COMPACT: COMPOSITIONAL 3D HUMAN-OBJECT INTERACTION SYNTHESIS

Anonymous authors
Paper under double-blind review

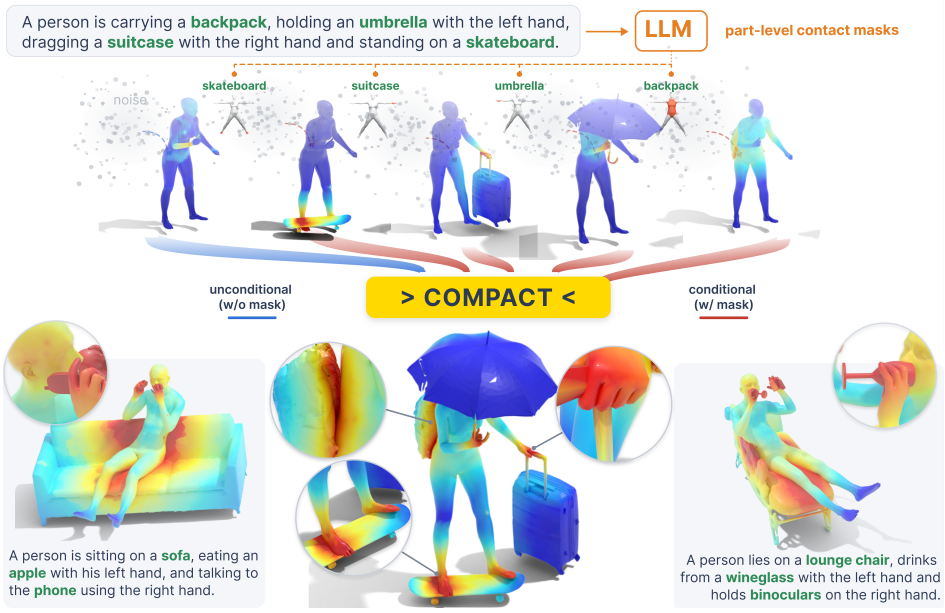

Figure 1: We develop CompAct, a novel model that generates *composite multi-object* interactions, given a text prompt and body and object shapes. Crucially, CompAct *trains only on single-object* interactions; it does *not* require new, costly data capture or augmentation for multi-object interactions. First, aided by LLM-estimated contacts, CompAct predicts proxy, conditional single-object interactions along with their interaction fields (InterFields). Then, it blends these proxies into a composite interaction, under a novel guidance term that exploits *per-part* InterFields, while regularization by an unconditional body prior helps realism. See three generated composite interactions at the bottom.

ABSTRACT

We rarely interact with a single object; we carry, hold, and manipulate several at once. Most methods on 3D interaction synthesis ignore multi-object interactions, as this needs new data capture or augmentation of many (ideally all) possible combinations; however, this does not scale. Our insight is that multi-object interactions are *compositions* of single-object ones. Thus, we represent interactions at the *body-part level* through interaction fields (InterFields) that encode contact and proximity. We exploit this in CompAct, a novel Flow-Matching model that, given a text prompt and body and object shape, jointly generates body pose, object pose, as well as body-part and object InterFields. At test time, CompAct predicts proxy single-object interactions and dynamically composes a multi-object one, using classifier-free guidance. To this end, composition exploits blending weights that naturally arise from the per-part InterFields of proxy interactions. Note that this is not a naive linear combination; instead, it is informed by a body prior arising from the classifier-free guidance, so that the composed interaction looks natural. Moreover, note that CompAct uses decoupled timesteps per input part; this supports interaction editing as a downstream application. CompAct is the first model to synthesize multi-object interactions by training only on single-object ones. Experiments on seven single-and multi-object benchmarks show that CompAct outperforms the state of the art in interaction fidelity, and produces multi-object interactions that are perceived as natural looking. Code and models will be released.

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.

1 INTRODUCTION

We rarely interact with one object at a time. A person sits on a sofa, eats an apple with one hand, and holds a phone to one ear with the other hand (Fig. 1); three interactions on one body at the same time. Generating such composite interactions matters for character animation, AR/VR, robotics and embodied agents. Yet most generative models of human-object interaction (HOI) still handle only one object at a time (Taheri et al., 2022; Wu et al., 2022; Kulkarni et al., 2024; Tendulkar et al., 2023; Paschalidis et al., 2025; Jiang et al., 2023; Li et al., 2023). Only a few multi-object methods exist, but these require new motion capture (Lv et al., 2024; Kim et al., 2025) for every combination of object shapes and positions, scene configurations, and intents, consequently, they do not scale.

Our insight is that a multi-object interaction is not a new task to be modelled on its own, but a *composition* of single-object ones, each anchored to different parts of the same body. Therefore, what stands in the way is not lacking data but an appropriate *representation*. Most methods encode an interaction at the level of the whole body. Even when the body and object are separate variables (Petrov et al., 2025), the body is one global pose and contact is one global map, consequently, the parts that drive an interaction cannot be considered separately from the ones at rest.

Therefore, we describe interaction at the *body-part level*. To this end, we build on interaction fields (InterFields) (Taheri et al., 2022; Fan et al., 2023; Zhang et al., 2021), i.e., maps of distances from the body to the object, v.v. We consider the distances as *signed*, so that penetration is distinct from contact. Importantly, we split the map into *one* field *per part*; each of these encodes contact and proximity as an independent quantity that can be learned, fixed, or edited. It can also be attended.

We exploit this by observing that single-object interactions involve only a *few* active parts, while multi-object interactions engage *all* active parts. Our key idea, is that *attending* active parts and their InterFields can effectively drive composition. However, naive attention does not consider the inter-dependence of parts in a skeletal chain, so it ignores useful structural information. To tackle this we define *contact chains* from each active part to the pelvis, and attend InterFields along these.

With the above, we develop *CompAct*, a novel Flow-Matching model that, given a text prompt and the body and object shapes, jointly generates body pose, object pose, the per-part body InterFields and the object InterField. A language model turns the prompt into per-part contact labels that assign an object only to body parts that engage with it. Then, the respective body parts and object, as well as their InterFields, are tokens of a transformer whose attention follows the contact chains.

At test time, CompAct first predicts *proxy* single-object interactions conditioned on object shape. Then, it blends these proxies into a multi-object interaction, through body-part weights that arise naturally from their InterFields, so that the body and object are "pulled" appropriately to satisfy contacts and proximal relationships. However, a naive linear combination results in unnatural-looking stitching. To tackle this, CompAct uses interaction-condition dropout during training, so it learns an *unconditional body prior*. Therefore, blending takes place "around" this prior using *classifier-free guidance*, so the composed interaction looks natural. Note that composition exploits signals extracted only from a model trained on single-object data; no multi-object training dataset is used.

Experiments on seven single- and multi-object benchmarks show that CompAct outperforms SotA methods in interaction fidelity and produces multi-object interactions that are perceived as natural.

As downstream application, we use CompAct also for *interaction editing*, namely first generating an interaction, and then changing it by selectively fixing some body parts and resampling others. This is enabled by the decoupled timesteps that let any subset of the interaction state condition the rest.

In summary, our contributions are:

1. A *formulation* of multi-object interactions as *part-level composition* of single-object interactions; this scales without requiring expensive new data capture or augmentation.
2. An efficient *representation* to guide composition, namely per-part signed InterFields.
3. CompAct, a *Flow-Matching model* with contact-chain attention, decoupled timesteps and interaction-condition dropout for learning joint, conditional and unconditional distributions in one network – thereby also supporting interaction editing.

Code and trained models for CompAct will be released for research purposes.

2 RELATED WORK

Single-Object HOI Synthesis: Most work handles *only one* object. GOAL (Taheri et al., 2022) and SAGA (Wu et al., 2022) predict a static SMPL-X grasp and use it as a boundary condition for motion; NIFTY (Kulkarni et al., 2024) learns an object-specific interaction field that guides diffusion sampling; FLEX (Tendulkar et al., 2023) and CWGrasp (Paschalidis et al., 2025) optimise body and hand priors against object and scene geometry. Beyond static grasps, OMOMO (Li et al., 2023) conditions the body on object trajectory, CHOIS (Li et al., 2024a) generates body and object motion from text and waypoints, and ChainHOI (Zeng et al., 2025) models joint- and chain-level interaction.

