}_v^\tau = \mathcal{K}_v(\mathbf{m}^\tau)$, and c_v^τ only has a spatial meaning through this position. DisContact keeps each modality in its own state space and couples their reverse processes through the generated hand geometry (Fig. 2).

3.2 Native-Space Forward Diffusion

We define a joint forward process over the interaction state $\mathbf{X}_t = (\mathbf{M}_t, \mathbf{C}_t)$. Since motion and contact belong to different state spaces, its transition kernel factorizes by modality,

$$q(\mathbf{X}_t \mid \mathbf{X}_0) = q_m(\mathbf{M}_t \mid \mathbf{M}_0)\, q_c(\mathbf{C}_t \mid \mathbf{C}_0). \tag{2}$$

Both modalities share the diffusion index t, but each is corrupted in its own space.

For continuous motion, we use a Gaussian DDPM,

$$q_m(\mathbf{M}_t \mid \mathbf{M}_0) = \mathcal{N}\Big(\mathbf{M}_t; \sqrt{\bar{\alpha}_t^m}\mathbf{M}_0, (1 - \bar{\alpha}_t^m)\mathbf{I}\Big), \tag{3}$$

with the cosine schedule of Nichol & Dhariwal (2021).

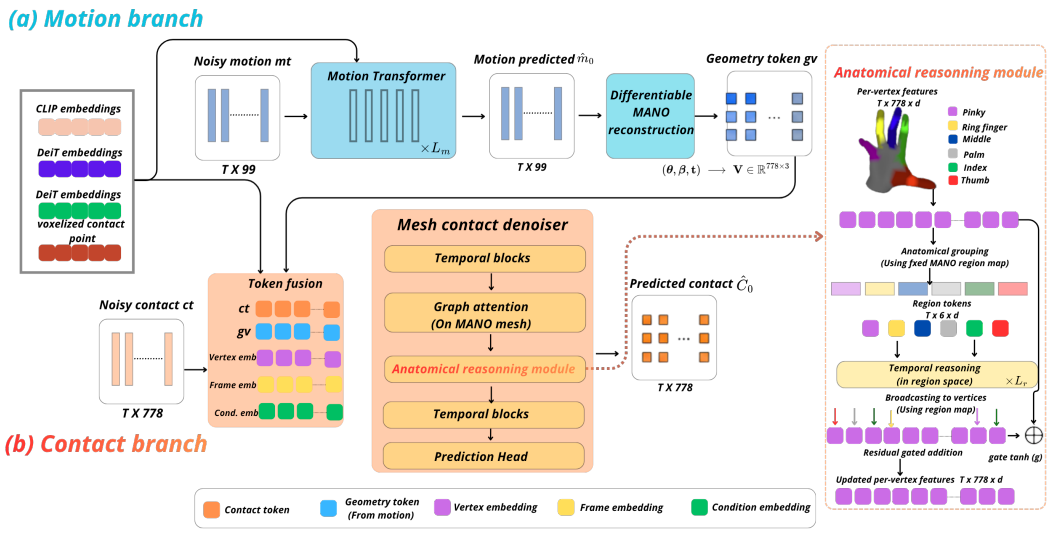

Figure 2: **Overview of DisContact.** Motion and contact are modeled in their native continuous and discrete state spaces. During reverse generation, the intermediate clean-motion estimate is mapped through MANO to vertex-aligned geometry that conditions contact denoising, while contact features are aggregated back to refine motion.

For binary contact, we use a marginal-preserving Bernoulli D3PM,

$$q_c(\mathbf{C}_t \mid \mathbf{C}_0) = \prod_{\tau=1}^{T_{\text{seq}}} \prod_{v=1}^{V} \text{Bern}\big(c_{t,v}^{\tau}; \bar{\alpha}_t^c c_{0,v}^{\tau} + (1 - \bar{\alpha}_t^c) p_v\big)\,, \tag{4}$$

where $p_v = \Pr_{\text{data}}(c_{0,v}^{\tau} = 1)$ is the empirical contact frequency of vertex v on the training set. This is the binary, per-vertex case of the marginal-preserving kernel of Vignac et al. (2022) (App. A.1): instead of drifting toward a uniform distribution, each vertex relaxes toward its own contact rate, so fingertips and palm keep their very different statistics at every noise level. As $\bar{\alpha}_t^m, \bar{\alpha}_t^c \to 0$, the process reaches the terminal distributions

$$\mathbf{M}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \mathbf{C}_T \sim \prod_{\tau,v} \text{Bern}(p_v). \tag{5}$$

Because the kernel factorizes, $\mathbf{M}_t$ and $\mathbf{C}_t$ are independent given $\mathbf{X}_0$. Corruption therefore introduces no cross-modal dependency, and all coupling is learned in the reverse process.