Text-driven generation (Wu et al., 2025a; Peng et al., 2025; Wu et al., 2025b; Song et al., 2024), physical or geometric priors (Diller & Dai, 2024; Xu et al., 2023; Li et al., 2024b; Xue et al., 2025), articulated objects and intent (Pi et al., 2025; Ghosh et al., 2023), distilled visual priors (Li & Dai, 2024; Xu et al., 2024), unseen geometry (Li et al., 2025), LLM affordance parsing (Zhang et al., 2025) and noise optimisation of a pretrained HOI model (Ron et al., 2026; Karunratanakul et al., 2024) extend the setting. Closest to us, TRiDi (Petrov et al., 2025) jointly denoises body, object and contact with modality-specific timesteps. Most closely related to our work, TRiDi (Petrov et al., 2025) jointly denoises the human body, object pose, and interaction representation using modality-specific timesteps. Building on Bao et al. (2023) and the reconstruction-guidance approach of Wang et al. (2026), it guides sampling with the gradient of a contact loss that penalizes the distance between body vertices predicted to be in contact and the object surface.

Across these methods interaction is encoded *globally* at the *whole-body level*: as inferred contact references (Xu et al., 2023), per-marker distances (Diller & Dai, 2024; Peng et al., 2026), binary contact maps (Petrov et al., 2025), separate human and object streams tied by an affordance module (Peng et al., 2025; Wu et al., 2025b; Song et al., 2024), or an object-specific field over poses (Kulkarni et al., 2024). Contact is stored as a binary map over body vertices in TRiDi and CG-HOI (Diller & Dai, 2024), or as proximity between body markers and object surface points (Peng et al., 2026). CompAct instead represents interaction *per body-part*, generating per-part signed InterFields together with body and object pose and an object InterField that stores, at every object point, its signed distance to the body and the owning part; both fields self-guide the sampler.

Compositional Synthesis: Diffusion models compose either scores, obtained from several conditions so that one sample satisfies all of them (Liu et al., 2022; Du et al., 2023), or space, generating an image as overlapping crops or graph factors that are averaged or summed where they overlap (Bar-Tal et al., 2023; Zhang et al., 2023b; Lee et al., 2023). In both cases the factors are *global concepts* and none is responsible for a part of the output; energy-based variants extend score composition to products, mixtures and negations and correct the sample with a few MCMC steps (Du et al., 2023).

For human motion, there is work that composes actions across time (Athanasiou et al., 2022) or body parts (Athanasiou et al., 2023), reuses a single-person prior for two people (Shafir et al., 2024), adds spatio-temporal and language control (Pinyoanuntapong et al., 2025; Huang et al., 2024), and conditions a part-aware model on per-part text (Li et al., 2026). Our work belongs to the score-composition family, but its factors are interactions grounded on body kinematic chains; per-part weights arise from the model's contact predictions, instead of being fixed scalars. During composition each body-part follows one interaction at a time, the seams between interactions are joint constraints rather than blends of poses, and the an unconditional prior fills in whatever no interaction claims, so the result is one coherent body; adding an interaction adds a branch, not a new network.

Multi-Object and Compositional HOI Synthesis: HIMO (Lv et al., 2024) and ParaHome (Kim et al., 2025) provide multi-object mocap of finite coverage. The work of Peng et al. (2026) zero-pads object slots to a fixed count and refines the body against a predicted contact distance field; Dex2HOI (Pratikaki et al., 2026) denoises each object in its own stream and fuses two streams with cross-attention. Both need multi-object supervision; the work of (Peng et al., 2026) trains on HIMO and ParaHome and Dex2HOI (Pratikaki et al., 2026) learns its cross-stream attention and fusion from two-object HIMO and HUMOTO (Lu et al., 2025a) sequences, and both fix the number of objects in the architecture. HOI-PAGE (Li & Dai, 2026) grounds LLM-planned part affordances on meshes without interaction data, and COINS (Zhao et al., 2022) composes atomic action-object pairs. Instead, CompAct recombines single-object predictions per body-part in one sampling pass, trained on single-object data only and without a fixed object-slot architecture.

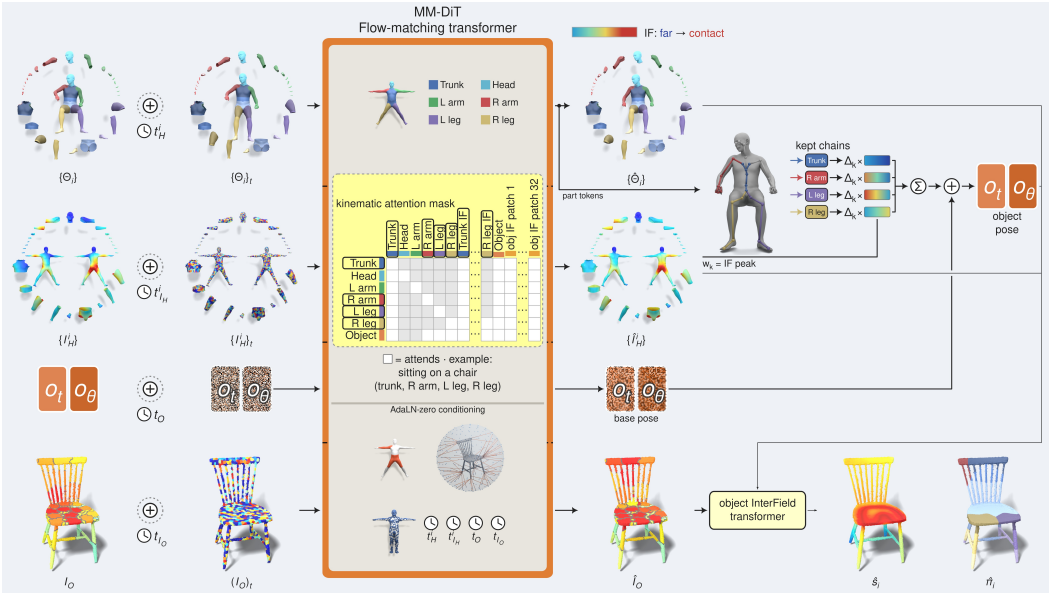

Figure 2: Pose and InterField tokens per body part, an object pose token and object surface-patch tokens, each at its own noise level, are denoised by an MM-DiT Flow-Matching transformer conditioned on the contact mask, body shape, object geometry and noise levels. A kinematic attention mask lets each part attend along its chain and the object only to the engaged chains (here trunk, right arm and both legs). The output gives clean per-part poses and InterFields and an object pose, a base pose plus a residual over the engaged chains weighted by their InterField peaks; the object InterField refiner sharpens the denoised patch-level object InterField into a per-point one and adds, for every point, the body part that owns it. InterField colours run from blue (far) to red (contact).

3 METHOD

CompAct takes as input a body and object shape, and a text prompt from which a language model derives a contact mask naming the body-parts involved. The model generates the pose of the body, the pose of the object and the body-part and object InterFields. In our model every body-part has its own pose, its own InterField and its own noise level. In composition Sec. 3.4 we run CompAct once per object and combine the resulting bodies part by part, giving each body part to the interaction that involves it, so that one body engages with several objects at once.

3.1 DATA REPRESENTATION

Body and object: We represent the body $\mathcal{H}$ by the 6D rotations of the 55 SMPL-X (Pavlakos et al., 2019) joints in a pelvis-centered frame, and group the joints into 26 parts $\Theta_i \in \mathbb{R}^{\mathcal{N}_i \times 6}$ (Fig. 7), six of them per hand. The subject's body shape is given by 1024 points sampled on its T-posed SMPL-X template, which describes body size directly as geometry rather than through shape coefficients. An object is described by 1024 surface points q_i, a BPS code Prokudin et al. (2019) and its pose in the pelvis frame, $\mathcal{X}_O = \{o_\theta, o_t\}$.

InterFields: For every SMPL-X vertex we take the distance to the nearest point of the object mesh and sign it with the outward object normal, so that a vertex outside the object is positive and a penetrating one negative. The sign gives contact, proximity and penetration distinct states. We split these distances into the 26 parts, subsample them by farthest-point sampling and map them to $[0, 1]$:

$$\phi(d) = \tfrac{1}{2}\left(1 + \operatorname{sign}(d)\left(1 - e^{-\omega|d|}\right)\right), \tag{1}$$

where $\omega = 5$. The unsigned term $e^{-\omega|d|}$ saturates at contact and cannot distinguish a penetrating vertex from a touching one. Instead, using Eq. (1) the object surface is placed at $\phi = \frac{1}{2}$, so a single bounded value tells both how close a vertex is and how deep it has penetrated; the values of part i form its InterField $\mathcal{I}^i_{\mathcal{H}}$. Which parts touch the object is stated by a 26-bit contact mask, written by a user or extracted from text by a language model (Sec. C).

The object has an InterField of its own, $\mathcal{I}_O$: for every object point q_i it stores the encoded signed distance s_i to the body and a distribution π_i over the corresponding active body-parts p_i.