3.3 Geometry-Coupled Reverse Diffusion

Both branches use an x_0 parameterization. At diffusion step t, the motion branch first predicts a base clean-motion estimate together with frame-level motion features,

$$\left(\hat{\mathbf{M}}_0^{\text{base}}, \mathbf{H}^m\right) = f_\theta(\mathbf{M}_t, t, \mathbf{y}), \qquad \mathbf{H}^m = \{\mathbf{h}^{m,\tau}\}_{\tau=1}^{T_{\text{seq}}}. \tag{6}$$

To give the contact branch geometry aligned with its frame–vertex layout, we decode this estimate through MANO and center the vertices frame by frame,

$$\mathbf{g}_{\tau,v}^t = \phi\left(s_g \left[\mathcal{K}_v(\hat{\mathbf{m}}_{0,\tau}^{\text{base}}) - \frac{1}{V}\sum_u \mathcal{K}_u(\hat{\mathbf{m}}_{0,\tau}^{\text{base}})\right]\right), \tag{7}$$

where ϕ is a shared MLP and $s_g = 10$ rescales metric coordinates. We write $\mathbf{G}_t = \Gamma(\hat{\mathbf{M}}_0^{\text{base}}) = \{\mathbf{g}_{\tau,v}^t\}_{\tau,v}$. Centering makes Γ invariant to global translation, so the tokens encode articulation rather than hand position; global position still reaches the contact branch through $\mathbf{H}^m$.

The contact branch combines the noisy contact state, the motion features, the geometry tokens, the timestep, and the conditioning,

$$(\boldsymbol{\ell}_t, \mathbf{H}^c) = h_\eta(\mathbf{C}_t, \mathbf{H}^m, \mathbf{G}_t, t, \mathbf{y}), \qquad \hat{p}_\eta(c^\tau_{0,v} = 1 \mid \mathbf{X}_t, \mathbf{y}) = \sigma(\ell^\tau_{t,v}). \tag{8}$$

The two motion signals are complementary. $\mathbf{H}^m$ is broadcast to all vertices and provides frame-level context, whereas $\mathbf{g}^t_{\tau,v}$ tells each contact variable where its own vertex currently is. Sec. 4.4 measures this difference by removing the geometry tokens while keeping $\mathbf{H}^m$.

The predicted clean-contact distribution parameterizes the D3PM reverse transition,

$$p_\eta(c^\tau_{t-1,v} \mid \mathbf{X}_t, \mathbf{y}) = \sum_{a \in \{0,1\}} q(c^\tau_{t-1,v} \mid c^\tau_{t,v}, c^\tau_{0,v} = a)\, \hat{p}_\eta(c^\tau_{0,v} = a \mid \mathbf{X}_t, \mathbf{y}). \tag{9}$$

The first factor is the analytic posterior of the forward process and acts on each variable separately. All cross-modal and spatio-temporal dependencies therefore enter through the learned prediction $\hat{p}_\eta$, which sees the full noisy state $\mathbf{X}_t$ and the conditioning.