3.2 CompAct: Part-wise Flow Matching

Flow matching: We train CompAct with Flow Matching (FM) (Lipman et al., 2023) on straight paths $x_t = (1-t)\,x_0 + t\,x_1$ from a noise sample x_0 to a clean sample x_1. Instead of regressing the velocity along the path, the network f_θ predicts the clean signal x_1 under an L_1 loss, and the velocity that the ODE solver integrates at inference, in 50 Euler steps from t=0 to t=1, is recovered from this prediction:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,\,x_0,\,x_1}\big\|f_\theta(x_t,t) - x_1\big\|_1, \tag{2}$$

$$v_\theta(x_t,t) = \big(f_\theta(x_t,t) - x_t\big)/(1-t) \tag{3}$$

Decoupled timesteps: The whole state $\big\{\{\Theta_i\}_{i=1}^{26}, \{\mathcal{I}_\mathcal{H}^i\}_{i=1}^{26}, \mathcal{X}_O, \mathcal{I}_O\big\}$ is generated jointly, and the only thing that tells a given variable from a generated one is its noise level. Each body-part pose and the object pose carry their own timestep, $t_\mathcal{H}^i$ and $t_\mathcal{O}$, which is zero when the variable is given. During training we noise a random subset of these variables at one shared level, so that a single network learns the joint distribution together with every conditional we use below, from a body posed around a given object to a hand resampled on a fixed body. Regarding the InterFields, they are deterministic functions of their body-part and of the object, so they inherit their clocks from the variables they depend on,

$$t_{\mathcal{I}_\mathcal{H}}^i = \max\big(t_\mathcal{H}^i,\, t_\mathcal{O}\big), \qquad t_{\mathcal{I}_O} = \max\big(t_\mathcal{O},\, \max_i t_\mathcal{H}^i\big). \tag{4}$$

Network: The backbone is an MM-DiT (Peebles & Xie, 2023; Esser et al., 2024) with four token streams: 26 part-pose tokens, 26 InterField tokens, two object-pose tokens for rotation and translation, and 32 object-patch tokens. The patch tokens carry the object InterField. Unlike the body, whose vertices share one fixed topology, objects differ in shape and topology, so their patches need a common ordering. We cluster the 1024 points of an object by position into 32 patches and, as a shared reference, cluster the directions of the BPS basis points once into 32 sectors. Each patch is mapped to a sector by the direction of its centroid, one patch per sector, so the patches of every object are ordered by the same 32 directions: the s-th patch token of every object holds the patch that faces the s-th direction, and a learned embedding per patch position can tell the network which side of the object it describes.

Each patch token encodes the noised InterField values of its points together with the patch centroid and normal, and a linear decoder shared across patches returns its clean values, a coarse object InterField. Attention follows the SMPL-X kinematic tree, a trunk (pelvis, back) and five chains (head, arms, legs) (see Fig. 2). Every token sees the trunk and the object; chains never see each other, and the object sees only the chains that its contact mask marks. This keeps each conditional local, so the arm that holds a phone is not shaped by what the legs happen to do. With probability 0.15 we drop object and contact mask together, which gives the unconditional body prior that composition rests on.

Outputs: The network returns the clean body-part poses and InterFields directly, while the object pose is anchored on the contacting body parts. This means that after the two object tokens decode a base pose, every body-part on a kinematic chain marked by the contact mask adds a correction, weighted by the maximum value of its predicted InterField, so that the parts in contact place the object. A small transformer (Vaswani et al., 2017) head then predicts $\mathcal{I}_O$ for all 1024 sampled points of the object. Each point q_i, with its position and the coarse value of its patch, attends to the clean body-part and object tokens, and the head returns the signed distance $\hat{s}_i$ from q_i to the body and a distribution $\hat{\pi}_i$ over the body-parts, the probability that q_i touches each of them. Together, $\hat{s}_i$ and $\hat{\pi}_i$ over all points form the predicted object InterField $\hat{\mathcal{I}}_O$. uring sampling we employ it as the clean estimate of the object InterField in Eq. (3), instead of the coarse output of the patch tokens. The head is differentiable in q_i and trained with an eikonal term, so that $q_i - \hat{s}_i \nabla_q \hat{s}_i$ lands on the body, which later lets us pull an object to the body point it should touch. Details and losses are given in Secs. A, A.2 and A.3.

3.3 SELF-INTERFIELD GUIDANCE

At every step the model returns a clean pose $\hat{x}_1$ and clean InterFields $\hat{\mathcal{I}}_{\mathcal{H}}$, and the two need not agree. An InterField can predict a contact on a hand while the pose has not yet brought the arm there, since moving an arm takes more steps than marking a contact. The InterFields therefore settle first, and we use them to guide the pose. Let $\hat{d}_{i,j}$ be the signed distance of sampled vertex j of part i to the posed object, and $d^*_{i,j}$ the distance that the model's own value $\hat{\mathcal{I}}^{i,j}_{\mathcal{H}}$ implies, obtained by inverting Eq. (1). Over the $\mathcal{V}_i$ vertices of every masked part we minimize

$$\mathcal{L}_{\text{self}}(\hat{x}_1) = \sum_i \frac{1}{\mathcal{V}_i} \sum_j \kappa_{i,j} \left(\hat{d}_{i,j} - d^*_{i,j}\right)^2, \qquad \kappa_{i,j} = 1 - \left|2\hat{\mathcal{I}}^{i,j}_{\mathcal{H}} - 1\right|, \tag{5}$$

where the difference between the two distances is weighted by the predicted InterField itself. Since the InterField encodes a signed distance see Eq. (1), with the surface at $\frac{1}{2}$ and far vertices at either end, its value cannot serve as a weight directly; the value $\kappa_{i,j}$ is 1 for a vertex the InterField predicts in contact and 0 for a vertex it predicts far from the object, outside or inside. The predicted InterField thus sets both where each vertex should be and how much it counts, so the vertices expected in contact pull the pose and the rest are left alone. The correction $\hat{x}_1 \leftarrow \hat{x}_1 - \eta \nabla_{\hat{x}_1} \mathcal{L}_{\text{self}}$ ($\eta = 20$) flows back through the SMPL-X kinematic chain and is applied before Eq. (3). It needs no oracle field and no external contact classifier, only what the model has already predicted (Sec. A).

The object is pulled similarly. For every object point q_i in the predicted pose, the head returns $\hat{s}_i$ and $\hat{p}_i$ against the predicted body, and $c_i = q_i - \hat{s}_i \nabla_q \hat{s}_i$ is the body point q_i corresponds to, snapped to the nearest vertex of part $\hat{p}_i$. Each point gets a weight $w_i = e^{-4|\hat{s}_i|/\text{diag}}$, with "diag" the object's bounding-box diagonal, which is 1 on the body and decays with distance. Over nearby points, $w_i > 0.5$, whose part $\hat{p}_i$ is in the contact mask and whose ownership is confident, $\max_k \hat{\pi}_{i,k} \geq 0.6$, we minimise:

$$\mathcal{L}_{\text{obj}}(o_\theta, o_t) = \frac{\sum_i w_i \left\| q_i(o_\theta, o_t) - c_i \right\|^2}{\sum_i w_i}. \tag{6}$$

We keep c_i detached and correct the clean object pose estimate along $-\nabla_{(o_\theta, o_t)} \mathcal{L}_{\text{obj}}$. While self-InterField guidance moves the body towards the object, this term moves the object towards the parts that own it, with CompAct-predicted correspondences; each branch guides its own object.

3.4 COMPOSITIONAL INFERENCE AND EDITING

Everything so far concerns one body and one object. To place several objects around one body we do not train a new model; we run the same model once per object and combine its predictions at inference (Fig. 1). For $\mathcal{M}$ interactions we keep one shared noisy body of one shape and denoise it $\mathcal{M} + 1$ times at every step, in one batch: once without any object, which gives the unconditional estimate $\hat{x}^u$, and once per interaction k with its own object and contact mask, which gives $\hat{x}^k$. We call each of these $\mathcal{M} + 1$ runs a branch. Let $p(x)$ be the distribution of bodies without any object, which the model estimates with $\hat{x}^u$, and $p(x \mid c_k)$ the distribution of bodies under interaction k, which it estimates with $\hat{x}^k$. Liu et al. (2022) showed for images that a sample satisfying several conditions at once can be drawn from $p(x) \prod_k \left[p(x \mid c_k)/p(x)\right]^{w^k}$; we apply the same idea to bodies and interactions. Each factor $p(x \mid c_k)/p(x)$ is what interaction k changes in the body, and w^k sets how strongly that change is enforced. In score form the product becomes a weighted sum of scores, and on our straight paths the clean estimate is an affine function of the score, so the same weighted sum holds for the clean estimates (Sec. B),

$$\hat{x}_1 = \hat{x}^u + \textstyle\sum_k w^k \left(\hat{x}^k - \hat{x}^u\right) \tag{7}$$

Claims from the masks, weights from the model: Liu et al. (2022) use one fixed weight per condition for the whole image. Instead, we give every body-part its own weight for every interaction, w^k_i, and read it at every step from the InterFields the model predicts, so an interaction weighs on a part as much as its InterField says the part is involved. A branch is allowed to act on a part if its contact mask marks any part of the chain that part belongs to (trunk, head, an arm or a leg): a phone held in the fingers thus claims the whole arm, since the arm must move to bring the fingers to the phone. We denote the branches allowed to act on part i with $\mathcal{C}_i$. For those branches the weight

is the maximum of the branch's own predicted InterField over the part, in the form κ of Eq. (5), $w_i^k = \max_j \kappa_{i,j}^k$. The weights are normalised over the branches acting on the part,

$$s_i^k = \frac{w_i^k}{\sum_{\ell \in \mathcal{C}_i} w_i^\ell}, \qquad \hat{x}_i = \begin{cases} \sum_{k \in \mathcal{C}_i} s_i^k \, \hat{x}_i^k, & \mathcal{C}_i \neq \emptyset, \\ \hat{x}_i^u, & \mathcal{C}_i = \emptyset, \end{cases} \tag{8}$$

This is Eq. (7) for part i with $w^k = s_i^k$; because the shares sum to one, a claimed part is a mix of the branches only, with no share left for the unconditional branch. A chain claimed by one branch follows it entirely, so an arm carrying an object follows its branch even though contact fires only on the fingers. A chain claimed by several branches, say the trunk by a sofa and by a phone at the ear, settles between their estimates, closer to the branch that predicts more contact. A part no branch claims keeps the unconditional estimate. As the weights are the model's own contact predictions, recomputed at every step, they follow the contacts as they form, with nothing to tune.

Shared body, separate objects: After every step the composed body is written back to all branches, so they keep denoising one body, without drifting apart. However, the objects are not mixed. Each branch predicts its object pose relative to the pelvis of its own estimate $\hat{x}^k$, and we re-express that pose in the frame of the composed pelvis, so the object stays where the branch placed it on the body.

Object freeze: Each branch denoises a body that carries interactions it was never trained with. Early in sampling this costs little, as x_t is still mostly noise. In final steps, however, x_t is nearly the composed clean body, which lies off every branch's single-object distribution, and each branch pulls the shared trunk and pelvis back towards its own mode; objects drift and contacts break. Thus, for more challenging compositions with more than 3 objects we freeze all objects at step 40 of 50, w.r.t. the composed pelvis, and keep guiding the body towards them as fixed targets. InterField guidance is essential until the end so that the body preserves the already established contacts with the objects.

Interaction editing: Editing comes for free from the same per-part conditionals. The parts to preserve are held clean and the rest is resampled, with no edit-specific training; see Fig. 8 in Sup. Mat.

4 Experiments

Datasets and protocol: We train CompAct on seven single-object datasets: GRAB Taheri et al. (2020), its mirror, BEHAVE Bhatnagar et al. (2022), InterCap Huang et al. (2022), OMOMO Li et al. (2023), CHAIRS Jiang et al. (2023), and NeuralDome Zhang et al. (2023a). We use official test splits, otherwise the InterAct Xu et al. (2025) split. For comparison, CompAct and TRiDi Petrov et al. (2025) use TRiDi's exact GRAB +BEHAVE subject split and best-of-three evaluation over three runs, taking the minimum per-frame distance and maximum accuracy.

For multi-object evaluation, baselines train on ParaHome Kim et al. (2025); all methods are tested on unseen HIMO Lv et al. (2024). At test time, an LLM skill derives body-part contact masks from text. In Sup. Mat. we show results with GT masks used as oracle.

Baselines: We use TRiDi's Petrov et al. (2025) code to build two baselines. For a fair comparison, sTRiDi is TRiDi trained with SMPL-X bodies in our pelvis-centred frame conditioned on body beta shapes. For multi-object interaction, mTRiDi applies the same changes and is trained with parallel object slots on the multi-object training data of ParaHome. Finally, mTRiDi +IK refines the generated body with inverse kinematics under a DPoser-X body prior Lu et al. (2025b). Note that, to favour this baseline, we retrain DPoser-X with ParaHome and HIMO as additional datasets, so it has seen the evaluation domains, whereas CompAct never observes a multi-object frame.

Metrics: We evaluate under three conditionals: body and interaction given the object $(H, I \mid O)$, object and interaction given the body $(O, I \mid H)$, and joint generation (H, O, I). For consistency with the ground truth, $(H, I \mid O)$ uses root-aligned MPJPE, PA-MPJPE, and body contact accuracy Acc_{sbj}; $(O, I \mid H)$ uses object vertex-to-vertex error E_{v2v}, centre error E_c, and Acc_{obj}; joint generation (H, O, I) uses all six. Distributional quality uses 1-NNA (optimal at 50%), coverage (COV), minimum matching distance (MMD), and sample distance to the training set (SD). Contact precision, recall, and F1 use a 5 cm rule on 690 decimated body vertices. For the multi-object evaluation we add the three contact metrics (Grady et al., 2021; Hasson et al., 2019) computed on body-parts that the contact mask names per object: the contact ratio (CR), the fraction of parts whose nearest vertex lies within 5 mm of the object; the contact distance (Gap), their mean nearest distance to the object; and the penetration depth (PD), the mean depth of body vertices inside the object.

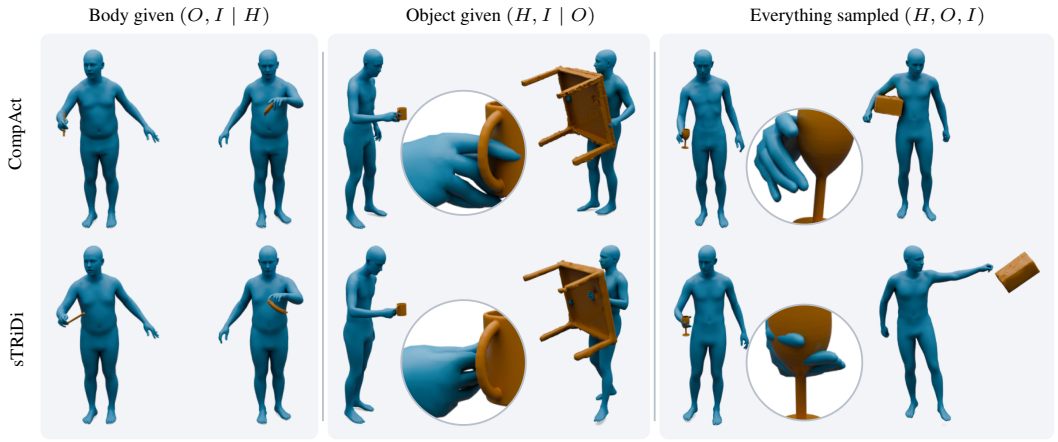

Figure 3: Single-object comparison: CompAct vs sTRiDi under three conditionals.

(a) $(H, I \mid O)$**, GRAB +BEHAVE**

	GRAB			BEHAVE		
Method	MPJPE↓	PA↓	Acc_{sbj} ↑	MPJPE↓	PA↓	Acc_{sbj} ↑
sTRiDi	10.14	7.14	98.57	14.27	8.11	96.57
CompAct	**8.31**	**6.41**	**98.95**	**11.90**	**8.02**	**97.47**

(b) $(O, I \mid H)$**, GRAB +BEHAVE**

	GRAB			BEHAVE		
Method	E_{v2v} ↓	E_c ↓	Acc_{obj} ↑	E_{v2v} ↓	E_c ↓	Acc_{obj} ↑
sTRiDi	7.11	5.26	**99.03**	28.27	14.75	95.75
CompAct	**5.30**	**3.51**	98.97	**23.01**	**11.19**	**96.04**

(c) (H, O, I)**, GRAB**

Method	MPJPE↓	PA↓	E_{v2v} ↓	E_c ↓	Acc_{sbj} ↑	Acc_{obj} ↑
sTRiDi	17.39	12.07	30.96	30.39	96.96	96.95
CompAct	**16.09**	**10.45**	**25.56**	**25.00**	**97.17**	**97.11**

(d) (H, O, I)**, BEHAVE**

Method	MPJPE↓	PA↓	E_{v2v} ↓	E_c ↓	Acc_{sbj} ↑	Acc_{obj} ↑
sTRiDi	20.42	**10.79**	45.35	36.18	94.44	93.67
CompAct	**18.49**	11.02	**34.96**	**25.11**	**95.48**	**94.46**

Table 1: Single-object comparison against sTRiDi. Both methods use guidance; CompAct uses LLM-derived contact masks. Errors: cm; accuracies: %. Bold: best per column.