Contact-conditioned motion refinement. In the other direction, contact features are averaged over the hand surface and used to correct the base motion estimate,

$$\mathbf{s}^{c,\tau} = \frac{1}{V} \sum_v \mathbf{h}^{c,\tau}_v, \qquad \hat{\mathbf{m}}^\tau_0 = \hat{\mathbf{m}}^{\text{base},\tau}_0 + r_\theta(\mathbf{h}^{m,\tau}, \mathbf{s}^{c,\tau}). \tag{10}$$

The asymmetry is intentional: motion reaches contact vertex by vertex, while contact returns to motion as a frame-level summary. The output layer of r_θ is zero-initialized, so the correction starts at zero and grows as contact features become informative. The corrected estimate then parameterizes the standard Gaussian posterior,

$$p_\theta(\mathbf{M}_{t-1} \mid \mathbf{X}_t, \mathbf{y}) = \mathcal{N}\Big(\mathbf{M}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{M}_t, \hat{\mathbf{M}}_0), \tilde{\beta}^m_t \mathbf{I}\Big)\,. \tag{11}$$

Full reverse step. Putting both directions together,

$$p_{\theta,\eta}(\mathbf{M}_{t-1}, \mathbf{C}_{t-1} \mid \mathbf{X}_t, \mathbf{y}) = p_\theta(\mathbf{M}_{t-1} \mid \mathbf{X}_t, \mathbf{y}) \prod_{\tau,v} p_\eta(c^\tau_{t-1,v} \mid \mathbf{X}_t, \mathbf{y}). \tag{12}$$

The two factors are conditionally independent given $\mathbf{X}_t$, but they are not estimated independently: the contact factor sees $\mathbf{M}_t$ through $\mathbf{H}^m$ and $\Gamma(\hat{\mathbf{M}}^{\text{base}}_0)$, and the motion factor sees $\mathbf{C}_t$ through $\mathbf{H}^c$. The coupling lives in the parameterization of the reverse conditionals, not in the forward kernels.

Motion denoiser. Noisy motion frames are embedded into $\mathbb{R}^D$ with temporal positional encoding, a fused condition–timestep token $\mathbf{u}_{y,t}$ is prepended, and a Transformer encoder produces the frame features $\mathbf{H}^m \in \mathbb{R}^{T_{\text{seq}} \times D}$, from which a linear head predicts $\hat{\mathbf{M}}^{\text{base}}_0$.

Structured contact denoiser. Contact is a spatio-temporal binary field defined over the hand topology. For every frame and vertex, its initial representation is

$$\mathbf{z}^\tau_v = \mathrm{LN}\big(\mathbf{b}^\tau_v + \mathbf{e}^v_v + \mathbf{e}^f_\tau + \mathbf{u}^\tau_m + \mathbf{u}_{y,t} + \mathbf{g}^t_{\tau,v}\big)\,, \tag{13}$$

where the terms respectively encode the noisy binary state, vertex identity, frame identity, broadcast motion information, shared conditioning and timestep, and vertex-aligned geometry.

These tokens are processed along three complementary structures. *Temporal reasoning* models the evolution of each vertex throughout the interaction. *Mesh reasoning* propagates information between neighboring vertices with sparse attention over the $|\mathcal{E}| \approx 4.6$k bidirectional MANO edges, instead of all $V^2 \approx 605$k vertex pairs. *Anatomical reasoning* complements local propagation by partitioning the hand into $R = 6$ functional regions—the palm and five fingers. Region features are pooled, processed temporally, and broadcast back to their vertices through a gated residual.

A linear head maps $\mathbf{H}^c$ to the logits $\boldsymbol{\ell}_t$ of Eq. equation 8, and its vertex average feeds the motion refinement of Eq. equation 10. Full architectural details are given in App. A.3.

3.4 Training and Generation

The model is trained end to end with an unweighted sum of five losses, $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{vb}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{reg}}$, averaged over valid frames. On the motion side, $\mathcal{L}_m$ supervises the clean motion estimate and $\mathcal{L}_{\text{pos}}$ the resulting MANO joint positions. On the contact side, $\mathcal{L}_{\text{vb}}$ is the D3PM variational term, $\mathcal{L}_{\text{aux}}$ a binary cross-entropy on the clean-state logits, and $\mathcal{L}_{\text{reg}}$ matches predicted and ground-truth contact density within each anatomical region. Full definitions are given in App. B.

Joint generation. Sampling starts from the terminal distributions of Eq. equation 5. At each of the $T = 1000$ steps, the motion branch predicts $\hat{\mathbf{M}}_0^{\text{base}}$ and its geometry $\mathbf{G}_t$; the contact branch predicts the clean-contact distribution and $\mathbf{H}^c$, which refines the motion estimate (Eq. equation 10); and $\mathbf{M}_{t-1}$ and $\mathbf{C}_{t-1}$ are sampled from Eqs. equation 11 and equation 9.