4.1 Quantitative evaluation

Single-object evaluation: Table 1 compares CompAct with sTRiDi under the three conditionals; both use guidance, and CompAct uses the LLM-derived test-time masks. Across both datasets, CompAct lowers 15 of 16 error cells and raises 7 of 8 contact accuracies. With the object given, MPJPE decreases from 10.14 to 8.31 cm on GRAB and from 14.27 to 11.90 cm on BEHAVE; under joint generation, the object vertex error decreases from 30.96 to 25.56 cm on GRAB and from 45.35 to 34.96 cm on BEHAVE. Distributional quality is reported in the Sup. Mat..

Multi-object evaluation: We evaluate composition on the HIMO test frames with the object poses fixed (Tab. 2) and under all three conditionals in the Sup. Mat.

4.2 Qualitative evaluation

Figure 3 compares CompAct and sTRiDi for single-object samples under the three conditionals. Figure 4 shows CompAct-composed multi-object scenes with objects drawn from all seven training datasets and no multi-object training. Figure 5 compares the composed scenes with mTRiDi on HIMO frames, and Fig. 6 compares them with mTRiDi +IK. In both, CompAct produces more natural poses and better grasps, although it has not been trained on multi-object data and, unlike mTRiDi +IK, has no access to a DPoser-X prior fitted to the HIMO body distribution.

4.3 Ablations and perceptual study

We ablate body-part granularity (1, 4, 8, 16 and 26 parts), self-InterField guidance against no guidance, and InterFields against binary contact targets; the table and comments are in Sec. D. Body error saturates at 8 parts, whereas contact realisation keeps improving up to 26, and every removed component costs contact rather than body accuracy. We conduct two perceptual studies. In the first, 20 participants viewed 40 interactions generated by CompAct and mTRiDi from the same objects



Figure 4: CompAct multi-object compositions assembled from conditionals over several objects.

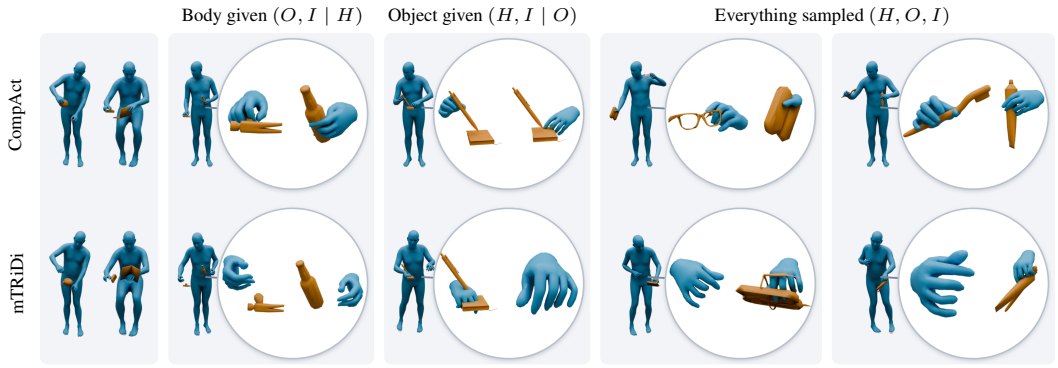

Figure 5: Multi-object comparison with mTRiDi on HIMO Lv et al. (2024) under three conditionals; each scene shows the body and hand–object contact close-ups.

Method	MPJPE↓	PA↓	Gap↓	CR↑	PD↓	F1↑	Acc_{sbj} ↑
mTRiDi	138.7	101.0	29.3	0.45	**12.3**	56.0	97.9
mTRiDi +IK	120.6	89.5	**3.7**	**0.87**	26.2	68.3	98.0
CompAct	**97.5**	**67.0**	15.3	0.46	23.9	**69.6**	**98.5**

Table 2: Multi-object evaluation on the HIMO dataset with fixed object poses. mTRiDi uses the best of three samples, CompAct a single sample. Errors, Gap and PD are mm, CR is a fraction, F1 and Acc_{sbj} are %. Best per column in bold.

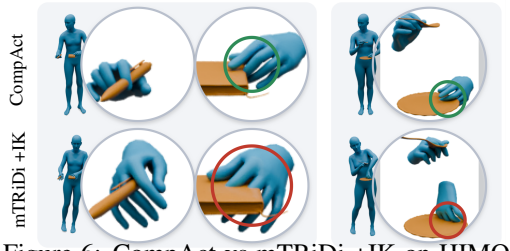

Figure 6: CompAct vs mTRiDi +IK on HIMO with fixed objects. IK closes the contact gap but often leads to penetrations.

and body shape, shown side by side in random order, and chose the one that looks more natural and physically plausible. In a separate pass they judged the hands alone, from close-ups of each hand-object contact. CompAct was preferred in 82.2% of the whole-body comparisons and 85.1% of the hand comparisons. The second study repeats the protocol on a different set of examples against mTRiDi +IK, where CompAct was preferred in 63.3% of the whole-body and 57.1% of the hand comparisons; the protocol is in Sec. F. Although mTRiDi +IK is expected to outperform mTRiDi, since it refines the body by optimisation under a privileged pose prior, CompAct still surpasses it without any training on multi-object data.

5 Conclusion

COMPACT generates static human–object interactions by coupling body-part pose variables, per-part InterFields, and object pose within one flow model. The experiments show that this representation supports all three single-object conditionals and composes single-object branches into multi-object scenes without multi-object training. The object InterField provides an object-side description of contact with which we drive the interaction geometrically, and facilitates the generation process in more challenging scenarios, such as the composition. Future work will extend the representation to dynamic interactions and denser object arrangements. Code and trained models will be released.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text and for helping to prepare the figures (rendering and layout scripts, and iterating on figure composition). We have not used generative AI tools for research ideation, for designing or running the experiments, or for producing any of the reported results. All AI-assisted text was reviewed and edited by the authors, and all AI-assisted figure code was checked by the authors against the underlying data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. Teach: Temporal action composition for 3D humans. In *International Conference on 3D Vision (3DV)*, pp. 414–423, 2022.

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *International Conference on Computer Vision (ICCV)*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning (ICML)*, pp. 1692–1717, Honolulu, HI, USA, 2023. PMLR.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning (ICML)*, 2023.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 15914–15925, New Orleans, LA, USA, 2022. IEEE.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 19888–19901, 2024.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Narain Sohl-Dickstein, A. Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning (ICML)*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, pp. 12606–12633, 2024.

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 *Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954, 2023.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. *Computer Graphics Forum (CGF)*, 42(2):1–12, 2023.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021.

Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 11799–11808, 2019.

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. CoMo: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision (ECCV)*, pp. 180–196, 2024.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pp. 281–299, 2022.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *International Conference on Computer Vision (ICCV)*, Paris, France, 2023. IEEE.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3D generative modeling of human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1816–1828, Nashville, TN, USA, 2025. IEEE.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David F. Fouhey, and Leonidas J. Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion: Part-level human motion generation and composition. In *Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *Transactions on Graphics (TOG)*, 42, 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*, pp. 54–72, 2024a.

Lei Li and Angela Dai. Genzi: Zero-shot 3D human-scene interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

Lei Li and Angela Dai. HOI-PAGE: Zero-shot human-object interaction generation with part affordance guidance. In *International Conference on Machine Learning (ICML)*. PMLR, 2026.

Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 3023–3032. IEEE, 2024b.

Shujia Li, Haiyu Zhang, Xinyuan Chen, Yaohui Wang, and Yutong Ban. GenHOI: Generalizing text-driven 4D human-object interaction synthesis for unseen objects, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D dataset of mocap human object interactions. In *International Conference on Computer Vision (ICCV)*, pp. 10886–10897, 2025a.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, and Ziwei Liu. DPoser-X: Diffusion model as robust 3D whole-body human pose prior. In *International Conference on Computer Vision (ICCV)*, pp. 9988–9997, 2025b.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision (ECCV)*, pp. 300–318, Milan, Italy, 2024. Springer.

Georgios Paschalidis, Romana Wilschut, Dimitrije Antić, Omid Taheri, and Dimitrios Tzionas. 3D Whole-body grasp synthesis with directional controllability. In *International Conference on 3D Vision (3DV)*, pp. 1497–1517, Singapore, 2025. IEEE.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

William S. Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

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 *Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2025.

Xiaogang Peng, Zeyu Han, Zichong Meng, Yiming Xie, Jihua Zhu, Gang Hua, and Huaizu Jiang. Surface keypoint representation for multi-object and articulated human-object interaction generation, 2026.

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Tridi: Trilateral diffusion of 3D humans, objects, and interactions. In *International Conference on Computer Vision (ICCV)*, Honolulu, HI, USA, 2025. IEEE.

Huaijin Pi, Zhi Cen, Zhiyang Dou, and Taku Komura. CoDA: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and S. Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2025.

Chrysa Pratikaki, Pablo Ruiz-Ponce, Jiankang Deng, Stefanos P. Zafeiriou, and Rolandos Alexandros Potamias. Dex2HOI: Dexterous bimanual two-object interaction generation, 2026.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient Learning on Point Clouds With Basis Point Sets. In *International Conference on Computer Vision (ICCV)*, 2019.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. HOIDiNi: Human-object interaction through diffusion noise optimization. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2026.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 811–820, 2024.

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 (ECCV)*, pp. 581–600, 2020.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 13263–13273, 2022.

Purva Tendulkar, Dídac Surís, and Carl Vondrick. FLEX: Full-body grasping without full-body grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 21179–21189, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Ziyin Wang, Sirui Xu, Chuan Guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Unleashing guidance without classifiers for human-object interaction animation. In *International Conference on Learning Representations (ICLR)*, 2026.

Lin Wu, Zhixiang Chen, and Jianglin Lan. Hoi-dyn: Learning interaction dynamics for human-object motion diffusion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025a.

Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. THOR: Text to human-object interaction diffusion via relation intervention. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–7, 2025b.

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 (ECCV)*, volume 13666, pp. 257–274, 2022.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In *International Conference on Computer Vision (ICCV)*, pp. 14882–14894, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3D dynamic human-object interaction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060, 2025.

Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 22714–22723, 2025.

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 12358–12369, 2025.

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *Transactions on Graphics (TOG)*, 40(4):121:1–121:14, 2021.

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. Interactanything: Zero-shot human object interaction synthesis via llm feedback and object affordance parsing. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 7015–7025, 2025.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023a.

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10188–10198, 2023b.

Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. COINS: Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*, pp. 311–327, 2022.

APPENDIX

A IMPLEMENTATION DETAILS

Network: CompAct is an MM-DiT with 8 layers of hidden width $\mathcal{W} = 256$, 8 attention heads and an MLP expansion ratio of 4, over four token streams: 26 part-pose tokens, 26 InterField tokens, two object-pose tokens for rotation and translation, and 32 object-patch tokens that carry the coarse object InterField; the 26 body-parts are shown in Fig. 7. The object BPS code and the 1024 Tpose points of the subject are each distilled into 64 read-only key-value tokens by a 4-head cross-attention distiller. The contact mask enters as a learned per-part embedding of three states (inactive, active, unconditional) added to the part's timestep embedding. With probability 0.15 the body-side queries read learned null object tokens and the mask takes its unconditional state, which yields the unconditional body prior. Each training row noises a random subset of the 26 parts and the object, of size uniform in 1 to 27, at one shared level $t \in \{1, \dots, 999\}$ of 1000.

A.1 OBJECT INTERFIELD HEAD

The head refines the patch stream's clean estimate at every step. Its queries are the 1024 object points posed by the predicted pose (o_θ, o_t), each encoded by 57 Fourier features of its position (3 coordinates and 9 sine and cosine bands per axis) concatenated with the patch value at that point, lifted to 32 values, and projected to width 256. The keys are 27 state tokens, one per body-part with its predicted rotation, InterField and a part embedding, and one with the object pose. Two self-attention blocks mix the state tokens and two cross-attention blocks let the queries read them, each 256 wide with 4 heads. Per query the head emits a signed distance $\hat{s}_i$ to the body, encoded as in Eq. (1) with the factor $4/\text{diag}$ of the object's bounding-box diagonal, and 26 ownership logits with softmax $\hat{\pi}_i$ and mode $\hat{p}_i$. The encoded distances replace the stream's clean estimate $\hat{v}^{\text{base}}$ in the solver. The head is trained jointly from scratch, with undetached inputs, by an L_1 on the encoded distance, uniform and closeness-weighted (weight 1), an L_1 on the metric distance truncated at 0.3 m (0.1), a cross-entropy on the ownership of points within 5 cm of the body (0.5), and an eikonal term driving $\|\nabla_q \hat{s}_i\|$ to one on 32 queries per row at least 5 mm off the surface (0.01).

At inference the correspondence $c_i = q_i - \hat{s}_i \nabla_q \hat{s}_i$, snapped to part $\hat{p}_i$ of the body, is kept when $e^{-4|\hat{s}_i|/\text{diag}} > 0.5$, $\hat{p}_i$ is in the mask and $\max_k \hat{\pi}_{i,k} \geq 0.6$, and Eq. (6) is minimised with c_i detached.

A.2 TRAINING OBJECTIVE

The 26 pose, 26 body-InterField and two object-pose components are supervised by the x_1 objective of Eq. (2), extended to the component-wise clocks,

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{t},\, \mathbf{x}_0,\, \mathbf{x}_1} \sum_{k \in \mathcal{K}} \left\| f_\theta^{(k)}(\mathbf{x}_\mathbf{t}, \mathbf{t}) - x_1^{(k)} \right\|_1, \tag{9}$$

each with unit weight over the rows in which its component is noised. The patch stream has no flow-matching term; its base decode is supervised directly against the encoded ground-truth field v, uniform and closeness-weighted,

$$\mathcal{L}_{\text{base}} = \frac{1}{N} \sum_i \left| \hat{v}_i^{\text{base}} - v_i \right| + \frac{\sum_i \kappa_i \left| \hat{v}_i^{\text{base}} - v_i \right|}{\sum_i \kappa_i}, \qquad \kappa_i = 1 - |2v_i - 1|, \tag{10}$$

over the $N = 1024$ object points; without it the base collapses to a constant field. $\mathcal{L}_{\text{v2v-kin}}$ is the mean squared error between the forward-kinematic vertices of the predicted and the ground-truth pose, and $\mathcal{L}_{\text{hot-v2v-kin}}$ the same error weighted per vertex by its ground-truth closeness $e^{-\omega |d_j^{\text{gt}}|}$ to the object; $\mathcal{L}_{\text{obj}}$ is the mean displacement of the object points under the predicted and the ground-truth pose. Four contact terms compare distances with the ground truth: $\mathcal{L}_{\text{IF}}$ the signed distances of the predicted body to the ground-truth object, $\mathcal{L}_{O \to H}$ the nearest distances of the predicted-posed object points to the ground-truth body, $\mathcal{L}_{\text{seat}}$ the predicted body against the detached predicted object under the ground-truth contact pattern, and $\mathcal{L}_{\text{con}}$ the closeness realised by the predicted body and object against the model's own detached InterFields, signed by the object normal within 5 cm. The head terms form $\mathcal{L}_{\mathcal{I}_O}$. The total objective is $\mathcal{L} = \mathcal{L}_{\text{FM}} + \sum_j \lambda_j \mathcal{L}_j$ over these terms, with $\lambda_{\text{kin}} = \lambda_{\text{hot}} = 500$, $\lambda_O = 25$, $\lambda_{\text{IF}} = \lambda_{\text{seat}} = \lambda_{\text{con}} = 2$, $\lambda_{O \to H} = 0.5$ and $\lambda_{\mathcal{I}_O} = \lambda_{\text{base}} = 1$.

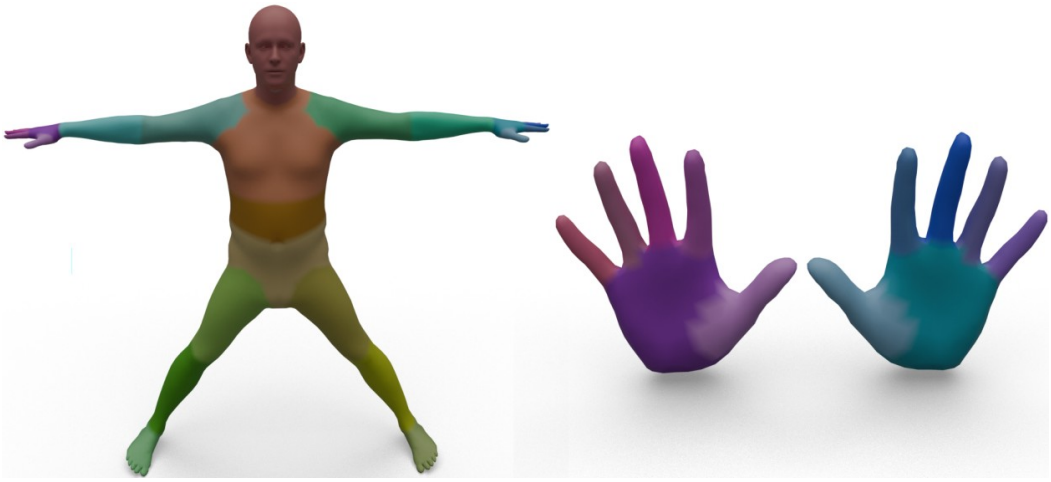

Figure 7: The 26 body parts of CompAct. Fourteen belong to the body: head, upper back, lower back and pelvis, and per side an upper leg, a lower leg, a foot, an upper arm and a lower arm. Twelve belong to the hands: per hand a palm and the thumb, index, middle, ring and little fingers, so half of the parts sit in the hands, where most contacts happen. Each part is one token of the diffusion model with its own noise clock, its own InterField and its own bit in the contact mask.