4 Experiments

We evaluate DisContact along five questions: *(i)* does it improve motion and contact generation over existing approaches (Sec. 4.2); *(ii)* does modeling contact in its native discrete state space matter (Sec. 4.3); *(iii)* does explicit geometry provide information beyond generic cross-modal coupling (Sec. 4.4); *(iv)* are the generated motion and contact mutually consistent; and *(v)* does structured reasoning over the hand surface help beyond per-vertex prediction? Questions *(iv)* and *(v)*, together with a longer generation horizon, are addressed in Sec. 4.5.

4.1 Experimental Setup

Protocol. We follow Prakash et al. (2025) and evaluate Forecasting and Interpolation, each with and without the hand visible in the input image (*Hand visible* / *No hands*). On HoloAssist, generalization is tested at the task, object, action, and scene levels; ARCTIC uses its official split. Unless stated otherwise, $T_{\text{seq}} = 30$, and controlled comparisons use HoloAssist No hands–Forecasting at the task level, with every variant retrained from scratch (App. C).

Sampling. Each model is trained once. For DisContact and the diffusion variants, Tabs. 1 and 3 report a single sampling run, while all other tables report the mean over 20 independent sampling runs; where shown, $\pm$ is the standard deviation across these runs and reflects sampling variability, not variation across training runs.

Baselines and metrics. We compare with the four methods evaluated by Prakash et al. (2025): HCTFormer, HCTDiff, LatentAct, and LatentAct-Diff. Rows marked † are taken from that paper; for the long-horizon experiment, we retrain LatentAct at $T_{\text{seq}} = 75$. We report MPJPE (M-PE) and Procrustes-aligned MPJPE (M-PA), both in cm, and vertex-level contact F1.

4.2 Interaction Trajectory Generation

HoloAssist. DisContact obtains the best contact F1 and MPJPE in all 16 settings (Tab. 1), with F1 gains of .03–.06 over the strongest baseline and MPJPE reductions of 0.13–0.78 cm. For MPJPE-PA, HCTDiff remains slightly better in three of the four No hands–Forecasting columns, by at most 0.07 cm. The margin holds in the No hands setting, where the input image does not show the hand configuration.

ARCTIC. On ARCTIC (Tab. 2), DisContact reduces MPJPE from 14.77 to 11.24 cm and raises contact F1 from .41 to .59 relative to LatentAct, showing that the gains carry over to a different dataset and set of objects.

Fig. 3 illustrates the difference in contact prediction, with further examples in App. D.

4.3 Native Continuous–Discrete Modeling

Block A of Tab. 3 varies how contact is represented. *Joint DDPM* concatenates motion and relaxed contact into a single Gaussian state with one denoiser, in the spirit of shared-representation approaches; *Factorized DDPM* keeps our two-branch architecture and coupling but diffuses contact as

Table 1: **Interaction generation on HoloAssist.** M-PE/M-PA are in centimeters (lower is better); higher F1 is better. DisContact results come from a single sampling run. †Reported from Prakash et al. (2025).

	Task			Object			Action			Scene		
Method	M-PE	M-PA	F1	M-PE	M-PA	F1	M-PE	M-PA	F1	M-PE	M-PA	F1
Hand visible — Forecasting												
HCTFormer†	8.32	3.06	.72	8.70	3.19	.72	8.59	3.25	.73	8.52	3.09	.72
HCTDiff†	8.42	2.75	.72	9.01	2.88	.73	9.11	3.10	.74	9.49	2.84	.71
LatentAct†	7.61	2.99	.75	8.09	3.06	.75	8.14	3.32	.77	7.87	3.04	.74
LatentAct-Diff†	7.90	3.03	.71	8.49	3.20	.72	8.29	3.34	.72	8.32	3.10	.72
DisContact	**6.83**	**2.60**	**.79**	**7.90**	**2.82**	**.81**	**7.98**	**3.04**	**.82**	**7.66**	**2.79**	**.78**
Hand visible — Interpolation												
HCTFormer†	7.52	3.02	.76	8.10	3.10	.76	7.77	3.29	.78	7.54	3.00	.76
HCTDiff†	8.30	2.72	.77	8.52	2.91	.78	8.38	3.16	.78	8.60	2.81	.77
LatentAct†	6.72	2.87	.80	7.48	3.02	.78	7.25	3.19	.80	7.27	2.93	.78
LatentAct-Diff†	6.53	2.62	.78	7.23	2.81	.78	7.11	3.10	.79	6.70	2.84	.78
DisContact	**6.40**	**2.43**	**.83**	**7.02**	**2.70**	**.82**	**6.81**	**2.98**	**.83**	**6.52**	**2.73**	**.82**
No hands — Forecasting												
HCTFormer†	8.34	2.95	.72	8.85	3.02	.73	8.54	3.21	.74	8.53	3.05	.72
HCTDiff†	8.85	**2.65**	.72	9.74	**2.81**	.72	9.50	3.11	.74	9.21	**2.74**	.71
LatentAct†	7.93	2.97	.76	8.26	3.01	.76	8.19	3.28	.77	7.90	3.00	.75
LatentAct-Diff†	7.82	2.89	.71	8.63	3.06	.71	8.19	3.15	.72	7.93	2.96	.73
DisContact	**7.55**	2.72	**.80**	**7.82**	2.83	**.80**	**7.96**	**3.03**	**.81**	**7.65**	2.78	**.79**
No hands — Interpolation												
HCTFormer†	7.30	2.95	.76	8.42	3.15	.76	8.01	3.30	.77	7.29	2.99	.77
HCTDiff†	8.27	2.76	.76	8.68	2.91	.76	9.33	3.10	.78	10.02	2.82	.76
LatentAct†	7.13	2.89	.79	7.24	2.97	.79	7.44	3.21	.80	7.12	2.89	.78
LatentAct-Diff†	6.60	2.65	.76	7.73	2.78	.76	7.42	3.04	.78	6.89	2.71	.77
DisContact	**6.46**	**2.56**	**.82**	**7.05**	**2.68**	**.82**	**7.14**	**2.95**	**.83**	**6.61**	**2.61**	**.81**

Table 2: **Results on ARCTIC.** DisContact results are averaged over 20 sampling runs. †Reported from Prakash et al. (2025).

Method	M-PE↓	M-PA↓	F1↑
HCTFormer†	15.76	3.61	.36
LatentAct†	14.77	3.76	.41
DisContact	**11.24**	**3.28**	**.59**

a continuous Gaussian variable; *Hybrid D3PM* switches to a Bernoulli process with a single global prior p. Each step improves both motion and contact, from .71 to .80 F1 and from 8.21 to 7.55 cm. Separating the two modalities accounts for .04 F1, diffusing contact as a binary variable for a further .03, and per-vertex rather than global marginals for the remaining .02. Since Joint DDPM also lacks geometry coupling, the comparison that isolates the state space alone is Factorized DDPM versus the Bernoulli variants.

4.4 Geometry-Mediated Coupling

Block B of Tab. 3 asks whether it matters that motion reaches contact *through geometry*. Without any coupling, F1 drops from .80 to .69. Restoring only the broadcast motion features (*w/o geometry tokens*) recovers just .03 of this gap, while the vertex-aligned geometry recovers the remaining .08. Motion benefits as well (8.07 → 7.95 → 7.55 cm), through the contact-to-motion refinement. Generic feature exchange between the branches is therefore not enough: most of the benefit of coupling comes from telling each contact variable where its vertex currently is.