A.3 Loss scheduling

The body-side auxiliary losses apply to every noised component over the whole noise range, never to clean components, and the object and head losses to every row. The one exception is the hands: their geometric terms apply only where the part's own clock is at most 500 of 1000, because otherwise these terms shape the fingers from pure noise and the hands close during sampling.

A.4 Predicting the clean signal

The network regresses x_1 and recovers the velocity by Eq. (3). A velocity loss weights clean-sample errors by $(1-t)^2$, so the late steps that resolve end effectors and contacts get almost no gradient, whereas the x_1 loss is stable across our heterogeneous state of chain rotations and bounded InterField values. Guidance and composition act on clean estimates, the network's native output.

Optimisation and inference: We train with AdamW Loshchilov & Hutter (2019), learning rate 10^{-5} for the denoiser and 5×10^{-4} for the object-InterField head, weight decay 10^{-2}, 5000 warmup steps, a $300{,}000$-step cosine schedule, a global batch of 6144, mixed precision and gradient-norm clipping at 5. Inference uses 50 Euler steps. Self-InterField guidance uses $\eta = 20$, over the final fifth of the schedule for a single object and over all steps in compositions; object-InterField guidance uses weight 0.5 and ownership confidence 0.6.

B Composition derivation

Let $c_1, \dots, c_{\mathcal{M}}$ be the interaction conditions, each an object with its contact mask. If they are conditionally independent given the body, $p(c_1, \dots, c_{\mathcal{M}} \mid x) = \prod_k p(c_k \mid x)$, Bayes' rule applied to each factor gives the composed density of composable diffusion Liu et al. (2022), $p(x \mid c_{1:\mathcal{M}}) \propto p(x) \prod_k \big[p(x \mid c_k)/p(x)\big]^{w^k}$, with exponents $w^k \in [0,1]$ that relax the conjunction. Taking $\nabla_x \log$ turns the product into a sum,

$$\nabla_x \log p(x \mid c_{1:\mathcal{M}}) = \nabla_x \log p(x) + \textstyle\sum_k w^k \big[\nabla_x \log p(x \mid c_k) - \nabla_x \log p(x)\big]. \tag{11}$$

Each bracket is the direction classifier-free guidance adds for one condition. On the straight path $x_t = (1-t)\, x_0 + t\, x_1$ with $x_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ the kernel is $x_t \mid x_1 \sim \mathcal{N}\big(t\, x_1, (1-t)^2 \mathbf{I}\big)$, so the posterior mean of the clean sample is:

$$\mathbb{E}[x_1 \mid x_t] = \frac{x_t + (1-t)^2\, \nabla_x \log p_t(x_t)}{t}, \tag{12}$$

	Single-object HOI					Multi-object HOI (HIMO)			
Variant	MPJPE↓	PA↓	Gap↓	CR↑	F1↑	MPJPE↓	Gap↓	CR↑	F1↑
Body-part granularity									
1 part, no contact mask	146.9	112.8	73.8	45.8	60.4	121.0	60.8	0.19	45.1
4 parts	143.0	113.5	55.3	63.6	67.4	224.5	165.9	0.08	27.4
8 parts	118.6	96.1	**9.4**	85.4	76.1	**102.1**	21.4	0.36	62.8
16 parts	**118.4**	91.9	13.4	66.0	73.8	119.4	32.3	0.25	56.8
Guidance									
No guidance	123.3	92.5	23.5	56.5	66.5	124.9	54.0	0.12	37.7
Interaction representation									
Binary contacts	130.0	101.8	17.3	74.0	67.8	109.1	44.2	0.23	50.6
CompAct	119.9	**91.5**	10.1	**96.0**	**76.4**	109.3	**18.7**	**0.44**	**66.2**

Table 3: Full ablation matrix. Each row changes one component of the base-supervised recipe; CompAct (last row, 26 parts, guided) is the reference. All arms are evaluated at their last checkpoint above 100k steps on the official test splits of the seven single-object datasets (123,510 frames, LLM-derived contact masks, body given the objects, guided unless stated; left block, frame-weighted over the seven splits) and on the HIMO composition evaluation with the objects fixed. MPJPE, MPJPE-PA (PA) and the contact gap (Gap) are in mm; CR is the contact ratio, in % of its ceiling for single-object and as a fraction for HIMO; F1 is the body contact F1 in %. The 1-part variant collapses the body into one denoising token and also drops the contact mask. Best per column in bold.

an affine function of the score whose slope $(1-t)^2/t$ and offset x_t/t depend on the noise level alone. All branches share x_t and t, so applying Eq. (12) to Eq. (11) keeps the offset once and cancels the slope in every difference, $\hat{x} = \hat{x}^u + \sum_k w^k(\hat{x}^k - \hat{x}^u)$, which is Eq. (7). No score is formed: the combination applies directly to the network's output. With the shares of Eq. (8) a claimed part is a convex combination of its claiming branches and an unclaimed part keeps $\hat{x}_i^u$. One step with $\mathcal{M}+1$ batched branches runs as follows. (1) One batched denoiser call on the shared body state returns $\hat{x}^u$, the $\hat{x}^k$, and each branch's object estimate and predicted InterFields. (2) The contact masks give the claiming set $\mathcal{C}_i$ of every part, and each claiming branch's predicted InterField gives its weight w_i^k. (3) The shares compose the body estimate by Eq. (8). (4) Each branch's object is carried from the pelvis of $\hat{x}^k$ to the composed pelvis; objects are never mixed. (5) Self-InterField guidance runs inside each branch against the composed body, on the chains that branch's mask touches, and the corrections are summed; object-InterField guidance, where enabled, moves only that branch's object. (6) Four-object scenes freeze every object at step 40 relative to the composed pelvis; guidance continues. (7) The composed body is written back into all branches, and the state advances with $x_{t+h} = x_t + h\,(\hat{x} - x_t)/(1-t)$.

C Contact masks from text

When the 26-part mask is derived from a text description, it is produced by an LLM that sees the prompt and nothing else, following a fixed instruction sheet. The sheet names the 26 parts, explains in physical terms which parts a described interaction puts on the object, and asks for a JSON list of part names per object; it carries no dataset statistics or test examples and forbids tools. Prompts are presented one at a time, shuffled, without the frame or the mesh. When no part is within 5 mm, the mask is set to the single closest part, and the sheet never returns an empty list. Against the ground-truth masks of the TRiDi test prompts the LLM masks reach a framewise F1 of about 0.80, agree on the body side in contact for 93% of frames and reproduce the full mask for about 27%.

D Ablations

Each ablation changes one factor of the full model and keeps the data and the learning-rate schedule (Tab. 3). **Body-part granularity:** The body is partitioned into 16, 8, 4 or 1 parts, unions of the 26: at 16 each hand is one part; at 8 the head, trunk, legs, arms and hands; at 4 head with trunk, legs, arms and hands; at 1 the whole body. Each part has one pose token, one InterField token, one

GRAB

Cond.	Method	1-NNA→ .5	COV↑	MMD↓	SD↓
$H, I \mid O$	sTRiDi Petrov et al. (2025)	.813	.433	1.102	**.526**
	CompAct	**.802**	**.477**	**1.035**	.652
$O, I \mid H$	sTRiDi Petrov et al. (2025)	.676	.475	**.259**	**.075**
	CompAct	**.665**	**.480**	.283	.080
H, O, I	sTRiDi Petrov et al. (2025)	.889	.322	**3.070**	**1.282**
	CompAct	**.887**	**.397**	3.381	1.600

BEHAVE

Cond.	Method	1-NNA→ .5	COV↑	MMD↓	SD↓
$H, I \mid O$	sTRiDi Petrov et al. (2025)	.660	.491	**1.298**	**.741**
	CompAct	**.616**	**.507**	1.769	1.059
$O, I \mid H$	sTRiDi Petrov et al. (2025)	.648	.511	**.152**	**.057**
	CompAct	**.636**	**.525**	.166	.072
H, O, I	sTRiDi Petrov et al. (2025)	.728	.445	**2.838**	**1.539**
	CompAct	**.717**	**.469**	4.043	2.274

Table 4: Generation and contact quality under guidance on TRiDi's feature scale (1-NNA is best at 0.5). CompAct uses the LLM-derived contact masks of Sec. C; sTRiDi is TRiDi retrained on our data. Best per column in bold.