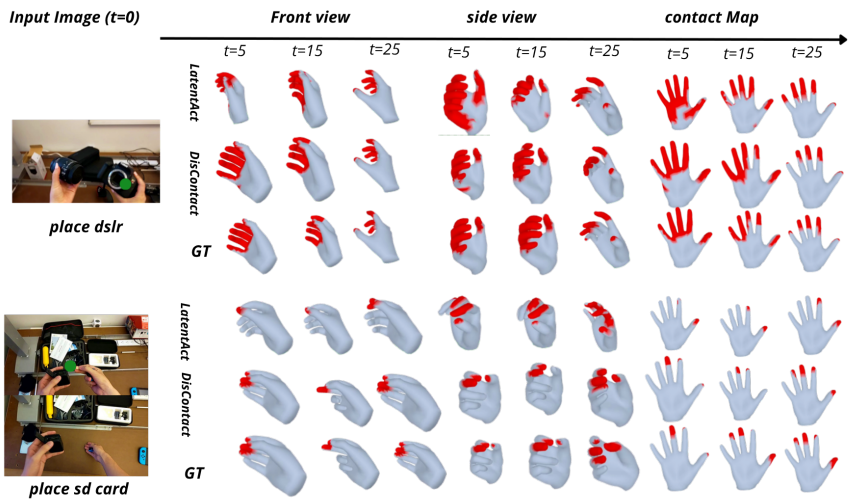

Figure 3: **Comparison with LatentAct** on two held-out HoloAssist interactions (task-level split). Front view, side view, and flattened MANO surface at $\tau = 5, 15, 25$; contact in red, input contact point in green.

Table 3: **Controlled comparisons** on HoloAssist No hands–Forecasting (task level, $T_{\text{seq}} = 30$). Each variant is retrained from scratch, changing only the listed component: the contact state space (A), the cross-modal coupling (B), or the contact denoiser (C). p is a single global contact rate, whereas DisContact uses per-vertex marginals p_v. Results come from a single sampling run.

Variant	M-PE↓	M-PA↓	F1↑
(A) State space of the contact process			
Joint DDPM (shared Gaussian state)	8.21	2.98	.71
Factorized DDPM (separate Gaussians)	7.91	2.87	.75
Hybrid D3PM, global prior p	7.69	2.78	.78
(B) Cross-modal coupling			
Independent processes (no coupling)	8.07	3.01	.69
w/o geometry tokens (latent only)	7.95	2.98	.72
(C) Structured contact denoising			
w/o mesh reasoning	7.73	2.82	.77
w/o anatomical reasoning	7.67	2.79	.78
w/o mesh & anatomical	7.82	2.87	.75
w/o region-density supervision	7.59	2.74	.79
DisContact (full)	**7.55**	**2.72**	**.80**

4.5 Consistency and Further Analysis

Motion–contact consistency. Per-modality accuracy does not show whether the generated motion and contact form a coherent interaction. Since HoloAssist has no 3D object geometry, we measure their consistency on the generated hand surface with four object-free diagnostics (Tab. 4; definitions in App. E): enclosure consistency E (fraction of contact vertices whose normals face the contact centroid), persistence P (IoU of consecutive contact supports), normalized rigidity $\widetilde{R}$ (residual of the best rigid fit of persistently contacting vertices), and coverage Q_R (fraction of frame pairs where $\widetilde{R}$ is defined). P and Q_R have no preferred direction and are read against the ground truth.

DisContact is the closest variant to this reference on all four diagnostics ($E = .75$ vs. $.82$, $\widetilde{R} = .10$ vs. $.06$, $Q_R = 88.0\%$ vs. 92.1%). With cross-modal coupling held fixed, Factorized DDPM is markedly less consistent, isolating the benefit of native discrete contact modeling. Removing the

Table 4: **Motion–contact consistency** on HoloAssist No hands–Forecasting (task level, $T_{\text{seq}} = 30$). GT is an empirical reference, not a strict bound. ±: std. over 20 sampling runs.

Method	$E \uparrow$	P	$\widetilde{R} \downarrow$	Q_R (%)
GT motion + contact (reference)	.82	.74	.06	92.1
Factorized DDPM	.63±.02	.57±.02	.19±.02	75.2
Independent processes	.55±.03	.48±.03	.27±.03	62.7
w/o geometry tokens	.60±.02	.54±.02	.22±.02	70.8
DisContact	**.75±.01**	.70±.01	**.10±.01**	88.0

Table 5: **Long-horizon generation** ($T_{\text{seq}} = 75$) on HoloAssist No hands–Forecasting (task level). All models are retrained at this horizon; LatentAct is deterministic and evaluated once.

Method	M-PE↓	M-PA↓	F1↑
Joint DDPM	9.12 ± 0.21	3.08 ± 0.08	$.69 \pm .02$
Factorized DDPM	8.54 ± 0.18	2.91 ± 0.07	$.75 \pm .02$
LatentAct	8.33	3.01	.73
DisContact	$\mathbf{7.62 \pm 0.15}$	$\mathbf{2.74 \pm 0.06}$	$\mathbf{.78 \pm .01}$

geometry tokens, or all coupling, degrades consistency further, with independent processes yielding the largest rigidity residual ($\widetilde{R} = .27$). These diagnostics also separate the variants more sharply than marginal metrics: removing the geometry tokens costs .08 in contact F1 but .15 in enclosure consistency, and more than doubles $\widetilde{R}$.

Structured contact denoising. Block C of Tab. 3 shows that mesh and anatomical reasoning contribute complementary spatial structure: removing either reduces contact F1 by .03 and .02, respectively, while removing both reduces it by .05 and yields the largest increase in motion error within the block (7.55 → 7.82 cm). Since the contact denoiser does not predict motion directly, this degradation propagates through the contact-to-motion refinement of Eq. (10), indicating that a better-structured contact representation also benefits motion. Region-density supervision has a smaller effect and is best interpreted as an auxiliary regularizer.

Long-horizon generation. We retrain all models at $T_{\text{seq}} = 75$ (Tab. 5). The ordering observed at the standard horizon is preserved, and DisContact outperforms LatentAct in both motion and contact. Relative to $T_{\text{seq}} = 30$, DisContact degrades only marginally (M-PE 7.55 → 7.62, F1 .80 → .78), whereas Joint DDPM degrades substantially (M-PE 8.21 → 9.12), suggesting that native-space modeling with geometry-mediated coupling remains stable over longer interactions.

5 Conclusion

We presented DisContact, a hybrid diffusion model that generates hand motion and per-vertex contact, each in its native state space: a Gaussian DDPM for articulated motion and a Bernoulli D3PM with per-vertex marginals for contact. The two reverse processes are coupled at every step through the hand geometry decoded from the current motion estimate, and a structured denoiser reasons over the mesh and anatomical regions of the hand. On HoloAssist and ARCTIC, DisContact improves contact F1 and motion accuracy over prior work in nearly all settings. Controlled comparisons show that most of the benefit of coupling comes from vertex-aligned geometry rather than generic feature exchange, and object-free diagnostics show that native discrete modeling and geometry-mediated coupling together produce motion and contact that are more consistent with each other.

The current model handles a single hand, requires a 3D contact point as input, and is slower than LatentAct because of its 1000-step sampling; our consistency diagnostics also do not assess physical compatibility with a specific object (App. F). Bimanual interaction, faster hybrid sampling, and object-aware evaluation are natural next steps. More broadly, coupling a continuous process with a discrete field defined on the geometry it generates is not specific to hands and could extend to body–scene contact or deformable-object manipulation.

REFERENCES

Josue Adossehoun and Hazem Wannous. Gcd: Geometry–constrained contact-aware diffusion for text–driven 3d hand–object motion synthesis. In *International Conference on Pattern Recognition*, pp. 404–418. Springer, 2026.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Guillem Capellera, Luis Ferraz, Antonio Romano, Alexandre Alahi, and Antonio Agudo. Jointdiff: Bridging continuous and discrete in multi-agent trajectory generation. In *International Conference on Learning Representations*, volume 2026, pp. 75912–75942, 2026.

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.

Sathvik Chereddy and John Femiani. Sketchdnn: Joint continuous-discrete diffusion for cad sketch generation. *arXiv preprint arXiv:2507.11579*, 2025.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19888–19901, 2024.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pp. 1–12. Wiley Online Library, 2023.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11807–11816, 2019.

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

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

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.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pp. 333–344. IEEE, 2020.

Elad Levi, Eli Brosh, Mykola Mykhailych, and Meir Perez. Dlt: Conditioned layout generation with joint discrete-continuous diffusion layout transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2106–2115. IEEE, 2023.

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, 2025a.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2779–2790. IEEE, 2025b.

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.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2878–2888, 2025.

Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7026–7036, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 560–576. Springer, 2023.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14928–14940, 2023.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27694–27705, 2025.

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.