Cond.	Method	MPJPE↓	PA↓	E_{v2v} ↓	E_c ↓	Gap↓	CR↑	PD↓	F1↑
$(O, I \mid H)$	mTRiDi (best of 3)	–	–	145.2	118.0	61.0	0.26	**12.7**	–
	CompAct	–	–	**111.6**	**88.9**	**11.0**	**0.61**	23.5	–
$(H, I \mid O)$	mTRiDi (best of 3)	138.7	101.0	–	–	29.3	0.45	**12.3**	56.0
	mTRiDi +IK	120.6	89.5	–	–	**3.7**	**0.87**	26.2	68.3
	CompAct	**97.5**	**67.0**	–	–	15.3	0.46	23.9	**69.6**
(H, O, I)	mTRiDi (best of 3)	**198.9**	**129.2**	**260.3**	**245.4**	41.2	0.58	**9.7**	**16.7**
	CompAct	238.5	142.7	347.6	338.7	**5.0**	**0.78**	26.1	5.0

Table 5: Multi-object evaluation on the HIMO dataset under the three conditionals. mTRiDi uses the best of three samples, CompAct one sample. MPJPE, PA-MPJPE, E_{v2v}, E_c, contact gap (Gap) and penetration depth (PD) are mm; CR is a fraction; F1 is %. Best per column and block in bold.

clock and the union of its mask bits. The 1-part variant also drops the mask. **No guidance:** The full model sampled without self-InterField guidance, a test-time correction. **Binary contacts:** Both fields become a diffused bit marking points within 5 mm, penetration included; $\mathcal{L}_{\text{con}}$, $\mathcal{L}_{\text{seat}}$ and the metric and eikonal head terms are removed, and $\mathcal{L}_{\text{IF}}$ and $\mathcal{L}_{O\to H}$ score a soft bit.

Findings: Body error saturates at about 8 parts (single-object MPJPE 118.4 to 119.9 mm for 8 to 26 parts), while contact realisation improves up to 26 parts (single-object CR from 85.4 to 96.0, HIMO CR from 0.36 to 0.44), because the hand parts make and hold the contacts. At 4 and 1 parts the model collapses (HIMO Gap 165.9 and 60.8 mm). Without guidance the contacts are not closed (Gap 23.5 against 10.1 mm, CR 56.5 against 96.0). Binary contact targets are worse than the InterFields on every column except a tie on HIMO MPJPE, because a binary map carries no distance for the guidance to pull on.

E Additional results

Generation metrics: In Tab. 4, CompAct is better on 1-NNA and COV in all 12 cells, and sTRiDi on MMD and SD in 11 of 12. MMD and SD are nearest-neighbour distances that favour samples close to individual real frames, whereas 1-NNA and COV measure whether the sample set is indistinguishable from and covers the real set.

Three conditionals on HIMO: In Tab. 5, with the objects given, IK closes the contact gap at the cost of body accuracy and more penetration. With everything sampled, CompAct makes the contacts (Gap 5.0 mm, CR 0.78 against 41.2 and 0.58), while the best of three mTRiDi samples stays closer to the recorded frame. mTRiDi uses the best of three samples, CompAct a single one.

Interaction editing: The user pins any subset of the variables at the clean endpoint and samples the rest; only the resampled parts return to a noisy state and guidance acts on their InterFields alone. Figure 8 shows local grasp edits and object replacement.



Figure 8: Downstream application: **Interaction editing**. Selected variables are resampled while the remaining body parts and object variables stay fixed.

F PERCEPTUAL STUDY PROTOCOL

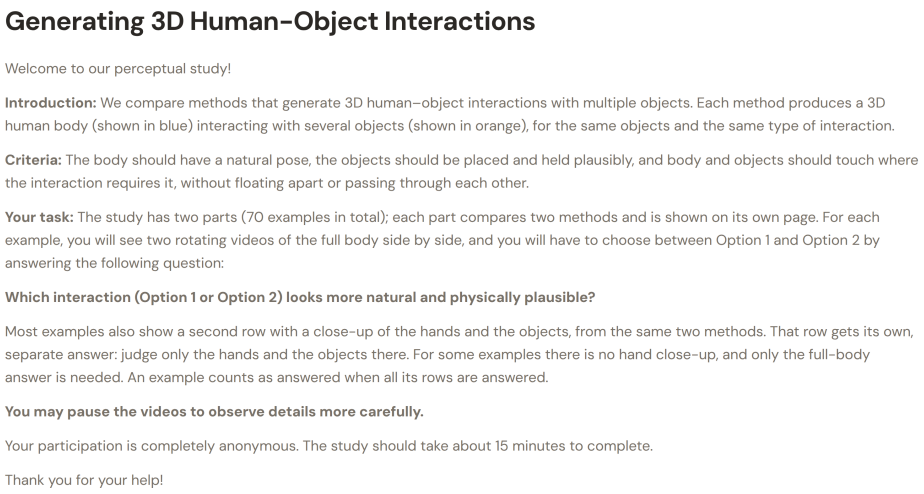

Figure 9: Perceptual study interface. The instructions shown to the 20 participants, who compare pairs of 3D interactions, CompAct against mTRiDi and CompAct against mTRiDi +IK, and pick the one that looks more natural and physically plausible, first for the whole body and then for the hand-object contacts alone.

G MORE QUALITATIVE RESULTS

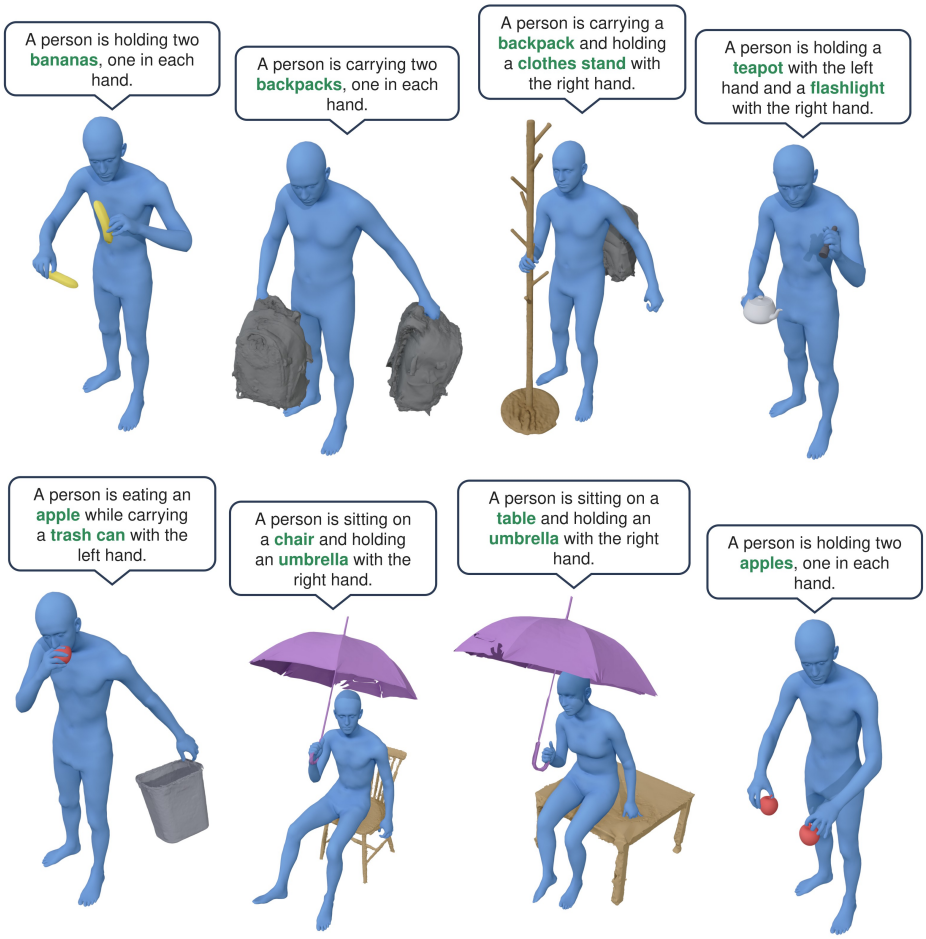

Figure 10: Extra qualitative results for multi-object cases.